Universidad Tecnológica de Panamá
Facultad de Ingeniería de Sistemas Computacionales
Licenciatura en Ingeniería de Software

Estadística con apoyo informático

Proyecto Final

Tema:
Análisis estadístico de bases de datos

Facilitador:
Ing. Juan Castillo, PhD

Nombre:
Yinela Bryant (3-749-2108)

Grupo:
1SF-131

Fecha de entrega:
26 de julio de 2023

Contenido

Introducción 3
Contenido 4
Victima Accidentes de Tránsito en Panamá (provincia)(BD4) 4
Importaciones por zona franca(BD5) 9
Cereales (BD9) 18
Cámaras (BD10) 25
Indicadores Económicos-Clasificación de Importaciones según el Uso o Destino Económico de los Bienes (BD6) 31
PIB anual por categoria economica (BD7) 37
Importaciones anuales (BD8) 42
Accidentes automovilísticos (conductores)(BD1) 49
Exportaciones por aranceles (BD2) 58
Exportaciones por países (BD3) 65
Conclusión 72
Anexo 73

Introducción

En este proyecto final de estadística, se llevarán a cabo diversos análisis y técnicas para comprender mejor un conjunto de datos. Se explorarán diferentes herramientas y conceptos estadísticos con el objetivo de obtener una visión detallada de la información presente en los datos.

Inicialmente, se utilizará la visualización de histogramas para examinar la distribución de los datos y obtener una idea general de su forma y dispersión. Además, se aplicará la técnica del diagrama de Pareto para priorizar y enfocar en los elementos más relevantes del conjunto de datos, resaltando las categorías o variables que contribuyen en mayor medida.

Posteriormente, se utilizarán pruebas estadísticas como la prueba t de dos muestras y la prueba de chi-cuadrado para examinar diferencias en medias y frecuencias observadas, respectivamente.

Además, se realizará un análisis de regresión lineal para investigar las relaciones entre variables y predecir valores futuros. Se calculará la pendiente e intercepto de la línea de regresión, así como el coeficiente de correlación para evaluar la fuerza y la dirección de la relación. Este análisis permitirá comprender mejor las interacciones entre las variables y su influencia en el fenómeno estudiado.

Finalmente, se calcularán medidas de centralidad y dispersión, como la media, varianza y desviación estándar, para caracterizar y resumir los datos.

En resumen, este proyecto explorará diferentes técnicas estadísticas y herramientas de visualización para comprender en profundidad un conjunto de datos. Los histogramas, diagramas de Pareto, pruebas estadísticas, análisis de regresión lineal y cálculo de medidas descriptivas jugarán un papel fundamental en la exploración y el análisis de los datos, brindando información valiosa para tomar decisiones informadas y obtener conclusiones fundamentadas.

Contenido

Victima Accidentes de Tránsito en Panamá (provincia)(BD4)

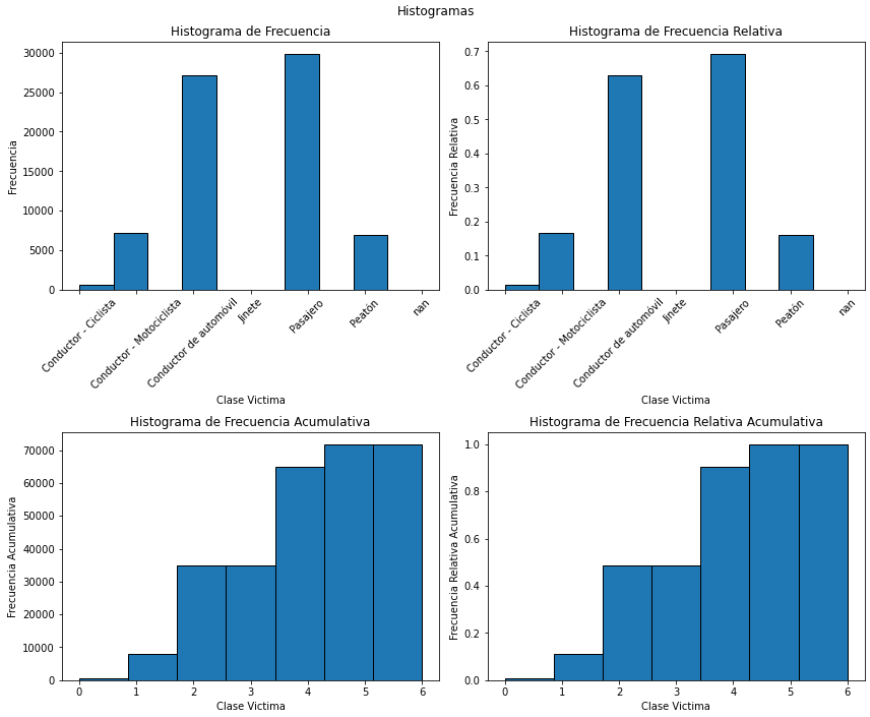
Composición del documento

- Año: valor categórico (cadena de números)
- Clase Victima: valor categórico (cadena de letras)
- Condicion Victima: valor categórico (cadena de letras)
- Sexo: valor categórico (cadena de letras)
- Edad: valor numérico

Plan de análisis

Analizaré la distribución de las víctimas según la clase (por ejemplo, conductor, pasajero, peatón) y la condición (por ejemplo, lesionado leve, lesionado grave, fallecido), si hay patrones o tendencias en relación con estas variables.
Además, exploraré posibles correlaciones entre variables, como la edad de la víctima y la gravedad de la condición, o la clase de víctima y el sexo.

Histograma de frecuencia de la columna "Clase de víctimas"



Observando este gráfico, podemos observar que los pasajeros son las víctimas de accidentes de tránsito más frecuentes.

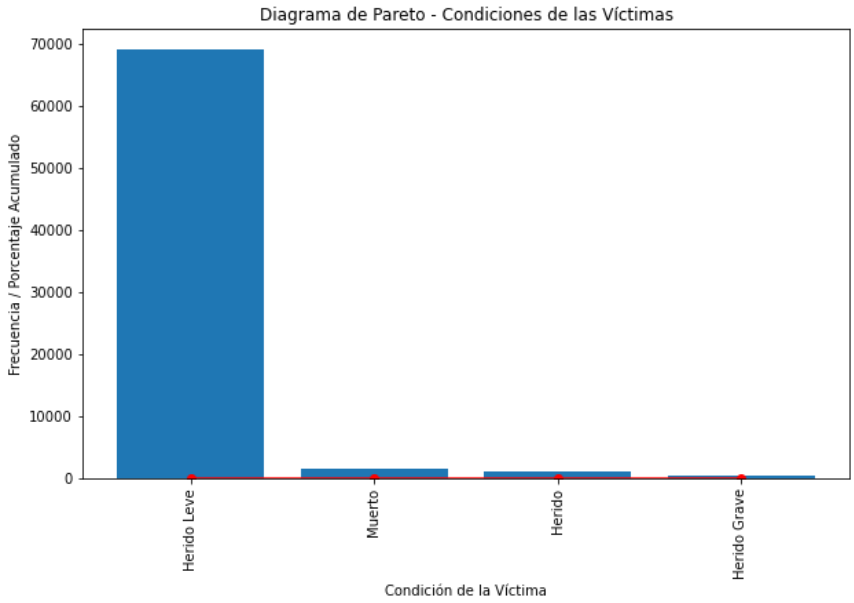
Medidas de dispersión de la columna "Edad"

Media de Edad: 33.495721811150695

Varianza de Edad: 237.44387719171385

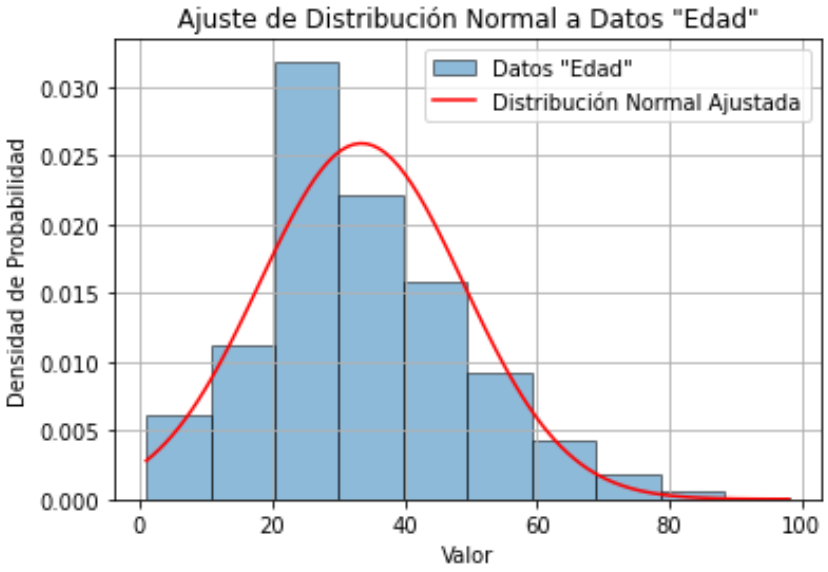
Desviación Estándar de Edad: 15.409214035495577

Diagrama de Pareto de la columna "Condiciones de las víctimas"



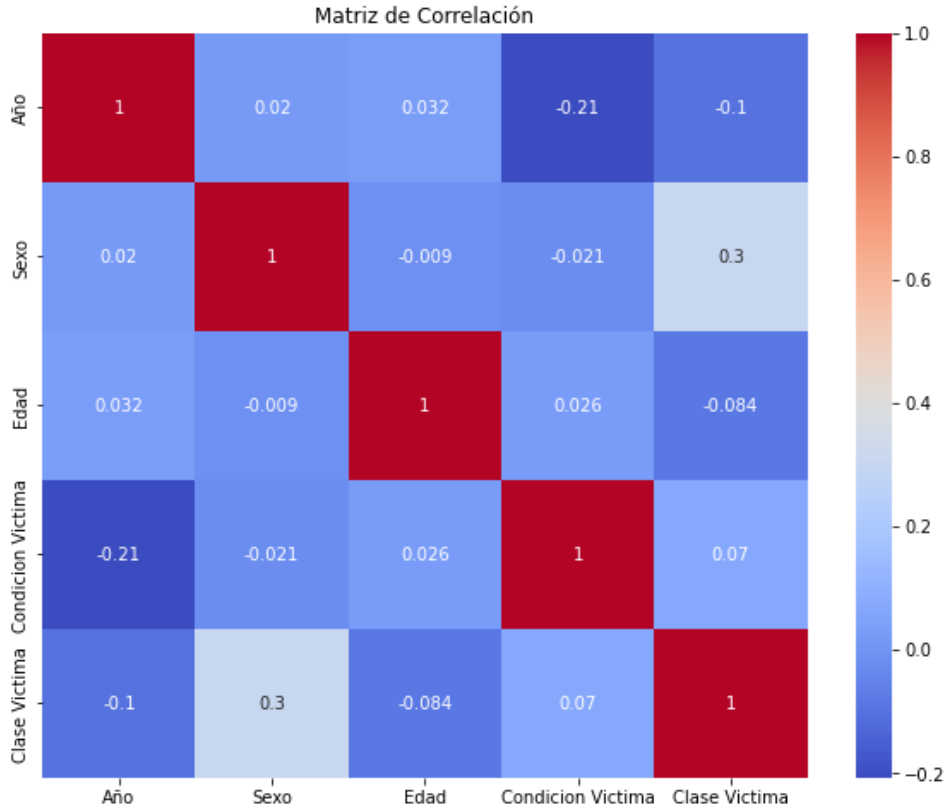
Este gráfico nos muestra que la diferencia entre la condición más frecuente es significativa, en este caso la mayoría de los accidentes registrados en el conjunto de datos son Heridos Leves.

Análisis de Conversión de distribución discreta a continua



El ajuste a la distribución normal se hizo utilizando máxima verosimilitud, como se ve en el gráfico, la técnica fue efectiva en la mayor parte.

Análisis de correlación



El dato más sorprendente de este gráfico es el hecho de que las variables sexo y clase de victima estén mucho más relacionadas entre sí que las demás.

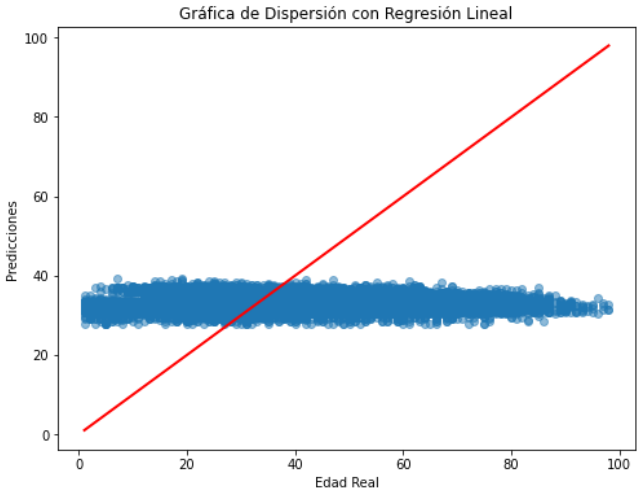
Análisis de regresión lineal

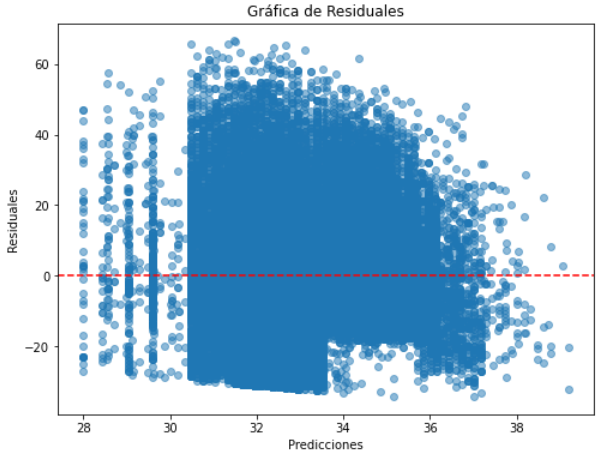
OLS Regression Results

==========

Dep. Variable: Edad R-squared: 0.009

(Resultados completos presentados en el documento con el código en Github)





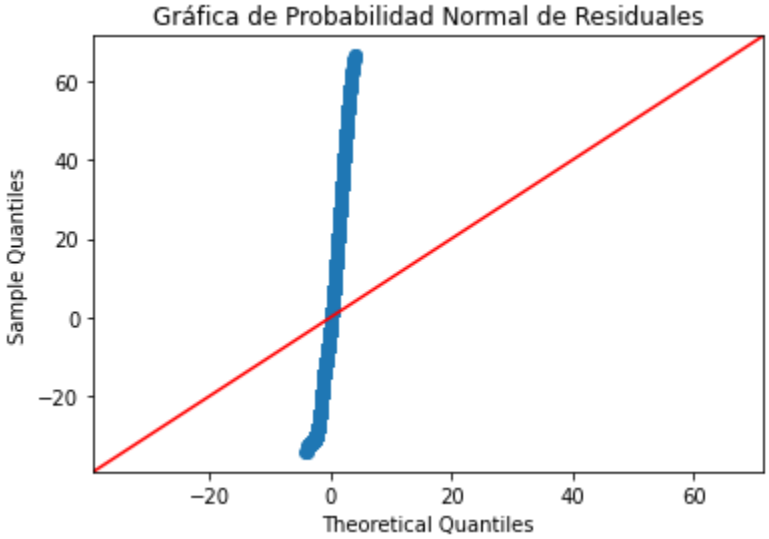


Tabla ANOVA:
<F test: F=153.87478853598662, p=2.844566314213748e-131, df_denom=6.57e+04, df_num=4>

R-cuadrado: 0.009284597231914438

Intervalos de confianza para coeficientes:

	0	1
const	30.724172	32.532332
Año	0.107797	0.182856
Sexo	0.316436	0.818333
Condicion Victima	1.634791	2.454399
Clase Victima	-1.144331	-0.954497

Valores de las constantes de la línea de regresión:
Intercept: 31.628252155270648
Slope: Año 0.145327
Sexo 0.567385
Condicion Victima 2.044595
Clase Victima -1.049414

El ajuste del modelo tiene un R-cuadrado de 0.009, lo que indica que solo alrededor del 0.9% de la variabilidad en la edad se explica por las variables independientes incluidas en el modelo. Esto sugiere que el modelo tiene un ajuste insuficiente para explicar las variaciones observadas en la edad.

Sin embargo, los resultados del modelo indican que las variables independientes tienen una influencia colectiva significativa en la edad, como se muestra en los intervalos de confianza para los coeficientes. Por ejemplo, la variable "Año" tiene un coeficiente estimado de 0.1453, lo que significa que, en promedio, por cada unidad adicional en "Año", la edad aumenta en 0.1453 unidades, manteniendo constantes las otras variables.

Asimismo, el coeficiente para la variable "Sexo" es de 0.5674, lo que sugiere que, en promedio, las personas de género masculino tienen una edad más alta que las personas de género femenino, manteniendo constantes las demás variables.

El coeficiente para la variable "Condicion Victima" es de 2.0446, lo que indica que las personas con ciertas condiciones de víctimas tienen una edad más alta que las que no tienen estas condiciones, manteniendo constantes las demás variables.

Por otro lado, el coeficiente para la variable "Clase Victima" es de -1.0494, lo que sugiere que ciertas clases de víctimas se asocian con una edad más baja en comparación con otras clases, manteniendo constantes las demás variables.

Importaciones por zona franca(BD5)

Composición del documento

El documento está formado por las siguientes columnas:

- Años: valor categórico (cadena de números)
- Continente: valor categórico (cadena de letras)
- Mes: valor categórico (cadena de letras)
- País: valor categórico (cadena de letras)
- Puerto de desembarque:valor categórico (cadena de letras)
- Vía: valor categórico (cadena de letras)
- Zona Franca: valor categórico (cadena de letras)
- Arancel:valor categórico (cadena de números)
- Año: valor categórico (cadena de números)
- MES: valor categórico (número)
- PaisCodigo:valor categórico (cadena de números)
- VIA: valor categórico (cadena de números)
- Zonas Francas: valor categórico (cadena de letras)
- Valor CIF: valor numérico

Plan de análisis

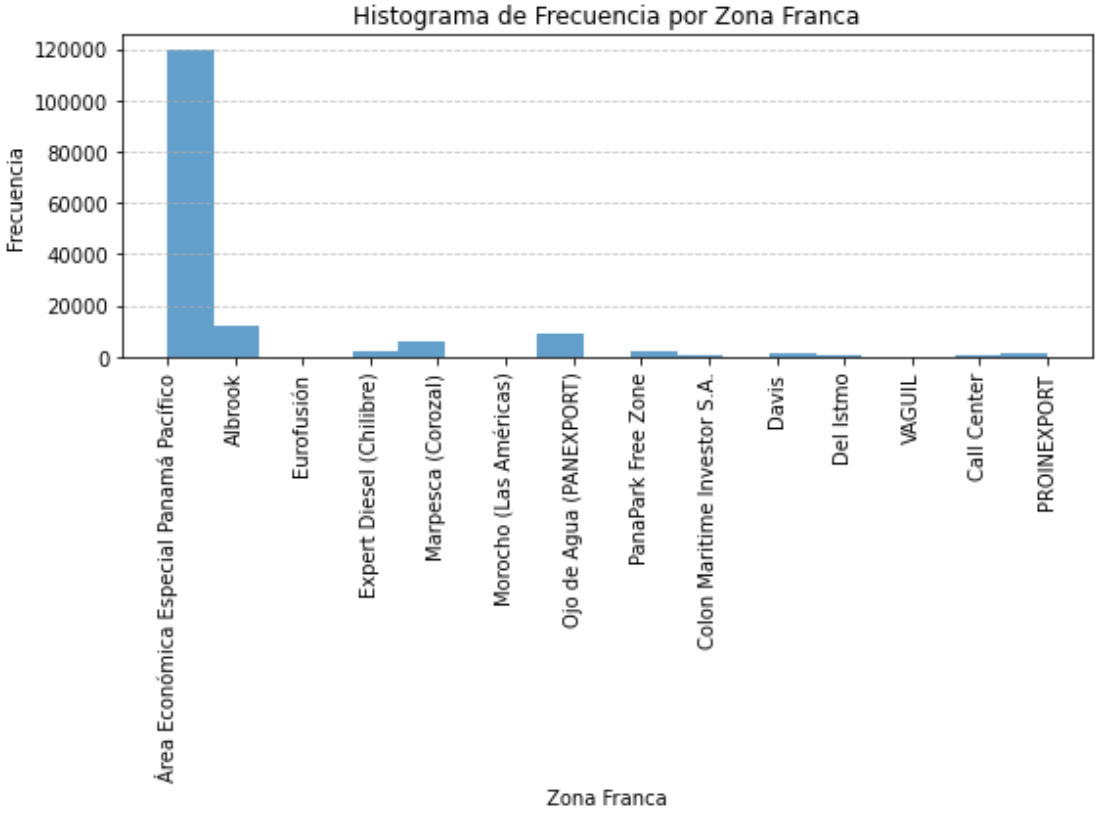
Examinaré las importaciones realizadas a través de diferentes zonas francas. Analizaré qué zonas francas son las más activas en términos de importaciones y cómo ha variado su participación a lo largo de los años y meses. Además,investigaré cómo los aranceles aplicados han afectado las importaciones en las zonas francas.

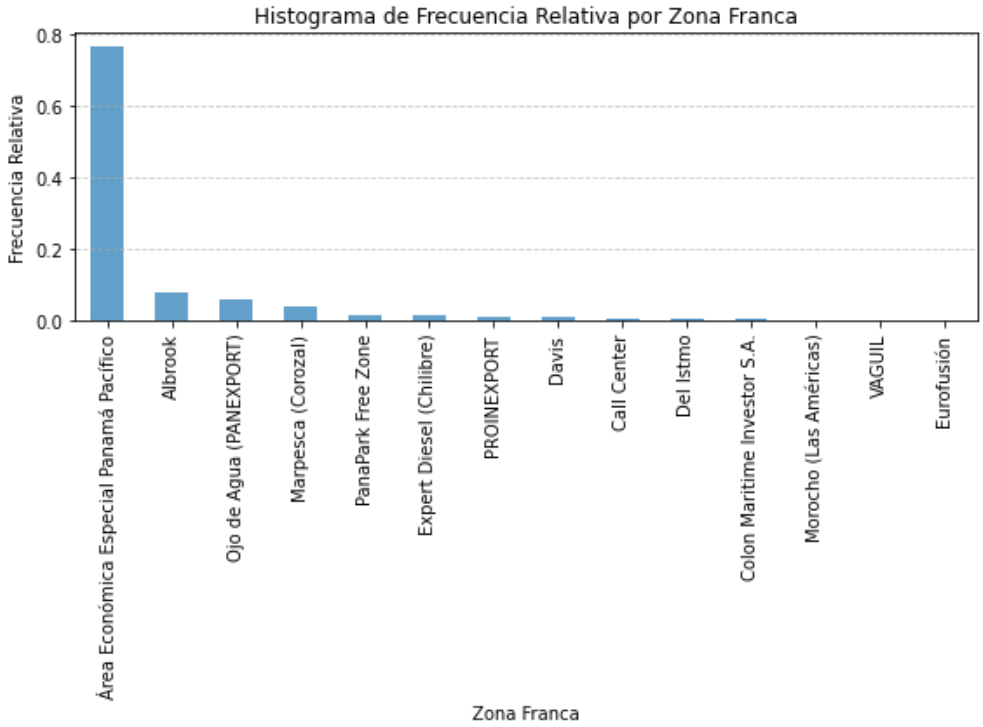
Medidas de variación de la columna Valor CIF

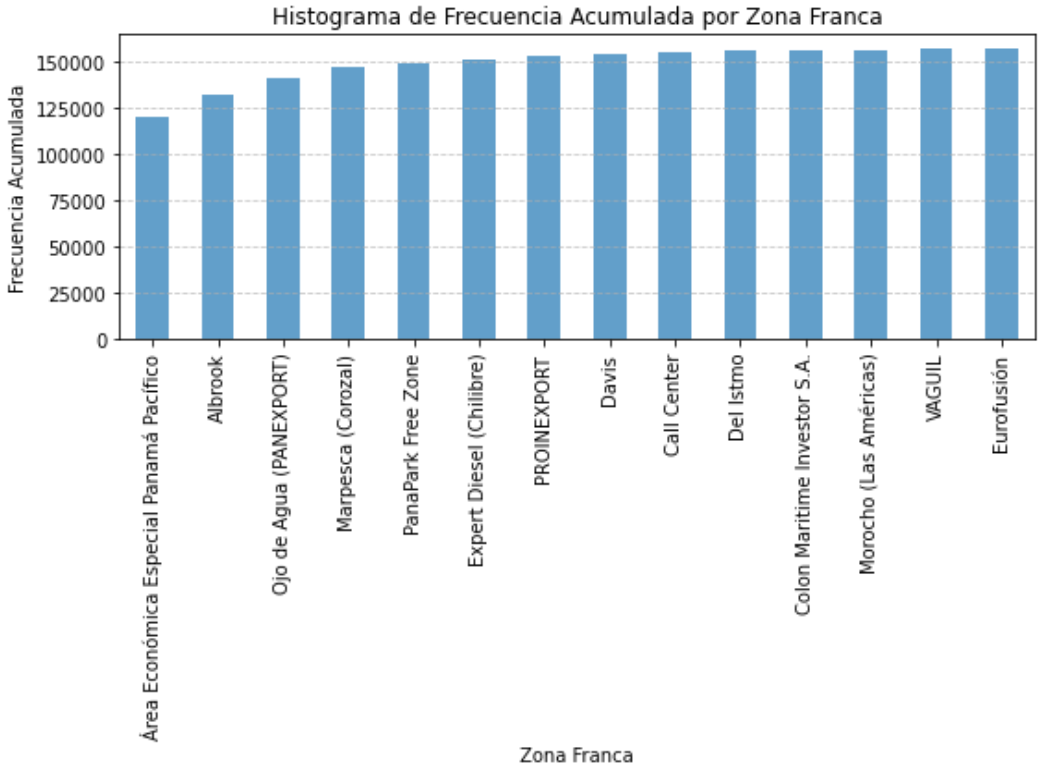
Media: 618096201.8571428

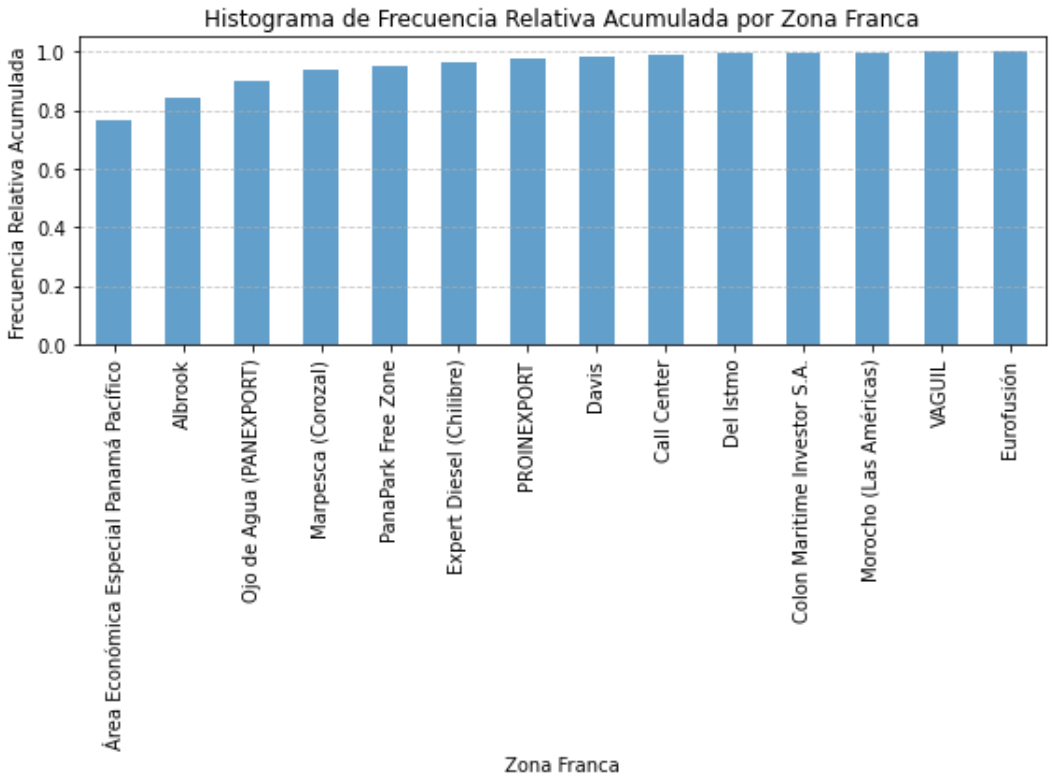
Varianza: 3.7578419980780805e+18

Desviación Estándar: 1938515410.843587

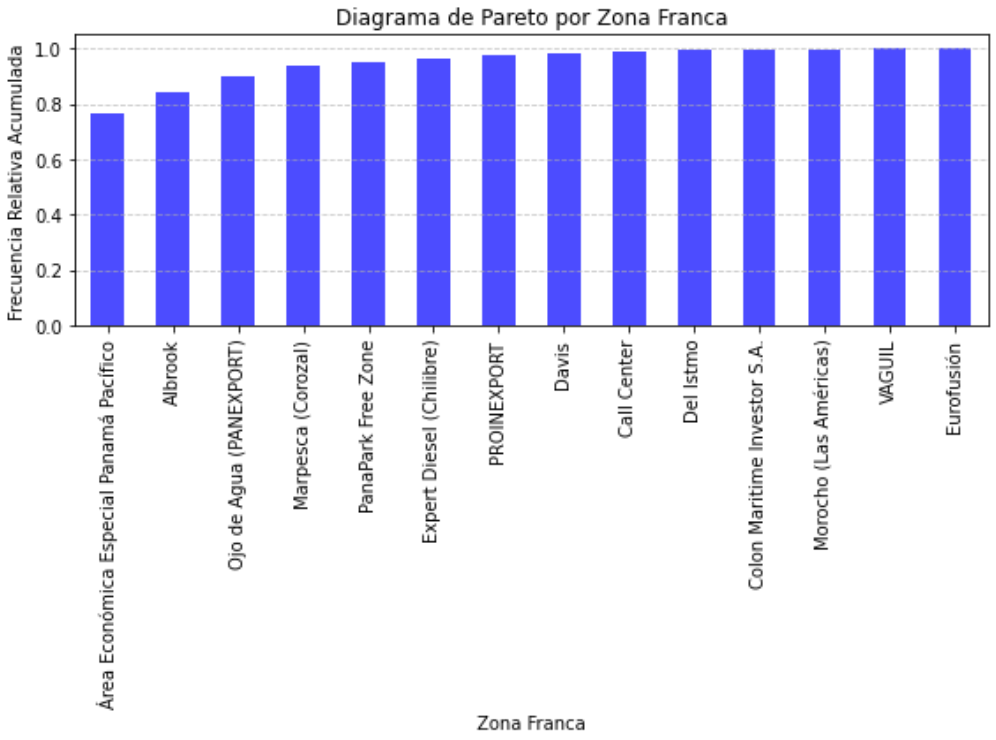

Histograma de Frecuencia por Zona Franca
Frecuencia
120000
100000
80000
60000
40000
20000
0
Área Económica Especial Panamá Pacífico
Albrook
Eurofusión
Expert Diesel (Chilibre)
Marpesca (Corozal)
Morocho (Las Américas)
Ojo de Agua (PANEXPORT)
PanaPark Free Zone
Colon Maritime Investor S.A.
Davis
Del Istmo
VAGUIL
Call Center
PROINEXPORT
Zona Franca


Histograma de Frecuencia Relativa por Zona Franca
Frecuencia Relativa
0.8
0.6
0.4
0.2
0.0
Área Económica Especial Panamá Pacífico
Albrook
Ojo de Agua (PANEXPORT)
Marpesca (Corozal)
PanaPark Free Zone
Expert Diesel (Chilibre)
PROINEXPORT
Davis
Call Center
Del Istmo
Colon Maritime Investor S.A.
Morocho (Las Américas)
VAGUIL
Eurofusión
Zona Franca

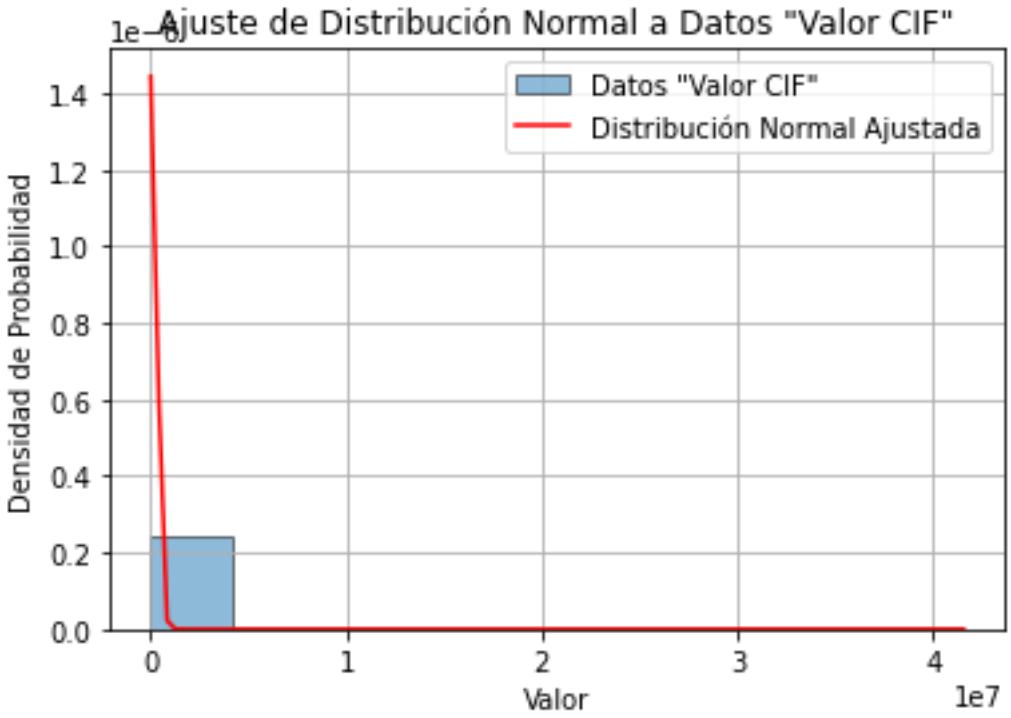




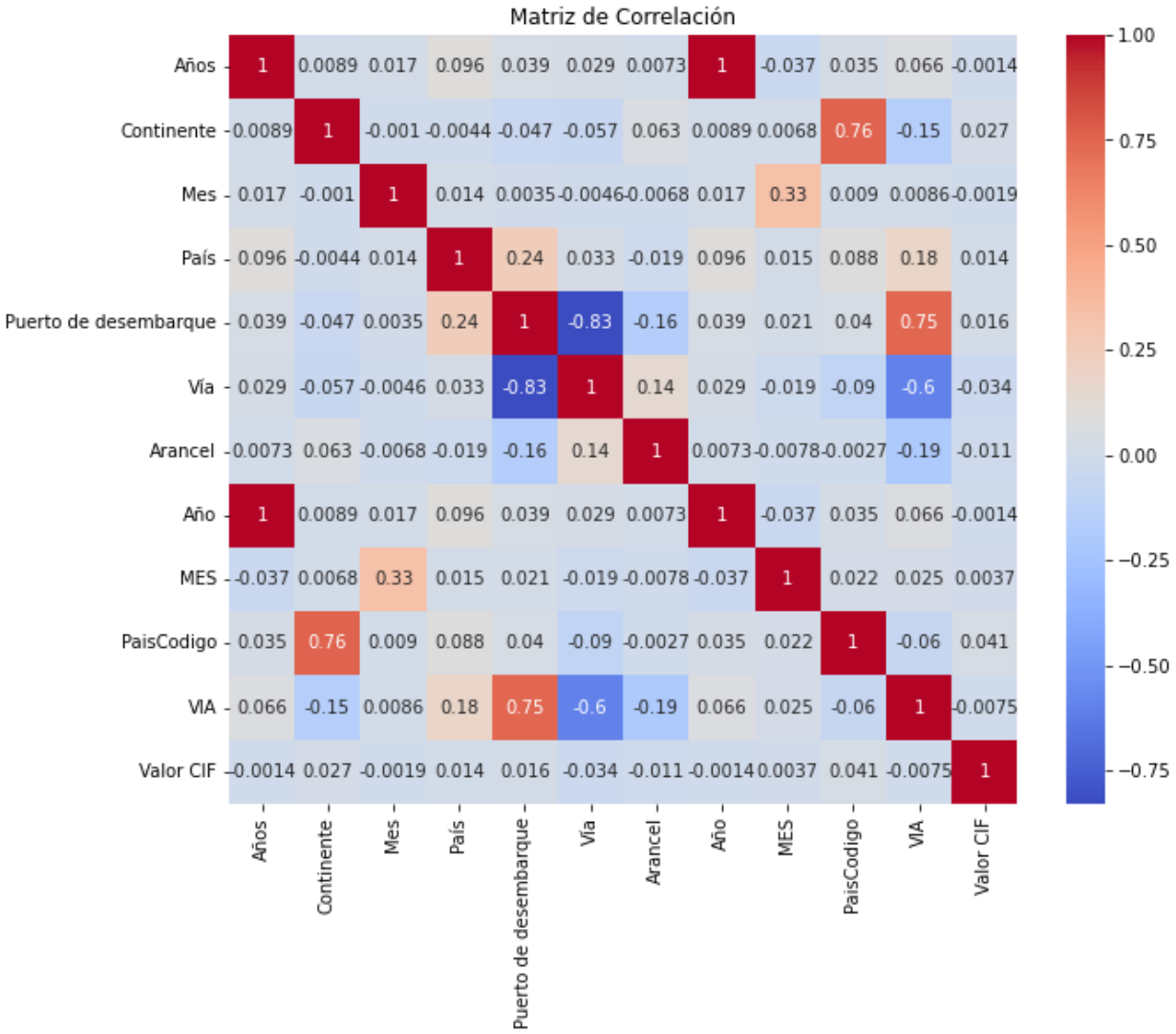
Los histogramas muestran que la zona franca más común es la de Panamá Pacífico, al momento de colocar estos datos en contexto, tendría completo sentido debido al auge que ha tenido esa zona en los últimos años.



Análisis de Conversión de distribución discreta a distribución continua



Análisis de Correlación



Según el gráfico, las variables que tienen una relación más alta es debido a la forma en la que está estructurada la base de datos.

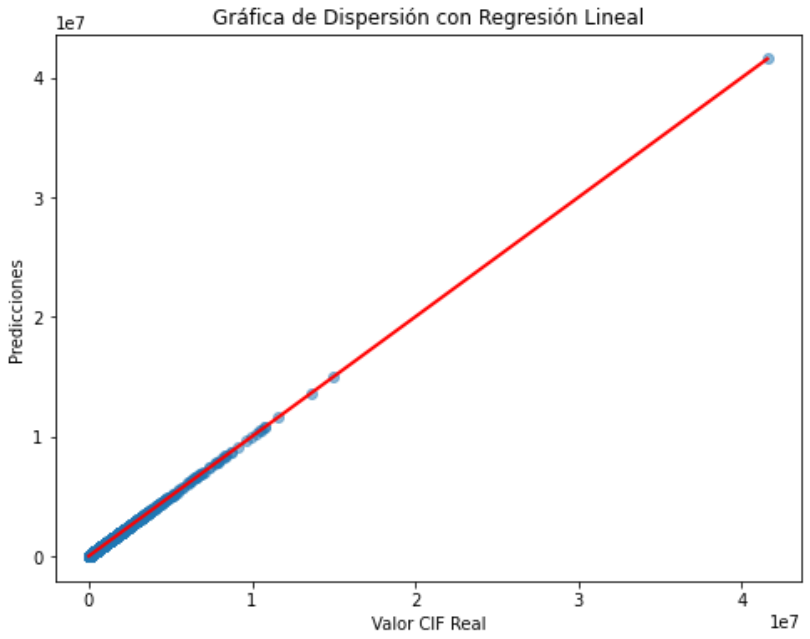
Análisis de Regresión Lineal

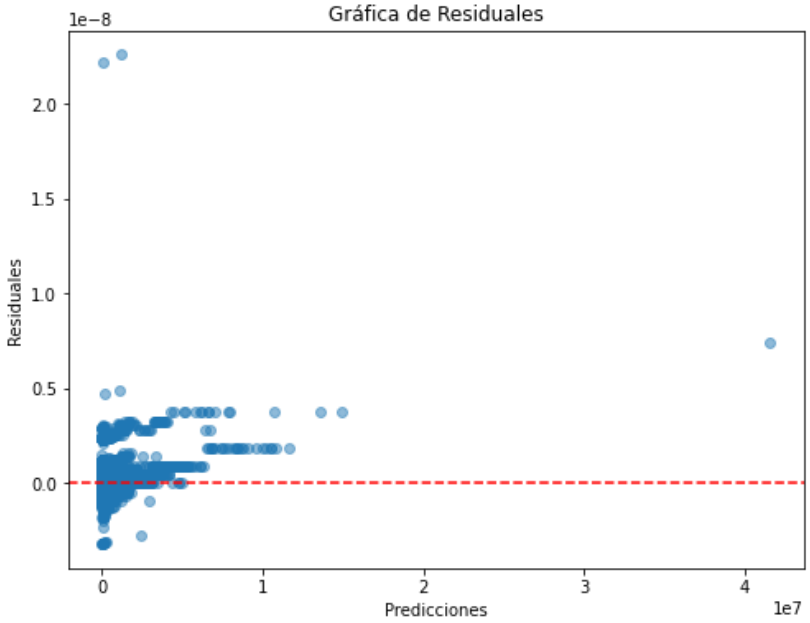
OLS Regression Results

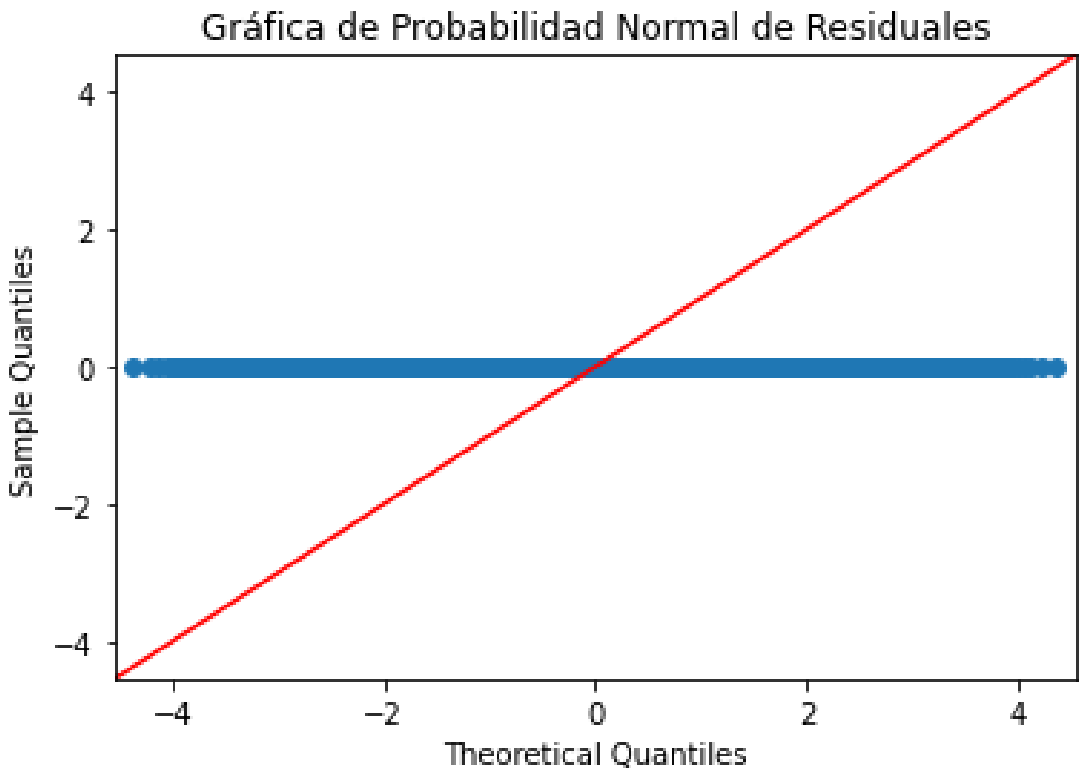
==========

Dep. Variable:	Valor CIF	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	7.649e+32
Date:	Mon, 24 Jul 2023	Prob (F-statistic):	0.00

Time: 20:25:37 Log-Likelihood: 3.2141e+06







Análisis de Varianza (ANOVA):

Valor F: 7.563511914350016e+32

P-valor (F): 0.0

Resumen de los resultados obtenidos del Análisis de Varianza (ANOVA) y pruebas de t:

- Valor F: 7.563511914350016e+32

- P-valor (F): 0.0

Pruebas de t y valores de P para los coeficientes individuales:

- Para el coeficiente c0 (constante):

 - Valor del coeficiente:-1.414e-10

 - Prueba de t:-5.182

 - Valor p: 0.000

- Para el coeficiente c0 (constante):

- Valor del coeficiente: 1.0000

- Prueba de t: 3.62e+17

- Valor p: 0.000

Intervalos de confianza para los coeficientes:

- Para la constante (c0):

- Límite inferior:-1.949137e-10

- Límite superior:-8.793920e-11

- Para el coeficiente Valor CIF:

- Límite inferior: 1.000000e+00

- Límite superior: 1.000000e+00

- Para los coeficientes de las variables dummy de País y VIA:

- Se muestra una tabla con los intervalos de confianza para cada variable dummy.

Valores de las constantes de la línea de regresión:

- Intercept:-1.4142642612569034e-10

- Slope (Valor CIF): 1.000000e+00

- Slope (Variables dummy de País y VIA): Se muestran los valores de los coeficientes para cada variable dummy. (lista completa en el código)

Estos resultados de Regresión OLS (Mínimos Cuadrados Ordinarios) muestran que el modelo tiene un ajuste casi perfecto a los datos, ya que el coeficiente de determinación R-cuadrado es 1.000.

Esto significa que el modelo explica el 100% de la variabilidad de la variable dependiente (Valor CIF) en función de las variables independientes incluidas en el modelo.

Además, el estadístico F es extremadamente alto, lo que indica que el modelo en su conjunto es altamente significativo. El valor p asociado con el estadístico F también es muy bajo (prácticamente cero), lo que sugiere que el modelo es altamente confiable y que la probabilidad de que los resultados sean aleatorios es prácticamente nula.

Sin embargo, al analizar los coeficientes de las variables independientes (países), algunos tienen valores muy pequeños, lo que sugiere que pueden no ser estadísticamente significativos para explicar la variabilidad de la variable dependiente. También hay algunos países con valores p elevados, lo que indica que no son estadísticamente significativos para el modelo.

Cereales (BD9)

Composición del documento

- Nombre: valor categórico (cadena de letras)
- mfr:valor categórico (cadena de letras)
- tipo: valor categórico (cadena de letras)
- calorías: valor numérico
- proteína: valor numérico
- grasa: valor numérico
- sodio: valor numérico
- fibra: valor numérico
- carbo: valor numérico
- azúcares: valor numérico
- potass: valor numérico
- vitaminas: valor numérico
- estante: valor categórico (cadena de números)
- peso: valor numérico
- tazas: valor numérico
- calificación: valor numérico

Plan de análisis

Para mi análisis estadístico, me gustaría comparar las características nutricionales de los cereales según el fabricante (mfr). Para ello, analizaré la media de calorías, proteína, grasa y otras variables para cada fabricante y determinaré si existen diferencias significativas entre ellos.

Además, exploraré la relación entre las variables nutricionales, como calorías, proteína, grasa, sodio y fibra. Utilizaré gráficos de dispersión o matrices de correlación para identificar posibles relaciones y determinaré si existe una asociación significativa entre estas variables.

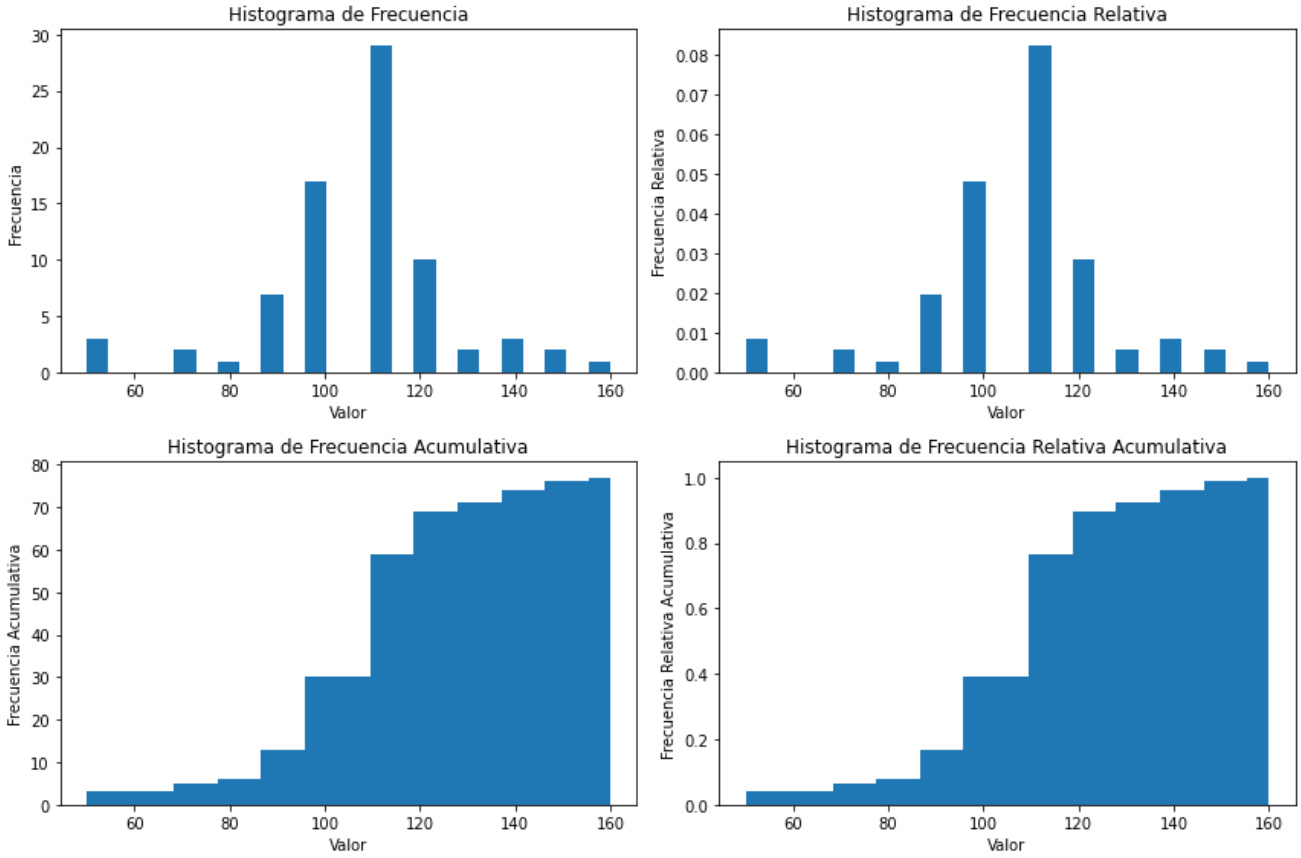
Estos análisis me permitirán comprender mejor las diferencias nutricionales entre los cereales fabricados por distintas empresas y evaluar si hay alguna relación significativa entre las variables nutricionales clave. De esta manera, podré obtener una visión más completa de la calidad nutricional de los cereales en función del fabricante y explorar posibles vínculos entre los diferentes nutrientes presentes en los cereales.

Estadísticas descriptivas de las columnas numéricas

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	106.883117	2.545455	1.012987	159.675325	2.151948	14.597403	6.922078	96.077922	28.246753	2.207792	1.029610	0.821039	42.665705
std	19.484119	1.094790	1.006473	83.832295	2.383364	4.278956	4.444885	71.286813	22.342523	0.832524	0.150477	0.232716	14.047289

Varianza de calorías, proteína y grasas

379.63089542036903, 1.1985645933014353, 1.0129870129870127



Histogramas de la columna calorías, podemos observar que la mayoría de los cereales están entre las 100 y 120 calorías. Necesitaríamos conocer el volumen por servida si quisiéramos tomar decisiones como cuál es el más saludable.

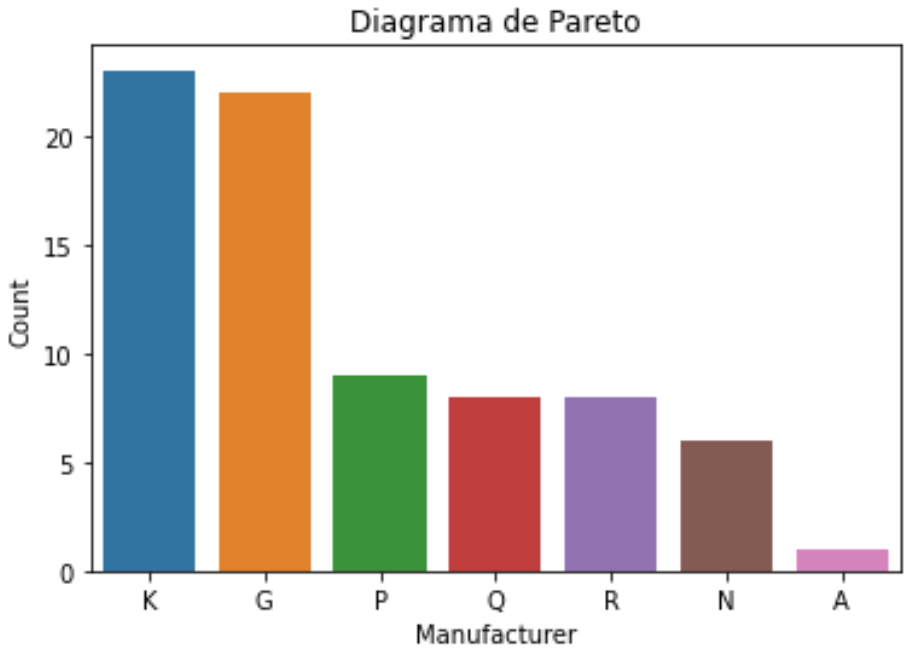
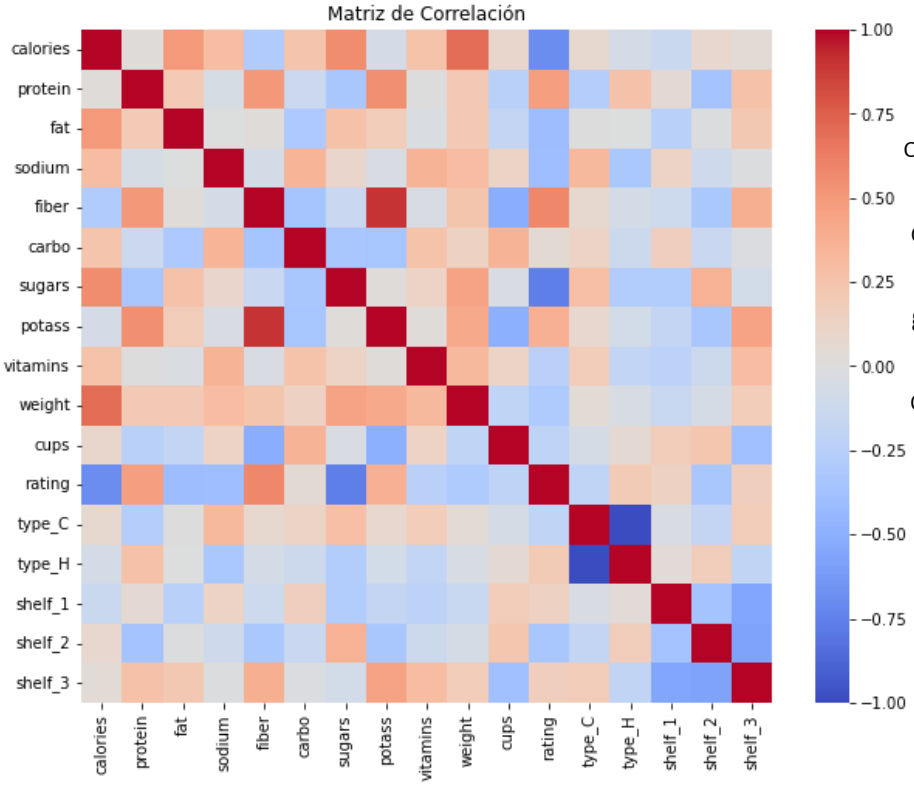


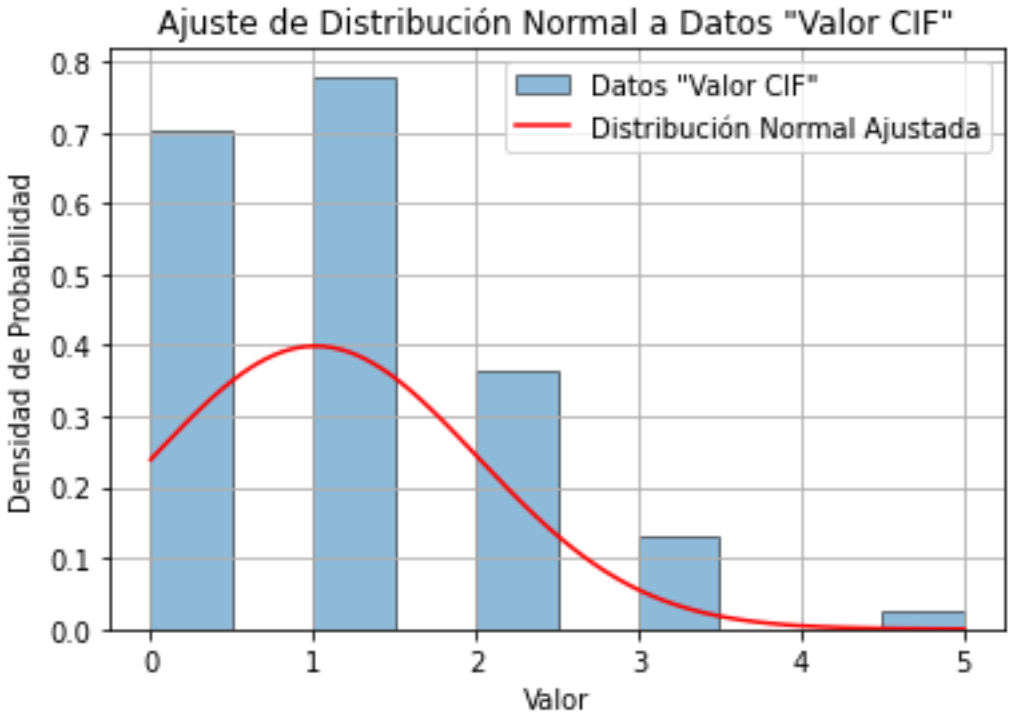
Diagrama de pareto de la frecuencia de aparición de las empresas de manufactura, podemos observar que el más frecuente es K que representa Kelloggs, como sabemos Kellogs es una de las empresas líder en la producción de cereales.

Análisis de correlación



En este gráfico podemos observar la relación entre los nutrientes de los cereales, por ejemplo, que el porcentaje de grasa está muy relacionado con la cantidad de calorías.

Análisis de Conversión de distribución discreta a distribución continua



Se hizo el ajuste de la distribución normal utilizando máxima verosimilitud, el gráfico no muestra una distribución normal, esto puede ser debido a que el valor CIF está compuesto por varios elementos, como el costo de la mercancía, los gastos de seguro y los gastos de transporte. Cada uno de estos componentes puede tener una distribución diferente, lo que contribuye a que el valor CIF en sí mismo sea una mezcla de distribuciones, en lugar de seguir una distribución normal única.

Análisis de Regresión Lineal

Intercepto: 1.69636257416548

Pendientes:

calories -0.008735

protein -0.065039

fat 0.310630

sodium -0.002228

fiber 0.052702

carbo 0.045898

sugars 0.029334

potass 0.002912

vitamins 0.013643

OLS Regression Results

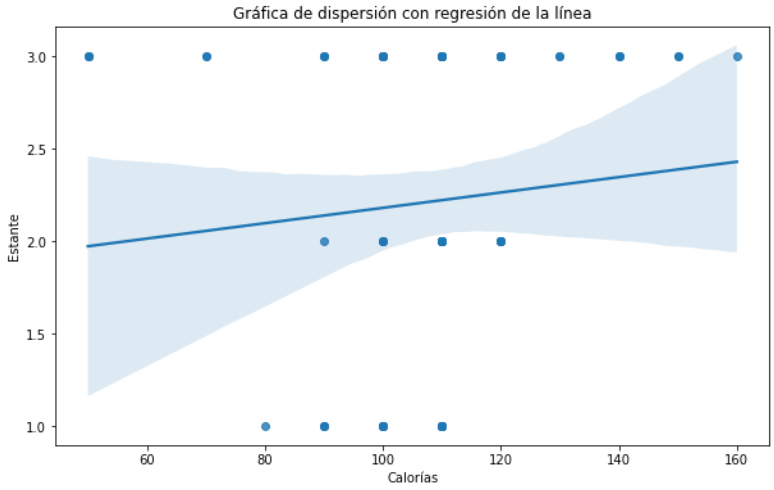
==========

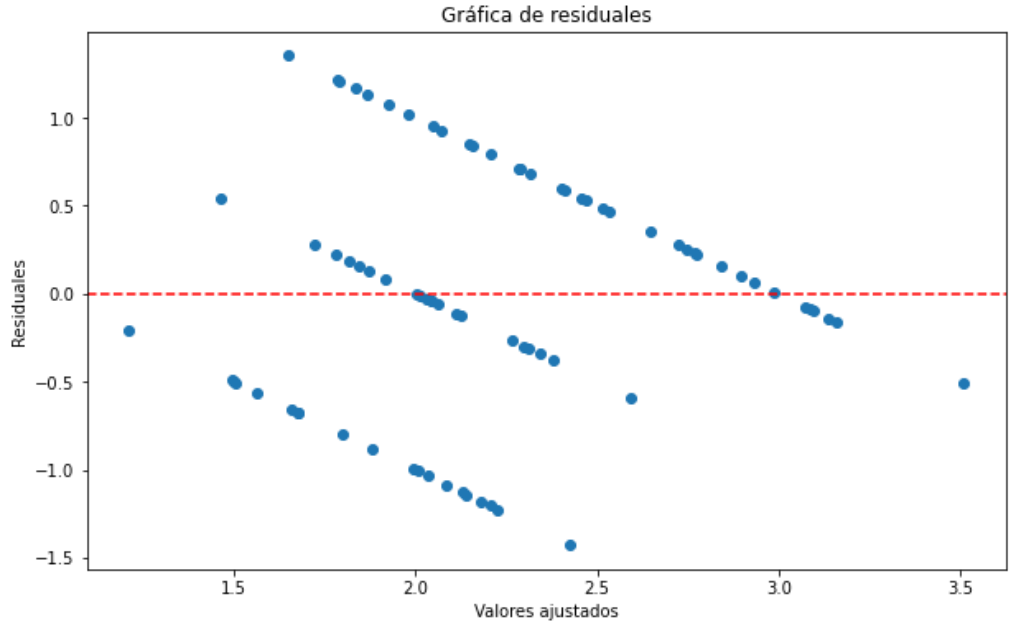
Dep. Variable: shelf R-squared: 0.319

Model: OLS Adj. R-squared: 0.228

Method: Least Squares F-statistic: 3.494

==========





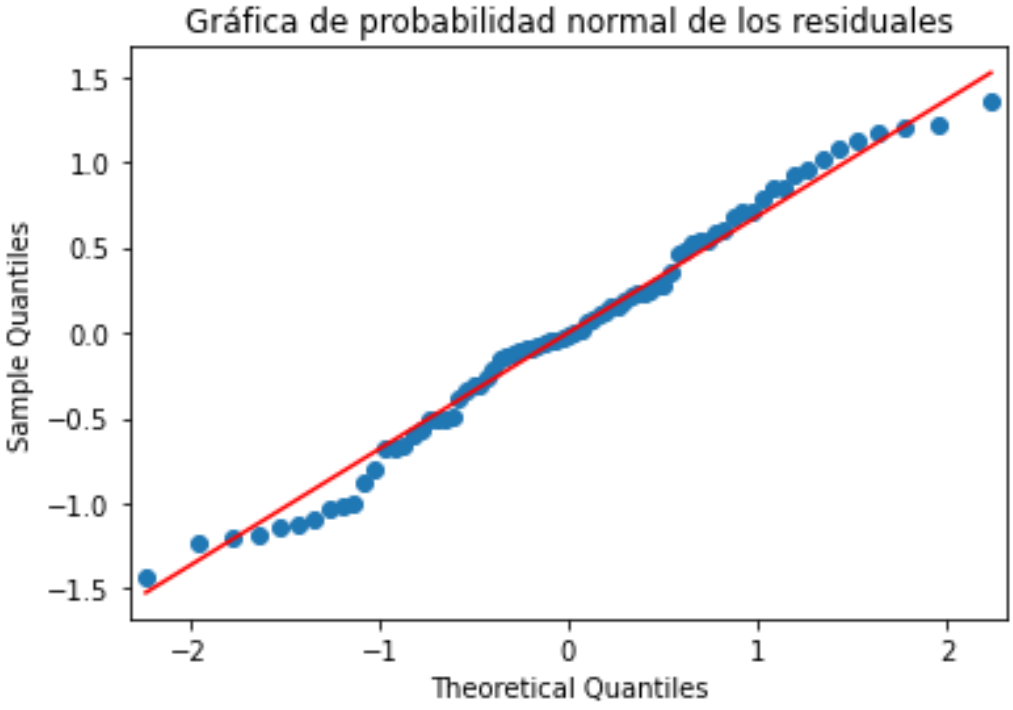


Tabla ANOVA para coeficientes:

<F test: F=3.494431118810747, p=0.0013611112998494234, df_denom=67, df_num=9>

R-cuadrado: 0.31945066918476506

El modelo de regresión lineal tiene un R-cuadrado de 0.319, lo que significa que aproximadamente el 31.9% de la variabilidad en la variable dependiente "estante" puede ser explicada por las variables independientes.

La tabla ANOVA muestra que el modelo de regresión es significativo (Prob (F-statistic): 0.00136), lo que indica que al menos una de las variables independientes tiene un efecto significativo en la variable dependiente.

Las pruebas de t y los valores de P para cada coeficiente individual muestran que solo las variables "fat" y "vitamins" tienen un efecto significativo en el estante, ya que sus valores P son menores que 0.05. Las demás variables no parecen tener un efecto significativo.

Los intervalos de confianza para los coeficientes muestran el rango de valores dentro del cual se espera que estén los coeficientes con un 95% de confianza.

Cámaras (BD10)

Composición del documento

- Modelo: valor categórico (cadena de letras)
- Fecha de lanzamiento: valor categórico (cadena de números)
- Resolución máxima: valor numérico
- Resolución mínima: valor numérico
- Píxeles efectivos:valor numérico
- Zoom angular (W): valor numérico
- Zoom telefoto (T): valor numérico
- Rango de enfoque normal: valor numérico
- Rango de enfoque macro: valor numérico
- Almacenamiento incluido: valor numérico
- Peso (incluidas las baterías): valor numérico
- Dimensiones: valor numérico
- Precio: valor numérico

Plan de análisis

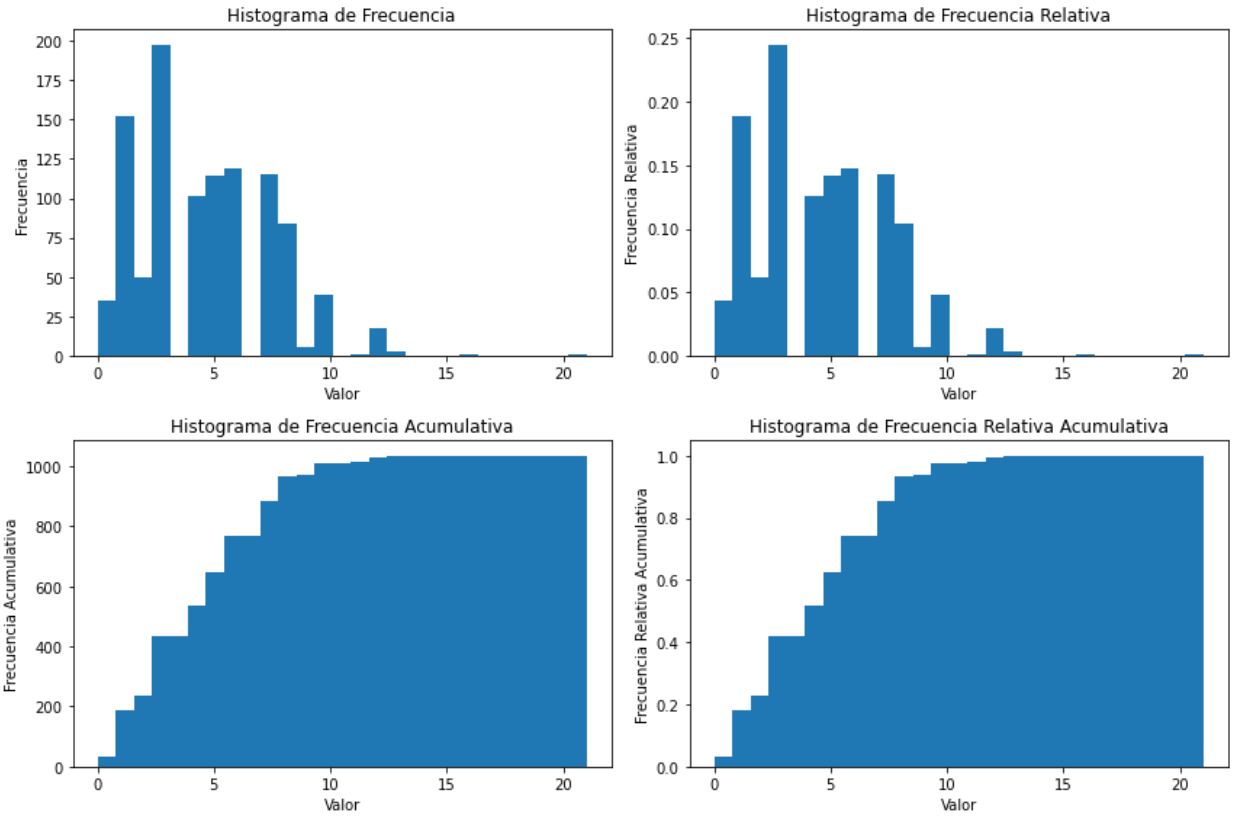
Investigar la relación entre el zoom angular y el zoom telefoto de las cámaras. Puedo calcular la diferencia entre los valores de zoom para cada modelo y determinar si existe una correlación significativa entre ellos. Comparar el peso y las dimensiones de las cámaras. Puedo calcular estadísticas descriptivas y visualizar estas características utilizando gráficos de dispersión o diagramas de caja para identificar posibles tendencias o patrones. Evaluar la capacidad de almacenamiento incluida en las cámaras. Puedo calcular la media y la mediana del almacenamiento y comparar estas medidas entre diferentes modelos para determinar si hay diferencias significativas.

Medidas de dispersión de la columna Zoom Wide (W)

Media: 32.955598455598455

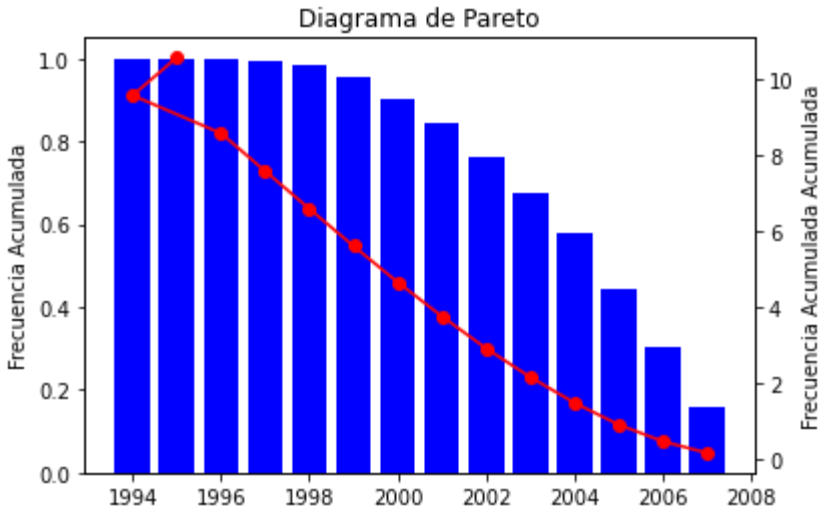
Varianza: 106.9410217671089

Desviacion estandar: 10.341229219348582

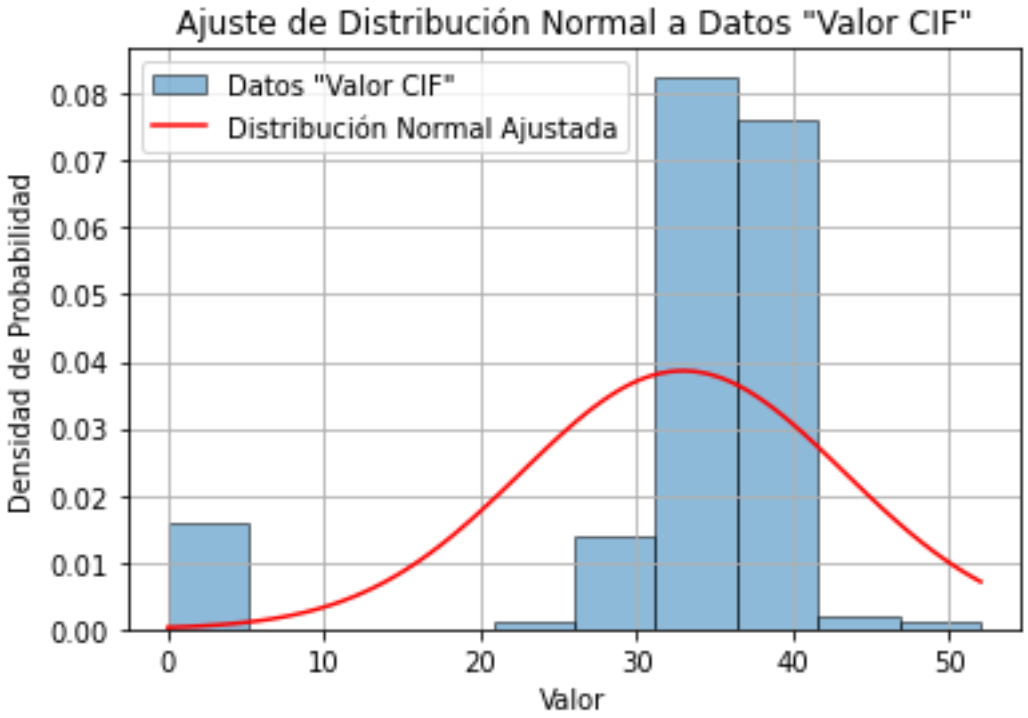


Histogramas de frecuencia de la columna de pixeles, la mayoría de las cámaras se encuentran entre 3 y 5 pixeles, estos valores son esperados debido a que los datos son de los años 90.

Diagrama de pareto de frecuencia de la apación de los años de manufactura de las cámaras. Elos datos se centran en las cámaras que surgieron en el mercado en la década de los 90 principalmente.

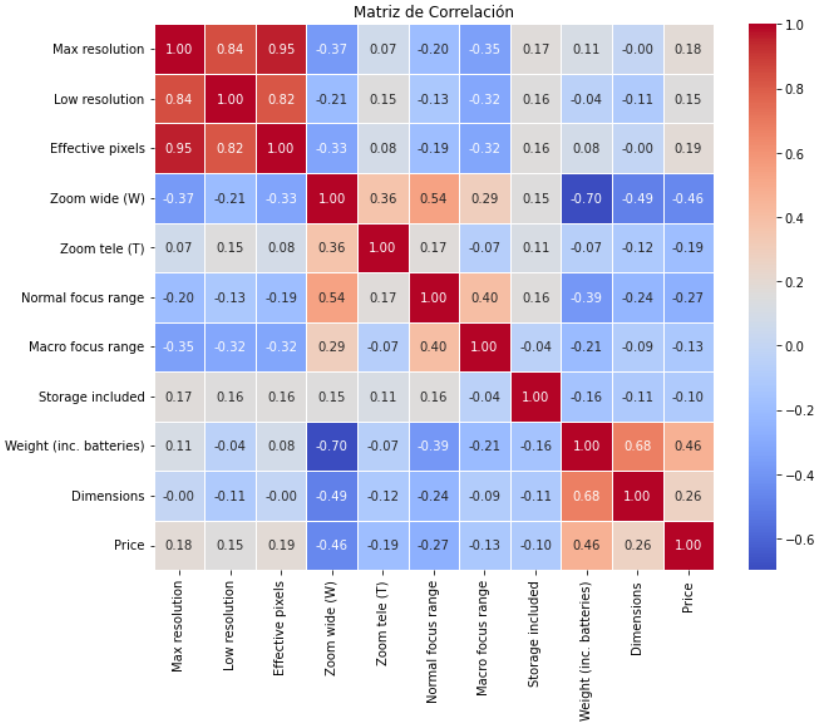


Análisis de Conversión de distribución discreta a distribución continua



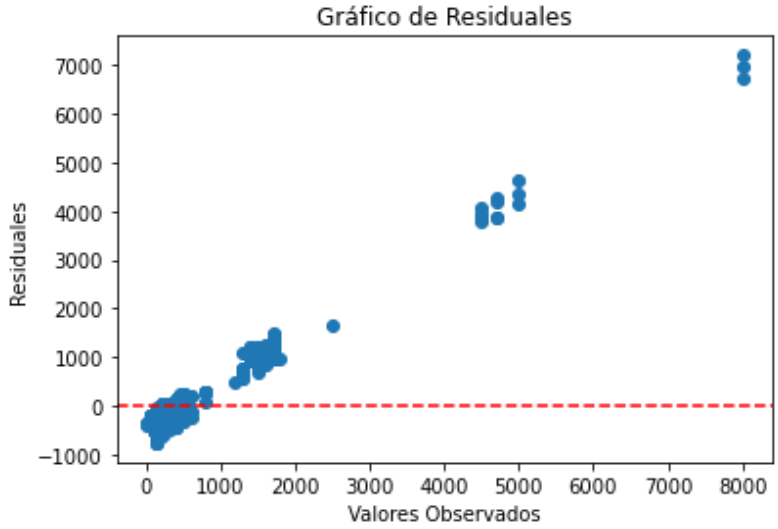
Se hizo el ajuste de la distribución normal utilizando máxima verosimilitud, el gráfico no muestra una distribución normal, esto puede ser debido a que El valor CIF está influenciado por múltiples factores, como la naturaleza de la mercancía, el país de origen y destino, las tasas de cambio, los aranceles y las condiciones del transporte. Estos factores externos pueden introducir variabilidad en el valor CIF, lo que dificulta que siga una distribución normal.

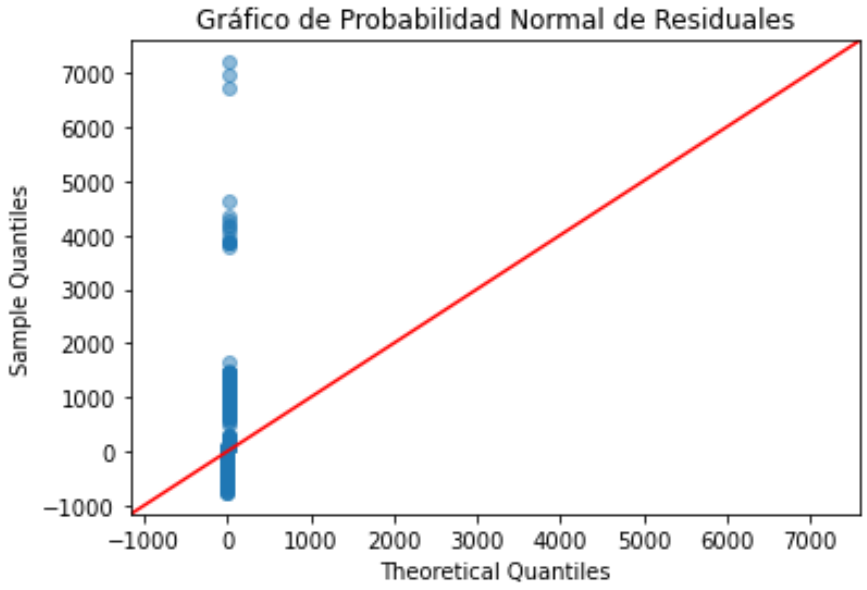
Análisis de Correlación

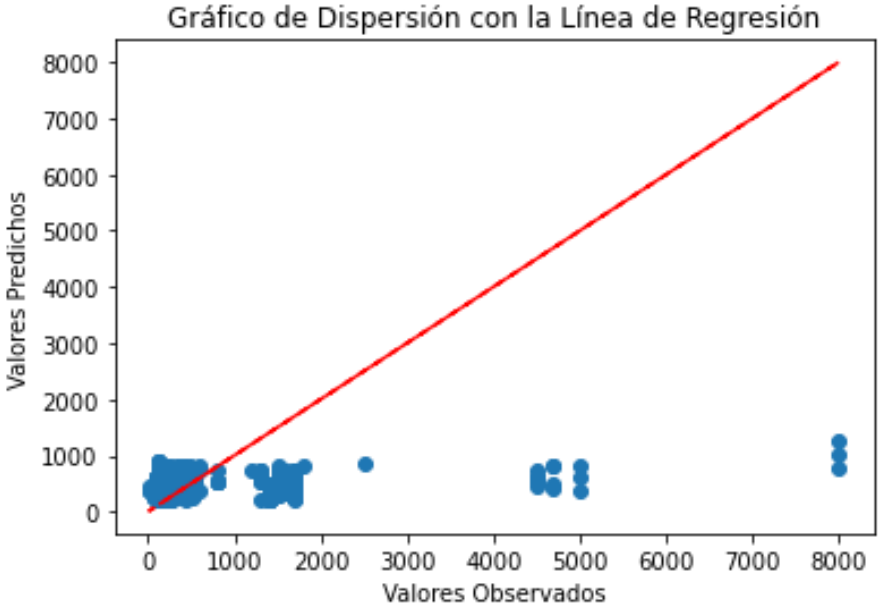


Observamos que las variables más relacionadas son los pixeles y la resolución máxima.

Análisis de Regresión Lineal







Las conclusiones basadas en los resultados son las siguientes:

1. Valor R cuadrado: El coeficiente R cuadrado es 0.036336445881477775. Esto significa que aproximadamente el 3.63% de la variabilidad en la variable objetivo (Un R cuadrado bajo como este sugiere que las variables predictoras tienen una capacidad limitada para explicar las variaciones en la variable objetivo.

2. Prueba de F: La prueba de F tiene un valor de 12.97105958797214. En este caso, el valor de F sugiere que el modelo de regresión es estadísticamente significativo, lo que significa que al menos una de las variables predictoras tiene un efecto significativo en la variable objetivo.

3. Pruebas de t: En este caso, vemos que algunas variables predictoras tienen valores de t significativos, mientras que otras no. Por ejemplo, la variable "Effective pixels" tiene un valor de t significativo (1.604097) debido a su intervalo de confianza que no incluye el cero.

4. Intervalos de Confianza: En este caso, los intervalos de confianza para todas las variables incluyen el cero, excepto para "Effective pixels". Esto indica que, excepto por "Effective pixels", no podemos estar seguros de que las variables tengan un efecto significativo en la variable objetivo.

5. Intercepto y Pendientes: El valor del intercepto es 181.7271092954884 y las pendientes son: "Max resolution": 0.038644, "Low resolution": -0.011828, "Effective pixels": 43.927844. Estos valores indican las contribuciones de cada variable a la predicción del valor objetivo cuando todas las demás variables predictoras son iguales a cero.

Indicadores Económicos-Clasificación de Importaciones según el Uso o Destino Económico de los Bienes (BD6)

Composición del documento

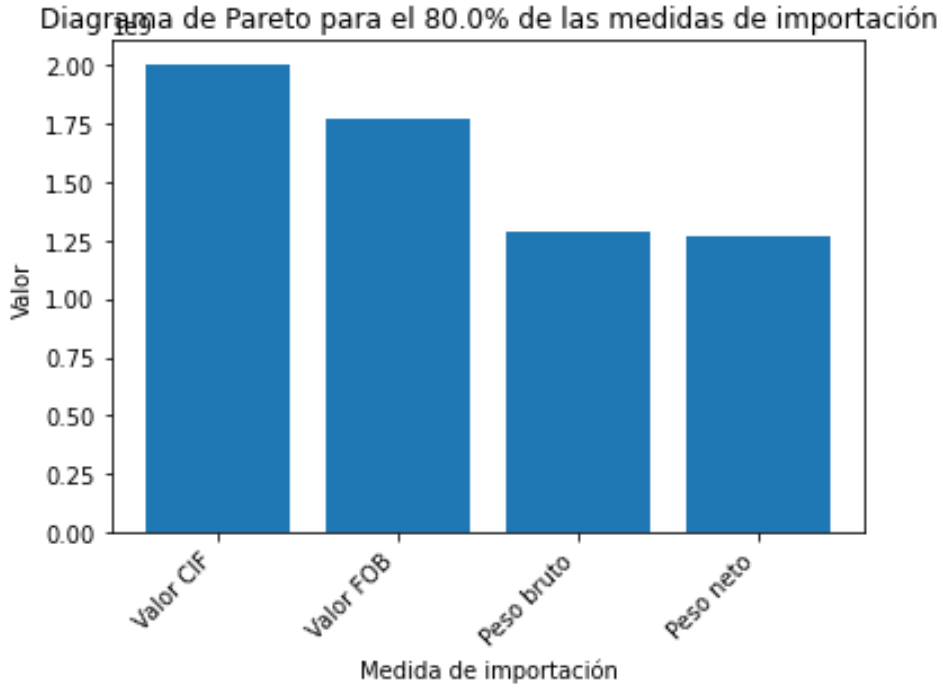
- Año: valor categórico (cadena de números)
- Trimestre: valor categórico (cadena de letras)
- Clase: valor categórico (cadena de letras)
- Nombres de medidas: valor categórico (cadena de letras) (Ej. peso neto, peso bruto, valor FOB, valor CIF)
- Valores de medidas:valor numérico

Plan de análisis

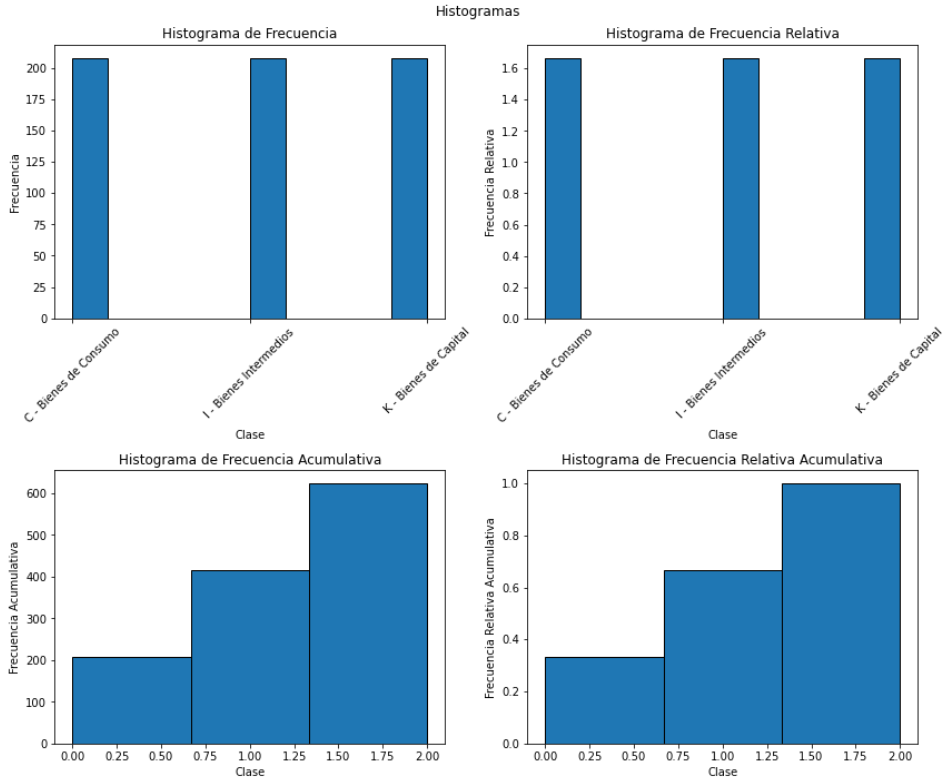
Porqué en este base de datos se organiza la información por trimestres?
Puedo comparar las medidas de importación entre diferentes trimestres para identificar patrones y tendencias. Puedo analizar si hay diferencias significativas en las medidas de importación en diferentes períodos y si existen relaciones con factores económicos o eventos específicos.

Medidas de dispersión de Valores de medidas

Media: 804036583.5576923
Varianza: 1.689471378170349e+17
Desviación Estándar: 411031796.6009867

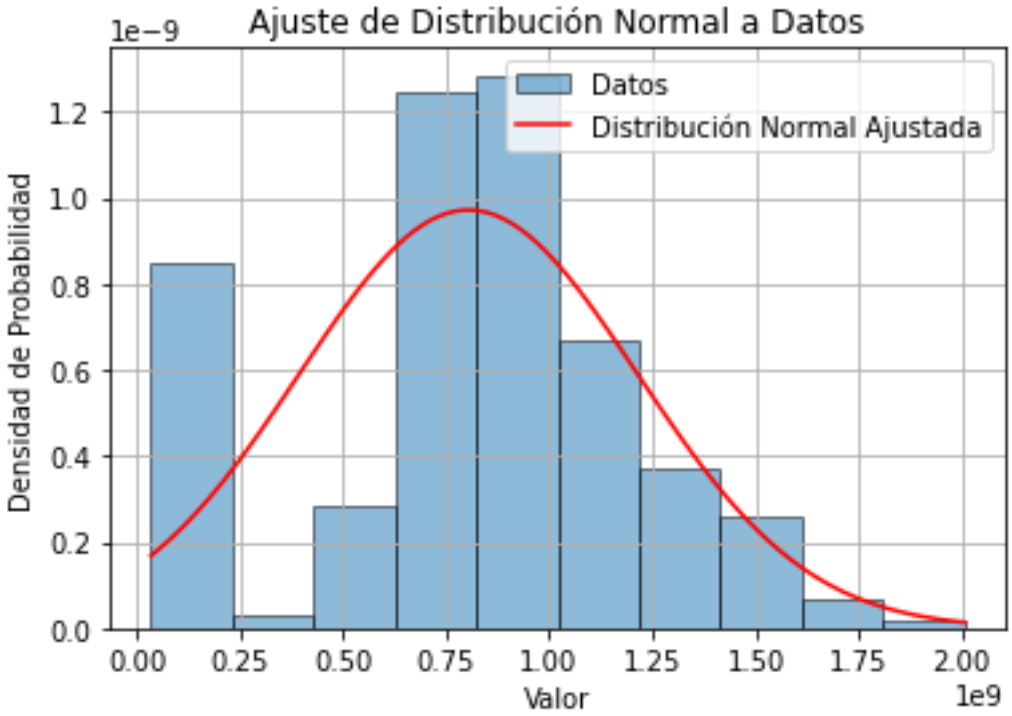


El valor CIF es la medida de importación más común, probablemente porque el valor CIF es el valor real de las mercancías durante el despacho aduanero.



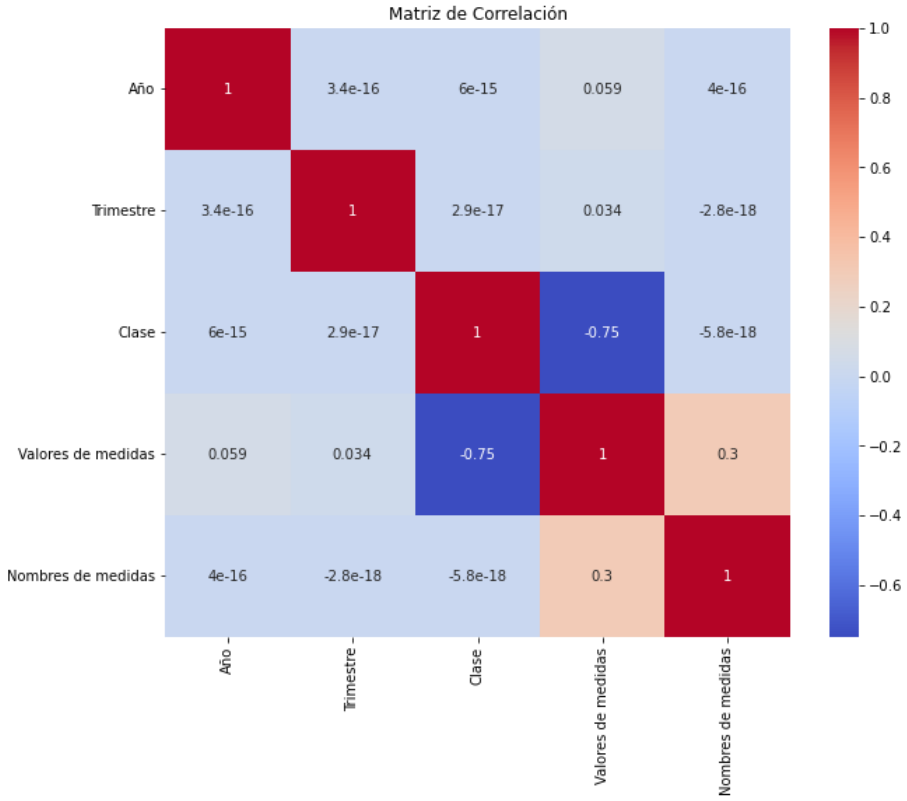
Los datos cuentan con la misma cantidad de registros por cada clase de bienes.

Análisis de Conversión de distribución discreta a distribución continua



En la columna de Valores de medidas hay valores atípicos, podremos analizar más profundamente esta situación en el análisis de correlación.

Análisis de Correlación



Según este resultado podemos deducir que no hay una relación lineal apreciable entre las dos variables en la muestra de datos analizada. Esto significa que, a medida que cambia una de las variables, no hay un cambio consistente y significativo en la otra variable.

Análisis de Regresión Lineal

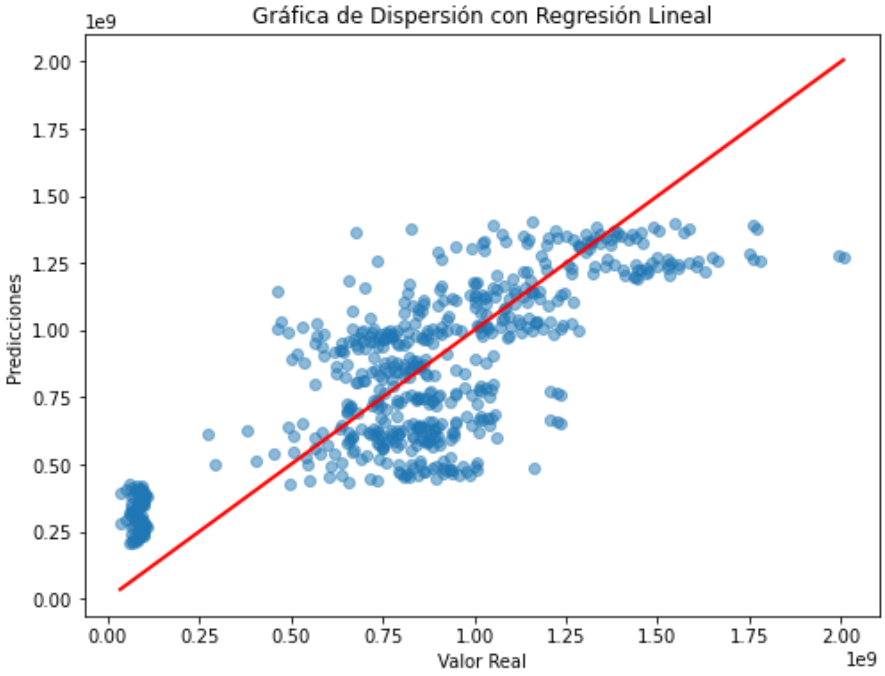
El modelo de regresión obtenido muestra un buen ajuste a los datos, ya que el valor del R-cuadrado es de 0.656, lo que significa que aproximadamente el 65.6% de la variabilidad en la variable dependiente se explica por las variables independientes incluidas en el modelo. El valor ajustado del R-cuadrado, que tiene en cuenta el número de variables independientes en el modelo, es de 0.654, lo que sigue indicando un ajuste razonablemente bueno.

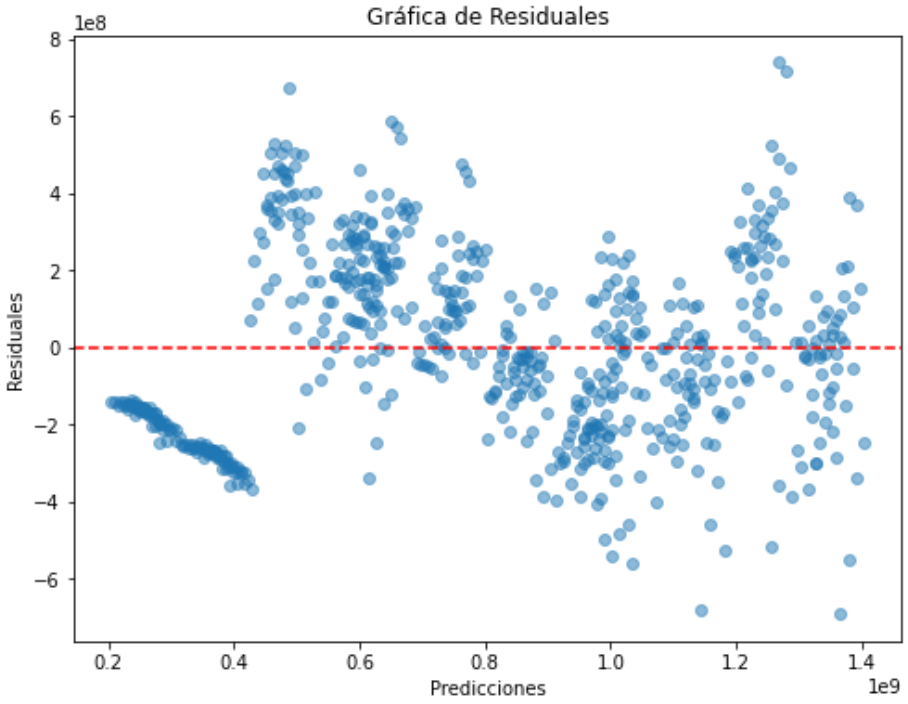
La prueba F-statistic tiene un valor de 295.0 y su correspondiente probabilidad (Prob (F-statistic)) es extremadamente baja (8.32e-142), lo que sugiere que al menos una de las variables independientes es estadísticamente significativa para explicar la variabilidad en la variable dependiente.

Al examinar los coeficientes de regresión (coef), podemos observar que todas las variables independientes tienen coeficientes distintos de cero, lo que implica que cada una de ellas está

contribuyendo de manera significativa al modelo. Sin embargo, el coeficiente para la variable "Trimestre" no es estadísticamente significativo al nivel de significancia del 0.05, ya que su valor P>|t| es 0.155.

Las constantes de la línea de regresión (intercepto) también tienen coeficientes significativos. Sin embargo, es importante señalar que el coeficiente para la constante "const" es extremadamente grande (-1.2e+10) y tiene un error estándar relativamente alto (5.22e+09), lo que sugiere que podría haber problemas de multicolinealidad u otros problemas numéricos en el modelo.





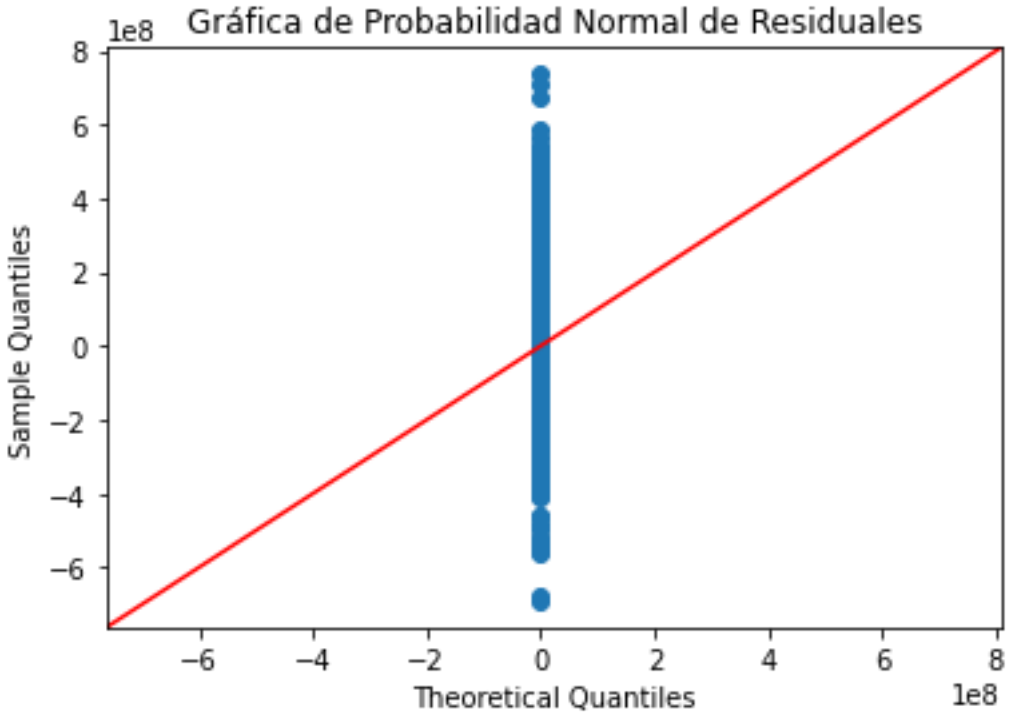


Tabla ANOVA:

<F test: F=294.9523458898949, p=8.317162791299655e-142, df_denom=619, df_num=4>

R-cuadrado: 0.6558834940169755

Intervalos de confianza para coeficientes:

	0	1
const	-2.225005e+10	-1.757603e+09
Año	1.365887e+06	1.153068e+07
Trimestre	-4.690405e+06	2.932751e+07
Clase	-3.998250e+08	-3.532440e+08
Nombres de medidas	9.377524e+07	1.277932e+08

Valores de las constantes de la línea de regresión:

Intercept:	-12003824988.506275
Slope: Año	6.448285e+06
Trimestre	1.231855e+07
Clase	-3.765345e+08
Nombres de medidas	1.107842e+08

Al examinar las pruebas de t y los valores de P para cada coeficiente individual, se observa que las variables "Año", "Clase" y "Nombres de medidas" tienen coeficientes estadísticamente significativos, ya que sus valores de P son muy cercanos a cero (menores que 0.05). Sin embargo, la variable "Trimestre" no es estadísticamente significativa en el nivel de significancia del 0.05, ya que su valor de P es 0.155.

Los intervalos de confianza para los coeficientes indican los rangos en los que es probable que se encuentren los verdaderos valores de los coeficientes con un cierto nivel de confianza. Por ejemplo, para la constante "const", el intervalo de confianza va desde aproximadamente-2.23e+10 hasta-1.76e+09. Esto sugiere que el valor verdadero del coeficiente de la constante tiene una alta incertidumbre.

PIB anual por categoria economica (BD7)

Composición del documento

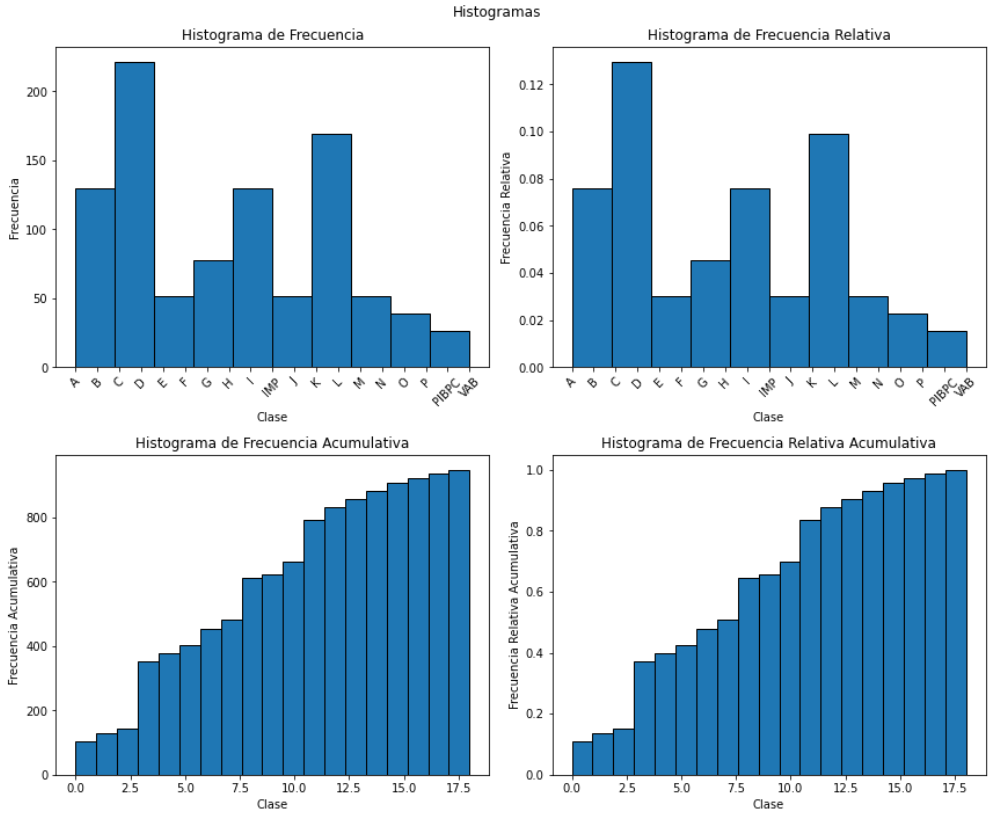
- Año: valor categórico (cadena de números)
- Categorías: valor categórico (cadena de letras)
- Codcategoria: valor categórico (cadena de letras)
- Aporte Absoluto Porcentual Corriente: valor numérico
- Valor Corriente: valor numérico
- Variación Absoluta Corriente: valor numérico
- Variación Porcentual Corriente: valor numérico

Plan de análisis

Con estos datos a mi disposición, puedo realizar diferentes análisis para comprender mejor el PIB anual por categoría económica. Por ejemplo, puedo calcular la contribución relativa de cada categoría económica al PIB total en diferentes años y examinar las tendencias a lo largo del tiempo. También puedo analizar las variaciones absolutas y porcentuales del PIB de cada categoría económica para identificar sectores que experimenten un crecimiento significativo o enfrenten desafíos.

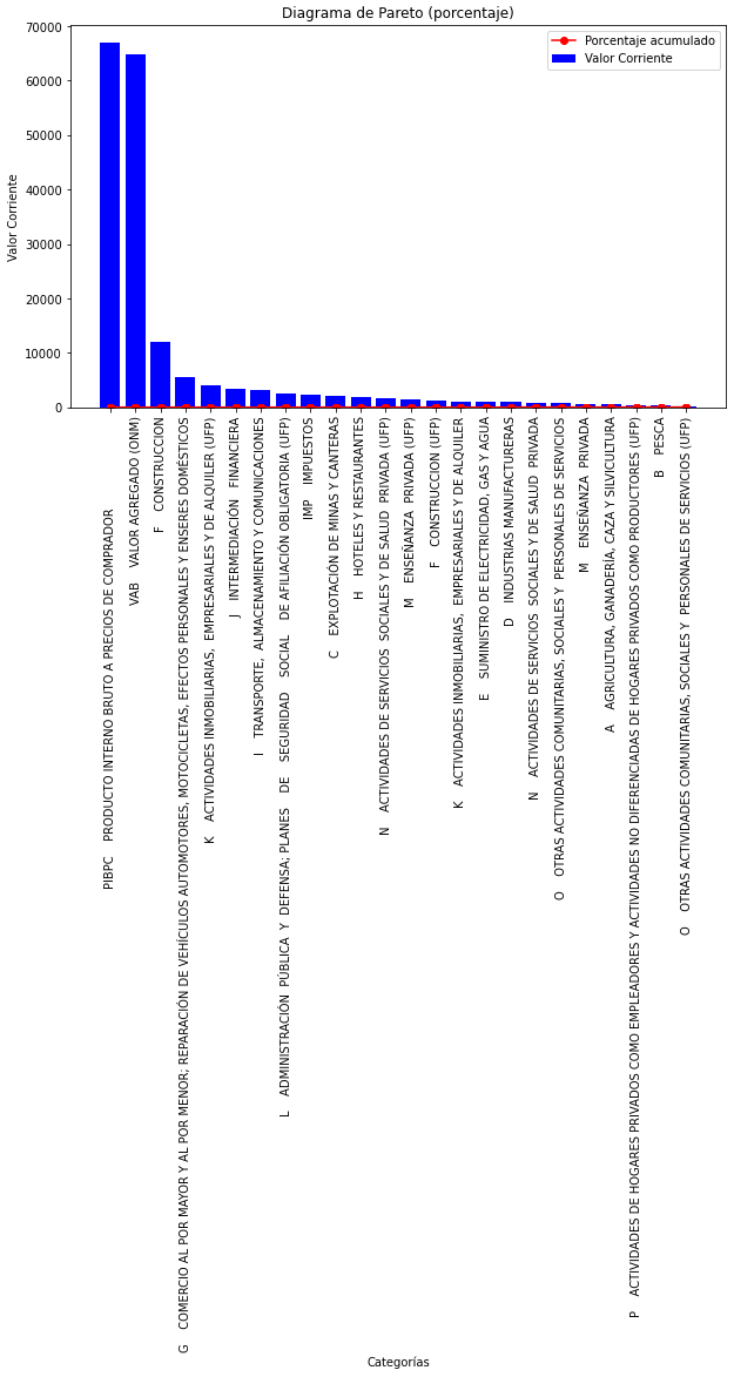
Media, varianza y desviación estándar del Valor Corriente

Media: 1911.9072708113797
Varianza: 61991661.76883948
Desviación Estándar: 7873.478378000379



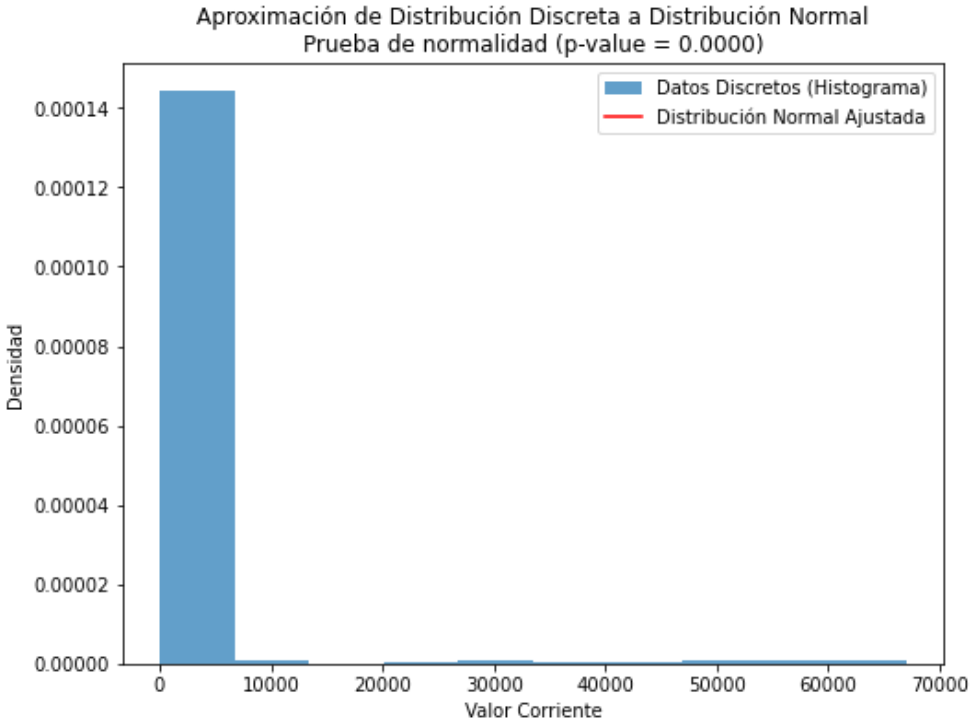
Concluimos que las categorías más comunes son C EXPLOTACIÓN DE MINAS Y CANTERAS y D INDUSTRIAS MANUFACTURERAS.

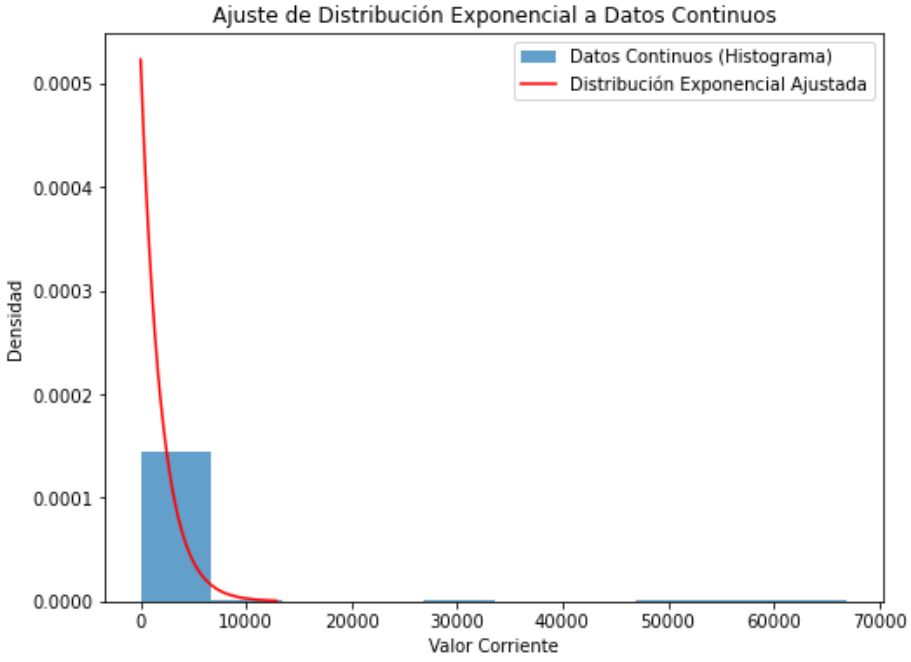
Diagrama de Pareto que relaciona la categoría y el valor corriente



Podemos observar que el mayor valor corriente lo toma la categoría de PIBPC, Producto Interno Bruto a precios de comprador

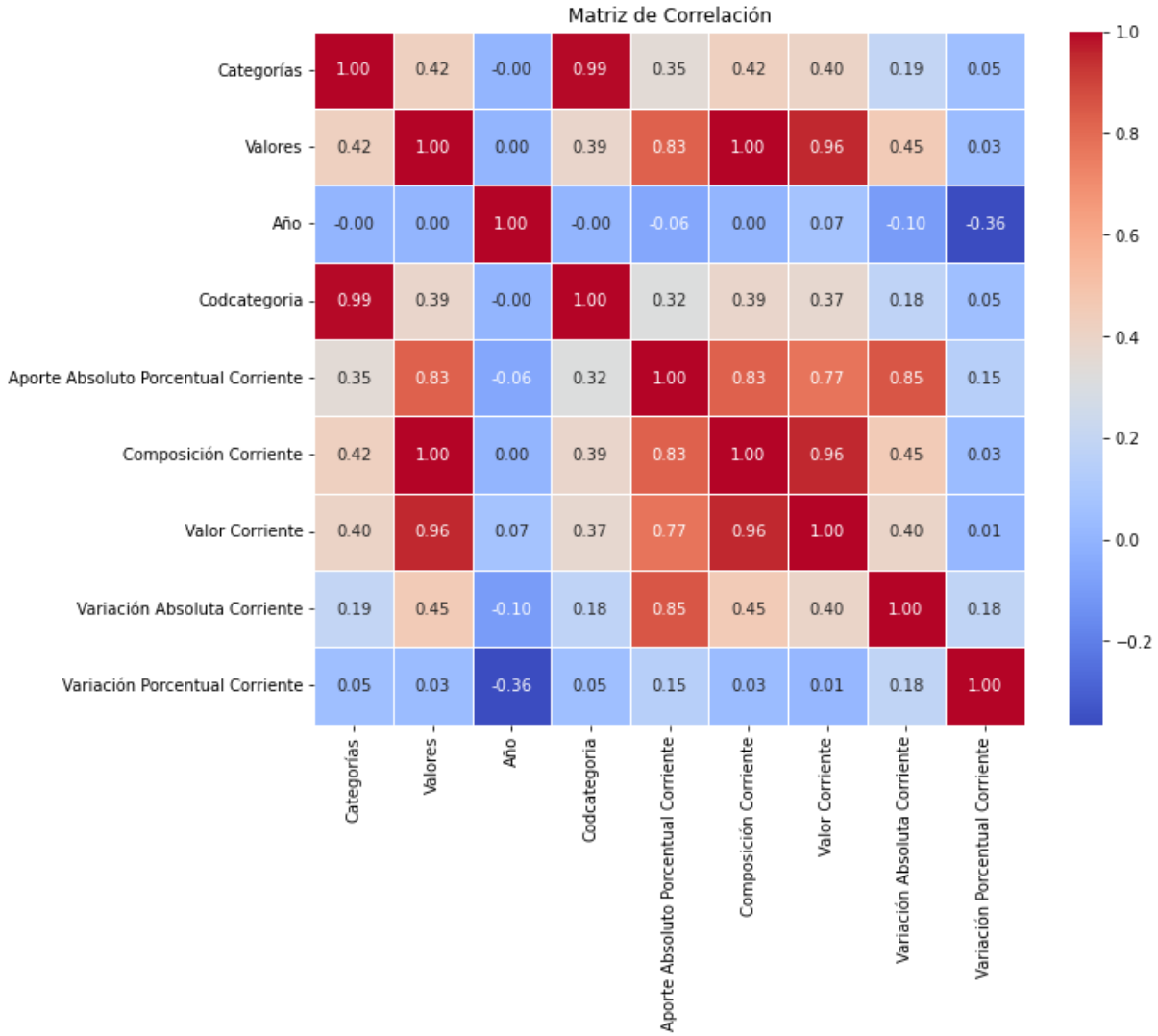
Análisis de Conversión de distribución discreta a distribución continua





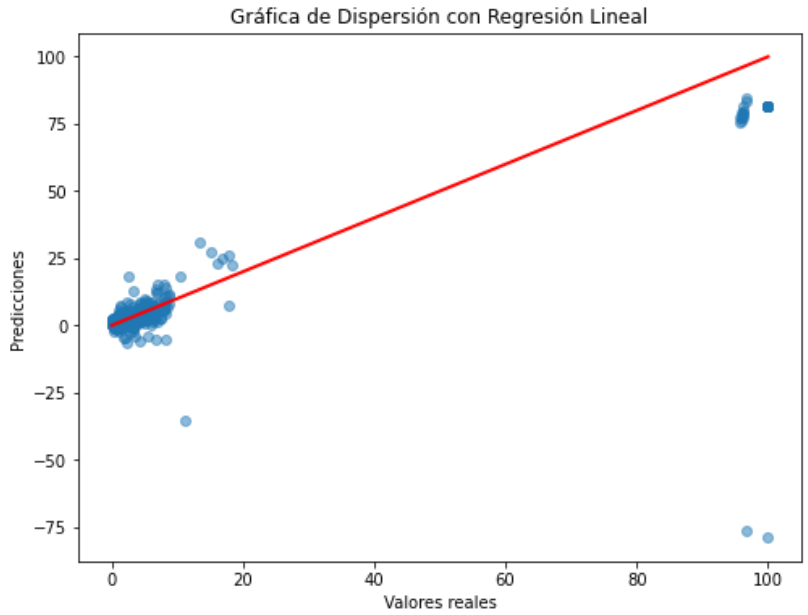
Según el resultado de estos gráficos, los datos no son normales y además no se ajustan a ninguna de las distribuciones más comunes.

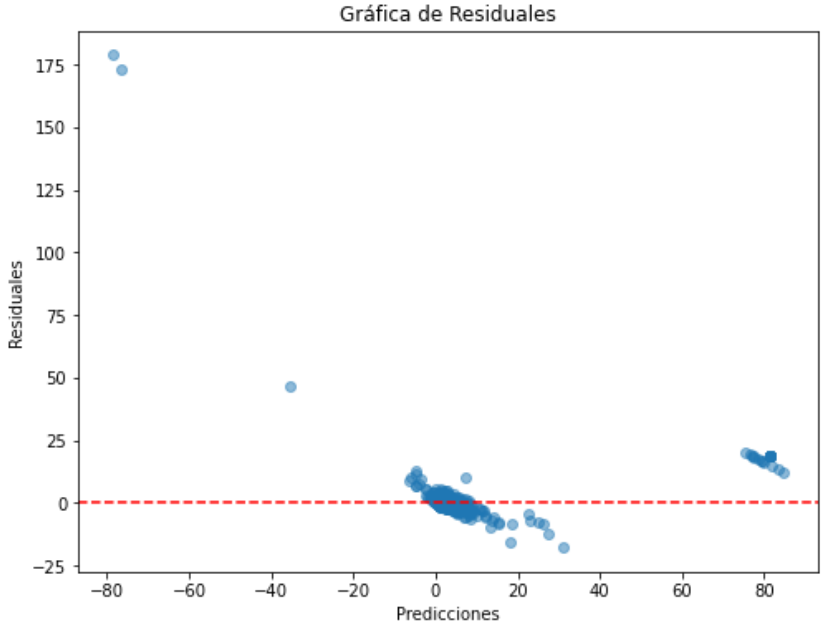
Análisis de Correlación

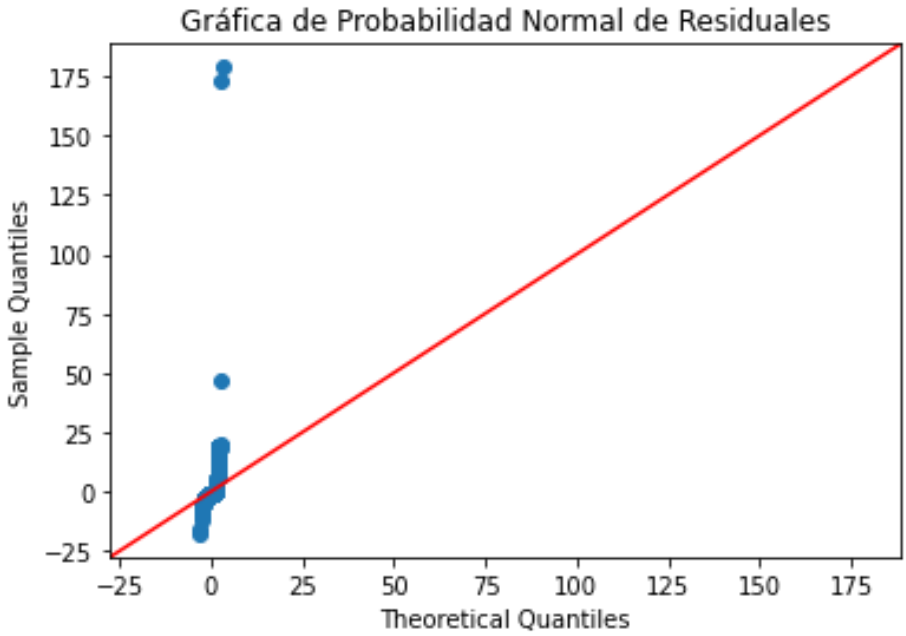


Debido a la forma en la que están organizados los datos, hay variables que son calculadas en base a otras y por esta razón están tan relacionadas.

Análisis de Regresión Lineal







Intercepto: 1.2969092817285286

Pendiente: 0.8003952837219531

R cuadrado: 0.684628960067893

Tabla ANOVA para coeficientes:

<F test: F=2055.812180230212, p=1.5411303050765463e-239, df_denom=947, df_num=1>

Basándonos en los resultados obtenidos del modelo de regresión lineal:

1. R cuadrado: El valor del coeficiente de determinación (R cuadrado) es aproximadamente 0.685. Esto significa que alrededor del 68.5% de la variabilidad de la variable dependiente ('Valores') puede explicarse por la variable independiente ('Aporte Absoluto Porcentual Corriente') en el modelo. Un R cuadrado cercano a 1 indica que el modelo se ajusta bien a los datos, aunque en este caso, podría haber espacio para mejorar la capacidad de predicción.

2. Coeficientes: El modelo de regresión lineal muestra que el intercepto (constante) es 1.2969 y la pendiente de la variable independiente 'Aporte Absoluto Porcentual Corriente' es 0.8004. Esto significa que, manteniendo constante el valor de la variable independiente, la variable dependiente

'Valores' aumentará en aproximadamente 0.8004 unidades por cada unidad que aumente 'Aporte Absoluto Porcentual Corriente'.

3. Prueba de Hipótesis: La tabla ANOVA muestra un valor F significativo (Prob (F-statistic) < 0.05), lo que indica que el modelo general es significativo en términos de predicción. Además, la prueba de t para el coeficiente de 'Aporte Absoluto Porcentual Corriente' también muestra un valor P muy bajo (P>|t| < 0.05), lo que indica que este coeficiente es significativo en el modelo. En conjunto, esto sugiere que 'Aporte Absoluto Porcentual Corriente' es una variable significativa en la predicción de 'Valores'.

Intervalos de confianza:

Intercepto: El intervalo de confianza para el intercepto (constante) está entre 0.713748 y 1.880070 con un 95% de confianza. Esto significa que podemos estar razonablemente seguros de que el valor real del intercepto de la población se encuentra en este rango.

Aporte Absoluto Porcentual Corriente: El intervalo de confianza para el coeficiente de 'Aporte Absoluto Porcentual Corriente' está entre 0.765752 y 0.835038 con un 95% de confianza. Esto indica que podemos estar razonablemente seguros de que el valor real del coeficiente de 'Aporte Absoluto Porcentual Corriente' en la población se encuentra en este rango.

Importaciones anuales (BD8)

Composición del documento

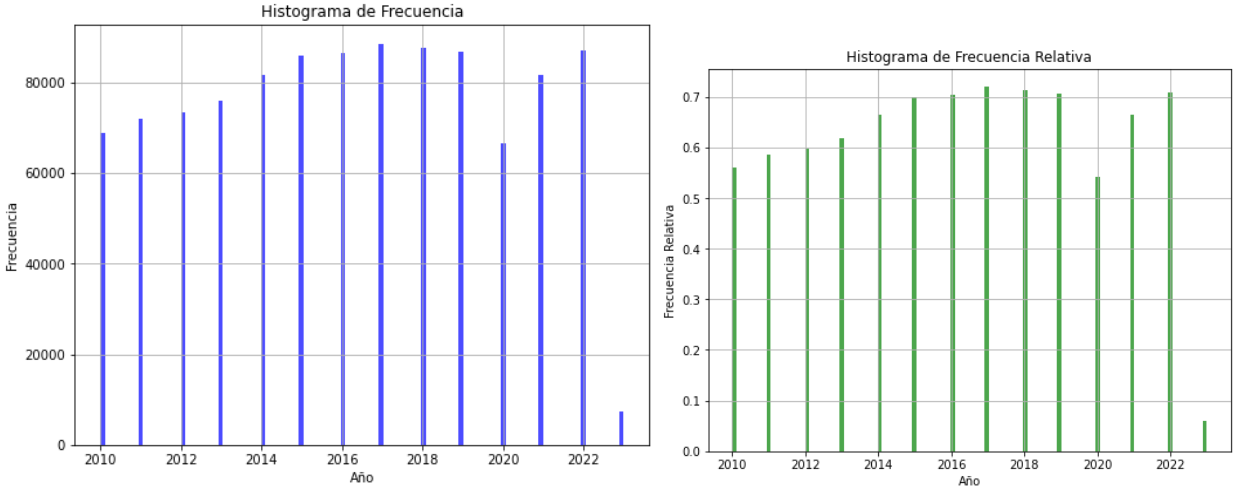
- Año: valor categórico (cadena de números)
- Trimestre: valor categórico (cadena de letras)
- Clase: valor categórico (cadena de letras)
- Mes: valor categórico (cadena de letras)
- Arancel:valor categórico (cadena de números)
- MES: valor categórico (número)
- Valor FOB: valor numérico
- Valor CIF: valor numérico
- Peso bruto: valor numérico
- Peso neto: valor numérico

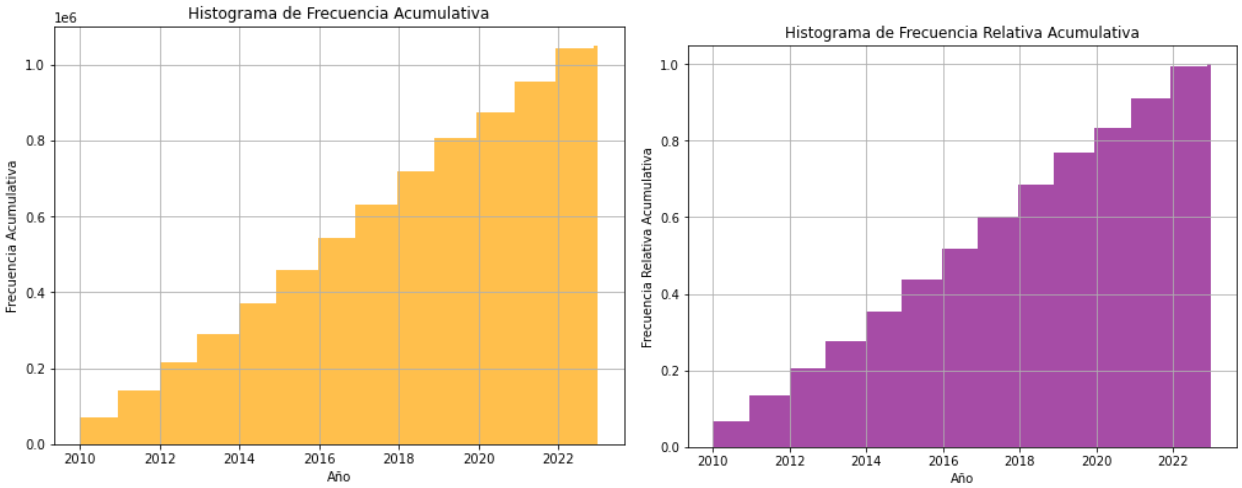
Plan de análisis

Con esta información, puedo llevar a cabo diferentes análisis para comprender mejor las importaciones anuales en relación con los meses, clases de bienes, aranceles y otras variables relevantes. Por ejemplo, puedo analizar la distribución de las importaciones a lo largo de los meses y trimestres del año, identificar las clases de bienes más importados y analizar las variaciones en los valores FOB y CIF de las importaciones a lo largo del tiempo.

Además, puedo realizar comparaciones entre los diferentes aranceles y su impacto en el valor FOB de las importaciones. También puedo examinar la relación entre el peso bruto y neto de las importaciones y realizar análisis descriptivos para comprender mejor las características físicas de los bienes importados.

Histograma de frecuencia





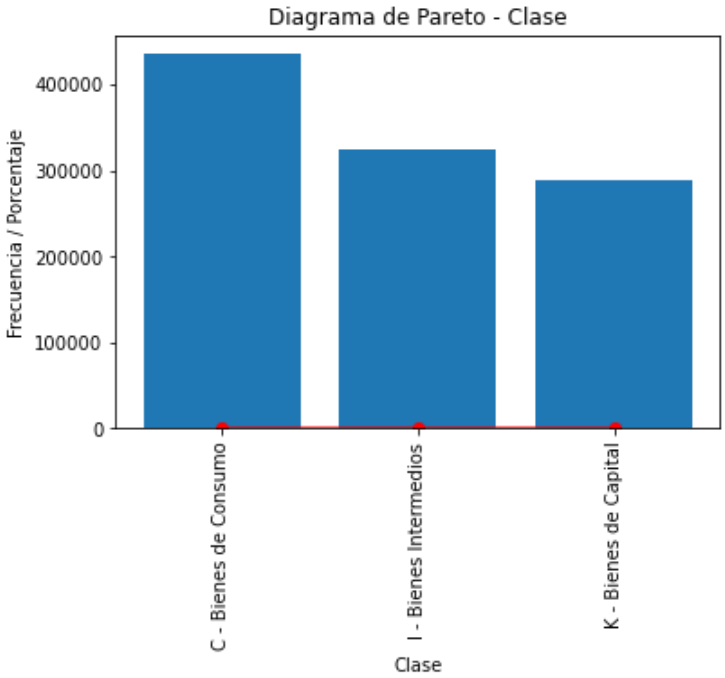
Este gráfico nos sirve para entender mejor las importaciones anuales a lo largo del tiempo

Medidas de variación de la columna Peso bruto

Media: 21594.165879407767

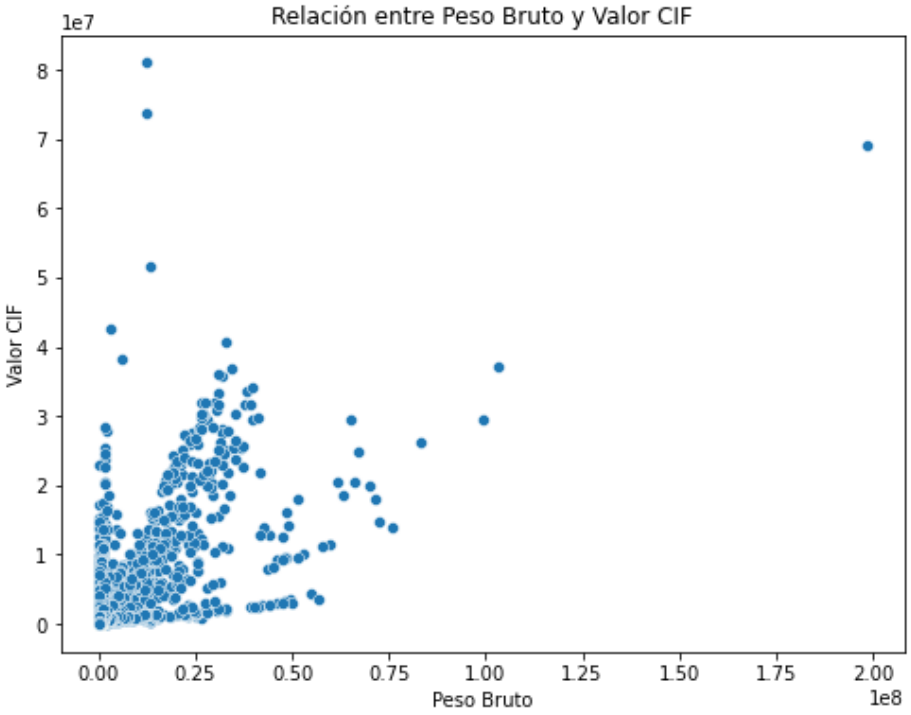
Varianza: 364856050731.78394

Desviación estándar: 604033.1536693858



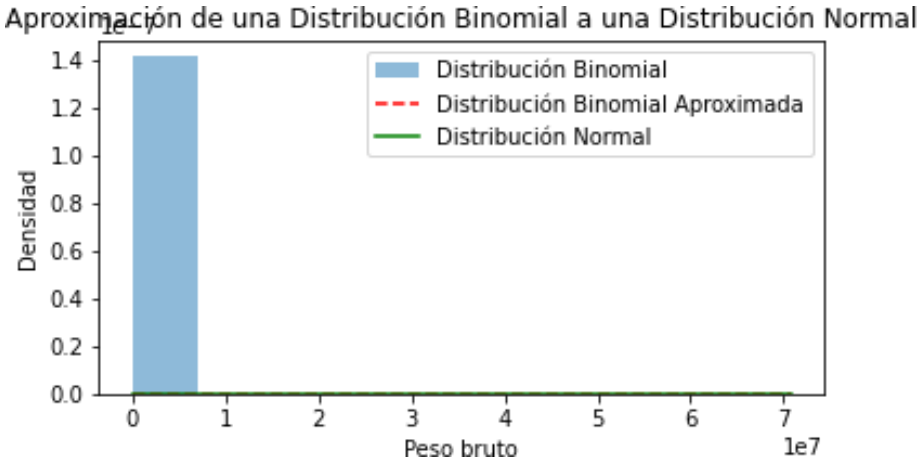
La clase más común de importación son bienes de consumo debido a las ventajas de diversidad, costo, calidad y acceso a productos internacionales que ofrece a los consumidores y empresas. Los bienes de consumo importados complementan y enriquecen la oferta de productos en los mercados locales, brindando más opciones y oportunidades para satisfacer las necesidades y deseos de los consumidores.

Comparación entre Preso bruto y valor CIF



El peso bruto es una medida física que representa el peso total de los bienes importados, incluyendo el embalaje y otros materiales de envío. El valor CIF, por otro lado, es el valor de los bienes importados, incluyendo el costo, el seguro y el flete (Cost, Insurance, and Freight). Al comparar el peso bruto y el valor CIF, evaluamos si existe una relación directa o indirecta entre el tamaño físico o peso de los bienes importados y su valor monetario.

Aproximación de distribución normal



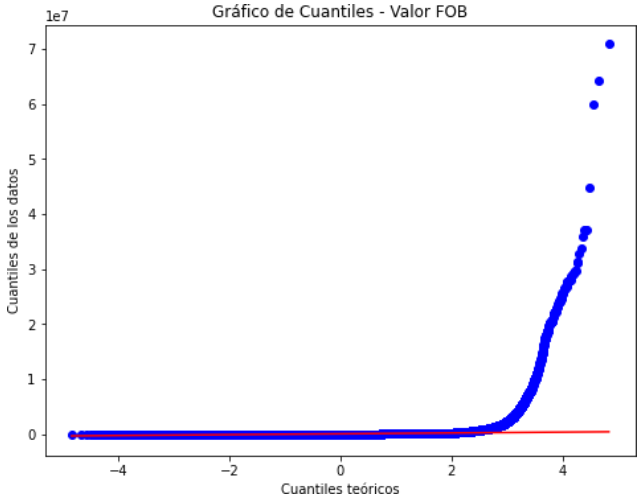
En la gráfica la distribución normal está demasiado separada de la distribución binomial, puede indicar que la aproximación de la distribución binomial a una distribución normal puede no ser adecuada para estos datos. La aproximación puede no ser precisa si los datos no cumplen con los supuestos necesarios para el teorema del límite central, como un tamaño de muestra suficientemente grande o variables aleatorias independientes. Además, si la distribución binomial en sí misma tiene características distintas, como una alta asimetría o valores extremos, la aproximación a una distribución normal puede no ser apropiada.

Análisis de regresión lineal

Pruebas estadísticas Prueba t de Student:

Estadística t:-0.2659230163636165 Valor p: 0.7902985247554513

En este caso, el valor p es 0.7902985247554513, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula y concluir que las medias de peso neto y peso bruto son significativamente diferentes.



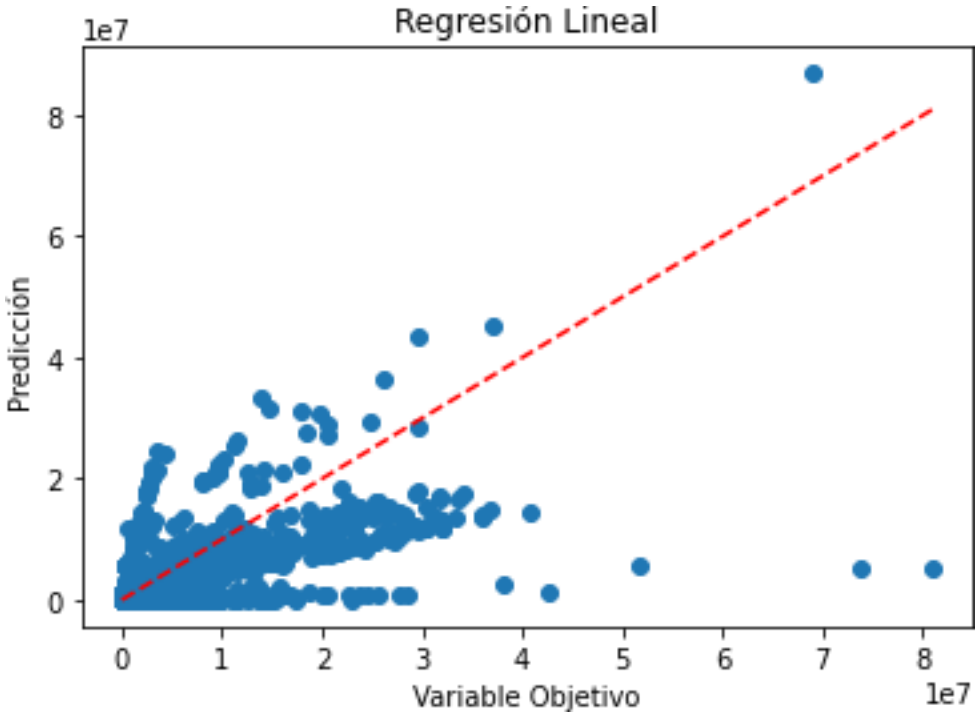
ANOVA

Valor F: 333.9226644291241

Valor p: 1.3622210136264743e-74

Estos valores son estadísticamente significativos y sugieren que al menos uno de los grupos tiene una media significativamente diferente de los otros grupos en términos de la variable "Valor FOB".

El valor extremadamente bajo del valor p (1.3622210136264743e-74) indica que hay una probabilidad muy baja de que los resultados de la prueba de ANOVA se deban al azar. Por lo tanto, podemos rechazar la hipótesis nula y concluir que existen diferencias significativas entre al menos dos de los grupos en términos del "Valor FOB".



Coeficiente de "Peso neto":-0.08388370342593596

Esto significa que, para cada unidad de aumento en la variable "Peso neto", se espera un decremento de aproximadamente 0.0839 en la variable objetivo (en la dirección opuesta).

Coeficiente de "Peso bruto": 0.5219135444533894

Esto indica que, para cada unidad de aumento en la variable "Peso bruto", se espera un incremento de aproximadamente 0.5219 en la variable objetivo.

Intercepto: 27697.88911138851

El intercepto (o término constante) es el valor estimado de la variable objetivo cuando todas las variables predictoras son cero. En este caso, indica que cuando las variables "Peso neto" y "Peso bruto" son cero, se espera que la variable objetivo tenga un valor de aproximadamente 27697.8891.

Accidentes automovilísticos (conductores)(BD1)

Composición del documento

- Provincia: valor categórico (cadena de letras)
- Material_Calle: valor categórico (cadena de letras)
- Año: valor categórico (cadena de números)
- Evento: valor categórico (cadena de letras)
- Conductores: valor numérico

Plan de análisis

Analizando las columnas de datos luego de haber eliminado las columnas con valores nulos, me interesa saber qué eventos se dan más frecuentemente dependiendo de las calles.
Hipótesis de relación entre el material de la calle y la gravedad de los accidentes: analizar si el tipo de material de la calle (por ejemplo, asfalto, concreto, grava) está relacionado con la gravedad de los accidentes. Por ejemplo, si los accidentes en calles con ciertos materiales tienden a ser más graves o tienen un mayor número de víctimas.

Tabla pivote de relación entre el material de la calle y el evento del accidente

Material_Calle	Evento	
Asfalto	Vuelco (caída en cuneta)	131
	Colisión	130
	Atropello	125
Concreto	Colisión	114
	Atropello	101
	Vuelco (caída en cuneta)	94
Grava	Colisión	76
	Vuelco (caída en cuneta)	42
	Atropello	21
Tierra	Colisión	121
	Vuelco (caída en cuneta)	107
	Atropello	63

Por efectos del análisis de datos, en la tabla solo se muestran los materiales especificados y los tres eventos más comunes por material.

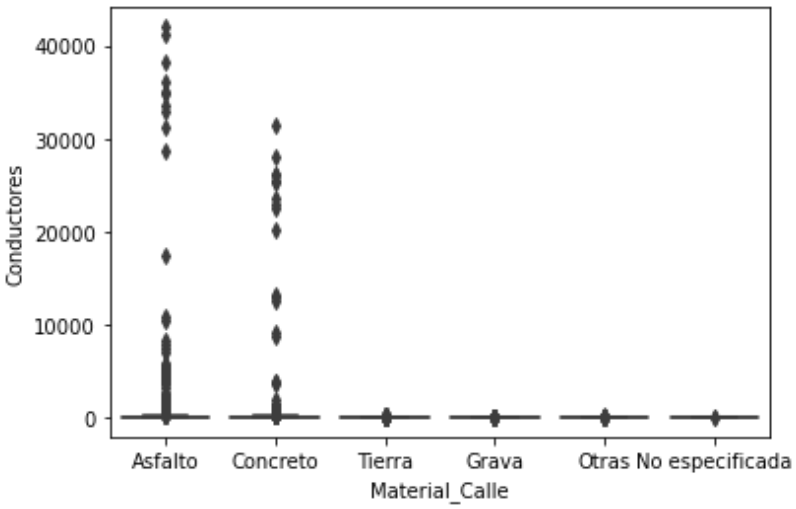
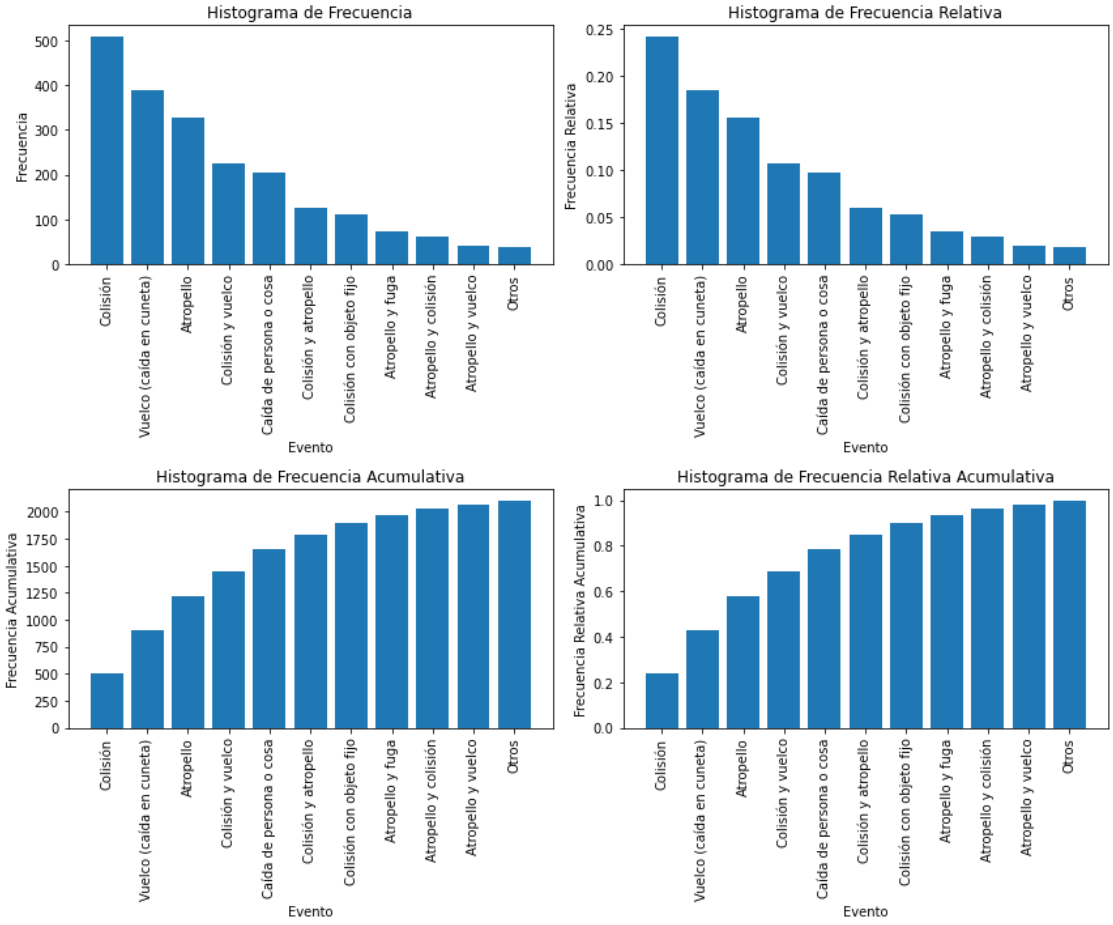


Gráfico de relación entre el material de la calle y la gravedad de los accidentes

Histograma de frecuencia de eventos



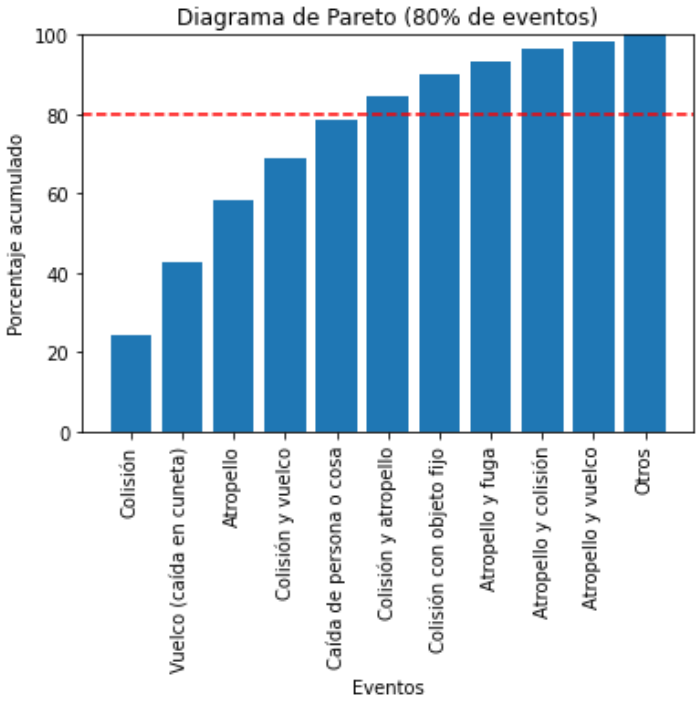
Podemos visualizar que colisión es el evento más frecuente, este resultado lo estudiaremos más profundamente más adelante

Medidas de dispersión de la columna Conductores

Media: 468.60095011876484

Varianza: 9807591.326424915

Desviación estándar: 3131.7074139237393



Este diagrama está hecho al 80% de los eventos

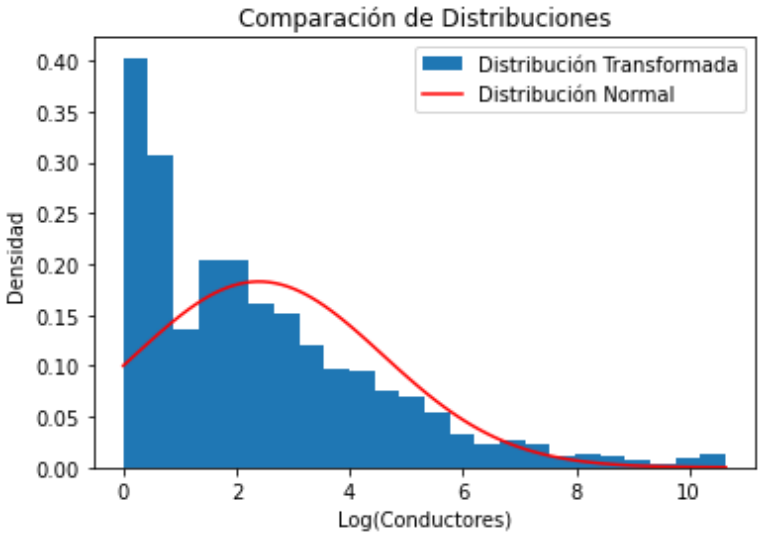
1. Características de los materiales de calle: El tipo de material de calle, ya sea asfalto o tierra, puede afectar la forma en que los vehículos se desplazan y cómo interactúan con su entorno. Por ejemplo, el asfalto generalmente ofrece una superficie más suave y mayor adherencia, lo que puede favorecer velocidades más altas y maniobras más rápidas. Por otro lado, la tierra puede ser más irregular y menos estable, lo que puede aumentar el riesgo de colisiones.

2. Uso y densidad del tráfico: La provincia más común podría tener características demográficas o geográficas que influyen en la forma en que se utiliza la infraestructura vial. El uso y la densidad del tráfico pueden variar en diferentes áreas y pueden afectar la frecuencia de eventos específicos. Por ejemplo, una provincia con una alta densidad de tráfico en vías asfaltadas puede aumentar el

riesgo de atropellos, mientras que en áreas rurales con caminos de tierra, las colisiones pueden ser más frecuentes debido a diferentes condiciones de conducción.

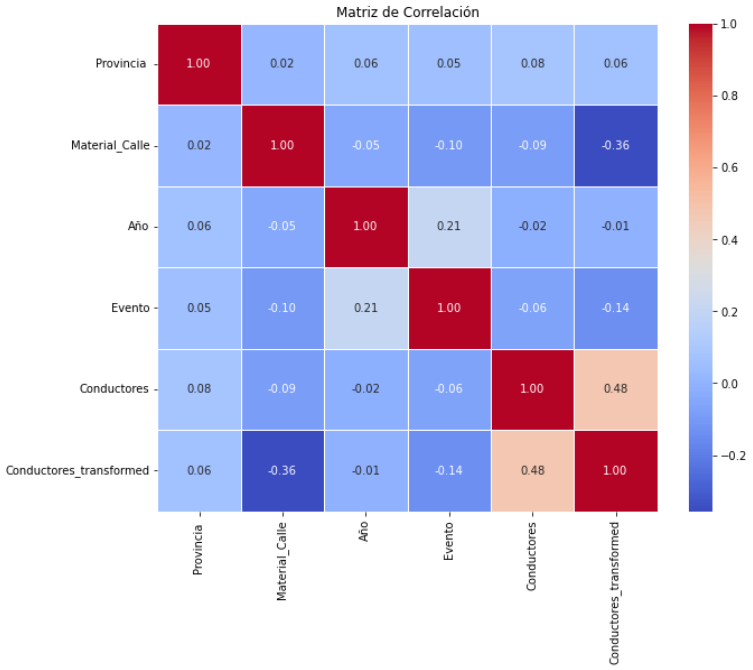
3. Infraestructura vial y diseño de carreteras: La infraestructura vial y el diseño de carreteras pueden diferir entre vías asfaltadas y caminos de tierra. Esto puede incluir características como señalización vial, iluminación, cruces peatonales, y la presencia de obstáculos u otros elementos que puedan influir en la ocurrencia de eventos específicos. Por ejemplo, una mayor presencia de cruces peatonales en vías asfaltadas puede aumentar la posibilidad de atropellos.

Análisis de Conversión de distribución discreta a distribución continua



La transformación se realizó utilizando logaritmos, como podemos observar la transformación fue acertada mayormente.

Análisis de Correlación



Las variables de esta base de datos no están relacionadas muy íntimamente.

Análisis de Regresión Lineal

Coeficientes:

const 30.722222

Asfalto 119.611111

Concreto -8.722222

Grava -28.722222

Otras -26.722222

Tierra -24.722222

Intercepto:

30.722222222222193

R cuadrado:

0.28382237759565976

Las conclusiones que se pueden obtener de los resultados de la regresión lineal son las siguientes:

1. Coeficientes: Los coeficientes indican cómo cambia el valor de la variable dependiente ("Conductores") en función de cada categoría de la variable independiente ("Material_Calle"). En este caso, los coeficientes nos indican lo siguiente:

- El coeficiente de la constante es 30.72. Esto significa que cuando la categoría de "Material_Calle" es "Otras" (que se toma como referencia), se espera que haya aproximadamente 30.72 conductores involucrados en el accidente.

- Cuando la categoría de "Material_Calle" es "Asfalto", se espera que haya un aumento de aproximadamente 119.61 conductores en comparación con la categoría de referencia "Otras".

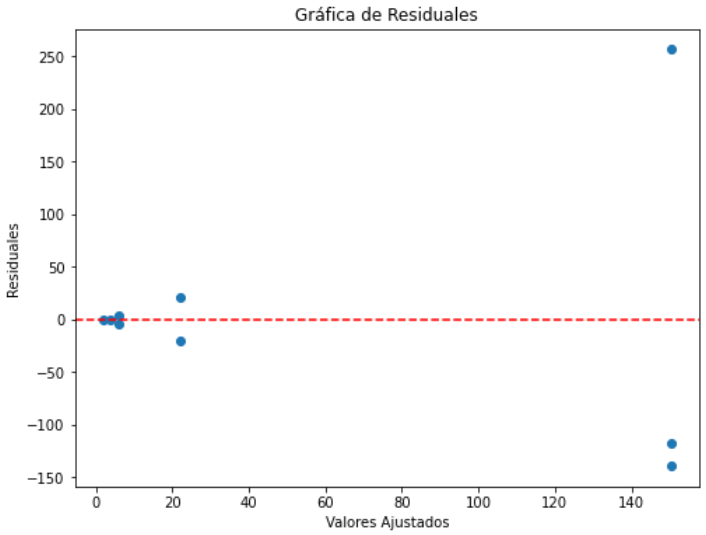
- Cuando la categoría de "Material_Calle" es "Concreto", se espera una disminución de aproximadamente 8.72 conductores en comparación con la categoría de referencia "Otras".

- Cuando la categoría de "Material_Calle" es "Grava", se espera una disminución de aproximadamente 28.72 conductores en comparación con la categoría de referencia "Otras".

- Cuando la categoría de "Material_Calle" es "Tierra", se espera una disminución de aproximadamente 24.72 conductores en comparación con la categoría de referencia "Otras".

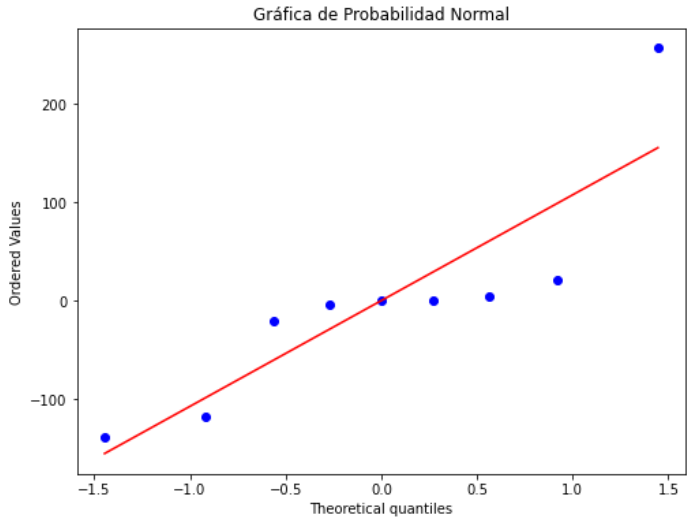
2. Intercepto: En este caso, cuando la categoría de "Material_Calle" es la de referencia "Otras", se espera que haya aproximadamente 30.72 conductores involucrados en el accidente.

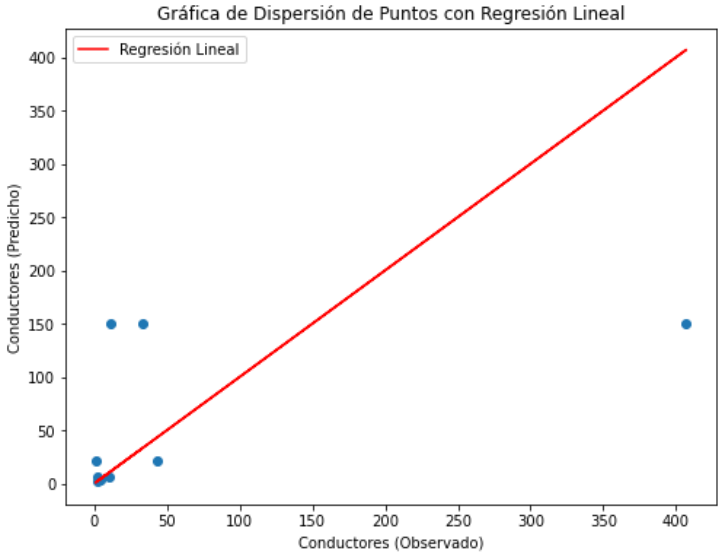
3. R cuadrado: El coeficiente de determinación R cuadrado es 0.28, lo que significa que aproximadamente el 28.38% de la variabilidad en la variable dependiente ("Conductores") puede explicarse por la variable independiente ("Material_Calle") en el modelo de regresión lineal. Esto sugiere que la variable "Material_Calle" tiene cierto poder predictivo para explicar las variaciones en la cantidad de conductores involucrados en los accidentes.



Prueba de normalidad (p-value):

0.014366894878093677





Prueba t (p-value):

0.9999999999999994

Prueba F (p-value):

1.0

De los resultados proporcionados, podemos obtener las siguientes conclusiones:

1. Prueba de normalidad: La prueba de normalidad (p-value = 0.0144) indica que los residuales del modelo no siguen una distribución normal. Esto puede sugerir que el modelo de regresión lineal puede no ser el más adecuado para ajustar los datos, y es posible que existan otros factores o relaciones no lineales que afecten la variable dependiente ("Conductores").

2. Prueba t: La prueba t (p-value = 1.0) sugiere que no hay diferencias significativas entre las medias de las categorías de "Material_Calle" en relación con la cantidad de conductores involucrados en los accidentes. Esto significa que el tipo de material de la calle no parece tener un impacto significativo en el número de conductores en los accidentes, según el modelo de regresión lineal utilizado.

3. Prueba F: La prueba F (p-value = 1.0) indica que no hay diferencias significativas entre las categorías de "Material_Calle" en general en relación con la cantidad de conductores involucrados

en los accidentes. Esto sugiere que el modelo de regresión lineal utilizado no es significativamente mejor que un modelo que no incluya la variable "Material_Calle".

4. Intervalos de confianza: Los intervalos de confianza para los coeficientes nos dan una idea de la incertidumbre asociada con las estimaciones de los coeficientes. Por ejemplo, para la variable "Asfalto", el intervalo de confianza (-126.67, 365.89) indica que el coeficiente para "Asfalto" puede variar entre esos valores con un nivel de confianza determinado. Estos intervalos amplios pueden deberse a la falta de significancia estadística en los resultados de las pruebas t y F.

En resumen, según el modelo de regresión lineal utilizado, no parece haber una relación significativa entre el tipo de material de la calle y la cantidad de conductores involucrados en los accidentes.

Exportaciones por aranceles (BD2)

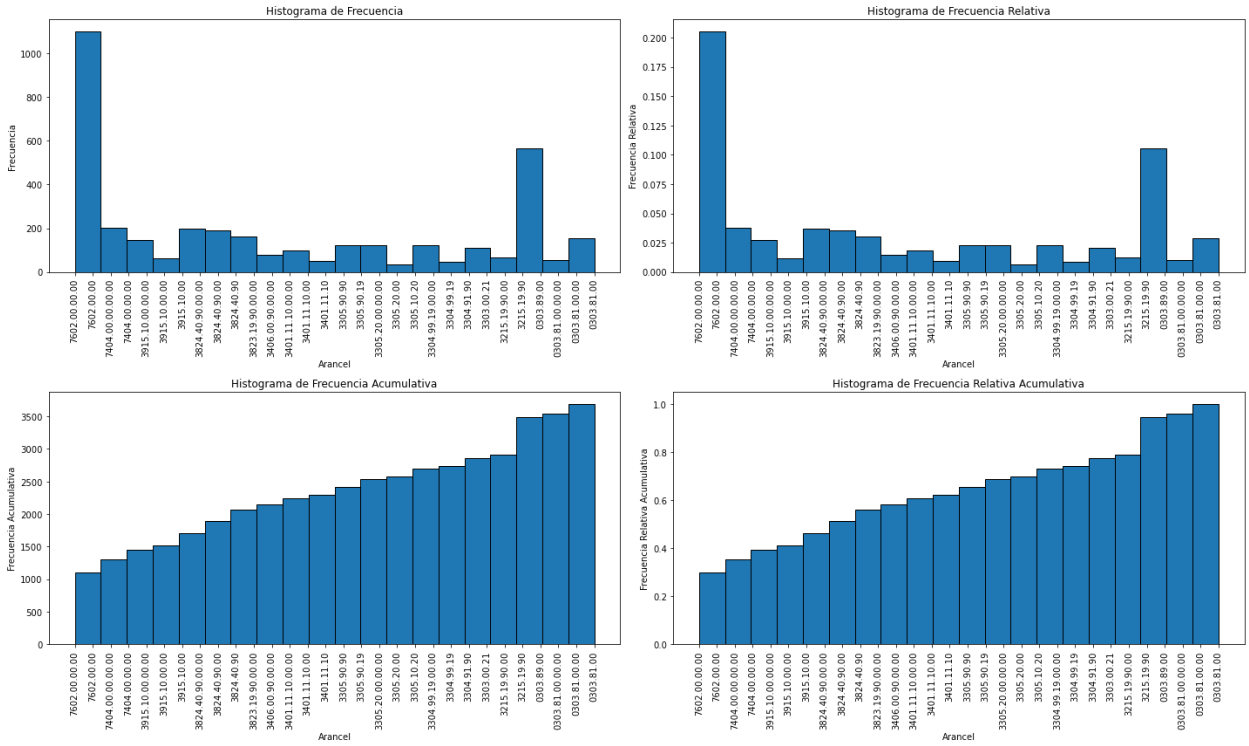
Composición del documento

- Mes: valor categórico (cadena de letras)
- Vía: valor categórico (cadena de letras)
- Arancel: valor categórico (cadena de números)
- Año: valor categórico (cadena de números)
- Codigopais: valor categórico (cadena de números)
- CodigoVia: valor categórico (cadena de números)
- MES: valor categórico (número)
- Valor FOB: valor numérico

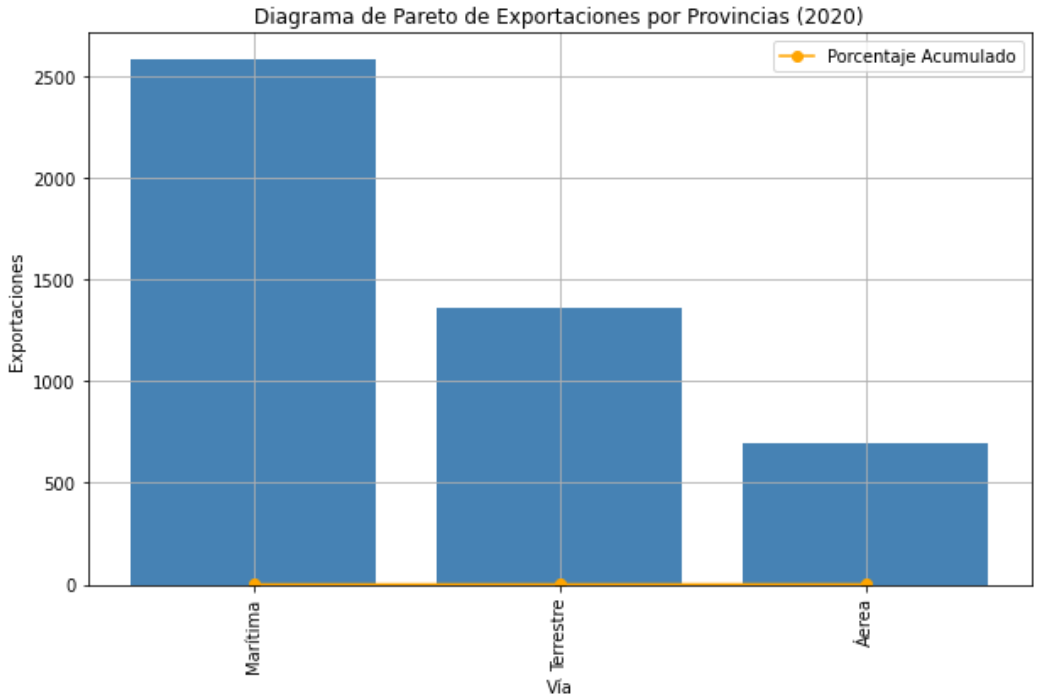
Plan de análisis

Como en el 2020 fue un año complicado para la industria del transporte, decidí centrarme en este y evaluar los productos más exportados asumiendo que los estos son los más importantes ya que los países recibidores abrían sus puertas a muchos riesgos al recibir dichos productos.

Histograma de frecuencia



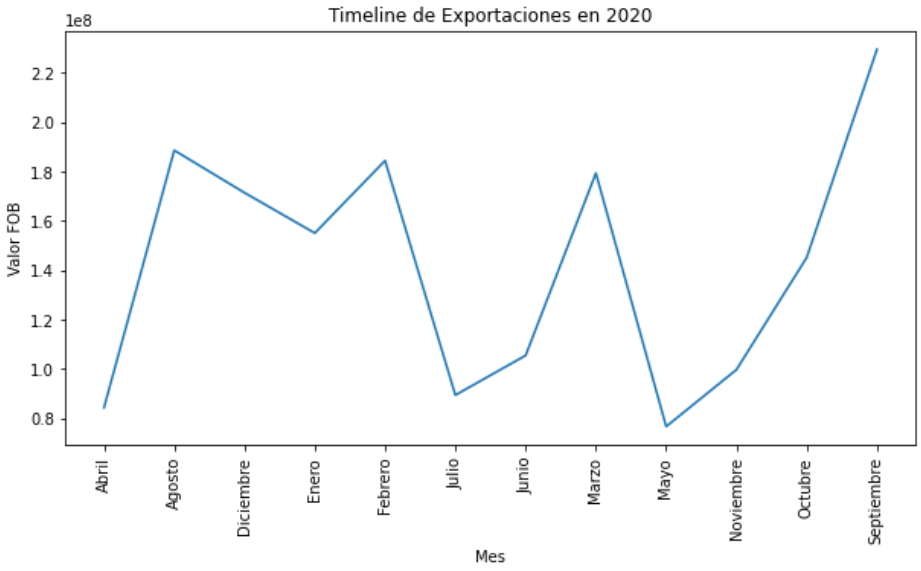
Cabe recalcar que para mejor entendimiento de los histogramas, solo se muestran los 30 aranceles más frecuentes.



La mayoría de mercancía entra por vía marítima, es un resultado esperado ya que el Canal de Panamá se encuentra en la provincia más común de los datos (Panamá).

Piñas tropicales (ananas) fueron el arancel más exportado durante el periodo 2010-2013 y volvieron a serlo en 2014-2017. Esto sugiere que las exportaciones de piñas fueron consistentes y dominantes en esos años, lo que indica una demanda estable o en crecimiento para este producto. Hubo un cambio en el arancel más exportado en 2018, donde las piñas fueron reemplazadas por café tostado sin descafeinar. Esto podría indicar una posible diversificación en los productos de exportación o un cambio en la demanda del mercado internacional. Sin embargo, en 2019, las piñas volvieron a ser el arancel más exportado. Esto podría indicar una recuperación en la demanda de este producto o una competencia más débil en comparación con otros aranceles. Desde 2020 hasta 2022, el arancel más exportado fue nuevamente el café tostado sin descafeinar. Esto sugiere una consolidación de la posición del café como un producto clave en las exportaciones, lo que puede estar relacionado con la calidad y la reputación del café producido en la región.

En resumen, podemos concluir que las exportaciones de piñas y café tostado sin descafeinar han sido importantes para el país durante estos años. Las piñas mantuvieron una posición destacada durante varios periodos, aunque hubo un cambio en 2018 cuando el café tomó el primer lugar. Estos datos pueden indicar la existencia de ventajas competitivas o fortalezas en la producción y exportación de estos productos en el país.



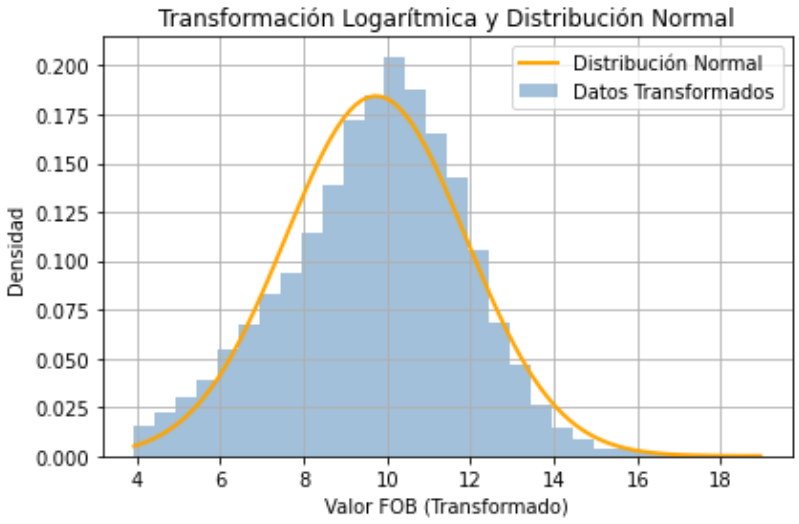
Medidas de dispersión de la columna Valor FOB

Media: 208587.90752205253

Varianza: 6499566378109.796

Desviación Estándar: 2549424.7151288455

Transformación a distribución normal



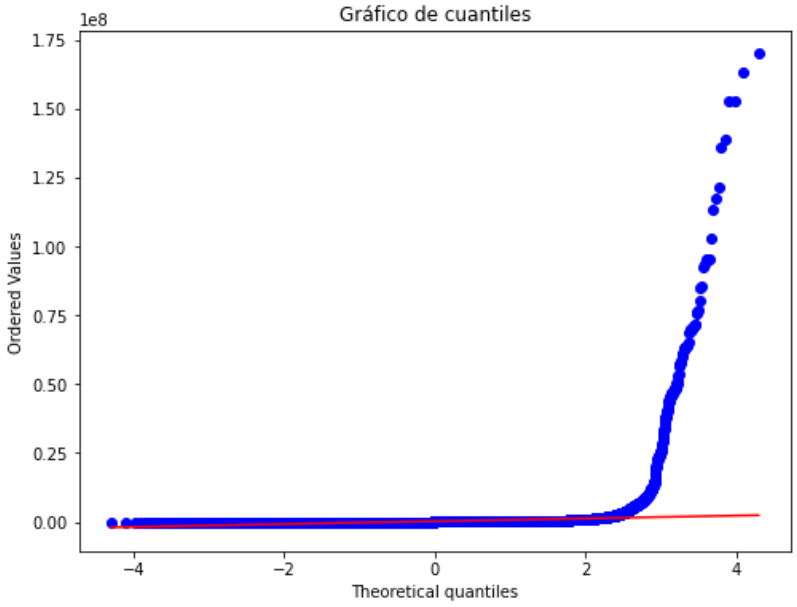
La transformación logarítmica resultó efectiva para este conjunto de datos.

Pruebas estadísticas

Estadística chi-cuadrado: 5905541.295915275

Valor p: 1.0

El resultado indica que no se encontró una asociación significativa entre las variables categóricas analizadas (por ejemplo, "Codigopais" y "Valor FOB"), debido a la alta estadística chi-cuadrado y al valor p igual a 1.0. Esto sugiere que las diferencias observadas en las frecuencias podrían deberse al azar y no a una relación real entre las variables.



Prueba t:

- Estadística t: 23.228429348802404

- Valor p: 5.794688283634502e-119

En este caso, la estadística t es de 23.228429348802404. Cuanto mayor sea el valor absoluto de la estadística t, mayor será la diferencia entre las medias de las muestras. En este caso, el valor p es extremadamente pequeño (5.794688283634502e-119), lo que sugiere que hay una diferencia significativa entre las medias de las muestras.

Prueba F:

- Estadística F: 539.363105029074

- Valor p: 4.0757830417075776e-119

En este caso, la estadística F es de 539.363105029074. Cuanto mayor sea el valor de la estadística F, mayor será la diferencia entre las varianzas de los grupos. En este caso, el valor p es extremadamente pequeño (4.0757830417075776e-119), lo que sugiere que hay una diferencia significativa en las varianzas entre los grupos.

Prueba de ANOVA:

- Valor F: 539.363105029074

- Valor p: 4.0757830417075776e-119

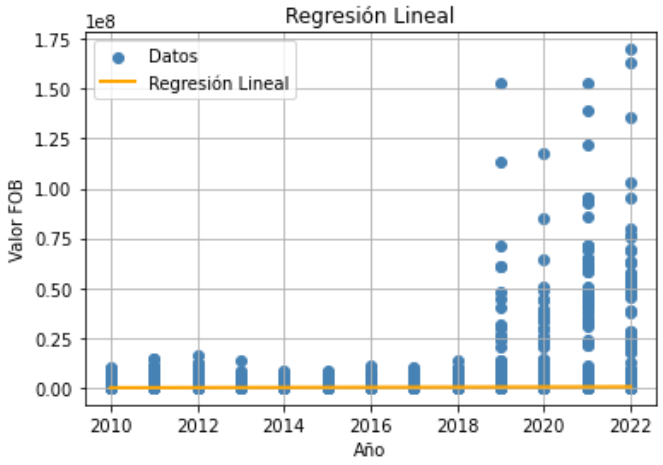
En este caso, el valor F es el mismo que en la prueba F, ya que se refiere a la misma comparación de varianzas. El valor p es también el mismo y representa la evidencia en contra de la hipótesis nula de igualdad de varianzas. Un valor p extremadamente pequeño (4.0757830417075776e-119) indica una diferencia significativa en las varianzas entre los grupos.

Regresión lineal

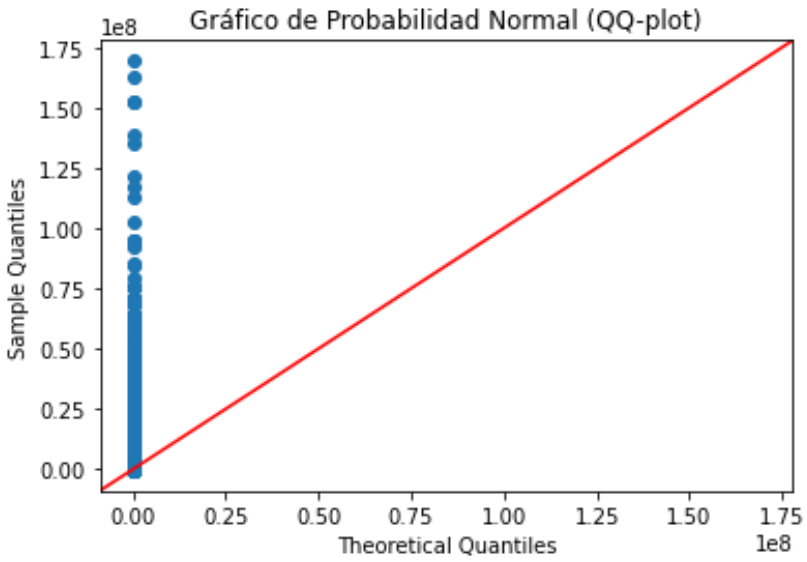
R-squared: 0.003

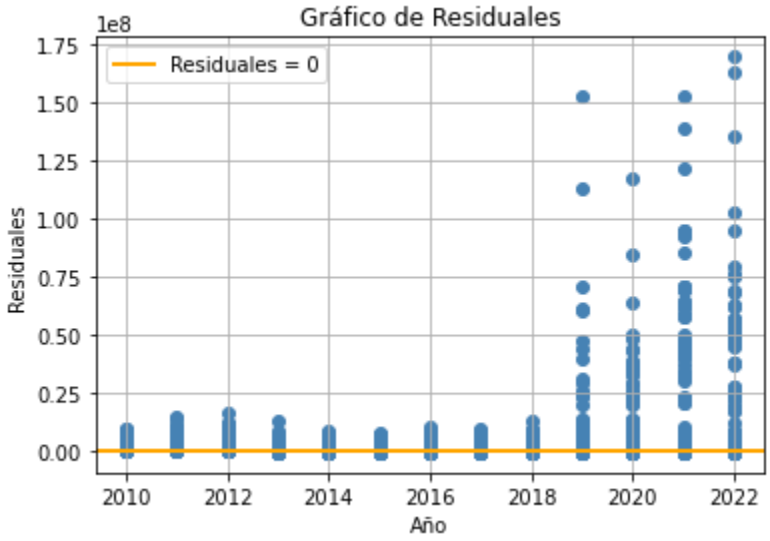
F-statistic: 254.7

Prob (F-statistic): 2.93e-57



El número de condición es grande, 1.09e+06. Esto podría indicar que existen problemas de multicolinealidad fuerte u otros problemas numéricos.





Con base en los intervalos de confianza para el intercepto y la pendiente del modelo de regresión lineal, se pueden obtener las siguientes conclusiones:

1. Intercepto (Valor constante):

- El intervalo de confianza para el intercepto se encuentra entre-8,651,477 y-6,754,499.

- Esto significa que con un 95% de confianza, el valor real del intercepto estará dentro de este rango.

- En el contexto del modelo, el intercepto representa el valor esperado de la variable dependiente (Valor FOB) cuando la variable independiente (Año) es igual a cero.

- Sin embargo, en este caso, dado que el valor de Año es una variable categórica con números, no tiene un significado real cuando se evalúa en cero, por lo que este resultado puede no ser interpretable de manera directa.

2. Pendiente (Coeficiente de Año):

- El intervalo de confianza para la pendiente está entre 33,616.005 y 43,027.862.

- Esto significa que con un 95% de confianza, el valor real de la pendiente estará dentro de este rango.

- La pendiente representa el cambio esperado en la variable dependiente (Valor FOB) por cada unidad de cambio en la variable independiente (Año).

- En este caso, el intervalo indica que, con un aumento de un año en la variable Año, se espera que el Valor FOB aumente en algún valor dentro de ese rango.

Exportaciones por países (BD3)

Composición del documento

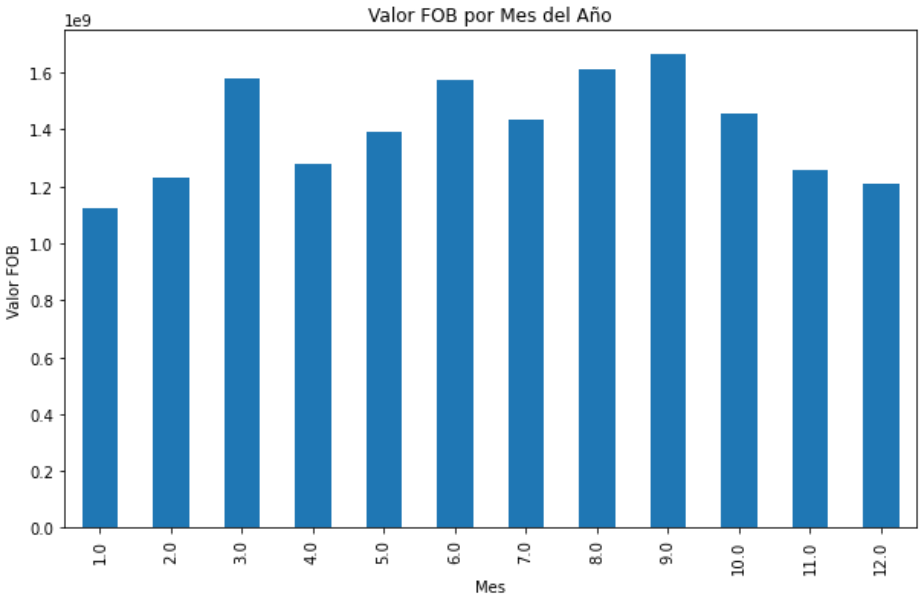
- Mes: valor categórico (cadena de letras)
- Años: valor categórico (cadena de números)
- Vía: valor categórico (cadena de letras)
- Arancel: valor categórico (cadena de números)
- Año: valor categórico (cadena de números)
- Codigopais: valor categórico (cadena de números)
- CodigoVia: valor categórico (cadena de números)
- MES: valor categórico (número)
- Valor FOB: valor numérico

Plan de análisis

Decidi analizar las tendencias temporales, es decir examina la evolución de las exportaciones a lo largo de los años y meses. Identificar patrones estacionales o tendencias a largo plazo en el valor FOB o en la cantidad de exportaciones.

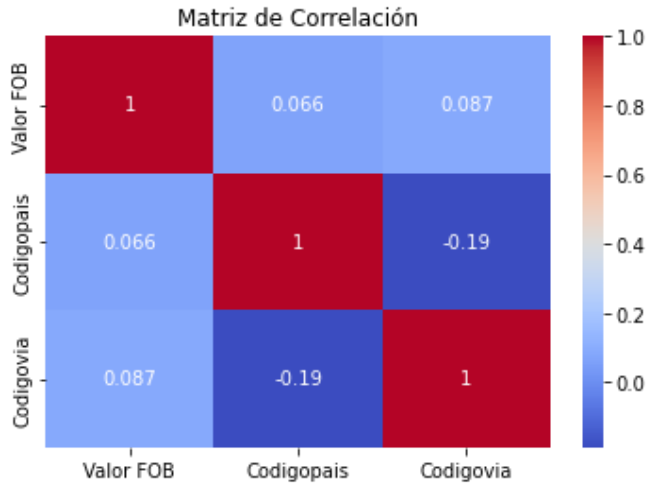
Además de explora posibles correlaciones entre variables. Por ejemplo, investigaré si hay una relación entre el valor FOB y el mes del año, el país de destino o el tipo de vía de exportación.

Análisis de variación del valor FOB



La temporada alta de exportaciones generalmente ocurre al final de verano, aproximadamente desde agosto hasta noviembre. En este gráfico confirmamos que en Panamá los meses de marzo, junio, agosto, septiembre y octubre son muy concurridos.

Análisis de correlación entre variables



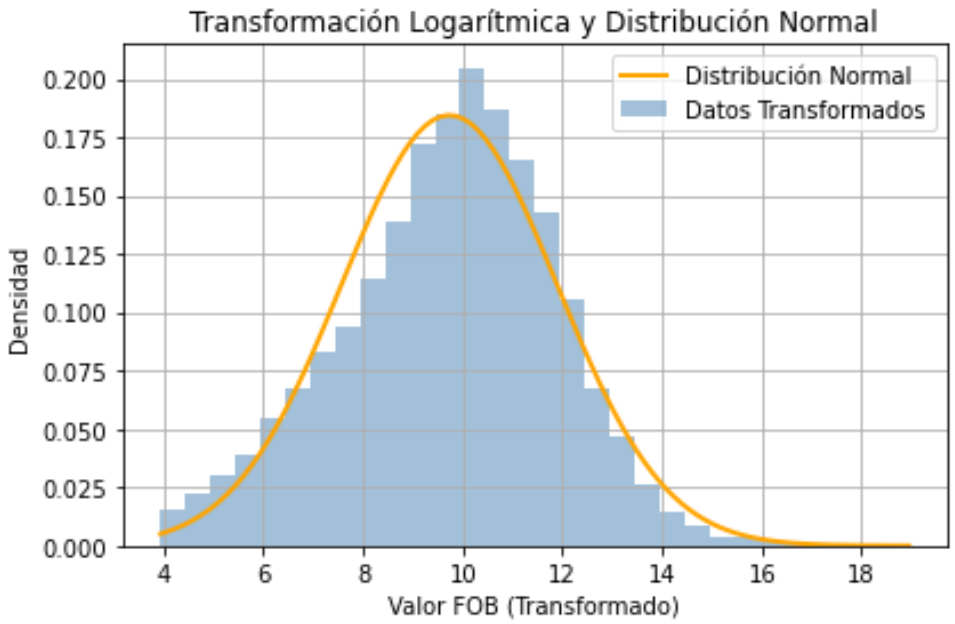
Medidas de dispersión del valor FOB

Media: 208463.63011627042

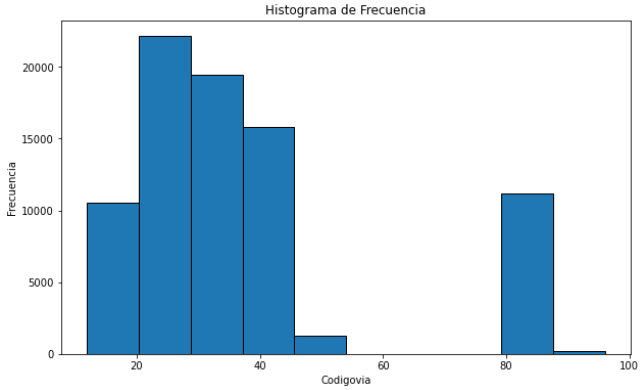
Varianza: 6493959025845.377

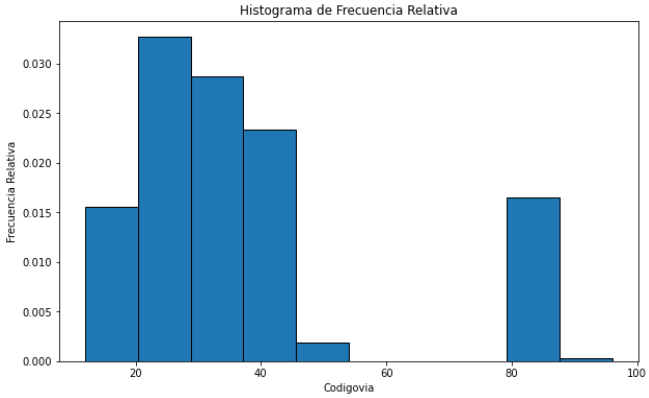
Desviación Estándar: 2548324.748897867

Transformación de los datos a distribución normal

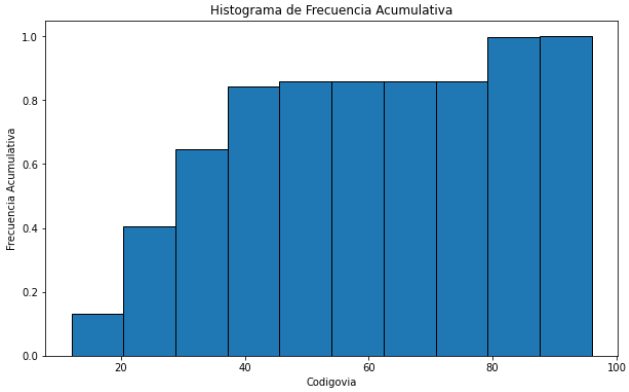


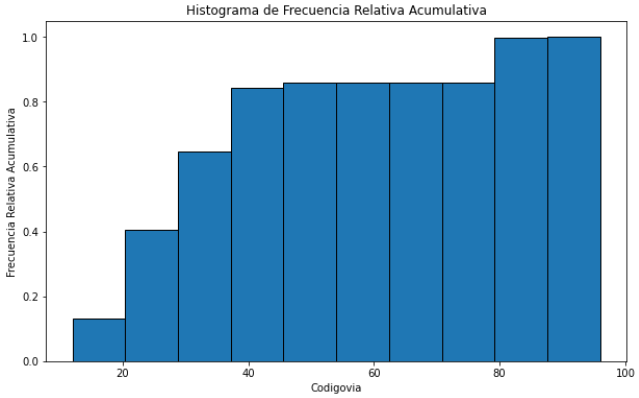
La transformación logarítmica resultó efectiva para este conjunto de datos.





Histogramas de frecuencia de código de vía





Regresión lineal

En la prueba t realizada, se obtuvo un estadístico t de 23.228429348802404 y un valor p extremadamente pequeño de 5.794688283634502e-119. Esto indica que hay una diferencia significativa en las medias de los grupos analizados. El estadístico t positivo sugiere que la media del primer grupo es mayor que la del segundo grupo.

Por otro lado, en la prueba F se obtuvo un estadístico F de 539.363105029074 y un valor p igualmente pequeño de 4.0757830417075776e-119. Esto también indica que hay una diferencia significativa en las varianzas entre los grupos.

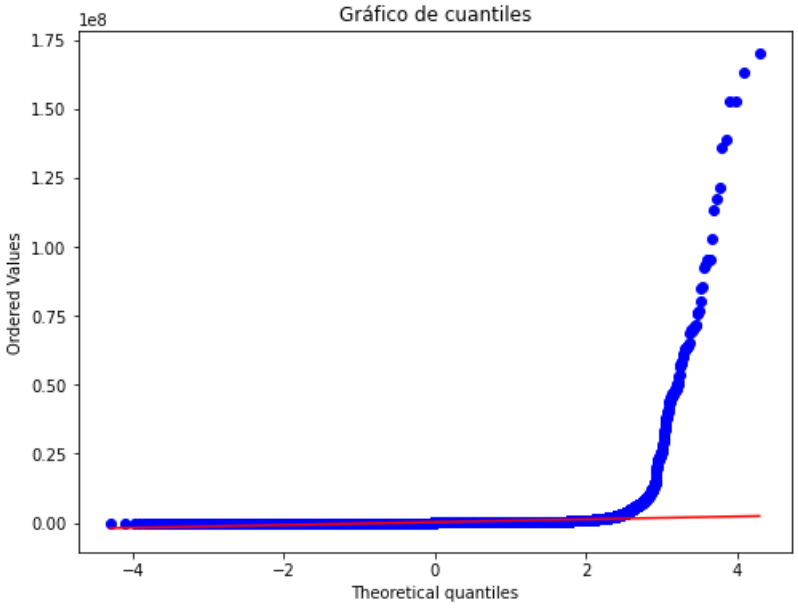
En ambos casos, los valores de p tan pequeños sugieren una fuerte evidencia en contra de las hipótesis nulas de igualdad de medias o igualdad de varianzas. Por lo tanto, podemos concluir que existen diferencias significativas en las medias y varianzas de los grupos analizados.

Estos resultados proporcionan información valiosa sobre las diferencias entre los grupos y pueden tener implicaciones importantes en la interpretación de los datos y la toma de decisiones basadas en ellos.

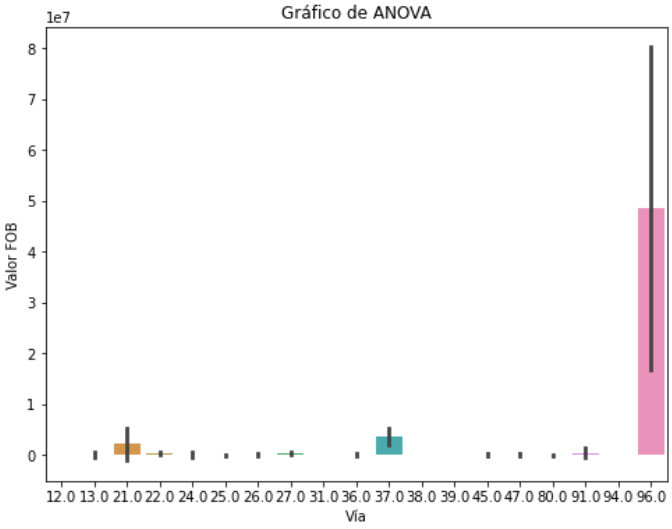
Estadístico Chi cuadrado: 7797.760578000269

Valor p: 4.755854653232651e-34

Los resultados indican que existe una asociación significativa entre los valores de 'Mes', 'Codigopais' y 'CodigoVia', ya que el valor del estadístico Chi cuadrado es alto y el valor p es muy pequeño. Esto implica que estas variables no son independientes entre sí y están correlacionadas en el conjunto de datos analizado.



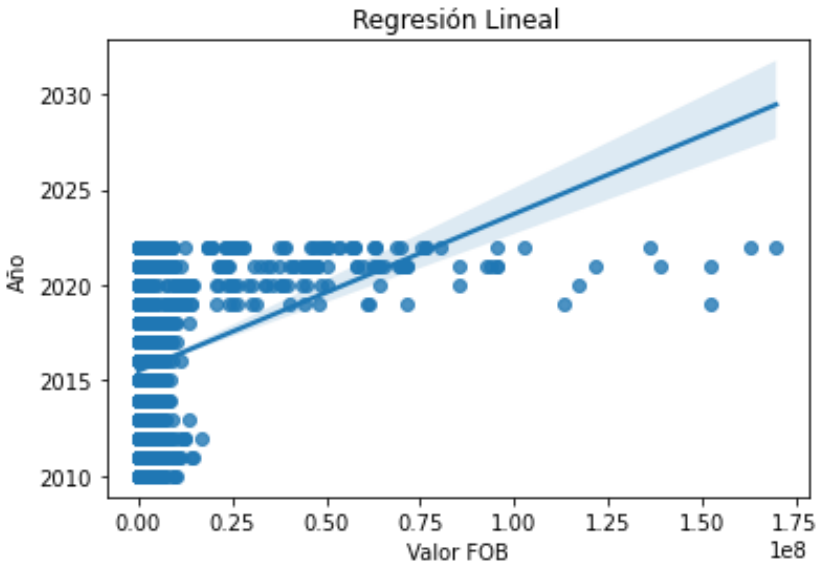
ANOVA



- Valor F: 539.363105029074. El valor F es el estadístico de la prueba de ANOVA y representa la relación entre la variabilidad entre los grupos y la variabilidad dentro de los grupos. Un valor F más grande indica una mayor diferencia entre las medias de los grupos en comparación con la variabilidad dentro de los grupos.

- Valor p: 4.0757830417075776e-119. El valor p es la probabilidad de obtener un valor de F igual o más extremo que el observado bajo la hipótesis nula de que no hay diferencias significativas entre las medias de los grupos. En este caso, el valor p es extremadamente pequeño, lo que sugiere que la diferencia observada entre las medias de los grupos es muy poco probable que se deba al azar. Por lo tanto, podemos rechazar la hipótesis nula y concluir que hay diferencias significativas entre las medias de los grupos.

En resumen, los resultados de la prueba de ANOVA indican que hay diferencias significativas en las medias entre los grupos analizados. El valor F alto y el valor p muy pequeño respaldan esta conclusión y sugieren que las diferencias observadas no son producto del azar. Estos resultados proporcionan evidencia sólida de que al menos una de las medias de los grupos es significativamente diferente de las demás.

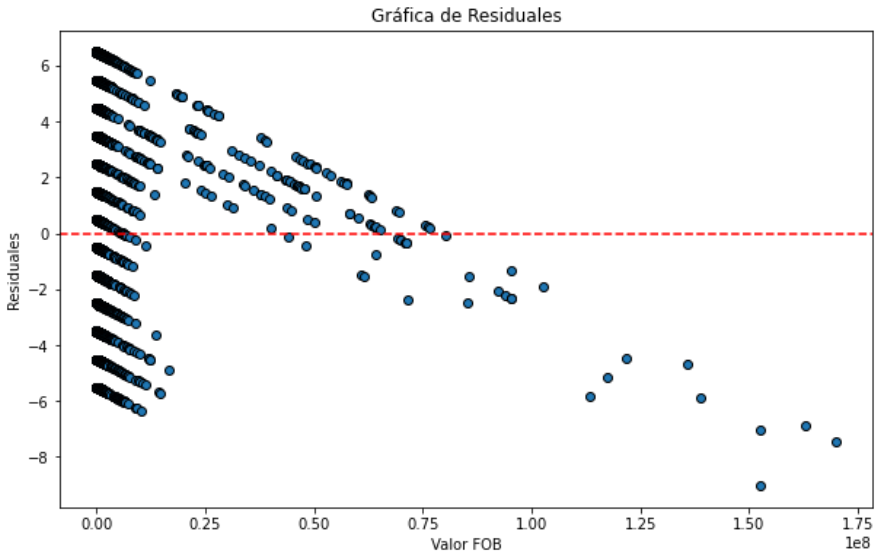


Pendiente: 8.209137155614173e-08. En este caso, la pendiente es un número muy pequeño y cercano a cero, lo que indica una relación muy débil entre las variables.

Intercepto: 2015.504504804134. Es el valor de la variable dependiente cuando la variable independiente es igual a cero. En este caso, indica el valor inicial de la variable dependiente.

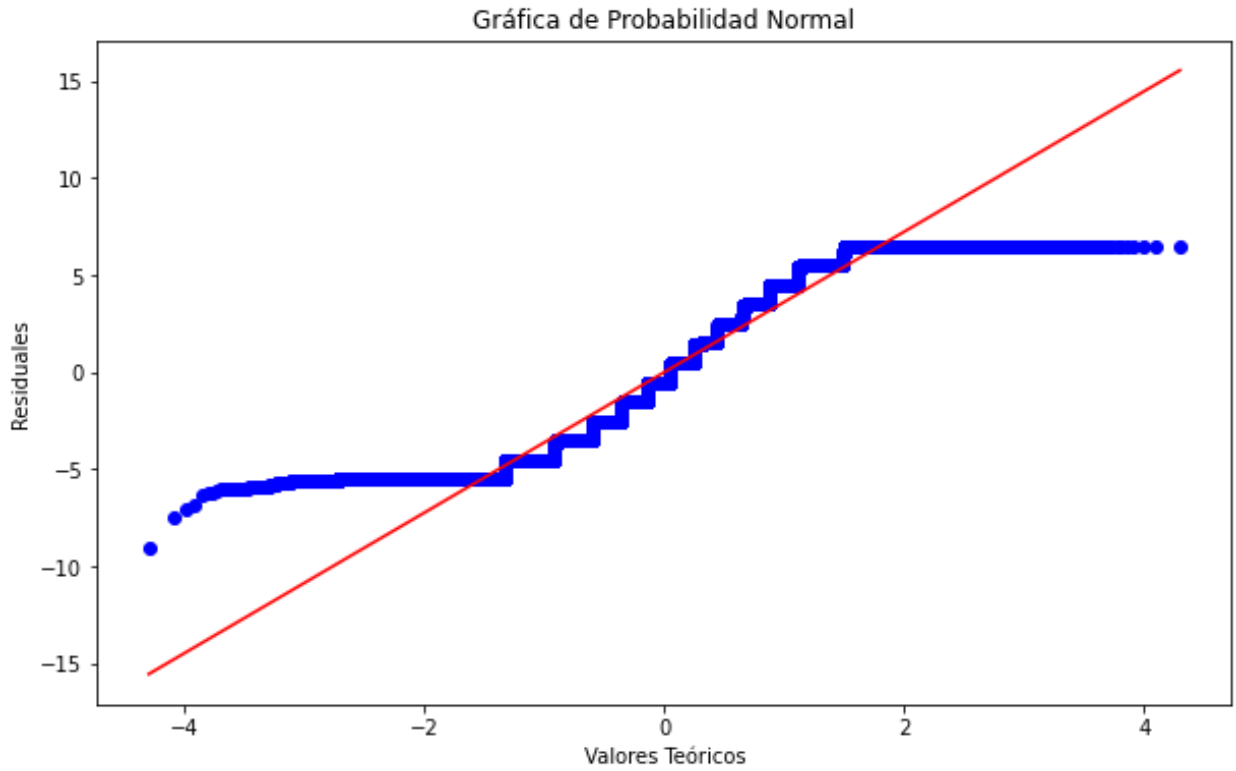
Coeficiente de Correlación: 0.05597180404319003. Un coeficiente de correlación cercano a cero indica una correlación débil.

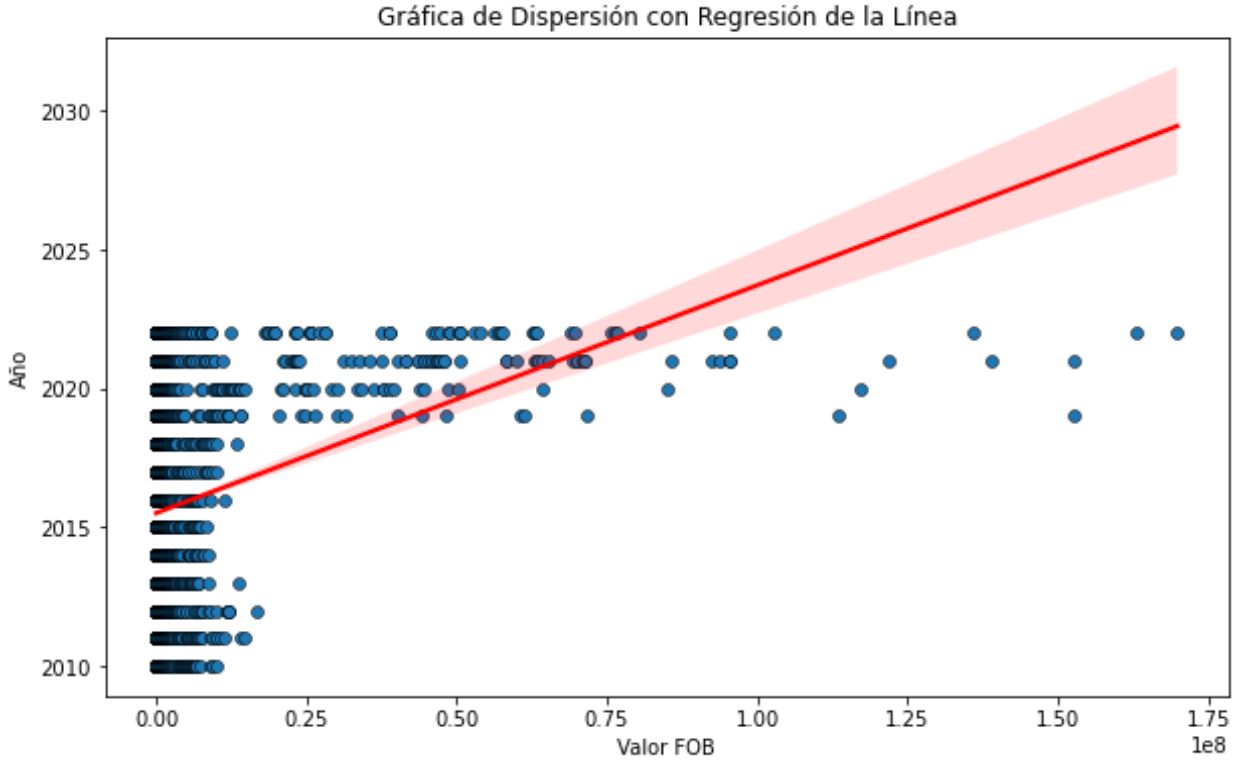
Valor p: 5.40204232671267e-57. Un valor p muy pequeño indica que la correlación observada no es probable que sea debido al azar, y por lo tanto, podemos concluir que hay una relación significativa entre las variables.



Intervalo de confianza para la pendiente: (7.198628481604866e-08, 9.21964582962348e-08)
Esto significa que podemos estar 95% seguros de que la pendiente de la regresión lineal está dentro de este rango. En otras palabras, el valor real de la pendiente tiene una alta probabilidad de estar entre 7.198628481604866e-08 y 9.21964582962348e-08. Este intervalo indica que hay una relación positiva entre las variables, es decir, a medida que aumenta el valor de 'Valor FOB', también se espera un aumento en el valor de 'Año'.

Intervalo de confianza para el intercepto: (2015.504504794029, 2015.5045048142392)
Esto indica que podemos estar 95% seguros de que el intercepto de la regresión lineal está dentro de este rango. En otras palabras, el valor real del intercepto tiene una alta probabilidad de estar entre 2015.504504794029 y 2015.5045048142392. El intercepto representa el valor de 'Año' cuando 'Valor FOB' es igual a cero. En este caso, el rango del intervalo está muy cerca, lo que sugiere que el año base para el valor de 'Valor FOB' es probablemente alrededor de 2015.


Gráfica de Probabilidad Normal
Residuales
Valores Teóricos
15
10
5
0
−5
−10
−15
−4
−2
0
2
4


Gráfica de Dispersión con Regresión de la Línea
Año
Valor FOB
1e8
2030
2025
2020
2015
2010
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

Conclusión

En conclusión, este proyecto de estadística ha utilizado una variedad de técnicas y herramientas para explorar y comprender a fondo un conjunto de datos. Mediante la visualización de histogramas, se ha examinado la distribución de los datos, lo que ha proporcionado una comprensión general de su forma y dispersión. Asimismo, al emplear el diagrama de Pareto, se han identificado y priorizado las categorías o variables más relevantes del conjunto de datos, destacando aquellas que tienen una contribución significativa.

Además, se han realizado pruebas estadísticas, tales como la prueba t de dos muestras y la prueba de chi-cuadrado, para evaluar las diferencias en las medias y frecuencias observadas, respectivamente. Estas pruebas han permitido analizar la significancia de las diferencias encontradas, brindando información valiosa acerca de las relaciones existentes entre las variables investigadas.

En cuanto al análisis de regresión lineal, se ha examinado la relación entre las variables y se han realizado predicciones de valores futuros. Mediante la determinación de la pendiente e intercepto de la línea de regresión, así como la evaluación del coeficiente de correlación, se ha logrado comprender mejor las interacciones entre las variables y su impacto en el fenómeno estudiado.

Por último, se han calculado medidas descriptivas, como la media, varianza y desviación estándar, con el propósito de caracterizar y resumir los datos. Estas medidas han proporcionado información clave sobre las características centrales y la dispersión de los datos.

En resumen, a través de la implementación de estas diversas técnicas y herramientas estadísticas, se ha obtenido una comprensión detallada del conjunto de datos analizados. Esto ha permitido tomar decisiones fundamentadas y obtener conclusiones sólidas, con implicaciones significativas en la toma de decisiones dentro del contexto del fenómeno estudiado.

Anexo

Enlace a mi cuenta de Github: https://github.com/yinelabryant/proyectoestadistica

Enlace a BD1: INEC - Dashboard. (n.d.). https://www.inec.gob.pa/DASHBOARDS/Sociales/TransitoConductores

Enlace a BD2: *INEC - Dashboard*. (n.d.-c). https://www.inec.gob.pa/DASHBOARDS/Comercio/ExportacionesporAranceles

Enlace a BD3: *INEC - Dashboard*. (n.d.-d). https://www.inec.gob.pa/DASHBOARDS/Comercio/ExportacionesPorPaises

Enlace a BD4: INEC- Dashboard. (n.d.-b). https://www.inec.gob.pa/DASHBOARDS/Sociales/Transito

Enlace a BD5: *INEC - Dashboard*. (n.d.-e). https://www.inec.gob.pa/DASHBOARDS/Comercio/ImportacionesporZonasFrancas

Enlace a BD6: *INEC - Dashboard*. (n.d.-h). https://www.inec.gob.pa/DASHBOARDS/IMAE/INDICADORES_ECONOMICOS_IMP_CUODE

Enlace a BD7: *INEC - Dashboard*. (n.d.-g). https://www.inec.gob.pa/DASHBOARDS/PIB/PIB_ANUAL_POR_RAMA

Enlace a BD8: *INEC - Dashboard*. (n.d.-f). https://www.inec.gob.pa/DASHBOARDS/Comercio/ImportacionesporPaises

Enlace a BD9: 80 Cereals. (2017, October 24). Kaggle. https://www.kaggle.com/datasets/crawford/80-cereals

Enlace a BD10: 1000 Cameras Dataset. (2017, October 24). Kaggle. https://www.kaggle.com/datasets/crawford/1000-cameras-dataset