INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

Sustentabilidad y Tecnología

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio

4J07 - Modelos de Predicción en Empresas y Gobierno Mediante Aprendizaje Estadístico

Analítica Avanzada de Operaciones de Comercio Exterior

PRESENTAN

Lic. en Ingeniería Financiera. Andrés Lares Barragán

Lic. en Ingeniería Financiera. Esteban Márquez Delgado

Lic. en Ingeniería Financiera. Paola Gómez Manzano

Lic. en Ingeniería Financiera. Rubén Hernández Guevara

Profesor PAP: Diana Paola Montoya Escobar

Tlaquepaque, Jalisco, Julio del 2022

ÍNDICE

Contenido

REPORTE PAP

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son experiencias socio-profesionales de los alumnos que desde el currículo de su formación universitaria- enfrentan retos, resuelven problemas o innovan una necesidad sociotécnica del entorno, en vinculación (colaboración) (co-participación) con grupos, instituciones, organizaciones o comunidades, en escenarios reales donde comparten saberes.

El PAP, como espacio curricular de formación vinculada, ha logrado integrar el Servicio Social (acorde con las Orientaciones Fundamentales del ITESO), los requisitos de dar cuenta de los saberes y del saber aplicar los mismos al culminar la formación profesional (Opción Terminal), mediante la realización de proyectos profesionales de cara a las necesidades y retos del entorno (Aplicación Profesional).

El PAP es un proceso acotado en el tiempo en que los estudiantes, los beneficiarios externos y los profesores se asocian colaborativamente y en red, en un proyecto, e incursionan en un mundo social, como actores que enfrentan verdaderos problemas y desafíos traducibles en demandas pertinentes y socialmente relevantes. Frente a éstas transfieren experiencia de sus saberes profesionales y demuestran que saben hacer, innovar, co-crear o transformar en distintos campos sociales.

El PAP trata de sembrar en los estudiantes una disposición permanente de encargarse de la realidad con una actitud comprometida y ética frente a las asimetrías sociales. En otras palabras, se trata del reto de “saber y aprender a transformar”.

El Reporte PAP consta de tres componentes:
El primer componente refiere al ciclo participativo del PAP, en donde se documentan las diferentes fases del proyecto y las actividades que tuvieron lugar durante el desarrollo de este y la valoración de las incidencias en el entorno.

En caso de requerirse alguna adecuación al nombre de las fases propuestas para este componente, se puede realizar siempre y cuando sea complementario a lo ya establecido.
El segundo componente presenta los productos elaborados de acuerdo con su tipología.

El tercer componente es la reflexión crítica y ética de la experiencia, el reconocimiento de las competencias y los aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

El objetivo general del *Proyecto de Aplicación Profesional* (PAP) "*Modelos de Predicción en Empresas y Gobierno Mediante Aprendizaje Estadístico*" es, a partir de información pública de operaciones de comercio exterior, proveniente de la *Secretaría de Hacienda y Crédito Público y del Servicio de Administración Tributaria* (SAT), y se conforma con metodologías de nivel profesional en materia de Ciencia de Datos, lo cual involucra la extracción de datos, limpieza, visualización e interpretación de estos con la finalidad de evaluar su comportamiento en KPIs y apoyar la toma de decisiones.

El proyecto se dividió en cuatro fases:

1. Exploración de datos
2. Limpieza de datos
3. Ingeniería de características
4. Creación de modelos

Cada una de estas fases forma parte de la metodología utilizada en el proyecto, el *Proceso de Ciencia de Datos en Equipos* (TDSP).

Los resultados fueron satisfactorios en las cuatro fases; la primera consistió en la familiarización de datos para lo cual se realizó un DQR, mapas de correlación, eliminación de outliers y gráficas para detectar variables de interés.

En la segunda fase se realizó la limpieza del texto con la ayuda de un mapa de palabras, el estudio de los precios y la categorización general de algunos artículos en base a su descripción, los cuales se identificaron en un 95% sobre la fase tres que incorporó la creación de variables de acuerdo con su localidad como lo es país y continente de importación.

En la última fase se desarrollaron modelos de regresión en los cuales se tuvo un buen desempeño.

1. Ciclo Participativo del Proyecto de Aplicación Profesional

El PAP es una experiencia de aprendizaje y de contribución social integrada por estudiantes, profesores, actores sociales y responsables de las organizaciones, que de manera colaborativa construyen sus conocimientos para dar respuestas a problemáticas de un contexto específico y en un

tiempo delimitado. Por tanto, la experiencia PAP supone un proceso en lógica de proyecto, así como de un estilo de trabajo participativo y recíproco entre los involucrados.

El proyecto siguió una metodología de Microsoft para ciencia de datos conocida como *Team Data Science Process* (TDSP), esta metodología moderna combina elementos centrales del ciclo de vida de ciencia de datos, ingeniería de software y procesos ágiles. Tiene cuatro fases principales:

1. Entendimiento del negocio
2. Adquisición y entendimiento de los datos
3. Modelado
4. Despliegue

Esto se realizó guiandonos con un cronograma de actividades que abarca estas cuatro fases de la metodología TDSP, el primer bloque del cronograma se compone de la exploración de datos y corresponde a la primera fase del TDSP, que en este caso se trata de las operaciones de comercio exterior en 2020 y 2021 de los Estados Unidos Mexicanos. El segundo es la limpieza exhaustiva de datos para encontrar posible correlación, mientras que el tercer bloque consiste en la generación de modelos para lo cual se realizaron 3: regresión, clasificación y clasificación multiclase donde cada equipo del PAP fue encargado para la realización de uno. Esto corresponde a la fase 2 y 3 del TDSP. El último bloque consiste en la entrega de resultados y reporte de proyecto lo cual se entiende como el despliegue, fase 4 del TDSP donde podría entonces plantearse su utilización para la toma de decisiones.

1.1. Entendimiento del ámbito y del contexto

El comercio es una actividad fundamental para México, según el ITC (International Trade Centre) en el año 2020 México se posicionó en el lugar número 11 de países con mayor número de importaciones, sumando una cifra total de 9.96B USD. Si bien la industria textil no es la que representa la mayor parte en las importaciones (está en el noveno lugar) por delante de sectores como productos alimenticios y productos animales, es un sector de suma importancia para el país y con un gran potencial de crecimiento.

Los principales países de donde provienen las importaciones textiles de México son China, Vietnam, Estados Unidos, Bangladesh e Indonesia; todos son países con mano de obra muy barata, a excepción de Estados Unidos, lo que explica la alta cantidad de importaciones de este país que colinda al norte más próximo con México, además de que tienen una muy buena relación comercial.

Bajo este contexto es que la información recopilada y almacenada por el Servicio de Administración Tributaria (SAT) de operaciones de comercio exterior es puesta a

disposición del público general, sin embargo, esta información no es rapidamente digerible por el público al que va dirigida y es por este motivo que el Proyecto de Aplicación Profesional *Modelos de Predicción en Empresas y Gobierno Mediante Aprendizaje Estadístico* tiene como objetivo el procesado de datos para posteriormente ser presentados de forma sencilla y utilizable por aquellos con relación directa o indirecta sobre las importaciones textiles en territorio nacional.

1.2. Caracterización de la organización

El escenario en el que se desarrolló este Proyecto de Aplicación Profesional fue en conjunto con el *Gobierno del Estado de Jalisco*, utilizando datos provenientes de la *Secretaría de Hacienda y Crédito Público* y del *Servicio de Administración Tributaria* (SAT), que fueron provistos por la profesora del PAP *Diana Paola Montoya Escobar* y *Angel Tomás Wong Dan*.

El proyecto fue realizado por un grupo de 11 estudiantes de la carrera de I*ngeniería Financiera* en el *Instituto Tecnológico y de Estudios Superiores de Occidente* (ITESO), quienes asumimos el rol de analistas de datos y entre todos nos encargamos del cumplimiento de los objetivos del PAP mediante entregas semanales realizadas por 3 equipos quienes habiendo realizado la misma tarea, entregamos diferentes hallazgos, presentando la información relevante sobre cada una de las fases en una reunión general el día sábado de cada semana.

Cada tarea semanal involucró la realización de distintos análisis, visualizaciones, cálculos y modelaje; esto fue presentado en distintos entornos de desarrollo integrado (IDEs) como Jupyter Notebook, Visual Code Studio y RStudio y con la utilización de distintos lenguajes de programación como Python y R.
Para su presentación posterior en los foros semanales, se resumió el contenido del trabajo colaborativo en los repositorios abiertos de Gitlab para su libre navegación.

1.3. Identificación de la(s) problemática(s)

La problemática principal del proyecto radica en la capacidad de identificar operaciones de comercio internacional que se consideran atípicas. Es decir, determinar actividades comerciales anormales dada una base de datos con información histórica de importaciones y exportaciones a nivel transacción.

Cabe resaltar que dichos datos son públicos. Sin embargo, estos son difícilmente interpretables por un público sin conocimientos de ciencia de datos debido a su alta

granularidad. Asimismo, una base de datos no homogénea afecta la capacidad de generar modelos estadísticos certeros y que resuelvan la problemática central. Debido a esto, se consideró como prioridad la limpieza de datos de forma que la base resultante se adecúe a la solución planteada. Como resultado se obtuvo una clasificación de entradas de acuerdo al artículo descrito en la descripción del producto, algunas de estas categorías son: pantalón, camisa, vestido, ropa de cama, calzado, etc.

Una vez resuelta esta primera problemática, la siguiente dificultad en el proceso fue la carencia de variables que expliquen la variable de interés: el precio unitario del artículo. Si bien, la base de datos cuenta con múltiples variables, se concluyó que estas no necesariamente aportan valor al momento de ajustar un modelo y se optó por . Consecuentemente, fue necesario realizar un análisis exploratorio de datos donde se identificara el comportamiento de cada variable respecto al precio (precios promedio, máximo y mínimo por tipo de prenda, por ejemplo) y poder así descartar aquellas con poca relación. Adicionalmente, se aplicó un proceso de ingeniería de características para generar nuevas variables útiles para el entrenamiento del modelo.

Por último, se trabajó el problema de estimaciones de precio al implementar diversas librerías para desarrollar modelos que utilicen las variables seleccionadas para determinar si el precio unitario de algún producto se considera atípico.

1.4. Planeación de alternativa(s)

La metodología de Microsoft para ciencia de datos "*Team Data Science Process*" (TDSP) sugiere el uso de herramientas para ciencia de datos de alto nivel, como Python o R e incluso excel en ciertos casos, por mencionar algunas, sin embargo no se contempló descartar ninguna de las herramientas puesto que cada una ofrece una amplia variedad de herramientas, funciones y características que fueron de gran utilidad a la hora de proponer soluciones para el problema y los fuertes de alguna pueden ser un a excelente alternativa a los puntos débiles de otras.
El lenguaje de programación "R" cuenta con librerías muy potentes en cuanto a procesamiento de lenguaje natural se refiere, algunas de estas librerías que utilizamos fueron:

- tm
- wordcloud
- quanteda
- SnowballC
- tidytext

Entre otras.

En su conjunto, estas librerías nos permitieron realizar minado de texto, procesar texto y comunicar de manera visual los más relevantes así como la realización de análisis entre pares, todo de manera eficaz y eficiente.
El resultado de todo lo realizado con R contribuyó a la parte de obtener información importante de los productos importados, gracias a la librería wordcloud se logró comunicar los productos más importados, pues a comparación de una simple gráfica de barras, esta permite diferenciar cada producto por color y el tamaño de acuerdo a la cantidad importada.

Sin embargo, la principal herramienta utilizada fue Python, el cual por su potencia y enorme cantidad de librerías facilitaron mucho la solución de los distintos problemas o retos presentados.
Las principales librerías y recursos utilizados en python fueron:

- pandas
- numpy
- plotly
- nltk
- re
- sklearn
- xgboost

Además de un script realizado en la materia “Ciencia de Datos e Inteligencia de Negocios”, el cual permite realizar el proceso completo de Exploratory Data Analysis (EDA) o Análisis Exploratorio de Datos, en español.
Estas librerías nos permitieron realizar un trabajo mucho más profesional y detallado pues por ejemplo, con plotly podemos realizar gráficas interactivas con gran nivel de detalle y personalización que difícilmente se puede lograr en excel, entonces plotly fue la principal alternativa a la hora de presentar gráficas. Por otra parte, al momento de realizar modelaje, específicamente en nuestro caso que fue realizar modelos de regresión, optamos por Python y sus librerías sklearn y xgboost pues son ya el estándar en la industria para estos tipos de modelos.

Al final utilizamos las distintas alternativas como complementos entre sí pues gran parte del rol de un analista de datos es el uso y diversificación de distintos recursos y el cómo son utilizados con sinergia para la óptima realización de la tarea y solución de problemas.

1.5. Desarrollo de la propuesta de mejora

Semana 1 Exploración de los datos

Introducción

Esta primera etapa del proyecto consiste en la familiarización con los datos, para ello, realizamos algunas técnicas que nos permitirán alcanzar esta familiarización mediante los hallazgos más importantes que estas nos aportan.

Desarrollo

- **Reporte de Calidad de Datos**
  Esto nos permite observar fácilmente el estado de los datos, es decir, datos faltantes, mínimos, máximos, valores únicos, y demás.
- **Mapa de Correlación**
  Para observar la relación que hay entre todas y cada una de las variables/columnas presentes.
- **Remover Outliers**
  Hay datos que podrían causar ruido al momento de modelar o graficar los datos, es por eso que estos datos deben ser removidos.

Resultados

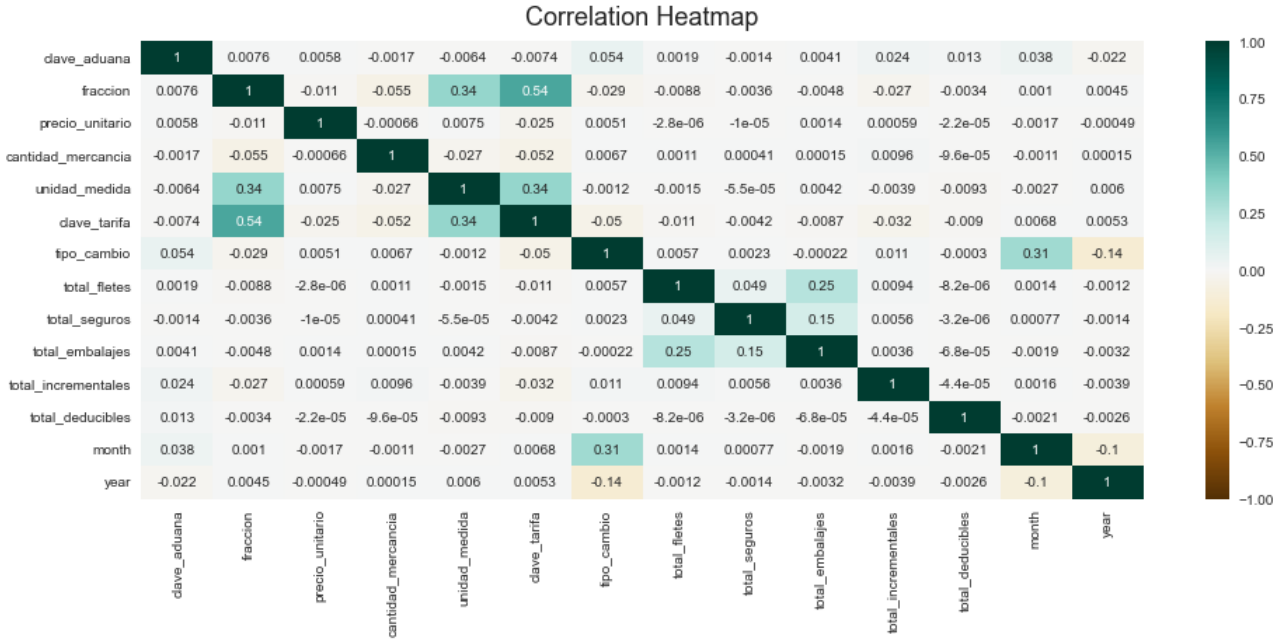
Data Quality Report

El DQR es realizado mediante un script que previamente desarrollamos, el cual permite ver cosas como tipo de dato, datos faltantes, valores únicos, y demás.

	Nombre	data_types	missing_values	present_values	unique_values	min	max
Unnamed: 0	Unnamed: 0	int64	0	1021082	112102	0	118643
clave_aduana	clave_aduana	int64	0	1021082	59	20	840
fraccion	fraccion	int64	0	1021082	1546	50040001	63109099
subdivision	subdivision	object	510433	510649	26	NaN	NaN
descripcion	descripcion	object	0	1021082	165975	"t-shirts", camisetas de punto de algodon	ã¿pantalon largo
precio_unitario	precio_unitario	float64	0	1021082	340232	0.0	14720000.0
cantidad_mercancia	cantidad_mercancia	float64	0	1021082	66177	0.001	47707792.0
unidad_medida	unidad_medida	int64	0	1021082	16	1	21
clave_tarifa	clave_tarifa	int64	0	1021082	5	1	9
valor_agregado	valor_agregado	float64	0	1021082	1	0.0	0.0
pais_origen_destino	pais_origen_destino	object	0	1021082	169	ABW	ZYA
tipo_cambio	tipo_cambio	float64	0	1021082	1197	9.3578	25.1185
total_fletes	total_fletes	float64	0	1021082	41776	0.0	1453186846.0
total_seguros	total_seguros	float64	0	1021082	7108	0.0	292964799.0
total_embalajes	total_embalajes	float64	0	1021082	1915	0.0	1314976.0
total_incrementales	total_incrementales	float64	0	1021082	21868	0.0	71089108.0
total_deducibles	total_deducibles	float64	0	1021082	21	0.0	1003824.0
month	month	int64	0	1021082	12	1	12
year	year	int64	0	1021082	2	2020	2021

Podemos observar que hay dos columnas/variables que no aportan información: “Unnamed: 0” y “valor_agregado”, la primera debido a que se trata simplemente de valores numéricos no asociados a nada en especial, la segunda, debido a que se compone únicamente de ceros.

Mapa de Correlación



Podemos observar que no hay ninguna correlación altamente positiva (color verde fuerte) ni otra altamente negativa (color naranja-café), es decir, los datos no están muy relacionados.

Remover Outliers

Para esta parte realizamos un método estadístico, utilizamos la media y la varianza para fijar un límite superior y límite inferior.

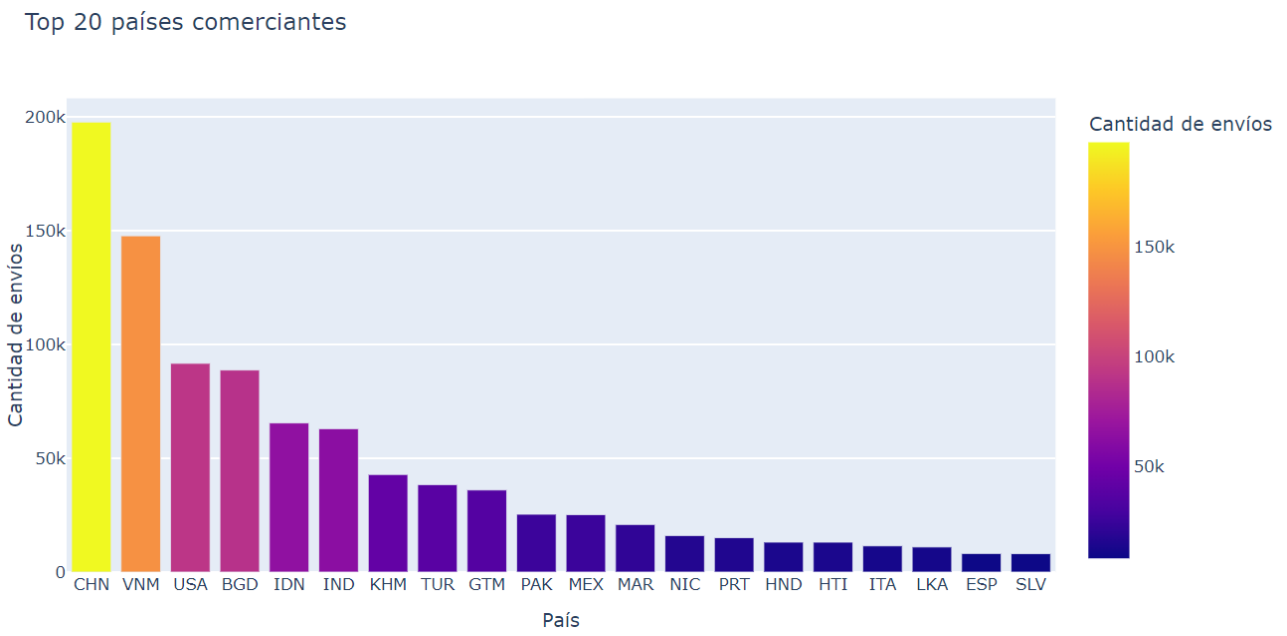
El límite inferior lo fijamos en 0.
El límite superior es de la media de los datos más tres veces la desviación estándar.

Realizamos un subset de los datos y creamos un nuevo dataframe sin outliers:

```
new_data = data[(data['precio_unitario'] < upper) &
                (data['precio_unitario'] > lower)].copy()
new_data = new_data.reset_index(drop=True)
```

En donde en la columna “precio_unitario” se mantienen únicamente los datos mayores al límite inferior y aquellos menores al límite superior.

Envíos por país



En la gráfica podemos observar el top 20 de países que tienen más registros de comercio, entre los principales están China, Vietnam, Estados Unidos, Bangladesh e Indonesia.

Semana 2 - Limpieza de texto

Introducción

Para esta 2° semana se trabajó en la limpieza del texto para lo cual se abordaron los registros de comercio mediante palabras clave que nos permitan ordenar por categorías los registros. Más adelante, la limpieza de texto nos permitirá realizar cálculos y obtener insights clave para cada tipo de prenda en particular y ya identificada.

Desarrollo

- **Mapa de Palabras**
  Esto nos permitirá elaborar palabras clave de las categorías, definirlas e identificar sinónimos para su integración y estudio posterior.

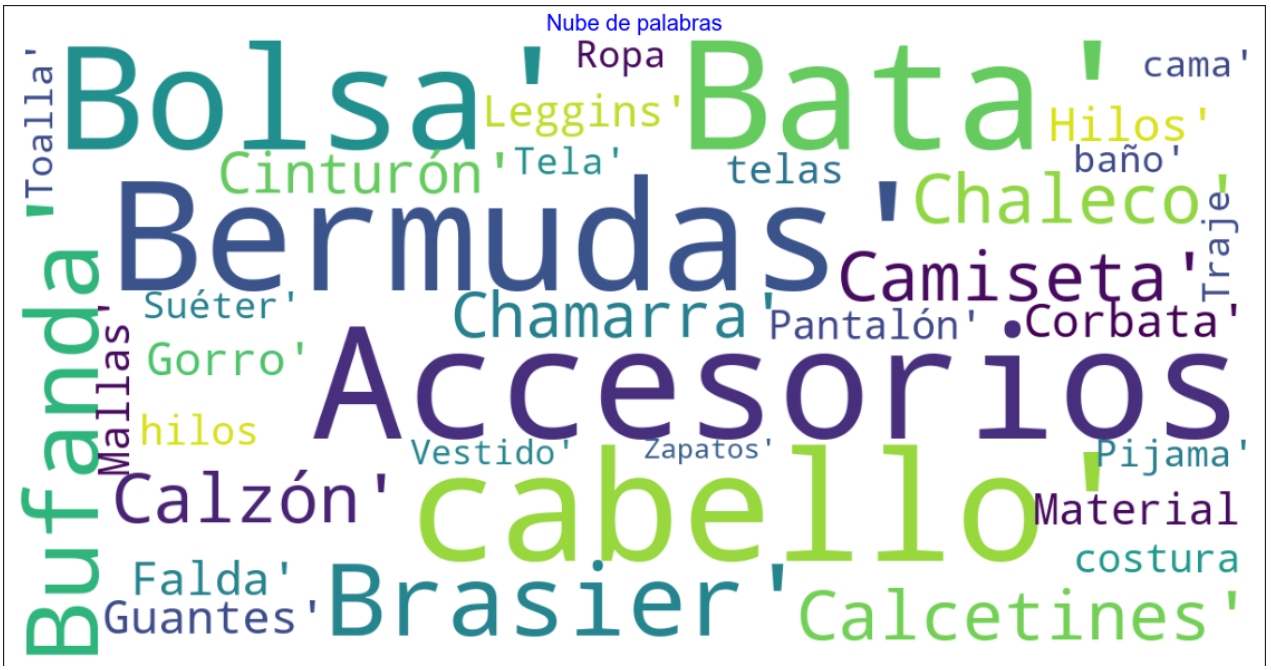
- **Categorización**
  Mediante la creación de una nueva columna llamada *Artículo* es que se agrupa por categorías únicas el apartado de las descripciones en los registros de comercio.

- **Estudio de Precios**: Se le otorga el nombre de *Sin descripción única* a aquellos registros que no están contenidos en ninguna categoría o que se repiten, posteriormente se hace un *groupby* de la columna creada de *Artículos* por la media de sus precios unitarios y se estudian los resultados para las categorías con mayor número de conteo.
  - *Camisetas*: Estudio de precios para los registros identificados como camisetas sin su límite superior (+95%).

- *Pantalones*: Estudio de los precios de los pantalones sin límite superior (+95%).

Resultados

Mapa de Palabras



Categorización

Se eligen y se crean las siguientes categorías y se asignan en la columna creada de *Artículos* en el dataframe original donde se agrupan los tipos de prenda definidos.

words
Accesorios para cabello
Bata
Bermudas
Bolsa
Brasier
Bufanda
Calcetines
Calzón
Camiseta
Chaleco
Chamarra
Cinturón
Corbata
Falda
Gorro
Guantes
Hilos
Leggins
Mallas
Material para costura
Pantalón
Pijama
Ropa de cama
Suéter
Tela
Toalla
Traje de baño
Vestido
Zapatos

	descripcion	Artículo
1780	tahalies de cordel con broche tubular, codigo ...	Accesorios para cabello
2182	broches para cabello	Accesorios para cabello
2183	diadema unicornio	Accesorios para cabello
8002	banda de mano, bolsa para pinzas, bolsa textil...	Cinturón
9958	camisa (t-shirt con broche y abertura) para bebe	Camiseta
...	...	...
947506	zapatos patucos para bebe	Vestido
949310	tenis	Vestido
949350	sandalia, zapatos	Vestido
953652	cordones de zapatos, cuerda	Vestido
970430	banda textil, benda, bolsas de malla para ropa...	Vestido

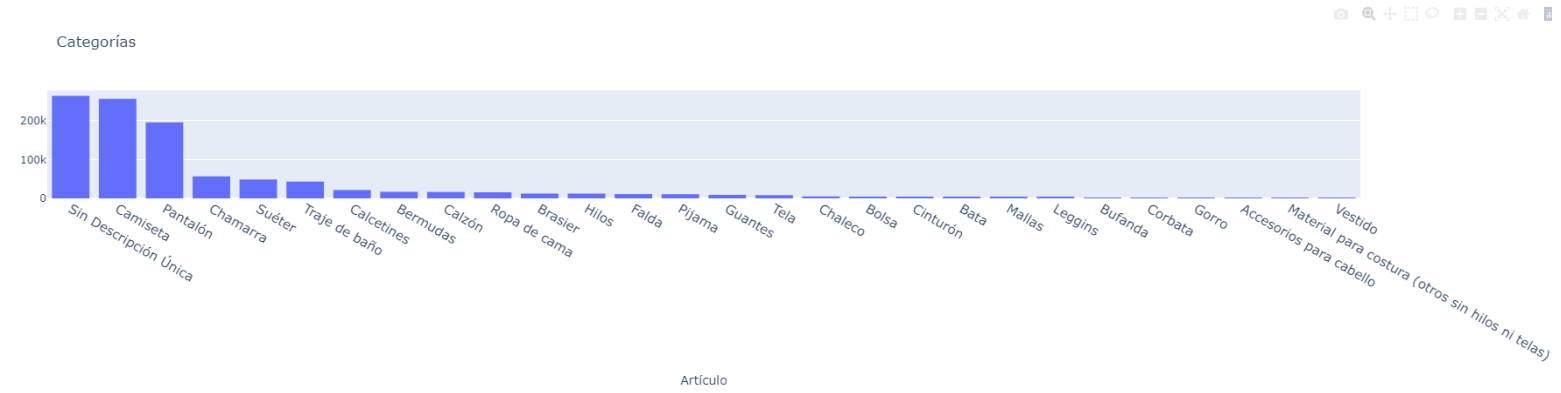
Se muestra en un dataframe el conteo de las categorías creadas para todos los registros según su descripción y se ponderan en una columna adjunta.

Categorías	Artículo	%
Sin Descripción Única	264019	0.258803
Camiseta	255993	0.250935
Pantalón	196532	0.192649
Chamarra	56595	0.055477
Suéter	48645	0.047684
Traje de baño	43140	0.042288
Calcetines	21527	0.021102
Bermudas	16851	0.016518
Calzón	15855	0.015542
Ropa de cama	15733	0.015422
Brasier	12436	0.012190
Hilos	12370	0.012126
Falda	11038	0.010820
Pijama	10731	0.010519
Guantes	8913	0.008737
Tela	8195	0.008033
Chaleco	5006	0.004907
Bolsa	2928	0.002870
Cinturón	2776	0.002721
Bata	2424	0.002376
Mallas	2380	0.002333
Leggins	2323	0.002277
Bufanda	1838	0.001802
Corbata	644	0.000631
Gorro	602	0.000590
Accesorios para cabello	307	0.000301
Material para costura (otros sin hilos ni telas)	263	0.000258
Vestido	91	0.000089

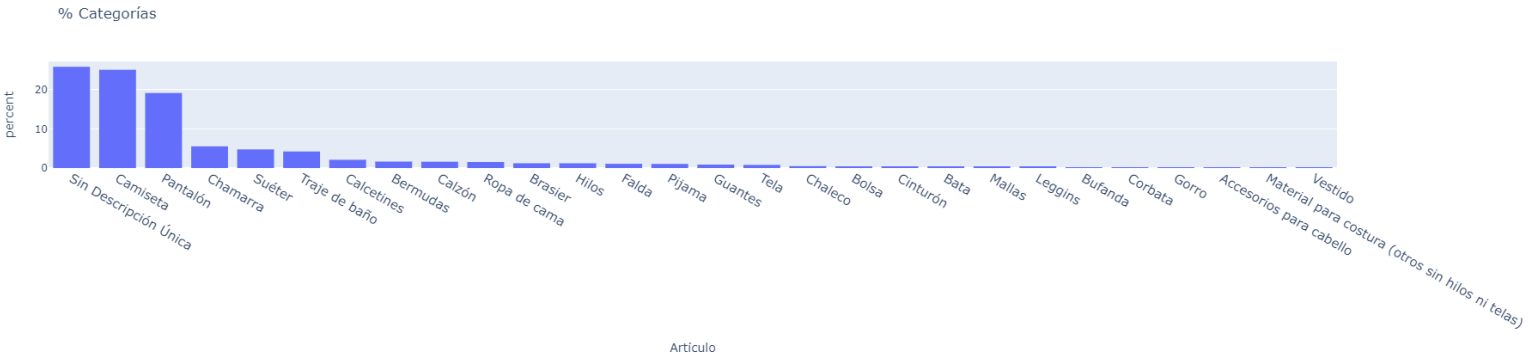
Se asigna la categoría de *Sin Descripción Única* a los registros de comercio que no están contenidos en ninguna de nuestras categorías seleccionadas o se repiten de manera que no son identificables.

```
0          Sin Descripción Única
1          Sin Descripción Única
2          Sin Descripción Única
3                          Bata
4                          Bata
                  ...
1020150    Sin Descripción Única
1020151    Sin Descripción Única
1020152                   Hilos
1020153    Sin Descripción Única
1020154                   Hilos
Name: Artículo, Length: 1020155, dtype: object
```

Se grafican los resultados obtenidos y se muestra el conteo para cada artículo de manera ordinal, se observa que los registros de comercio que se agrupan bajo las categorías de camisas y pantalones son alrededor del 45% de las observaciones mientras el 25% permanece no categorizado de manera automatizada debido a que las descripciones no contienen las palabras propuestas o estas se repiten, por otro lado las demás categorías componen de manera individual menos del 5% de la base de datos cada una.



Se procede a visualizar el histograma normalizado porcentualmente para hacer el cálculo anterior más atinado.

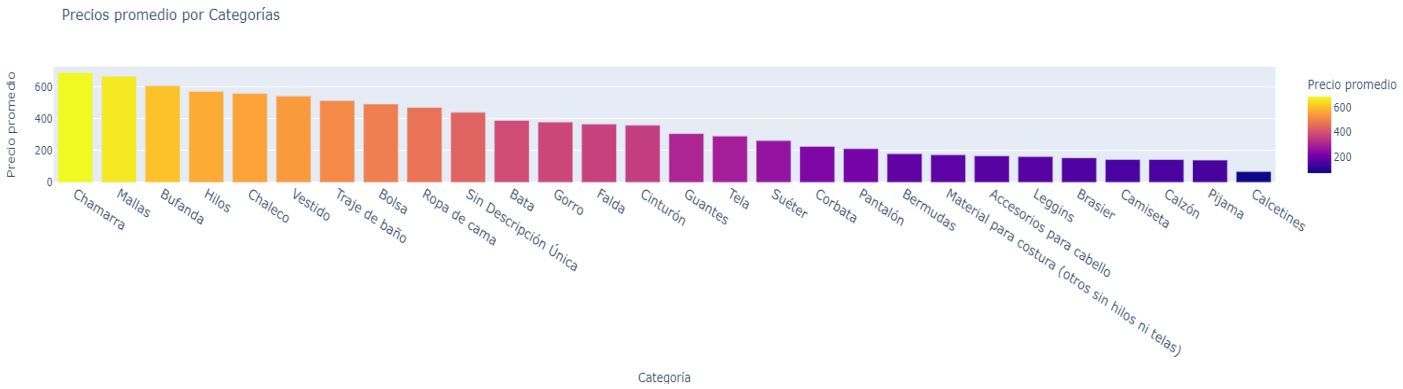


Estudio de Precios

Se agrupa la columna creada de Artículos o Categorías propuestas y se realiza un *groupby* de la media de sus precios unitarios de manera que se observa que las categorías que se declaran con un mayor precio promedio son las de Chamarras, Mallas, Bufanda, e Hilos y contrariamente al conteo de registros, las camisas y pantalones se encuentran por debajo de la media de las categorías por precio.

	Categoría	Precio promedio
10	Chamarra	689.214424
18	Mallas	666.420418
5	Bufanda	606.958423
16	Hilos	571.023091
9	Chaleco	557.976414
27	Vestido	542.297114
26	Traje de baño	513.866649
3	Bolsa	492.059388
22	Ropa de cama	470.195692
23	Sin Descripción Única	440.236486
1	Bata	389.147560
14	Gorro	379.227583
13	Falda	366.568700
11	Cinturón	359.355170
15	Guantes	306.383026
25	Tela	291.303131
24	Suéter	262.948487
12	Corbata	226.329058
20	Pantalón	211.905429
2	Bermudas	180.669471
19	Material para costura (otros sin hilos ni telas)	174.046781
0	Accesorios para cabello	167.111912
17	Leggins	162.630786
4	Brasier	154.809042
8	Camiseta	144.205082
7	Calzón	143.983882
21	Pijama	140.601317
6	Calcetines	68.803551

Se realiza a continuación un histograma de los precios promedio para nuestras categorías propuestas e identificadas en la base de datos.



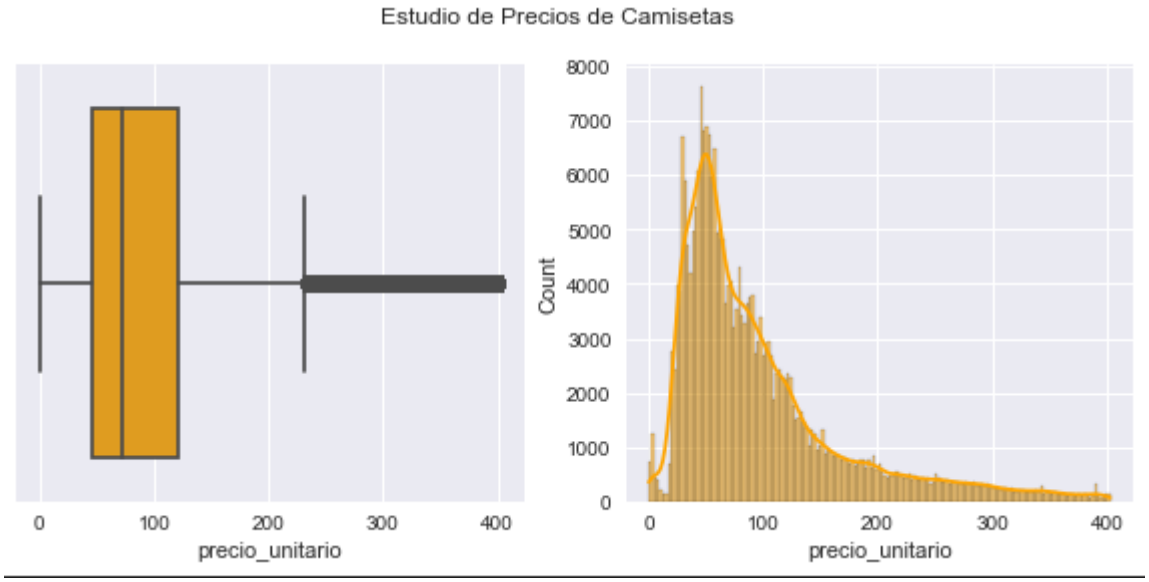
Se procede a estudiar la distribución e histograma de las categorías con mayor número de registros (pantalones y camisetas) en la base de datos y para poder remover de la visualización los valores extremos, se toma el 95% de los datos mediante la construcción de un arreglo de veintiles.

Camisetas

Se hacen veintiles del precio de las camisetas y se estudia la distribución y el diagrama de caja sin su límite superior.

```
array([  0.20167 ,  26.013836,  31.5     ,  37.75    ,  43.5     ,
        48.      ,  52.425242,  57.      ,  62.5     ,  69.      ,
        77.33333 ,  85.5     ,  94.313246, 104.230128, 117.      ,
       132.5     , 157.44444 , 197.      , 262.42857 , 403.      ])
```

A pesar de que camisetas tiene un precio medio de $144, al remover el límite superior (95-100%) tenemos que la media disminuye hasta $96.81 lo cual es un indicio de que las playeras en los registros de comercio son muy baratas y además que el límite superior de ellas es cara, tanto así que sesga la distribución que puede considerarse decreciente con una máxima de frecuencias sobre alrededor de $40.

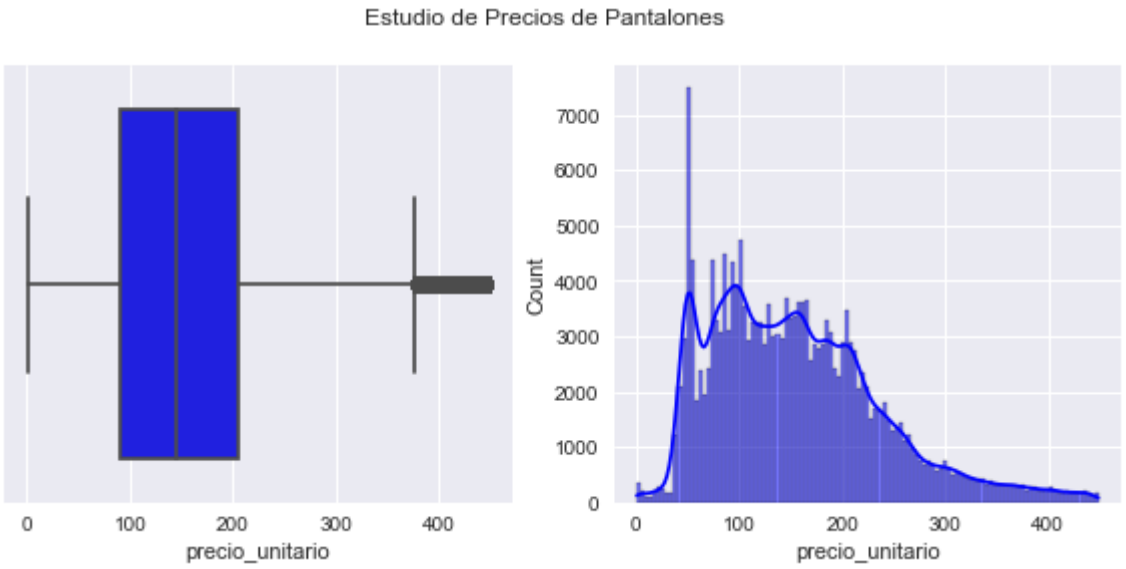


Pantalones

Se realiza el mismo procedimiento para la categoría de los pantalones y se estudia sin su límite superior al 95% que resulta estar en $449.

Considerando que la media original del precio de los pantalones es de $212.13, al remover el límite superior (+95%: 9,781 obs.) de la base de datos obtenemos que la nueva media es de $155.38 lo cual nos índica que del 19.17% ó 195,631 de los registros que contienen las palabras *pantalón, pantalón, jean, vaquero* en su descripción, quedan 185,850 (18.22%). Lo anterior nos indica que los precios extremos de las categorías identificadas sesgan sus distribuciones y la cantidad de observaciones con precios extremos para camisetas y pantalones compone más del 2% de los registros totales.

Sin embargo, a diferencia que para las camisetas, no se identifica puramente decreciente y la máxima de frecuencias es alrededor de $50. Los resultados obtenidos demuestran que las frecuencias máximas de los precios de los registros de comercio identificadas como camisetas y pantalones es muy baja ($40 y $50) y estas categorías tienen valores extremos en su límite superior y estos corresponden a grandes cantidades.



Semana 5 - Ingeniería de características

Introducción

Se propuso mediante la limpieza de texto de la entrega anterior, obtener insights clave para cada tipo de prenda en particular. En esta entrega se elaboraron las columnas de artículos, sexo, el nombre del país, su continente y código oficial en base al estándar internacional *ISO 3166*.

Desarrollo

- **Outliers**: Se propuso un nuevo modelo para remover outliers en virtud de su precio y este modelo sugirió pasar de la condición *>(μ+3σ)* a *>99%* para la clasificación de datos como outliers, lo cual redujo nuestro límite superior de $63,481 (927, .09%) a $4,220 (10,311, 1.01%).

- **Recategorización:** Mediante el estudio de las descripciones por categoría, se logró identificar al 94% de los artículos y lo restante se clasificó como *Sin Descripción Única lo cual indica que no

puede formar parte de ninguna categoría por que su descripción está contenida en más de un artículo o no está contenida en ninguna, se visualizan los resultados en data frame e histograma normalizado.

- **Estudio de Precios:** Se obtiene la media de los todos los artículos identificados, se visualizan en dataframe e histograma y como ejemplo se analizan las densidades y distribuciones de camisas, pantalones, calzado, chamarras, sueters, overalls, shorts, chalecos sin cola derecha en su distribución (*98%*) para efectos de visualización.

- **Relación de Precios con Países**: Se hace mediante un groupby un dataframe de los precios promedio por país y se grafican los 20 países con mayor precio promedio de manera ordinal.

- **Relación de Precios con Género de Artículo:** Se identifica en una columna llamada *Sexo* si las prendas son para hombre, mujer, unisex o no se especifica y se muestra en un histograma el precio promedio para cada género.

- **Origen**: Se crean las sig. columnas que especifican el origen de los registros de comercio:

Nombre País: El nombre del país del cual se originan los registros.

Código de País: Código oficial en base al estándar internacional *ISO 3166*.

Continente: Se identifica el continente al cual pertenecen los países para cada registro de comercio.

Resultados

1. **Outliers**:

Se recalculan los precios outliers y pasan de la manera convencional *L.S = >(μ+3σ)* (L.I = 0) a *L.S = > .99%*. Esto se realizó debido a que en la Recategorización se identificó que las Categorías tenían pocos valores cercanos a su L.S lo cual sesgó considerablemente la distribución de los datos y se optó por reducir dicho sesgo.

```
Outliers Removed(μ+3σ): 927 Outliers %: 0.090786 , Old Data: 1021082 , Usable Data: 1020155 , Upper Lvl (μ+3σ): $ 63481.71
```

Mediante esta propuesta se remueven 10,311 (1.01%) outliers en vez de 927 (.09%) y se reduce nuestro espacio de trabajo a X : [$0, $4,220.0).

```
Outliers Removed 99%: 10311 Outliers %: 1.009811 , Old Data: 1021082 , Usable Data: 1010771 , Upper Lvl (+99%): $ 4220.0
```

2. **Recategorización**:

Mediante el estudio de las descripciones por categoría, se logró identificar al 94% de los artículos y lo restante se clasificó como *Sin Descripción Única* lo cual indica que no puede formar parte de ninguna categoría porque su descripción está contenida en más de un artículo o no está contenida en ninguna, se visualizan los resultados en dataframe e histograma normalizado.

Se añaden o se modifican las siguientes prendas para nuestra columna de Artículo y se añaden posibles descripciones contenidas en nuestros artículos propuestos anteriormente:

Recategorización
Cubrebocas / Material médico: Cubrebocas, mascarillas, respirador, etc.
Overall: Mandil, impermeable, vestido, overalls, etc.
Correa/Cinta: Correa, cordones, cinta, arnés, etc.
Artículos de baño: Cotonete, jabón, pañuelo, etc.
Artículos de limpieza: Trapo, escoba, trapeador, filtrante, etc.
Artículos del hogar: Tapete, lampara, ventilador, cortina, alfombra, mantel, etc.

De esta manera, la columna de artículo según su descripción queda identificada correctamente y se reducen los artículos *Sin Descripción Única* de 26% (entrega anterior) a 5.99% (identificación pre-modelo).

	descripcion	Artículo
58	dona para la sujeccion del cabello	Accesorios para cabello
76	coletero para el cabello	Accesorios para cabello
248	tela sintetica de terciopelo por trama cortado...	Material para costura
355	lazos de cuello de seda no de punto partidas: ...	Material para costura
433	bandas para la cabeza, muâ¥equeras	Accesorios para cabello
...	...	...
1010726	alfombra numero de partes: 75130-52r10-r3f	Artículos del hogar
1010729	tapete	Artículos del hogar
1010730	cortina para baâ¥o	Artículos del hogar
1010738	alfombras para vehiculos	Artículos del hogar
1010753	cortina de fibras sinteticas de seguridad	Artículos del hogar

1158922 rows × 2 columns

Finalmente se obtiene que las categorías se comportan de la siguiente manera:

Categorías	Artículo	%
Camiseta	231488	0.229021
Material para costura	216393	0.214087
Pantalón	178358	0.176457
Sin Descripción Única	60573	0.059928
Suéter	54026	0.053450
Overall	45988	0.045498
Chamarra	40209	0.039781
Correa/Cinta	22460	0.022221
Calcetines	19442	0.019235
Ropa de cama	18459	0.018262
Pijama	14565	0.014410

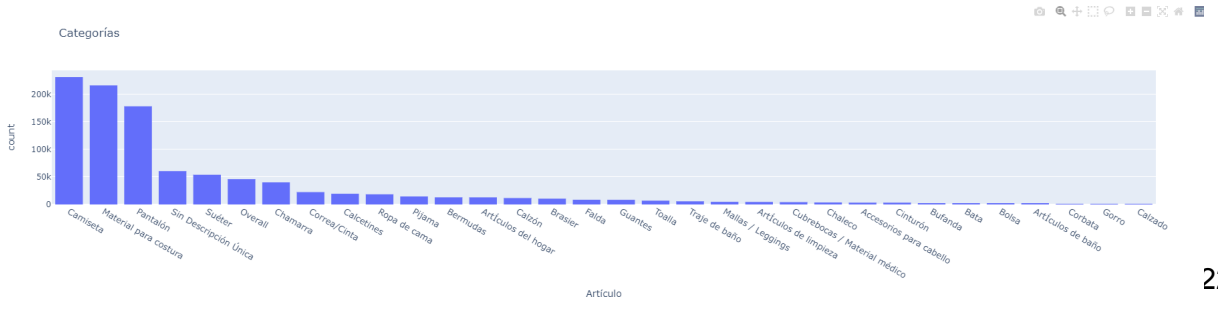
Para comprobar, se verifica intuitivamente que las descripciones de los registros pertenecientes a la categoría *Sin Descripción Única* (60,000 datos) no puedan formar parte de forma automatizada de ningún artículo o si debe crearse una nueva categoría (se repiten los procesos de ser así).

descripcion	Artículo	precio...
manga de proteccion np: 108m55	Sin Descripción Única	94
ancla	Sin Descripción Única	161.5
articulo de uso tecnico (m1356215-1t...	Sin Descripción Única	3.5915
armaduras motorizadas	Sin Descripción Única	453.9971
prendas de vestir para mujeres o niâ¥...	Sin Descripción Única	355.4492
lona np: 5wtk7	Sin Descripción Única	1605
indicador de viento np: 3lwe7	Sin Descripción Única	130.28571
manga absorvente np: 5fz78	Sin Descripción Única	392.25
acollador np: 5npp0	Sin Descripción Única	94.2619
talegas partidas: 6,7,8,9,10,11,14,17,1...	Sin Descripción Única	22.54121
monos para mujer de algodon no de ...	Sin Descripción Única	233.20814
tirantes np: 6efe3	Sin Descripción Única	602
lona np: 5wtp2	Sin Descripción Única	182
sello automotriz, acojinamiento auto...	Sin Descripción Única	2.71265
sello automotriz	Sin Descripción Única	3.05833
palazzo	Sin Descripción Única	1080.2549
649831100.750na elastico	Sin Descripción Única	74.71475
enterizos	Sin Descripción Única	94.66667
red elastica coffinet	Sin Descripción Única	145
paâ¥o para celulares	Sin Descripción Única	26.4
toreras	Sin Descripción Única	142.6
discos pulidores	Sin Descripción Única	120.048
panos np: 32kl19	Sin Descripción Única	784.7
rollo de filtro np: 4wz66	Sin Descripción Única	1076
rollo de filtro np: 4wz62	Sin Descripción Única	590
cuellera.	Sin Descripción Única	228.5
ruedas de algodon para pulir	Sin Descripción Única	231.2
velas	Sin Descripción Única	2634.529...
manga protectora	Sin Descripción Única	95.26667
articulos confeccionados.	Sin Descripción Única	64.88889
algodon sin pepita sin cardar ni peinar	Sin Descripción Única	30.48385

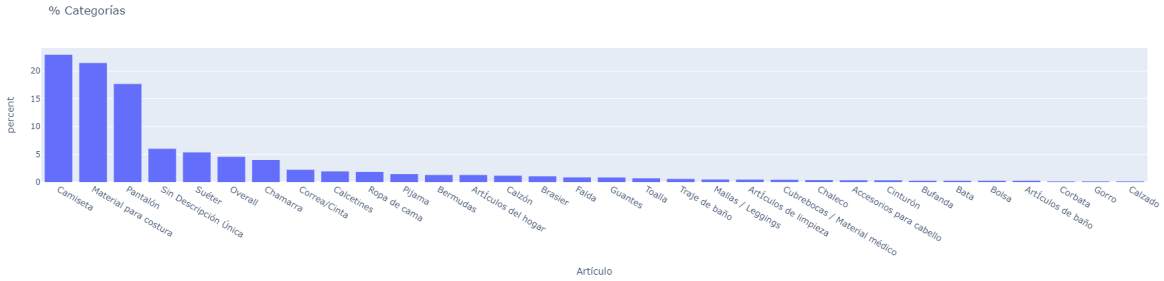
Para terminar la recategorización se comprueba de forma intuitiva en las bases de datos de cada artículo que sus descripciones pertenecen efectivamente a su categoría (posible modificación de espacios de búsqueda y repetición de procesos de ser así).

index	descripcion	Artículo	precio...
14514	short para hombre de algod...	Bermudas	94
14515	short para hombre de algod...	Bermudas	88.75
14541	short para hombre de algod...	Bermudas	94
14542	short para hombre de algod...	Bermudas	89
14549	short para hombre de algod...	Bermudas	89
14637	short para hombre de algod...	Bermudas	123
15227	bermuda	Bermudas	76.39474
15242	bermuda	Bermudas	156.44
15255	bermuda	Bermudas	68
15269	bermuda	Bermudas	201.02564
15278	short	Bermudas	121.34722
15294	bermuda	Bermudas	125.35294
15299	bermuda	Bermudas	127.53623
15314	shorts	Bermudas	158
15319	shorts	Bermudas	127
15339	shorts	Bermudas	98.52381
15660	short	Bermudas	126.12963
15838	short para hombre (de punt...	Bermudas	39.04167
15850	short para niâ¥a	Bermudas	76.66146
15914	short para beb/	Bermudas	35.9967
16114	bermuda	Bermudas	163
16164	bermuda	Bermudas	192
16438	short para hombres.	Bermudas	49.11607
16752	bermuda	Bermudas	135.41667
16763	bermuda	Bermudas	114
16855	short	Bermudas	135.66667
17205	shorts (sinteticos)	Bermudas	67
17230	shorts (sinteticos)	Bermudas	67
17257	bermuda	Bermudas	191.84

Se muestra en histogramas la categorización pre-modelo de los registros



Se normaliza la recategorización pre-modelo.

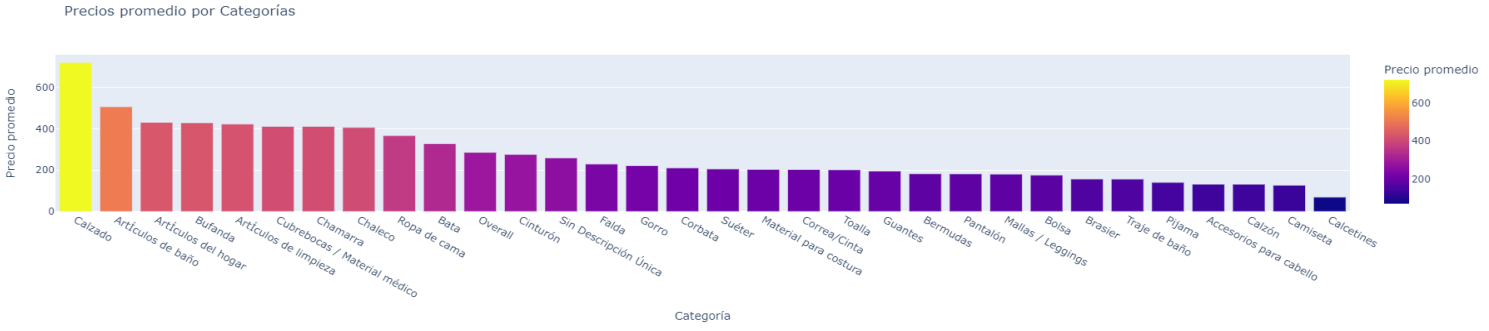


3. Estudio de Precios

	Categoría	Precio promedio
10	Calzado	720.969100
1	Artículos de baño	507.187560
3	Artículos del hogar	431.497672
8	Bufanda	429.218801
2	Artículos de limpieza	423.485261
18	Cubrebocas / Material médico	411.727136
14	Chamarra	411.695107
13	Chaleco	406.914978
27	Ropa de cama	367.440400
4	Bata	328.336803
24	Overall	286.152260
15	Cinturón	276.300403
28	Sin Descripción Única	259.184491
19	Falda	230.003025
20	Gorro	221.930337
16	Corbata	211.180324
29	Suéter	205.759404
23	Material para costura	203.616636
17	Correa/Cinta	203.102888
30	Toalla	201.694713
21	Guantes	195.905705
5	Bermudas	182.619589
25	Pantalón	182.177548
22	Mallas / Leggings	181.061521
6	Bolsa	175.994963
7	Brasier	157.146605
31	Traje de baño	156.740134
26	Pijama	140.825283
0	Accesorios para cabello	132.220889
11	Calzón	132.125892
12	Camiseta	127.105436

Se agrupan los precios promedio por artículo y se muestran en un dataframe de forma descendente.

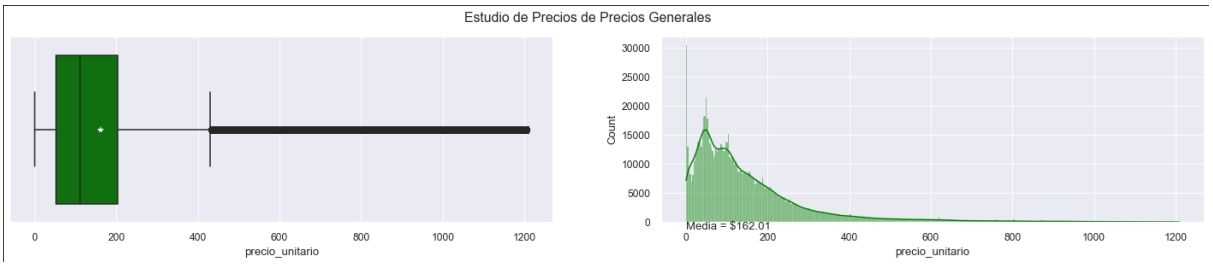
Los resultados obtenidos se muestran en el siguiente histograma.



Se muestran los percentiles de los precios generales.

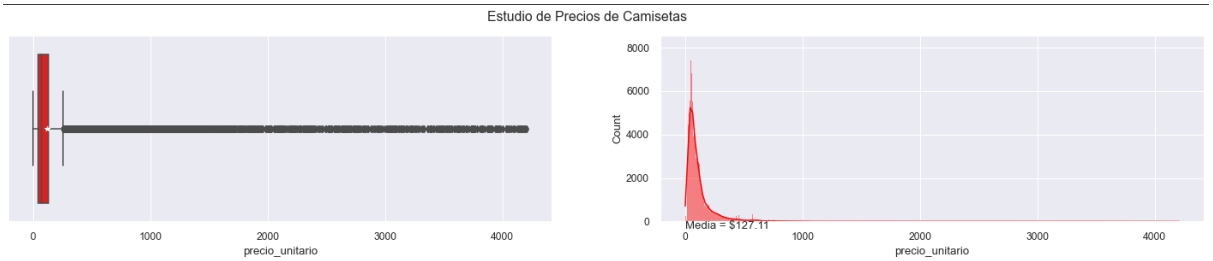
```
array([   0.00055 ,    1.336738,    5.248328,   11.996534,   19.723716,
         25.01336 ,   29.83333 ,   34.      ,   38.740902,   42.5     ,
         46.      ,   49.      ,   52.      ,   55.54167 ,   59.985514,
         64.54444 ,   70.      ,   75.      ,   79.750476,   84.58333 ,
         89.125   ,   94.25    ,   99.      ,  103.33333 ,  108.      ,
        113.33333 ,  119.      ,  125.      ,  132.      ,  139.      ,
        146.      ,  153.25    ,  160.5     ,  168.5     ,  177.33333 ,
        186.25    ,  196.      ,  206.      ,  217.      ,  230.25    ,
        244.74194 ,  261.5     ,  282.      ,  308.773826,  343.      ,
        389.5     ,  457.      ,  565.      ,  747.666662, 1209.44928 ])
```

Se crea función de visualización por diagrama de caja y distribución de precios por histograma y estimación de densidad por kernel y se grafican los precios generales.

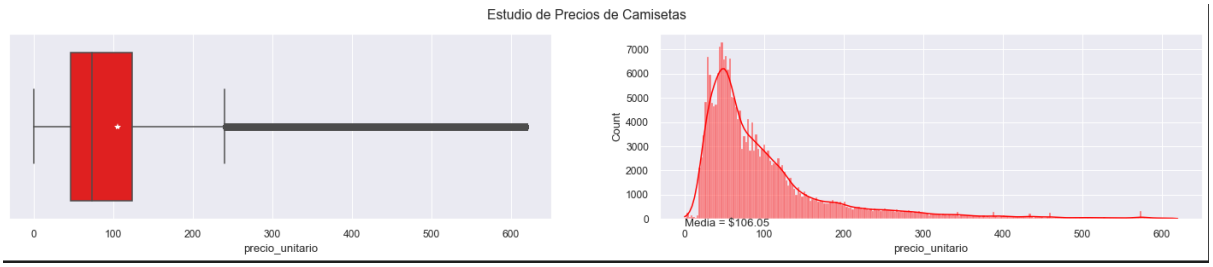


- *Camisetas:*

Se ejemplifica porque son necesarias las visualizaciones al 98% para todas las categorías (aún después de remover los *Outliers: [$4,220, $63,481)* de forma general)

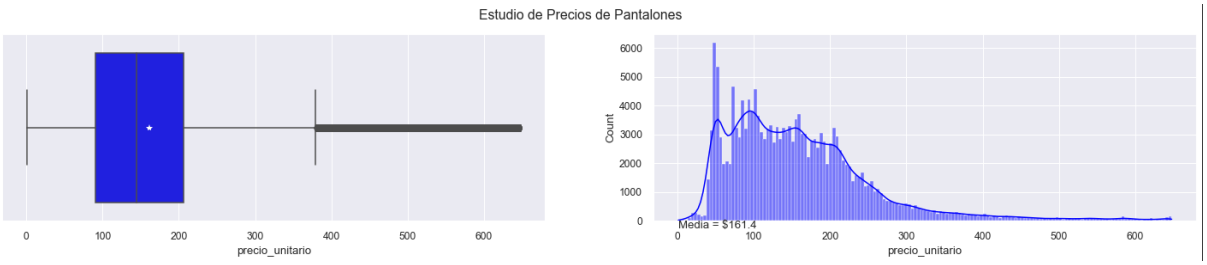


Para efectos de visualización, se grafica Camisetas sin percentiles >98%.



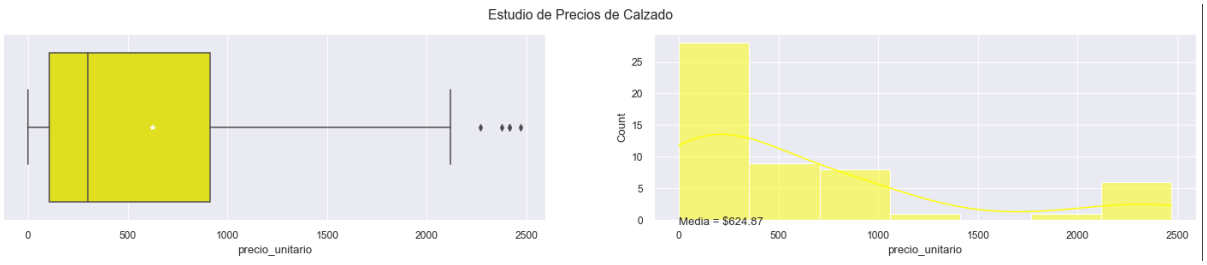
- *Pantalones:*

Se grafica el artículo Pantalones sin precios >98%.



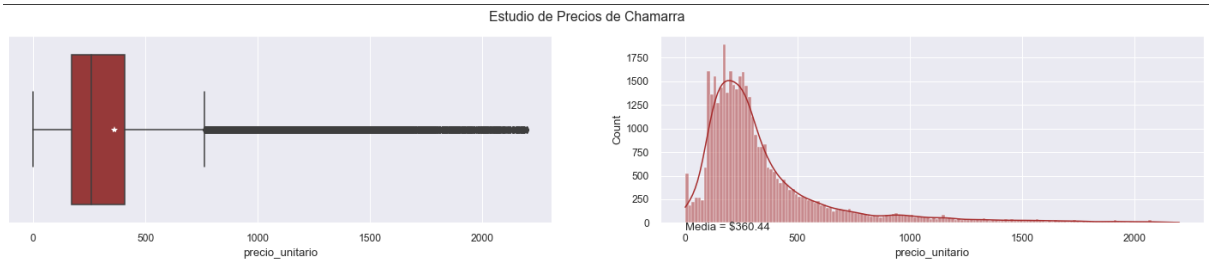
- *Calzado*:

Se grafican los registros que se reconocen como Calzado sin sus precios >98%.



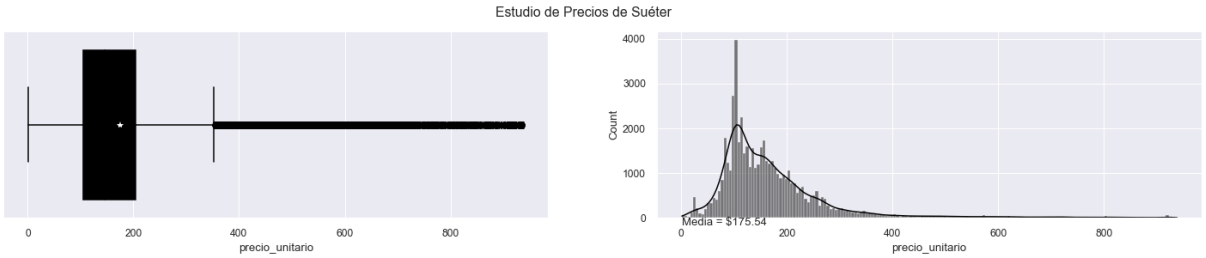
- *Chamarra*:

Se grafica Chamarra sin sus precios >98%.



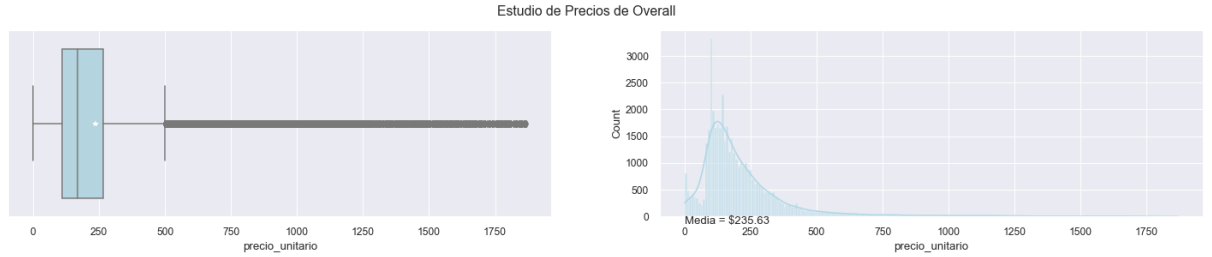
- *Suéter*:

Se grafica Suéter sin datos con precios >98%.



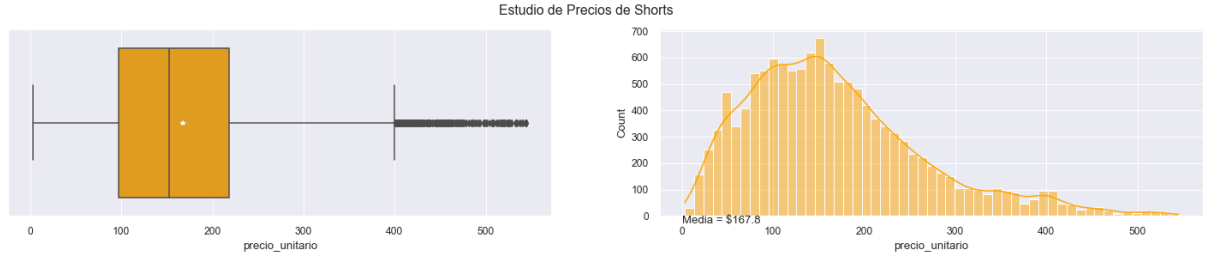
- *Overall*:

Se grafica Overall sin precios >98%.



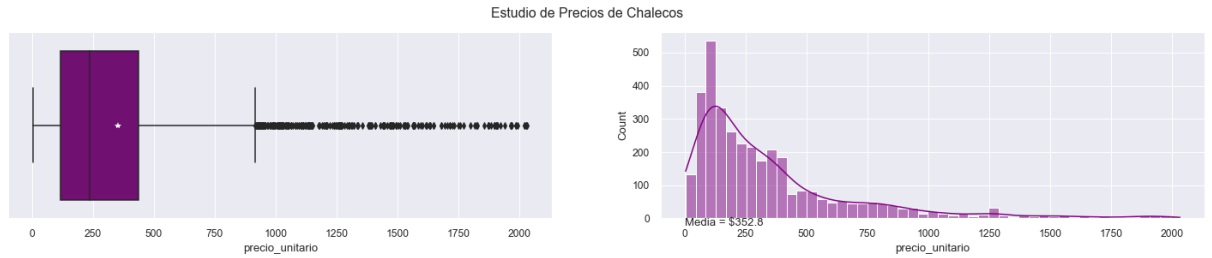
- *Shorts*:

Se grafican los Shorts sin precios >98%.



- *Chaleco*:

Se grafican los Chalecos sin precios >98%.



Para finalizar el análisis de precios se obtiene la desviación std. de los artículos anteriores:

Artículo	σ
camisetas	207.718871
pantalones	201.684839
calzado	878.572636
chamarra	492.422442
suéter	268.731583
overall	434.213826
chaleco	519.908993

Habiendo identificado que los datos originales (sin outliers pero con precios extremos) de pantalones y camisetas tienen la menor desviación estándar en precios sobre la muestra, se puede también identificar su dispersión sobre precios extremos con el cambio que sufrió μ de los datos originales a aquellos utilizados para sus visualizaciones (sin outliers y sin precios extremos).

Como se puede observar Pantalones tiene coincidentemente el menor *Avg Δ%* y la menor *σ*, lo cual significa que tienen la menor dispersión en sus precios y que los pantalones con precios extremos no son causa de una de una mayor clientela, por lo que hay pocos de estos.

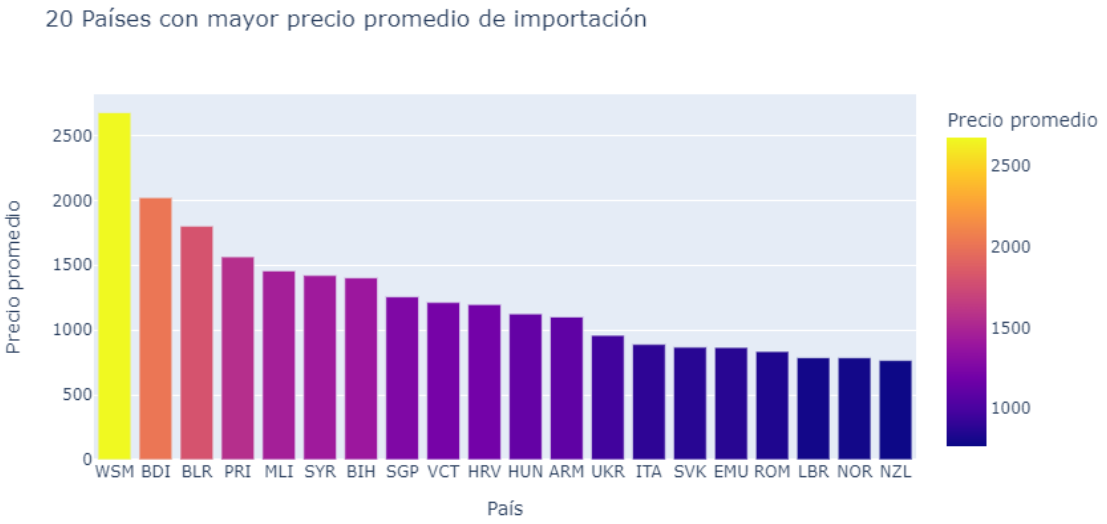
Por otro lado, Camisetas tiene la 2° desviación estándar *σ* más baja y muy cercana a Pantalones, sin embargo su *Avg Δ%* es el 2° mayor en nuestra muestra. Esto signfica que por ser la prenda más comerciada (231,488: 22.90%) tiende a tener menor dispersión en sus precios, aunado a que la distribución de sus precios se concentran de $20 a $150. Pero también, que de considerarse el límite 98%: [$619.80, $4208.88], a pesar de que pudieran ser pocas las camisetas fuera de la concentración de registros de $20 a $150, se obtiene aquello que puede apreciarse en el gráfico de su distribución, es decir, que tras dicha concentración las apariciones de nuevos registros se mantienen más constantes que decrecientes por lo cual de 231,488, el L.S al 98% (4,630) sesga la media poblacional en 20%, y esto es porque en el extremo del Límite Superior hay bastante data cuyo precio es mayor a $2,063.80 y la media poblacional es de $127.10.

Artículo	Avg Δ%
camisetas	0.198543
pantalones	0.128711
calzado	0.153784
chamarra	0.142193
suéter	0.172161
overall	0.214421
chaleco	0.153402

4. Relación de Precios con Países:

	País	Precio promedio
163	WSM	2676.000000
14	BDI	2021.000000
20	BLR	1801.166705
126	PRI	1564.139384
104	MLI	1456.000000
146	SYR	1420.687295
19	BIH	1402.225041
135	SGP	1256.244974
158	VCT	1212.500000
71	HRV	1196.938107
73	HUN	1124.788692
8	ARM	1100.546418
155	UKR	958.928814
80	ITA	889.271569
141	SVK	867.464934
49	EMU	865.436102
131	ROM	834.269006
90	LBR	787.000000
117	NOR	786.565822
119	NZL	765.367274

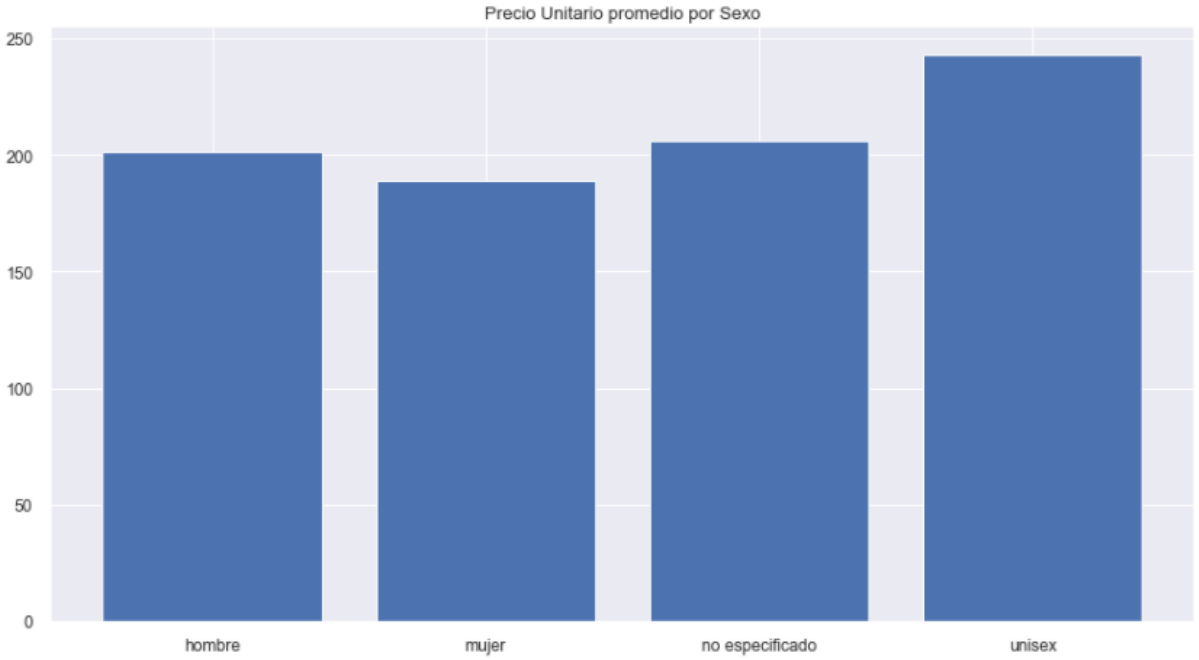
Con groupby se organizan en un dataframe los precios promedio por país y se muestran de forma descendente para posteriormente graficar los 20 países con mayor precio promedio en los registros de comercio. Se visualiza que Samoa (WSM), Burundi (BDI), Belarus (BLR), Puerto Rico (PRI) y Mali (MLI) son aquellos países con mayor precio promedio, mientras Qatar (QAT), Islas Faroe (FRO), Nigeria (NGA), Oman (OMN) y Turkmenistan (TKM) son son aquellos países con menor precio promedio.



	País	Precio promedio
42	DMA	90.695556
44	DOM	86.628821
116	NIC	83.018056
72	HTI	81.520439
114	NER	73.576925
159	VEN	70.586134
156	URY	67.455374
106	MNE	65.060000
67	GTM	63.247481
63	GMB	62.862500
34	COD	44.000000
107	MNG	43.800000
134	SEN	42.968000
130	REU	37.800000
128	PSE	37.583330
149	TKM	35.998100
120	OMN	25.586450
115	NGA	9.920000
56	FRO	5.975850
129	QAT	5.793323

5. **Relación de Precios con Género de Artículo:**

A través de un buscador en las descripciones de los registros se determina en una nueva columna el *Sexo* de las prendas (hombre, mujer, unisex o no especificado). Posteriormente, se muestra en un histograma el precio promedio para cada género.



6. **Origen:**

Nombre País: Se modifican las abreviaturas de algunos países de acuerdo a estándares internacionales para su reconocimiento posterior.

```
array(['CHN', 'USA', 'SLV', 'PHL', 'ESP', 'MEX', 'PAK', 'NIC', 'HND',
       'HTI', 'IDN', 'IND', 'BGD', 'MAR', 'TUR', 'VNM', 'ITA', 'PRT',
       'MMR', 'ROU', 'FRA', 'TWN', 'KHM', 'JOR', 'DEU', 'LKA', 'GTM',
       'ISR', 'BRA', 'BEL', 'COL', 'BGR', 'KOR', 'JPN', 'DNK', 'LAO',
       'PER', 'AUT', 'MYS', 'CAN', 'THA', 'ALB', 'CZE', 'HUN', 'DOM',
       'SRB', 'GRC', 'EGY', 'SAU', 'TUN', 'UKR', 'BIH', 'SVK', 'ECU',
       'SGP', 'SVN', 'GBR', 'CHE', 'MDG', 'MUS', 'LSO', 'ZAF', 'ASM',
       'HKG', 'ETH', 'PAN', 'CRI', 'GEO', 'LVA', 'MDA', 'LTU', 'RUS',
       'POL', 'AUS', 'NA', 'FIN', 'KEN', 'NLD', 'OMN', 'ARG', 'TZA',
       'NZL', 'NOR', 'MAC', 'IOT', 'IRL', 'CHL', 'TKM', 'BHR', 'ARE',
       'LUX', 'MRT', 'MKD', 'URY', 'SWE', 'PRI', 'HRV', 'NPL', 'AFG',
       'MOZ', 'LBN', 'GUY', 'GHA', 'DMA', 'SWZ', 'ARM', 'REU', 'BLR',
       'ZMB', 'BRN', 'EST', 'JAM', 'QAT', 'SYC', 'CMR', 'MNE', 'ISL',
       'CPV', 'SYR', 'CAF', 'BOL', 'TCD', 'TTO', 'NER', 'BLZ', 'DZA',
       'BMU', 'SMR', 'ALA', 'AND', 'AGO', 'MNG', 'VGB', 'GMB', 'GLP',
       'MLI', 'ERI', 'ATG', 'MTQ', 'LBR', 'BDI', 'GIN', 'FSM', 'NGA',
       'VEN', 'LIE', 'ABW', 'GNB', 'VCT', 'VIR', 'BRB', 'BWA', 'PSE',
       'SUR', 'FRO', 'MNP', 'IRQ', 'COD', 'DJI', 'SLE', 'SEN', 'HMD',
       'GRL', 'CYM', 'PRY', 'WSM'], dtype=object)
```

Código de País: Se identifica de acuerdo a las abreviaturas de los países, su nombre completo y su código, entre otras variables.

```
Country(name='Afghanistan', alpha2='AF', alpha3='AFG', numeric='004', apolitical_name='Afghanistan')
Country(name='Åland Islands', alpha2='AX', alpha3='ALA', numeric='248', apolitical_name='Åland Islands')
Country(name='Albania', alpha2='AL', alpha3='ALB', numeric='008', apolitical_name='Albania')
Country(name='Algeria', alpha2='DZ', alpha3='DZA', numeric='012', apolitical_name='Algeria')
Country(name='American Samoa', alpha2='AS', alpha3='ASM', numeric='016', apolitical_name='American Samoa')
Country(name='Andorra', alpha2='AD', alpha3='AND', numeric='020', apolitical_name='Andorra')
Country(name='Angola', alpha2='AO', alpha3='AGO', numeric='024', apolitical_name='Angola')
Country(name='Anguilla', alpha2='AI', alpha3='AIA', numeric='660', apolitical_name='Anguilla')
Country(name='Antarctica', alpha2='AQ', alpha3='ATA', numeric='010', apolitical_name='Antarctica')
Country(name='Antigua and Barbuda', alpha2='AG', alpha3='ATG', numeric='028', apolitical_name='Antigua and Barbuda')
Country(name='Argentina', alpha2='AR', alpha3='ARG', numeric='032', apolitical_name='Argentina')
Country(name='Armenia', alpha2='AM', alpha3='ARM', numeric='051', apolitical_name='Armenia')
Country(name='Aruba', alpha2='AW', alpha3='ABW', numeric='533', apolitical_name='Aruba')
Country(name='Australia', alpha2='AU', alpha3='AUS', numeric='036', apolitical_name='Australia')
Country(name='Austria', alpha2='AT', alpha3='AUT', numeric='040', apolitical_name='Austria')
Country(name='Azerbaijan', alpha2='AZ', alpha3='AZE', numeric='031', apolitical_name='Azerbaijan')
Country(name='Bahamas', alpha2='BS', alpha3='BHS', numeric='044', apolitical_name='Bahamas')
Country(name='Bahrain', alpha2='BH', alpha3='BHR', numeric='048', apolitical_name='Bahrain')
Country(name='Bangladesh', alpha2='BD', alpha3='BGD', numeric='050', apolitical_name='Bangladesh')
Country(name='Barbados', alpha2='BB', alpha3='BRB', numeric='052', apolitical_name='Barbados')
Country(name='Belarus', alpha2='BY', alpha3='BLR', numeric='112', apolitical_name='Belarus')
Country(name='Belgium', alpha2='BE', alpha3='BEL', numeric='056', apolitical_name='Belgium')
Country(name='Belize', alpha2='BZ', alpha3='BLZ', numeric='084', apolitical_name='Belize')
Country(name='Benin', alpha2='BJ', alpha3='BEN', numeric='204', apolitical_name='Benin')
Country(name='Bermuda', alpha2='BM', alpha3='BMU', numeric='060', apolitical_name='Bermuda')
Country(name='Bhutan', alpha2='BT', alpha3='BTN', numeric='064', apolitical_name='Bhutan')
Country(name='Bolivia, Plurinational State of', alpha2='BO', alpha3='BOL', numeric='068', apolitical_name='Bolivia, Plurination
al State of')
Country(name='Bonaire, Sint Eustatius and Saba', alpha2='BQ', alpha3='BES', numeric='535', apolitical_name='Bonaire, Sint Eusta
tius and Saba')
Country(name='Bosnia and Herzegovina', alpha2='BA', alpha3='BIH', numeric='070', apolitical_name='Bosnia and Herzegovina')
Country(name='Botswana', alpha2='BW', alpha3='BWA', numeric='072', apolitical_name='Botswana')
Country(name='Bouvet Island', alpha2='BV', alpha3='BVT', numeric='074', apolitical_name='Bouvet Island')
Country(name='Brazil', alpha2='BR', alpha3='BRA', numeric='076', apolitical_name='Brazil')
Country(name='British Indian Ocean Territory', alpha2='IO', alpha3='IOT', numeric='086', apolitical_name='British Indian Ocean
Territory')
Country(name='Brunei Darussalam', alpha2='BN', alpha3='BRN', numeric='096', apolitical_name='Brunei Darussalam')
Country(name='Bulgaria', alpha2='BG', alpha3='BGR', numeric='100', apolitical_name='Bulgaria')
Country(name='Burkina Faso', alpha2='BF', alpha3='BFA', numeric='854', apolitical_name='Burkina Faso')
Country(name='Burundi', alpha2='BI', alpha3='BDI', numeric='108', apolitical_name='Burundi')
Country(name='Cambodia', alpha2='KH', alpha3='KHM', numeric='116', apolitical_name='Cambodia')
Country(name='Cameroon', alpha2='CM', alpha3='CMR', numeric='120', apolitical_name='Cameroon')
Country(name='Canada', alpha2='CA', alpha3='CAN', numeric='124', apolitical_name='Canada')
Country(name='Cabo Verde', alpha2='CV', alpha3='CPV', numeric='132', apolitical_name='Cabo Verde')
Country(name='Cayman Islands', alpha2='KY', alpha3='CYM', numeric='136', apolitical_name='Cayman Islands')
Country(name='Central African Republic', alpha2='CF', alpha3='CAF', numeric='140', apolitical_name='Central African Republic')
Country(name='Chad', alpha2='TD', alpha3='TCD', numeric='148', apolitical_name='Chad')
Country(name='Chile', alpha2='CL', alpha3='CHL', numeric='152', apolitical_name='Chile')
Country(name='China', alpha2='CN', alpha3='CHN', numeric='156', apolitical_name='China')
Country(name='Christmas Island', alpha2='CX', alpha3='CXR', numeric='162', apolitical_name='Christmas Island')
Country(name='Cocos (Keeling) Islands', alpha2='CC', alpha3='CCK', numeric='166', apolitical_name='Cocos (Keeling) Islands')
Country(name='Colombia', alpha2='CO', alpha3='COL', numeric='170', apolitical_name='Colombia')
Country(name='Comoros', alpha2='KM', alpha3='COM', numeric='174', apolitical_name='Comoros')
Country(name='Congo', alpha2='CG', alpha3='COG', numeric='178', apolitical_name='Congo')
Country(name='Congo, Democratic Republic of the', alpha2='CD', alpha3='COD', numeric='180', apolitical_name='Congo, Democratic
Republic of the')
Country(name='Cook Islands', alpha2='CK', alpha3='COK', numeric='184', apolitical_name='Cook Islands')
Country(name='Costa Rica', alpha2='CR', alpha3='CRI', numeric='188', apolitical_name='Costa Rica')
Country(name="Côte d'Ivoire", alpha2='CI', alpha3='CIV', numeric='384', apolitical_name="Côte d'Ivoire")
Country(name='Croatia', alpha2='HR', alpha3='HRV', numeric='191', apolitical_name='Croatia')
Country(name='Cuba', alpha2='CU', alpha3='CUB', numeric='192', apolitical_name='Cuba')
Country(name='Curaçao', alpha2='CW', alpha3='CUW', numeric='531', apolitical_name='Curaçao')
Country(name='Cyprus', alpha2='CY', alpha3='CYP', numeric='196', apolitical_name='Cyprus')
Country(name='Czechia', alpha2='CZ', alpha3='CZE', numeric='203', apolitical_name='Czechia')
```

Continente: Se les atribuye un continente de origen a los países.

igen_destino	tipo_cambio	total_fletes	total_seguros	total_embalajes	total_incrementales	total_deducibles	month	year	Codigo_Pais	Nombre_Pais	Continente
CHN	19.6113	0.0	0.0	0.0	0.0	0.0	1	2020	156	China	Asia
CHN	19.6113	0.0	0.0	0.0	0.0	0.0	1	2020	156	China	Asia
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
SLV	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	222	El Salvador	North America
SLV	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	222	El Salvador	North America
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
PHL	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	608	Philippines	Asia
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
CHN	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	156	China	Asia
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
CHN	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	156	China	Asia
PHL	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	608	Philippines	Asia
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America
USA	18.8727	0.0	0.0	0.0	11324.0	0.0	1	2020	840	United States of America	North America

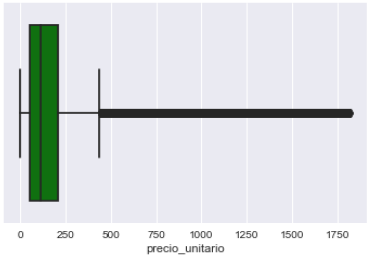
Semana 7- Modelos de regresión

Introducción

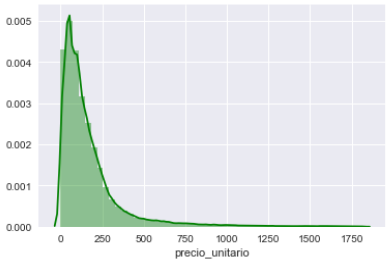
Para esta parte del proyecto se desarrollaron varios modelos de regresión por los miembros del equipo, para estos modelos se usó la variable de precio unitario como variable objetivo. Es decir, se entrenaba al modelo con las otras variables diferentes al precio unitario y se buscaba predecir el precio del nuevo artículo con esas características.

Se inició con una nueva limpieza de la base de datos enfocada a dejar las variables más significativas para la predicción del precio. Después se dividió el resultado entre train (90%) y test (10%), buscando que la distribución de ambos conjuntos fuera parecida, esto se logró de manera gráfica como se puede observar en las siguientes gráficas:

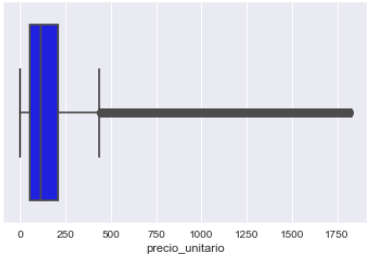
```
In [32]: sns.boxplot(test['precio_unitario'], color='green')
Out[32]: <matplotlib.axes._subplots.AxesSubplot at 0x1c0d1c50e88>
```



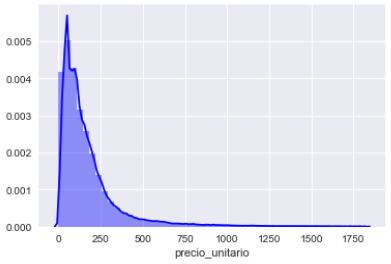

```
In [34]: sns.distplot(test['precio_unitario'], color='green')
Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x1c0cb482dc8>
```




```
In [33]: sns.boxplot(train['precio_unitario'], color='blue')
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1c0cb414048>
```




```
In [35]: sns.distplot(train['precio_unitario'], color='blue')
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x1c0cc377dc8>
```



Finalmente se realizaron diversos modelos de regresión para modelar los precios.

Desarrollo

- **Limpieza de la base de datos**: Se realizó una nueva limpieza de la base de datos para dejar las variables más significativas, así como separar la variable “precio_unitario” para ser la variable a predecir. También se transformaron variables de texto existentes en numéricas o dummies para poder ser utilizadas en los modelos predictivos.

- **División en train y test:** Mediante una semilla random se buscó dividir la base de datos en “train” y “test” para poder tener una muestra que permita evaluar el desempeño de cada modelo. Ambas muestras deben tener distribuciones similares lo cual se pudo observar de manera gráfica.

- **Modelo de regresión lineal/polinomial:** La función group_by es usada para calcular los precios promedio mensuales, después usando la función PolynomialFeatures y usando un ciclo for se hacen

pruebas para encontrar el mejor grado que minimiza el error cuadrático medio y maximiza el R2 score sin caer en overfitting.

- **Regresión Gradient Boosting**: Se utiliza la función XGBRegressor() con el set de train definido previamente para poder modelarlo. Se obtuvo un MSE de 0.07 y un RMSE de 0.27.

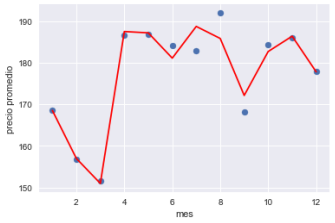
- **Regresión Random Forest:** Se realiza una separación aleatoria de datos entre entrenamiento y prueba de 90% - 10%. Posteriormente, se optimiza la cantidad de árboles del modelo utilizando un proceso iterativo y escogiendo aquel con menor RMSE calculada con validación cruzada. Por último, se obtienen las estimaciones de las primeras 50 observaciones del set de entrenamiento. Se obtuvo un RMSE de 335.55.

Resultados

El primer modelo realizado fue una regresión polinomial simple que tenía como objetivo analizar los cambios entre los precios promedio de cada mes, se buscaba detectar si había inflación de un mes a otro. Como eran 12 meses de información se ajustaron desde grado 0 hasta grado 12 y se analizó el r^2 y el error cuadrático medio de cada uno. El que dio los mejores resultados sin caer en overfitting fue el de grado nueve, que finalmente fue graficado.

7]:

	0	mean_sq
0	0.000000	0.000000
1	-1.963762	114.793736
2	-0.398164	89.507951
3	-0.393151	89.373857
4	0.123844	71.696012
5	0.625337	41.843364
6	0.626725	41.730496
7	0.626849	41.720420
8	0.719698	33.612150
9	0.941648	8.464657
10	0.962906	5.491432
11	0.965864	5.067977
12	0.962454	5.562443




```
Error cuadratico medio: 8.464657216946897
R2_score: 0.9416480919901024
La ecuacion lineal para la regresion es: x*[[ 4717.53145315 -4673.55687158  2405.48066616  -724.19206116
    134.54775486   -15.62579289     1.10445099    -0.04343428
      0.00072842]]+[-1676.74989283]
```

Analizando los resultados del modelo polinomial no podemos ver una tendencia clara de un aumento sostenido de los precios, es decir no muestra que haya inflación.

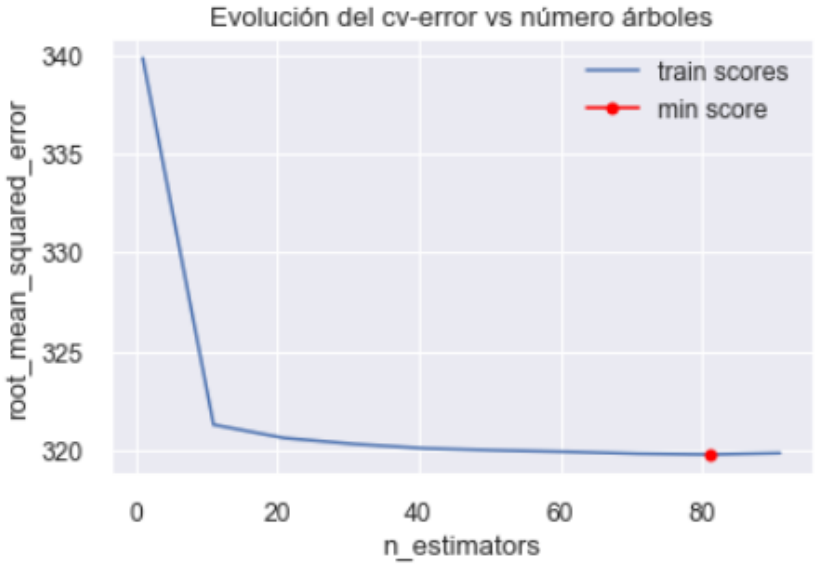
El segundo modelo está basado en gradient boosting, de igual manera se tiene el precio como variable objetivo. Para este modelo se tomaron en cuenta las siguientes variables: 'Artículo', 'fraccion', 'Codigo_Pais' ,'cantidad_mercancia', 'tipo_cambio', 'month' y 'precio_unitario'. Se transformó la variable artículo de texto a númerica con el siguiente diccionario para que fuera más significativa para el modelo:

```
{'Accesorios para cabello': 1,
 'Bata': 2,
 'Bermudas': 3,
 'Bolsa': 4,
 'Brasier': 5,
 'Bufanda': 6,
 'Calcetines': 7,
 'Calzón': 8,
 'Camiseta': 9,
 'Chaleco': 10,
 'Chamarra': 11,
 'Suéter': 12,
 'Cinturón': 13,
 'Corbata': 14,
 'Falda': 15,
 'Gorro': 16,
 'Guantes': 17,
 'Mallas / Leggings': 18,
 'Pantalón': 19,
 'Pijama': 20,
 'Ropa de cama': 21,
 'Toalla': 22,
 'Traje de baño': 23,
 'Calzado': 24,
 'Material para costura ': 25,
 'Cubrebocas / Material médico': 26,
 'Overall': 27,
 'Correa/Cinta': 28,
 'Artículos de baño': 29,
 'Artículos de limpieza': 30,
 'Artículos del hogar': 31}
```

Finalmente se obtuvo el siguiente modelo:

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='ls', max_depth=3,
                          max_features='auto', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=10,
                          n_iter_no_change=None, presort='auto',
                          random_state=123, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
```

Se propuso un tercer modelo basado en árboles aleatorios (Random Forest Regressor) el cual utiliza las variables ´cantidad_mercancia´ y ´tipo_cambio´ para estimar ´precio_unitario´. Una implementación adecuada de este modelo requiere de una optimización de hiperparámetros. Utilizando la metodología random train-test split en una proporción de .90 y .10 respectivamente, se ajustó el modelo a los datos de entrenamiento en 100 iteraciones variando la cantidad de árboles (parámetro ´n_estimators´) en cada una de estas. A continuación se presenta la métrica de desempeño ´rmse´, la cual mide el error de ajuste entre los datos de entrenamiento y las predicciones del modelo, de cada iteración.



Cabe resaltar que la métrica fue calculada utilizando la metodología de validación cruzada para reflejar más acertadamente el desempeño de los modelos. Marcado por el punto rojo se encuentra el modelo que obtuvo el menor error de ajuste; aquel con 82 árboles como predictores. Por otro lado, el criterio del codo indica que el valor óptimo del hiper parámetro se encuentra cercano a 12; puesto que la mejora marginal en desempeño comienza a decrecer a partir de este punto y no justifica el incremento de esfuerzo computacional. Finalmente, un análisis cualitativo de la gráfica indica que una asíntota horizontal se comienza a formar a partir de la iteración 40. Esto se puede interpretar como un segundo punto de inflexión (o “codo”) que otorga un desempeño bastante parecido al del mejor modelo con una carga computacional significativamente menor. Consecuentemente, se le asigna 40 al parámetro del modelo ´n_estimators´.

A continuación se presentan las 50 primeras estimaciones del modelo definido comparadas contra su valor real y sus métricas de desempeño:


```
{'mape': 7.942610633413545,
 'me': 0.37454683511591685,
 'mae': 153.86775814974004,
 'mpe': 7.683169597844277,
 'rmse': 335.5495913283836}
```

De la gráfica se concluye que el comportamiento del artículo 29 asemeja al de un outlier. Sin embargo, no es posible aseverar que este lo es puesto que la diferencia puede ser atribuíble a una mala estimación por parte del modelo.

Tras realizar una comparación entre los tres modelos propuestos, se determinó que el modelo Gradient Boosting se ajusta de mejor manera y, por ende, presenta una mayor capacidad de realizar estimaciones certeras.

1.6. Valoración de productos, resultados e impactos

La ciencia de datos es un campo con demasiado potencial en México, día a día se generan inmensas cantidades de datos las cuales, desafortunadamente, muchas veces no son aprovechadas por el gobierno y las dependencias que los generan. Y cuando son utilizadas y puestas a disposición del público, estas son difíciles de interpretar.
Por eso uno de los objetivos de todo lo que se realizó en el PAP fue realizar una limpieza profunda de datos, con la cual fue posible eliminar tanto variables como datos que no aportaran información; otra parte importante del proyecto fue la ingeniería de características con lo cual se obtuvieron variables de alta importancia a partir de las ya existentes; entre las variables que se crearon fueron el nombre del artículo principal, el país y el continente de proveniencia. Los resultados de estas variables indican que principalmente se importan camisas, pantalones, blusas, vestidos, entre otras prendas. En cuanto a los principales países de donde se importa, estos son: China, India, Vietnam, Pakistán y Turquía, todos del continente asíatico.

Con estas variables nos fue posible generar distintos modelos de regresión con el objetivo de realizar predicciones de precio dependiendo de las características del producto.
Algunos aspectos que no fueron contemplados en este periodo de PAP fue la verificación de transacciones legítimas, es decir, que podamos encontrar patrones o indicios de movimientos fraudulentos, así mismo un factor muy importante previo a realizar esto es considerar la composición de los artículos importados, es decir, si el artículo se compone de distintos materiales, definir qué peso tiene este factor en el precio del artículo.

1.7. Bibliografía y otros recursos

Base de Datos 2020, Secretaría de Hacienda y Crédito Público y del Servicio de Administración Tributaria (SAT). Información Pública de Operaciones de Comercio Exterior.. *BD_Comercio_Exterior_2020*. SHCP. Recuperado de: https://iteso01-my.sharepoint.com/:x:/r/personal/angel_wong_iteso_mx/_layouts/15/Doc.aspx?sourcedoc=%7BF3512EBC-B4C0-4AC4-AD5E-5FE966E89F6D%7D&file=BD_Comercio_Exterior_2020.csv&action=default&mobileredirect=true

Descripción de Campos, Secretaría de Hacienda y Crédito Público y del Servicio de Administración Tributaria (SAT). Información Pública de Operaciones de Comercio Exterior. SHCP. Recuperado de: https://iteso01-my.sharepoint.com/personal/angel_wong_iteso_mx/_layouts/15/onedrive.aspx?ct=1653338979610&or=OWA-NT&cid=e7155230-df90-2ac8-ac58-fdbee1a5d42c&ga=1&id=%2Fpersonal%2Fangel_wong_iteso_mx%2FDocuments%2FPAP2022%2FDescripcion_de_Campos_28062021%2Epdf&parent=%2Fpersonal%2Fangel_wong_iteso_mx%2FDocuments%2FPAP2022

Gradient Boosting con Python, Amat, J. (2020). Ciencia de datos.net. Recuperado de https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

Team Data Science Process. Docs.microsoft.com. 2022. Azure Architecture Center. Recuperado de: https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

Planeación. PAP MMD Verano ITESO 2022. Modelos de Predicción en Empresas y Gobierno Mediante Aprendizaje Estadístico. Recuperado de: https://iteso01-my.sharepoint.com/:p:/g/personal/diana_montoya_iteso_mx/ETCG5ZdAf-VCmazBM8v880gBAfDEgEQxLvumISWEqvZUNw?rtime=H2S7U_Fl2kg

Wordcloud Function. Rdocumentation.org. 2022. RDocumentation. Recuperado de: https://www.rdocumentation.org/packages/wordcloud/versions/2.6/topics/wordcloud

Plotly.graph_objects.Figure. Plotly.com. 2022. — *5.9.0 documentation*. Recuperado de: https://plotly.com/python-api-reference/generated/plotly.graph_objects.Figure.html

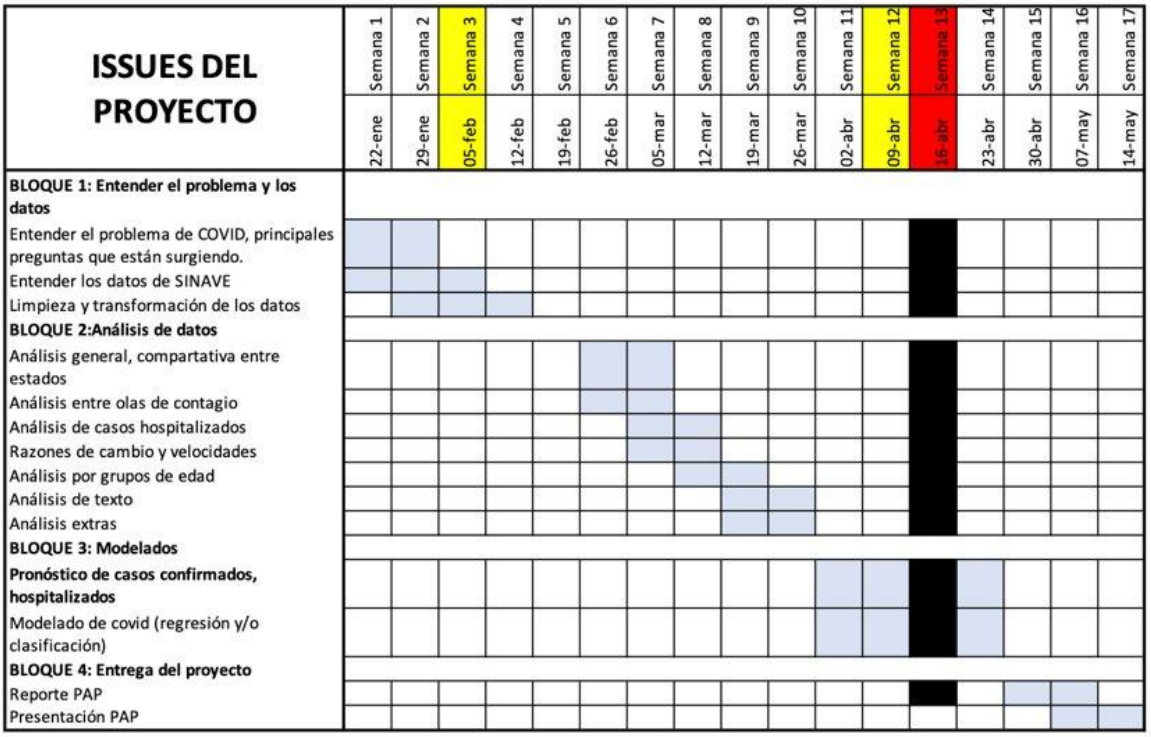
XGBoost Documentation. Xgboost.readthedocs.io. 2022. — *xgboost 1.6.1 documentation* Recuperado de: https://xgboost.readthedocs.io/en/stable/

1.8. Anexos generales

Anexo 1. Repositorio con scripts y notebooks utilizados en el proyecto.

Gitlab Issues. PAP MMD V2022, Equipo 1 (2022). Recuperado de: https://gitlab.com/dpmontoy1/pap-mmd-v2022/-/issues/?sort=created_date&state=closed&first_page_size=20

Anexo 2. Cronograma del proyecto



2. Productos

Los productos generados en este proyecto son virtuales, están contenidos en el repositorio del **anexo 1**, se trabajó principalmente con Python, pero R fue utilizado para tareas específicas como análisis de texto. El producto principal fue una base de datos limpia, con nuevas variables generadas a partir de las ya existentes y la realización de modelos de regresión con diferentes técnicas para la predicción de precios, esto último buscando hacer modelos que puedan ser utilizados nuevamente en cualquier momento del futuro.

3. Reflexión crítica y ética de la experiencia

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

3.1. Sensibilización ante las realidades

Rubén:

El comercio forma parte fundamental de la actividad económica del país, y la importación de artículos textiles tiene un peso que antes de cursar este PAP no era capaz de dimensionar; día a día los comerciantes buscan artículos de moda, baratos y de calidad para ofrecer y subsistir, muchas veces sin éxito pues la competencia es árdua y compañías grandes tienen en sus manos la mayor parte del mercado, pues cuentan con recursos, conocimiento privilegiado y experiencia así como equipos dedicados a obtener los mejores artículos. Creo que una aplicación ética y correcta de todo lo aprendido y obtenido en este proyecto es la democratización de la información, si bien esta información es pública, es difícil de encontrar y sacar provecho de ella sin conocimiento técnico, es por eso que creo que sería una excelente idea la creación de dashboards totalmente públicos, con información previamente procesada y precisa que pueda otorgar a cualquier persona interesada en importar las mejores opciones, en cuanto a precio, distancia y características para realizarlo y ser más competitivos en este sector.

Esteban:

El PAP despertó mi interés por trabajar en equipo y me resaltó la importancia de estar familiarizado con distintas plataformas de desarrollo de software y control de versiones de Git y además me permitió tener un acercamiento con una extensa base de datos gubernamental. También me permitió reconocer que en la práctica las características del material de trabajo, es decir, las bases de datos pueden no ser habituales, más sin embargo se debe estar preparado para lo cual es indispensable el trabajo en equipo y de forma remota para eficientar resultados.

Paola:

El análisis de datos es una herramienta que puede volverse muy poderosa en el futuro, al estar trabajando en este PAP me percaté que se puede sacar una enorme cantidad de información de este tipo de bases. En este caso en particular era una base de datos de precios y descripción de los artículos pero este mismo proceso puede aplicarse a una base de datos que contenga información sensible como una base de datos médica. Sin una legislación adecuada y una formación ética de los profesionales que realizan estos trabajos

podría haber afectaciones a las personas u organizaciones cuya información aparece en la base de datos.

Andrés:

Haber trabajado en este proyecto fue un parteaguas en mi forma de relacionar los aprendizajes que he adquirido durante el transcurso de la carrera con las funciones que se realizan en un ambiente profesional. Al inicio del PAP, consideraba que el campo principal donde eventualmente aplicaría mis conocimientos era necesariamente finanzas bursátiles. Sin embargo, ahora me doy cuenta que finanzas es sólo una de bastantes áreas donde mis habilidades pueden aportar valor. Trabajar con datos de comercio internacional para generar insights oportunos para el gobierno amplió mi perspectiva de oportunidades profesionales. Una reflexión que me tomó por sorpresa fue el hecho que nuestros productos de trabajo tienen la capacidad de mejorar la situación económica del país al identificar actividades con sospechas de lavado de dinero. Es un sentimiento gratificante saber que, utilizando sólo una computadora y dedicándole un poco de tiempo día con día, es posible tener un impacto positivo en la economía.

3.2. Aprendizajes logrados

Andrés:

Indudablemente, el mayor reto al cual me enfrenté fue la implementación de aprendizajes adquiridos en un entorno académico a un entorno profesional. Esta transición resultó personalmente retadora debido a las diferencias en las metodologías de trabajo. Un ejemplo de esto es el hecho que, durante el transcurso del PAP, cada uno de los participantes formó parte de un equipo con el mismo objetivo pero no necesariamente las mismas tareas por hacer. Fue la primera vez que trabajé en una estructura de equipo conformada por equipos de menor tamaño y diferentes responsabilidades. Debido a esto, resultó sumamente importante mantener un canal de comunicación constante con los demás grupos del equipo y estar al tanto de sus avances para poder así colaborar de manera eficiente. Ahora, puedo reconocer un desarrollo en mis capacidades de comunicación y organización, habilidades que me permiten ser un integrante de un equipo en un ambiente profesional.

Formar parte del proyecto representó una oportunidad de poner en práctica el autoaprendizaje. Si bien se iban definiendo los avances esperados de manera semanal, uno era responsable de obtener esos resultados sin contar con la asistencia total de un docente; como es el caso cuando se cursa una asignatura. Es claro que la habilidad de aprender de manera autodidacta es indispensable para el éxito de una carrera profesional. El momento en

el que uno obtiene su título universitario es cuando esta habilidad se vuelve aún más importante, puesto la responsabilidad de mantenerse actualizado en temas de su profesión recae en su misma persona. Me considero mejor preparado para dar ese paso después de haber ejercitado esta capacidad a lo largo del verano.

Por último, los entregables del proyecto requieren de un uso intensivo de herramientas de programación y análisis. El uso repetido de programación para realizar distintas tareas conlleva a una mejora en la capacidad de desarrollar algoritmos. Poder interpretar los resultados y hacer conclusiones con base en ellos también resultó un factor que impulsó mi entendimiento. Me quedo contento con el desarrollo que tuve en estas dos competencias específicamente.

Esteban:

Aprendí que la información de interés público puede resultar de gran utilidad tanto para su uso profesional como para proyectos personales y de emprendimiento. También, que está por lo general disponible en la web y no hace falta de recabar grandes y costosas bases de datos pero sí de saber utilizarlos eficientemente sin importar el rubro de estos. En este PAP se despertó bastante mi interés por trabajar para empresas que manejen grandes bases de datos y me motiva a aprovechar la facilidad y los beneficios del trabajo en equipo.

Rubén:

Considero que la experiencia del Proyecto de Aplicación Profesional es una de las mejores experiencias que he tenido durante mi estancia en la universidad, pues es una experiencia muy cercana a lo que nos podremos encontrar en el mercado laboral del mundo real. Durante el proyecto aprendí qué pasos seguir al momento de empezar por la parte más básica de un proceso profesional de ciencia de datos, la exploración y limpieza de datos con técnicas eficientes y automatizadas, así como los pasos siguientes como lo son la ingeniería de características y creación de diferentes modelos. Aprendí que hay que prestar mucha atención a los datos, esto significa no simplemente ver qué tipo de datos son, su rango, conocer algún estadístico, sino comprender de dónde viene el dato, qué quiere indicar y con qué otras variables se relaciona. Me he dado cuenta de lo que soy capaz y de mis puntos débiles; este proyecto me ha permitido darme cuenta que me apasiona este campo de aplicación de mi carrera.

Paola:

En la sesión informativa de este PAP se comentó que la forma de trabajo era principalmente autodidacta, se tenía una reunión semanal de dos horas en la cual se explicaba el trabajo realizado durante la semana y el trabajo a realizar para la próxima semana. Esta forma de trabajo fue el principal reto y también la principal ventaja de este proyecto, tener esa libertad de no trabajar en un horario predefinido requiere de mucha autogestión para dedicarle el tiempo necesario y lograr el mejor aprovechamiento del proyecto. Debido a que usualmente trabajábamos de forma independiente el saber buscar la información en las fuentes adecuadas fue clave para el desarrollo de los entregables. También tuve que investigar cómo usar librerías de los lenguajes de programación que nunca había usado antes, definitivamente fue un reto debido a que no siempre funcionaban como uno espera y fueron momentos de estar prueba y error hasta que encontráramos la solución, sin embargo una vez que la solución era encontrada seguía un sentimiento muy agradable de superación. Finalmente, este proyecto me hizo reafirmar que cuando egrese de la universidad me gustaría trabajar en el área de ciencia de datos, creo que es un área muy interesante que está llena de retos y podría ser muy gratificante para mí dedicarme a ella.