Logotipo

Descripción generada automáticamente con confianza mediaImagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza bajaUniversidad 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:

Juan Marcos Castillo PhD.

Nombre:

José Luis Camarena (8-990-353)

Grupo:

1SF-131

Fecha de entrega:

26-7-2023

# INTRODUCCIÓN

En este trabajo, nos embarcaremos en un análisis profundo de diez bases de datos diversas, aplicando los conocimientos fundamentales impartidos en el curso de estadística. Nuestro objetivo es explorar y comprender patrones, relaciones y tendencias ocultas en estos conjuntos de datos para obtener valiosas conclusiones y tomar decisiones informadas. A través de una metodología rigurosa y herramientas estadísticas adecuadas, examinaremos detalladamente cada base de datos, destacando la relevancia y aplicabilidad de los conceptos estadísticos aprendidos en nuestro recorrido académico. Con este enfoque analítico, esperamos obtener una perspectiva clara y precisa sobre la naturaleza de los datos y su impacto en diferentes contextos, fomentando así una mayor comprensión de la estadística como una herramienta esencial para la toma de decisiones en diversas áreas.

# Producto Interno Bruto Anual por Provincia

Fuente: https://www.inec.gob.pa/DASHBOARDS/PIB/PIB\_ANUAL\_POR\_PROVINCIA

Composición del documento:

* Código Provincia: valor categórico
* Año: valor categórico
* Provincia: valor categórico
* Composición constante: valor numérico
* Composición corriente: valor numérico
* Per Cápita constante: valor numérico
* Per Cápita corriente: valor numérico
* Valor constante: valor numérico
* Valor Corriente: valor numérico

Plan de análisis:

Analizaré el valor constante por provincia a lo largo de los años, tomando una muestra desde el 2014 en adelante, ya que la misma incluye la provincia de Panamá Oeste.

Histograma de frecuencia de “Valor Constante” por provincia



Con la observación de este gráfico, podemos observar que Panamá tiene un valor constante mucho mayor a lo largo de los años que las demás provincias del país.

Analizando con más detalle la provincia de Panamá, al ser la que más destaca sobre el valor constante, tenemos lo siguiente:

Media de Valor Constante: 24735.89506

Varianza de Valor Constante: 3535581.53

Desviación Estándar de Valor Constante: 1880.314211

Histogramas de frecuencia

Análisis a la base de datos

1. ¿Qué provincia, sin tomar en cuenta a Panamá, tiene mayor valor constante?

Para hallar solución a esta pregunta, debemos tomar el gráfico presentado previamente y eliminar a Panamá de la lista, para así poder llegar a una respuesta.

Gráfico de Frecuencia de Valor Constante por Provincia sin tomar en cuenta Panamá

Notamos que Colón es la provincia más prominente en lo que a Valor Constante respecta.

1. ¿En qué año Panamá tuvo su pico de Valor Constante?

Una pregunta que surge, al ver más a fondo el valor Constante de Panamá es cuando alcanzó su punto más alto en los distintos años, la respuesta se puede apreciar en la siguiente gráfica.

Podemos observar que el punto más alto del valor constante fue en 2019, sin embargo, algo que también resulta llamativo es que en 2020 fue su punto más bajo, probablemente debido a los efectos de la pandemia del COVID-19.

1. ¿En qué escenario, Panamá se encuentra más igualado con otras provincias?

Revisando las distintas gráficas, me surgió la curiosidad de ver en qué caso, Panamá se encontraba con valores más cercanos a otras provincias.

Aquí podemos notar que en el valor Per Cápita Corriente, la provincia de Panamá y Colón, se encuentran bastante igualadas, pero aún, muy alejadas de las demás provincias.

1. ¿Cuál es la relación entre composición corriente y composición constante?

A continuación, se muestra una gráfica para representar la regresión lineal realizada entre la composición corriente y la composición constante.

|  |  |
| --- | --- |
| *Estadísticas de la regresión* | |
| Coeficiente de correlación múltiple | 0.99993539 |
| Coeficiente de determinación R^2 | 0.99987078 |
| R^2 ajustado | 0.99986888 |
| Error típico | 0.21505224 |
| Observaciones | 70 |

Estos resultados indican que la regresión lineal entre la composición corriente y la composición constante es altamente significativa y que el modelo de regresión explica prácticamente toda la variabilidad en la composición corriente. Es probable que la composición constante sea un predictor muy fuerte de la composición corriente, y el modelo tiene un alto poder predictivo en función de los valores ajustados y ajustados del coeficiente de determinación (R^2).

A continuación, los intervalos de confianza

|  |  |
| --- | --- |
|  | Coeficientes |
| Intercepción | 0.020456863 |
| Composición Corriente | 1.00048838 |

Gráficos de Regresión Lineal

ANOVA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Grados de libertad | Suma de cuadrados | Promedio de los cuadrados | F | Valor crítico de F |
| Regresión | 1 | 24334.52228 | 24334.52228 | 526180.6793 | 5.8739E-134 |
| Residuos | 68 | 3.144827585 | 0.046247464 |  |  |
| Total | 69 | 24337.6671 |  |  |  |

El valor F calculado es 526180.6793 y el valor crítico de F es 5.8739E-134. Dado que el valor F calculado (526180.6793) es significativamente mayor que el valor crítico de F (5.8739E-134), podemos rechazar la hipótesis nula y afirmar que la regresión es altamente significativa.

En otras palabras, hay evidencia estadística suficiente para concluir que existe una relación significativa entre la composición corriente y la composición constante en base a este análisis de varianza (ANOVA). La regresión entre ambas variables es estadísticamente significativa y, por lo tanto, el modelo de regresión es adecuado para explicar la relación entre ellas.

# Cáncer-Muertes por Estado

Fuente:<https://data.world/nrippner/cancer-linear-regression-modeltutorial/workspace/file?filename=death+.csv>

Composición del documento:

* County: valor categórico
* FIPS: valor categórico
* Met Objective of 45.5: valor categórico
* Age-Adjusted Death Rate: valor numérico
* Lower 95% Confidence interval for Death Rate: valor numérico
* Upper 95% Confidence Interval for Death Rate: valor numérico
* Average Deaths per Year: valor numérico
* Recent Trend (2): valor categórico
* Recent 5-Year Trend (2) in Death Rates: valor numérico
* Lower 95% Confidence Interval for Trend: valor numérico
* Upper 95% Confidence Interval for Trend: valor numérico

Plan de análisis:

Analizaré la siguiente base de datos par descubrir cuáles valores están fuertemente relacionados, además de ver cuál es la tendencia de las enfermedades en los distintos Death Rates en Estados Unidos.

Media de Age-Adjusted Death Rate: 53.18514042

Varianza de Age-Adjusted Death Rate: 196.9530793

Desviación Estándar de Age-Adjusted Death Rate: 14.03150255

Histogramas de frecuencia

Análisis a la base de datos

1. ¿Cuál es la tendencia general entre todos los estados?

Para hallar solución a esta pregunta, realizamos un conteo de la categoría de tendencia y obtenemos el siguiente gráfico:

Gráfico de la suma de las distintas tendencias

Podemos resaltar que “stable” es la tendencia que más se repite en los distintos estados, por una gran mayoría.

1. ¿La mayoría de los estados lograron llegar a la meta de bajar la tasa de mortalidad al 45.5?

Para responder esta pregunta, realizamos algo similar al punto anterior, contando todas las ocasiones en las que se cumplió la meta.

Nos percatamos de que la mayoría de los estados no lograron llegar a la meta de tener una tasa de mortalidad de 45.5.

1. ¿Cuál es la relación entre Age-Adjusted Death Rate y Upper 95% Confidence Interval for Death Rate?

A continuación, se muestra una gráfica para representar la regresión lineal realizada entre dichas variables.

|  |  |
| --- | --- |
| *Estadísticas de la regresión* | |
| Coeficiente de correlación múltiple | 0.91501503 |
| Coeficiente de determinación R^2 | 0.8372525 |
| R^2 ajustado | 0.8371946 |
| Error típico | 7.51207879 |
| Observaciones | 2813 |

El valor del coeficiente de correlación múltiple es 0.91501503. Esto indica una relación positiva bastante fuerte entre el "Age-Adjusted Death Rate" y el "Upper 95% Confidence Interval for Death Rate". Un valor cercano a 1 sugiere que ambas variables están estrechamente relacionadas y que la tasa de mortalidad ajustada por edad tiende a aumentar a medida que el intervalo de confianza superior de la tasa de mortalidad también aumenta.

A continuación, los intervalos de confianza

|  |  |
| --- | --- |
|  | *Coeficientes* |
| Intercepción | 1.70959342 |
| Age-Adjusted Death Rate | 1.21387111 |

Gráficos de Regresión Lineal

ANOVA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Grados de libertad | Suma de cuadrados | Promedio de los cuadrados | F | Valor crítico de F |
| Regresión | 1 | 24334.52228 | 24334.52228 | 526180.6793 | 5.8739E-134 |
| Residuos | 68 | 3.144827585 | 0.046247464 |  |  |
| Total | 69 | 24337.6671 |  |  |  |

El análisis de varianza (ANOVA) indica que existe una relación estadísticamente significativa entre la tasa de mortalidad ajustada por edad y el intervalo de confianza superior de la tasa de mortalidad. El modelo de regresión es adecuado y explica una gran parte de la variabilidad observada en la tasa de mortalidad ajustada por edad a través de su relación con el intervalo de confianza superior.

# Exportaciones principales mercaderías

Fuente:<https://www.inec.gob.pa/DASHBOARDS/Comercio/ExportacionesPrincipalesMercaderias>

Composición del documento

* Grupo: Datos Categórico
* Suprimida: Dato Categórico
* Año: Dato categórico
* Mes: Dato categórico
* Arancel: Dato categórico
* Código país: Dato categórico
* Continente: Dato categórico
* Peso Bruto: Dato numérico
* Peso Neto: Dato numérico
* Valor FOB: Dato numérico

Plan de Análisis:

El análisis estadístico de esta base de datos se basa en un enfoque impulsado por la curiosidad con relación al comercio internacional de Panamá.

Análisis de las Bases de Datos

1. ¿Qué continente exporta la mayor cantidad de peso neto?

En los siguientes gráficos mostrados se muestra una gran prevalencia del peso neto en las exportaciones en el continente africano, seguido de América en segundo lugar y Oceanía en tercer lugar.

|  |  |
| --- | --- |
| **Etiquetas de fila** | **Suma de Peso neto** |
| Africa | 102409444 |
| América | 32253078 |
| Asia | 4115655 |
| Oceania | 22028278 |
| **Total general** | **160806455** |

Gráfico, Gráfico de barras

Descripción generada automáticamente

A continuación, podemos observar que las exportaciones de la mercadería 1511.10 se dan con mayor frecuencia en América y representan más de la mitad del total de exportaciones de dicha mercadería durante los últimos años. Debido a este hecho, el análisis realizado a continuación se centrará exclusivamente en el continente americano, sin considerar a los demás continentes.

|  |  |
| --- | --- |
| **Etiquetas de fila** | **Recuento de Continente** |
| Africa | 75 |
| América | 140 |
| Asia | 4 |
| Oceania | 8 |
| **Total general** | **227** |

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico circular

Descripción generada automáticamente

1. ¿Cómo se ha comportado la exportación de esta mercadería? ¿Cómo se comporta el Peso Neto y el valor FOB de este arancel?

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Al examinar los gráficos de las exportaciones de la mercadería 1511.10, se puede observar que tanto el Peso Neto exportado como el Valor FOB alcanzan sus niveles más altos en el año 2018. Sin embargo, a partir del año 2019, se observa una disminución en ambos valores. Esta tendencia descendente se acentúa en el año 2020, siendo este último año el punto más bajo tanto para el Peso Neto como para el Valor FOB. Estos datos sugieren una disminución en las exportaciones de la mercadería 1511.10 a partir del año 2019, con una notable caída en el año 2020.

1. ¿Promedio de valor FOB pagado en el periodo estudiado?

A continuación, se presenta un análisis de los datos numéricos del valor FOB en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| **Media** | **Mediana** | **Varianza** | **Desviación Estándar** |
| 454436.537 | 174125 | 4.87872E+11 | 698478.13 |

Distribución normal de la columna Valor FOB.

Gráfico, Histograma

Descripción generada automáticamente

Al observar el gráfico, se puede apreciar que los datos estudiados del valor FOB siguen la distribución de campana de una distribución normal.

1. ¿Promedio de Peso Bruto importado en el periodo estudiado?

A continuación, se presenta un análisis de los datos numéricos del peso bruto en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| **Media** | **Mediana** | **Varianza** | **Desviación Estándar** |
| 713163.242 | 250750 | 1.34314E+12 | 1158937.60 |

Distribución normal de la columna Peso Bruto.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

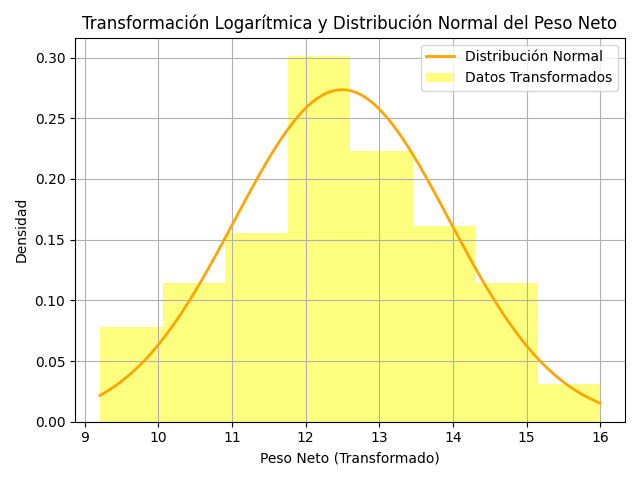
Al observar el gráfico, se puede apreciar que los datos estudiados del Peso Bruto siguen la distribución de campana de una distribución normal.

1. ¿Promedio de Peso Neto importado en el periodo estudiado?

A continuación, se presenta un análisis de los datos numéricos del peso neto en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| **Media** | **Mediana** | **Varianza** | **Desviación Estándar** |
| 708398.48 | 250000 | 1.32922E+12 | 1152916.85 |

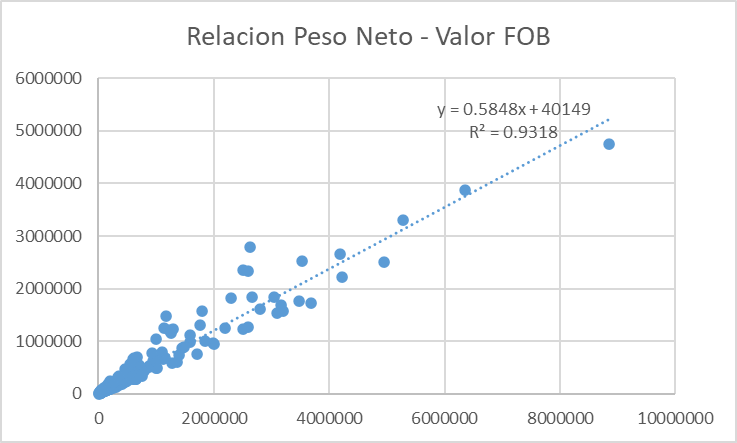
Distribución normal de la columna Peso Neto.



Al observar el gráfico, se puede apreciar que los datos estudiados del Peso Neto siguen la distribución de campana de una distribución normal.

1. ¿Cuál es la relación entre el Peso Neto y valor FOB?

En el gráfico presentado a continuación se muestra la representación gráfica de la regresión lineal que relaciona el Peso Neto y el Valor FOB.



|  |  |
| --- | --- |
| *Estadísticas de la regresión* | |
| Coeficiente de correlación múltiple | 0.96531609 |
| Coeficiente de determinación R^2 | 0.93183515 |
| R^2 ajustado | 0.93153219 |
| Error típico | 182766.256 |
| Observaciones | 227 |

En este caso, se obtuvo un coeficiente de correlación de 0.965 lo que sugiere una correlación fuerte y positiva entre el Peso Neto y el Valor FOB. Además, el valor de R^2 de 0.932 es muy cercano a 1, lo que indica una correlación significativa entre el Peso Neto y el Valor FOB. Esto implica que alrededor del 93.2% de las fluctuaciones en el Valor FOB pueden ser atribuidas a las variaciones en el Peso Neto lo que significa que el Peso Neto es un muy buen indicador o predictor del Valor FOB dentro del marco de la regresión lineal utilizada.

A continuación, los intervalos de confianza para la intersección y la variable x1.

|  |  |
| --- | --- |
|  | *Coeficientes* |
| Intercepción | 40148.887 |
| Variable X 1 | 0.5848229 |

|  |  |  |
| --- | --- | --- |
| Intervalo de confianza B0 | | |
| 12226.4107 | <B0 | 68071.3632 |

|  |  |  |
| --- | --- | --- |
| Intervalo de confianza B1 | | |
| 0.56415483 | <B1 | 0.60549096 |

El intervalo de confianza al 95% para la intercepción (β0) y el coeficiente de x1 (β1) nos proporciona una estimación que sugiere que, en caso de tomar numerosas muestras distintas y calcular el intervalo de confianza para cada una de ellas, aproximadamente el 95% de dichos intervalos abarcarían los valores reales de la población correspondientes a la intercepción y el coeficiente de x1.

También obtenemos el grafico de residuales y el grafico de probabilidad normal para analizar mejor la regresión.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

ANOVA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Grados de libertad* | *Suma de cuadrados* | *Promedio de los cuadrados* | *F* | *Valor crítico de F* |
| Regresión | 1 | 1.0274E+14 | 1.0274E+14 | 3075.82143 | 3.282E-133 |
| Residuos | 225 | 7.5158E+12 | 3.3404E+10 |  |  |
| Total | 226 | 1.1026E+14 |  |  |  |

El análisis de varianza realizado en el modelo de regresión reveló una fuerte correlación positiva entre el Peso Neto y el Valor FOB, indicando que a medida que el Peso Neto aumenta, también tiende a aumentar el Valor FOB de las importaciones.

Se rechaza la hipótesis nula debido a que se obtuvo un valor de la estadística F de 3075,82 y este es mayor que el valor crítico de F, esto significa que hay evidencia estadística suficiente para afirmar que existe una relación significativa entre el Valor FOB (variable dependiente) y el Peso Neto (variable independiente), y por tanto el modelo tiene una capacidad significativa para predecir el Valor FOB en función del Peso Neto.

Estos hallazgos respaldan la relevancia del peso neto como factor influyente en el valor FOB de las importaciones realizadas en las zonas francas.

# Exportaciones zonas francas

Fuente:<https://www.inec.gob.pa/DASHBOARDS/Comercio/ExportacionesporZonasFrancas>

Composición del documento:

* Año: Dato categórico
* Continente: Dato categórico
* Vía: Dato categórico
* Zonas francas: Dato categórico
* Puerto de desembarque: Dato categórico
* Mes: Dato categórico
* Arancel: Dato categórico
* Valor FOB: Dato numérico
* Peso Neto: Dato numérico

Plan de análisis:

El enfoque del análisis estadístico de esta base de datos se basa en una perspectiva curiosa acerca del comercio internacional de Panamá, con un enfoque específico en el estudio de aspectos relevantes sobre las zonas francas y su influencia en el panorama exportador del país.

Análisis de las Bases de Datos

1. ¿Qué zona franca exporta la mayor cantidad de peso neto?

En el gráfico adjunto se observa que la Zona Franca Panamá Pacífico presenta la mayor cantidad de peso neto en las exportaciones del arancel analizado en los últimos años, lo cual destaca como una diferencia significativa en comparación con las demás zonas francas.

|  |  |
| --- | --- |
| **Etiquetas de fila** | **Suma de Peso neto** |
| Área Económica Especial Panamá Pacífico | 215237 |
| Chilibre | 1039 |
| PanaPark Free Zone | 25 |
| VAGUIL | 6 |
| **Total general** | **216307** |

Gráfico, Gráfico de barras

Descripción generada automáticamente

El siguiente gráfico muestra que las exportaciones del arancel 7117.19.00.00 en la Zona Franca de Panamá Pacífico representan el 94% del total de exportaciones de dicho arancel durante los años 2017-2021. Debido a este hecho, el análisis a continuación se centrará exclusivamente en esta Zona Franca, sin considerar las demás.

|  |  |
| --- | --- |
| **Etiquetas de fila** | **Recuento de Zonas Francas** |
| Área Económica Especial Panamá Pacífico | 169 |
| Chilibre | 6 |
| PanaPark Free Zone | 1 |
| VAGUIL | 3 |
| **Total general** | **179** |

Gráfico, Gráfico circular

Descripción generada automáticamente

1. ¿Por cuál vía de transporte se suele exportar este arancel?

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico circular

Descripción generada automáticamente

Los gráficos revelan que las vías más utilizadas para la exportación de este arancel son, en primer lugar, la vía terrestre, seguida de la vía marítima. Además, el Peso Neto de las exportaciones se concentra prácticamente al 100% en estas dos vías, lo cual indica una clara predominancia. Por lo tanto, el análisis que se presenta a continuación no tomará en cuenta la vía aérea.

1. ¿Cómo se ha comportado la exportación de este arancel? ¿Cómo se comporta el Peso Neto y el valor FOB de este arancel?

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Al analizar el comportamiento de las exportaciones del arancel basándonos en los gráficos, durante los años estudiados se observa una notable disminución tanto en el Peso Neto exportado como en el Valor FOB a partir del año 2018. Sin embargo, el Valor FOB experimenta un aumento en el año 2019, pero vuelve a caer en el año siguiente. Esta fluctuación posiblemente se deba a los efectos de la pandemia, no obstante, también se tendrían que estudiar otros factores como cambios en la demanda del mercado, fluctuaciones económicas, entre otros, pues también pueden haber influido en estas tendencias observadas en los datos.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Dado al pico en 2019 en el valor FOB, también se realizó un análisis en el comportamiento del valor FOB entre peso neto y podemos ver en el grafico previo que presenta una tendencia creciente, podíamos decir que esta tendencia se vio afectada en el 2020, esto podría ser producto de la pandemia, pero de todas maneras se mantiene la tendencia de que el valor FOB por peso neto va en aumento de este arancel (bisuterías de metal) cada año.

1. ¿Promedio de valor FOB pagado en el periodo estudiado?

A continuación, se presenta un análisis de los datos numéricos del valor FOB en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| Media | Mediana | Varianza | Desviación Estándar |
| 16364.49 | 4982 | 2084508073 | 45656.41 |

Distribución normal de la columna Valor FOB.

Gráfico, Histograma

Descripción generada automáticamente

Al analizar los datos del valor FOB en el estudio, se aprecia que los datos siguen una la distribución normal, sin embargo, cabe mencionar que la gran mayoría de ellos se encuentran concentrados hacia la izquierda del pico de la distribución normal. Esta disposición sugiere que existe una asimetría hacia la derecha, lo que indica un sesgo positivo. Es decir, la cola de la distribución se extiende hacia los valores más altos, lo que indica que hay algunas observaciones con valores considerablemente más elevados que influyen en esta asimetría en los datos.

1. ¿Promedio de Peso Neto exportado en el periodo estudiado?

A continuación, se presenta un análisis de los datos numéricos del peso neto en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| Media | Mediana | Varianza | Desviación Estándar |
| 1208.42 | 180 | 35336621 | 5944.46 |

Distribución normal de la columna Peso Neto.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Al examinar los datos del Peso Neto en el estudio, notamos que el histograma sigue una forma similar a la típica distribución en forma de campana que caracteriza a las distribuciones normales. Esto sugiere la posibilidad de que estos datos puedan seguir una distribución normal. No obstante, para asegurar y confirmar esta suposición, sería recomendable realizar otros análisis estadísticos que permiten validar de manera más rigurosa si la distribución se ajusta realmente a una distribución normal. Es una distribución bastante simétrica, sin embargo, aun presenta un peque sesgo positivo.

1. ¿Cuál es la relación entre el Peso Neto y valor FOB?

En el gráfico presentado a continuación se muestra la representación gráfica de la regresión lineal que relaciona el Peso Neto y el Valor FOB.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

|  |  |
| --- | --- |
| *Estadísticas de la regresión* | |
| Coeficiente de correlación múltiple | 0.88629878 |
| Coeficiente de determinación R^2 | 0.78552553 |
| R^2 ajustado | 0.78431381 |
| Error típico | 21203.764 |
| Observaciones | 179 |

En este caso, se obtuvo un coeficiente de correlación de 0.886 lo que indica que hay una correlación fuerte y positiva entre el Peso Neto y el Valor FOB. Además, un valor de R^2 de 0.785 indica una correlación significativa y fuerte entre el Peso Neto y el Valor FOB. Esto implica lo que significa que aproximadamente el 78.55% de la variabilidad en el Valor FOB puede ser explicada por el Peso Neto y por lo tanto el Peso Neto es un buen indicador o predictor del Valor FOB dentro del marco de la regresión lineal utilizada.

A continuación, los intervalos de confianza para la intersección y la variable x1.

|  |  |
| --- | --- |
|  | *Coeficientes* |
| Intercepción | 8138.51877 |
| Variable X 1 | 6.80721447 |

|  |  |  |
| --- | --- | --- |
| Intervalo de confianza B0 | | |
| 3178.38099 | <B0 | 13098.6566 |

|  |  |  |
| --- | --- | --- |
| Intervalo de confianza B1 | | |
| 6.16643898 | <B1 | 7.44798996 |

El intervalo de confianza al 95% para la intercepción (β0) y el coeficiente de x1 (β1) nos proporciona una estimación que sugiere que, en caso de tomar numerosas muestras distintas y calcular el intervalo de confianza para cada una de ellas, aproximadamente el 95% de dichos intervalos abarcarían los valores reales de la población correspondientes a la intercepción y el coeficiente de x1.

También obtenemos el grafico de residuales y el grafico de probabilidad normal para analizar mejor la regresión.

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

ANOVA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Grados de libertad* | *Suma de cuadrados* | *Promedio de los cuadrados* | *F* | *Valor crítico de F* |
| Regresión | 1 | 2.9146E+11 | 2.9146E+11 | 648.273047 | 4.5273E-61 |
| Residuos | 177 | 7.9579E+10 | 449599606 |  |  |
| Total | 178 | 3.7104E+11 |  |  |  |

El análisis de varianza realizado en el modelo de regresión reveló una fuerte correlación positiva entre el Peso Neto y el Valor FOB, indicando que a medida que el Peso Neto aumenta, también tiende a aumentar el Valor FOB de las importaciones.

Dado que el valor obtenido de la estadística F es mayor que el valor crítico de F, se rechaza la hipótesis nula, mostrando que hay suficiente evidencia estadística para afirmar que existe una relación significativa entre el Valor FOB y el Peso Neto, y por tanto el modelo tiene una capacidad significativa para predecir el Valor FOB (variable dependiente) en función del Peso Neto (variable independiente).

Estos hallazgos respaldan la relevancia del peso neto como factor influyente en el valor FOB de las exportaciones realizadas en las zonas francas.

# Malls más grandes del mundo

Fuente: <https://www.kaggle.com/datasets/drahulsingh/largest-malls>

* Composición de la base de datos a analizar:
* País: valor categórico
* Ciudad: valor categórico
* Año de apertura: valor numérico
* Superficie Bruta Alquilable: valor numérico
* Tiendas: valor numérico

Plan de análisis:

Se analizarán los centros comerciales más grandes de distintos países y la cantidad de tiendas que los conforman.

Promedio de tiendas: 1807.5

Desviación Estándar: 2353.913155

Varianza: 863723.8095

Análisis de la base de datos

Gráfica Cantidad total de tiendas vs Países

Como china es el país con más tiendas entre los malls más grandes del mundo, se segmentó la muestra. En base a este dato, se hizo el siguiente análisis numérico:

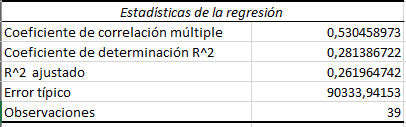
Una vez se observó que China es el país con mayor cantidad de tiendas, se decidió analizar cómo es la dispersión entre las áreas de terreno de cada mall vs la cantidad de tiendas que tiene.

Para este análisis se tomó como base la siguiente pregunta:

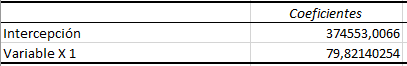
* ¿Cómo es la dispersión de los datos entre el terreno vs las tiendas que hay?

Gráfico 3. Cantidad de tiendas ocupadas en los malls en China vs Terreno disponible

Con estos datos veremos la relación entre el área y las tiendas.

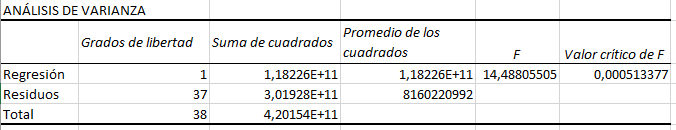


A continuación, los intervalos de confianza:



ANOVA

Análisis de Varianza



A continuación, las gráficas de la probabilidad normal.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Se observó que Filipinas es quien más espacio tiene, pero no quien más tiendas tiene. Podríamos decir que China aprovecha mejor el espacio colocando más tiendas.  
Con esto se puede concluir que China es el país con mayor cantidad de tiendas en los malls y quien mejor aprovecha sus terrenos a la hora de colocar tiendas dentro de estos.

# Indicadores de Fecundidad en la República de Panamá

Fuente:<https://www.inec.gob.pa/DASHBOARDS/Vitales/EV_INDICADORES_FECUNDIDAD>

Composición de la base de datos a analizar:

Ubicación: valor categórico

Año: valor numérico

Edad Media de la fecundidad: valor numérico

Tasa Bruta de Natalidad: valor numérico

Tasa Bruta de reproducción: valor numérico

Tasa de Fecundidad General: valor numérico

Tasa global de fecundidad: valor numérico

Plan de análisis:

Se decidió evaluar y comparar la fecundidad de una de las comarcas del país en relación con una provincia y compararlas a través de los años 2010-2019. El objetivo de este análisis se basó en ver cómo influye la taza de la fecundidad en el crecimiento de la población.

Análisis de la base de datos

1. ¿Cómo varía la tasa de fecundidad en áreas de fácil acceso vs áreas de difícil acceso a través de los años?

A partir de los datos presentados se obtuvo el siguiente resultado:

Se puede atribuir la diferencia presentada a la falta de educación sexual en las áreas de difícil acceso.

1. ¿A qué se deberá el aumento en la tasa global de fecundidad en Panamá durante el año 2019?

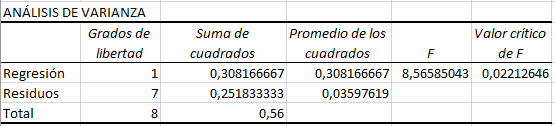
Además, se observó un valor atípico en el año 2019 en la provincia de Panamá donde su tasa global de fecundidad aumentó abruptamente en comparación con los años anteriores.

Se utilizarán los datos de los años y la tasa de fecundidad de panamá para hacer la tabla ANOVA y sus gráficos de regresión lineal.

|  |  |
| --- | --- |
| *Estadísticas de la regresión* | |
| Coeficiente de correlación múltiple | 0,741820476 |
| Coeficiente de determinación R^2 | 0,550297619 |
| R^2 ajustado | 0,486054422 |
| Error típico | 0,189673906 |
| Observaciones | 9 |

Gráficos de Regresión Lineal

ANOVA



# Nacimientos Vivos en la República de Panamá

Fuente: <https://www.inec.gob.pa/DASHBOARDS/Vitales/EV_NACIMIENTOS>

Descripción de Variables:

* Año: valor categórico
* Edad de las madres: valor categórico
* Provincia: valor categórico
* Meses: valor categórico
* Cantidad de nacidos: valor numérico

Plan de análisis:

Se decidió evaluar y comparar la fecundidad de una de las comarcas del país con relación una provincia y compararlas a través de los años 2010-2019. El objetivo de este análisis se basó en ver cómo influye la taza de la fecundidad en el crecimiento de la población.

Análisis de base de datos

1. ¿Qué provincia posee la mayor cantidad de Nacimientos vivos en la república de Panamá?

En la siguiente tabla podemos ver los valores de nacimientos vivos por cada provincia en la república de Panamá.

|  |  |
| --- | --- |
| Provincias | Nacimientos vivos |
| 1 | 13387 |
| 2 | 12303 |
| 3 | 15727 |
| 4 | 24714 |
| 5 | 3244 |
| 6 | 4744 |
| 7 | 3374 |
| 8 | 73104 |
| 9 | 12540 |
| 10 | 31208 |

En el siguiente gráfico podemos observar que la provincia con la mayor cantidad de nacimientos vivos en la república de Panamá en la provincia #8 “Panamá”, su capital.

1. ¿Cuál mes es el que contiene la mayor cantidad de nacimientos vivos en la república de Panamá?

En la siguiente tabla podemos observar la similitud de la cantidad de nacimientos por mes en la república de Panamá. Se puede notar a su vez que el mes que contiene menos días es de igual forma el mes con menos nacimientos.

|  |  |
| --- | --- |
| Todas las Provincias | |
| Mes | Total |
| 1 | 16680 |
| 2 | 14463 |
| 3 | 15737 |
| 4 | 15480 |
| 5 | 15966 |
| 6 | 15395 |
| 7 | 16304 |
| 8 | 16652 |
| 9 | 17244 |
| 10 | 17473 |
| 11 | 16745 |
| 12 | 16206 |

En la siguiente gráfica de frecuencia nacimientos/meses se puede observar mejor el comportamiento de los nacimientos a lo largo de los 12 meses del año notando que su mayor pico de encuentra entre los meses 9 y 10 (septiembre y octubre).

En esta gráfica de frecuencia relativa encontramos los porcentajes de nacimientos en cada mes notan una diferencia de 0.0155 del mes con mayor tasa de natalidad contra el menor.

Por último, nuestro análisis numérico quedaría de la siguiente manera:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Media | Mediana | Varianza | Desviación Estándar | Coeficiente de Varianza |
| 16195.4167 | 16255 | 718682.99 | 847.7517 | 0.05235 |

1. ¿Cuál es la correlación que existe entre la edad de la madre y la cantidad de hijos?

|  |  |
| --- | --- |
| *Estadísticas de la regresión* | |
| Coeficiente de correlación múltiple | 0.410009309 |
| Coeficiente de determinación R^2 | 0.168107634 |
| R^2 ajustado | 0.168071505 |
| Error típico | 1.422358027 |
| Observaciones | 23028 |

Como se puede observar en la siguiente tabla el análisis obtenido del coeficiente de correlación es muy bajo siendo este de 0.4100 por lo cual no podemos hacer una regresión lineal adecuada.

En esta gráfica podemos observar el comportamiento de los datos con respecto a la edad de la madre y la cantidad de hijos que tiene.

Puedo concluir que realmente la edad no es un factor determinante para la cantidad de hijos que tenga la madre. La cantidad hijos se mantiene bastante similar a lo largo de la vida de la madre.

# Consumo de combustible en New York

Fuente: <https://www.kaggle.com/code/thetoxicguy/visualizaci-n-de-datos>

Composición de la base de datos a analizar:

* Name: valor categórico
* mpg: valor numérico
* cylinders: valor numérico
* displacement: valor numérico
* horsepower: valor numérico
* weight: valor numérico
* acceleration: valor numérico
* model\_year: valor numérico
* origin: valor numérico

Plan de análisis:

La aproximación del análisis estadístico de esta base de datos se fundamenta en un enfoque guiado por la curiosidad para saber si el cilindraje influye en las millas por galón que consume un automóvil.

Análisis de la Base de Datos

Gráficas de Frecuencia

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Diagrama de Pareto

Gráfico, Histograma

Descripción generada automáticamente

Media: 23.5145

Varianza: 61.08961

Desviación Estándar: 7.8156

Coeficiente de correlación entre mpg y displacement: -0.80420

Análisis de Regresión Lineal

En la siguiente gráfica podremos observar el comportamiento de los datos entre MPG vs Displacement.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

ANOVA

Texto

Descripción generada automáticamente

R cuadrado: 0.6467421834257859

Imagen que contiene texto, marcador, exterior, firmar

Descripción generada automáticamente



Tras llevar a cabo este análisis, podemos llegar a la conclusión de que existe una relación directa entre el cilindraje de un automóvil y su consumo de galones de gasolina por milla (mpg). En otras palabras, a medida que aumenta el cilindraje del vehículo, su consumo de gasolina por milla también aumenta.

# Cantidad de calorías en comidas rápidas

Fuente: <https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition>

Composición de la base de datos a analizar:

* restaurant: Nombre del restaurante (texto).
* Item: Nombre del platillo (texto).
* calories: Cantidad total de calorías (valor numérico).
* cal\_fat: Calorías provenientes de las grasas (valor numérico).
* cholesterol: Cantidad de colesterol presente en el platillo, medido en mg (valor numérico).
* total\_carb: Cantidad de carbohidratos presentes en el platillo, medido en g (valor numérico).
* fiber: Cantidad de fibra presente en el platillo, medido en g (valor numérico).
* protein: Cantidad de proteína presente en el platillo, medido en g (valor numérico).

Plan de análisis:

La aproximación del análisis estadístico de esta base de datos se fundamenta en un enfoque guiado por la curiosidad de conocer alternativas de comida rápida que no sean dañinas para nuestro cuerpo.

Análisis de la Base de Datos

Gráficas de Frecuencia:

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Diagrama de Pareto

Gráfico, Histograma

Descripción generada automáticamente

Media: 530.9126

Varianza: 79770.1771

Desviación Estándar: 282.4361

Coeficiente de correlación entre calories y cholesterol: 0.76246

Análisis de Regresión Lineal

En la siguiente gráfica se puede observar el nivel de correlación que tienen los datos para determinar cuales variables hacen más dañino los platillos

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

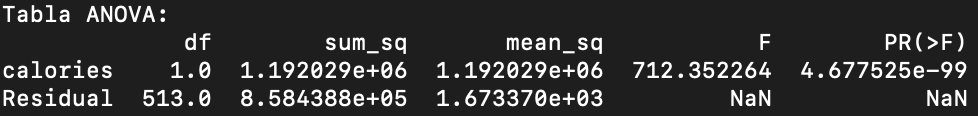
Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

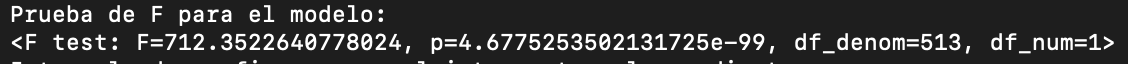
ANOVA



R cuadrado: 0.58134487931429

Texto

Descripción generada automáticamente



Los resultados del análisis de Regresión Lineal señalan que la variable independiente "calories" tiene un efecto significativo en la variable dependiente "cholesterol" en los platillos de los restaurantes de comida rápida en Estados Unidos. Según el coeficiente de determinación (R cuadrado), aproximadamente el 58.13% de la variabilidad en el contenido de colesterol puede ser explicada por las variaciones en las calorías del platillo.

Además, al emplear intervalos de confianza al 95%, se ha concluido que el valor del intercepto se sitúa con una probabilidad del 95% entre -25.613794 y -10.521950. Esto indica que tenemos un alto grado de certeza de que el valor real del intercepto se encuentra dentro de ese rango.

# Precios de Laptops

Fuente: <https://www.kaggle.com/datasets/muhammetvarl/laptop-price>

En la siguiente base de Datos de Kaggle se evaluarán diferentes laptops con sus componentes para determinar si estos influyen en su precio final.

Descripción de Variables:

* Laptop: Nombre completo de la laptop (texto).
* Status: Estado de la laptop, puede ser nuevo o reconstruido (texto).
* Brand: Marca de la laptop (texto).
* Model: Modelo de la laptop (texto).
* CPU: Nombre del procesador con el que cuenta la laptop (texto).
* RAM: Cantidad de memoria RAM que tiene la laptop (valor numérico).
* Storage: Cantidad de almacenamiento que tiene la laptop (valor numérico).
* GPU: Modelo de tarjeta gráfica con la que cuenta la laptop, puede ser nula (valor numérico).
* FinalPrice: Precio final de la laptop (valor numérico).

Plan de análisis:

La aproximación del análisis estadístico de esta base de datos se fundamenta en un enfoque guiado por la curiosidad de conocer el comportamiento de los diferentes componentes que poseen las laptops lo cual influye en su precio final.

Análisis de la Base de Datos

Gráficas de Frecuencia:

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Media: 15.41388888888889

Varianza: 97.37377643970974

Desviación Estándar: 9.86781518066232

Coeficiente de correlación entre RAM y Final Price: 0.7249463124103899

Análisis de Regresión Lineal

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

ANOVA

Texto

Descripción generada automáticamente

R cuadrado: 0.5255471558774225

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Los resultados del análisis revelan que la variable "RAM" muestra una correlación positiva significativa con la variable dependiente "FinalPrice". El coeficiente de correlación entre "RAM" y "FinalPrice" es aproximadamente 0.525, lo que sugiere una relación positiva entre la cantidad de RAM y el precio final.

Además, los intervalos de confianza al 95% para el intercepto se sitúan entre 231.337 y 329.644, y para la pendiente se encuentran entre 64.276 y 69.648. Esto implica que podemos tener un 95% de confianza en que el valor real del intercepto y la pendiente se encuentran dentro de estos intervalos.

# CONCLUSIÓN

En conclusión, este análisis de diez bases de datos ha sido un desafío emocionante y enriquecedor. A lo largo del proceso, nos enfrentamos a la dificultad del tamaño y complejidad de los conjuntos de datos, lo que requería una infraestructura y recursos adecuados para llevar a cabo la asignación. Sin embargo, gracias a la utilización de distintas plataformas y herramientas, logramos superar estas dificultades y completar con éxito el análisis estadístico.