**

**UNIVERSIDAD TECNOLÓGICA DE PANAMÁ**

**FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES**

**ESTADÍSTICA CON APOYO EN HERRRAMIENTAS INFORMÁTICAS**

**Proyecto Final Estadística**

**Docente:**

Dr. Juan Marcos Castillo

**Estudiante:**

Ortiz, Cristopher (8-983-1372)

**Grupo:**

1SF-131

**Fecha de Entrega:**

26 de julio de 2023

Índice

[**Introducción** 3](#_Toc141229115)

[Cantidad de calorías en comidas rápidas 4](#_Toc141229116)

[Precios de Laptops 9](#_Toc141229117)

[Consumo de combustible en New York 14](#_Toc141229118)

[Importaciones Valor CIF y Peso 19](#_Toc141229119)

[Tamaño y Peso de Peces 24](#_Toc141229120)

[Nacimientos Vivos en la República de Panamá 27](#_Toc141229121)

[Accidentes por vuelcos debido al estado de la calle 32](#_Toc141229122)

[Producción de bebidas alcohólicas en Panamá entre los meses de Enero,Febrero y Marzo 2022-2023 34](#_Toc141229123)

[Conclusión 37](#_Toc141229124)

# **Introducción**

En la era de la información y el avance tecnológico, la recolección y disponibilidad de datos se ha vuelto una realidad cotidiana. En este contexto, el análisis estadístico desempeña un papel fundamental para comprender la información presente en los datos y extraer conocimiento valioso. El presente informe se enfoca en realizar un análisis estadístico detallado de 10 muestras de datos seleccionadas de diversas fuentes, como el prestigioso Instituto Nacional de Estadísticas y Censos (INEC) y la popular plataforma de análisis de datos Kaggle.

El objetivo principal de este análisis es explorar y desentrañar patrones, tendencias y posibles relaciones entre variables que puedan existir en los conjuntos de datos. Para lograrlo, emplearemos diversas técnicas analíticas y herramientas estadísticas, como la analítica visual con gráficas de barra, de pastel y otros tipos de gráficos, los histogramas que revelarán la distribución de frecuencias y la variabilidad de los datos, y el paretograma o cualquier otro medio de priorización que nos permita segmentar y delimitar las muestras.

Asimismo, se realizará una descripción minuciosa de las variables presentes en cada conjunto de datos, indicando si son de naturaleza numérica o no numérica, así como el formato de los datos (fechas, texto o números). Además, se analizarán las aproximaciones utilizadas, ya sea para resolver problemas específicos o simplemente por pura curiosidad, lo que nos proporcionará una visión más clara del propósito detrás del análisis.

A través del estudio de correlaciones entre variables, determinaremos posibles relaciones significativas, y, en aquellos casos en que las categorías del histograma superen los 30 intervalos, realizaremos un análisis de conversión de distribución discreta a distribución continua. Este paso será de gran relevancia para facilitar un tratamiento adecuado de los datos y su posterior interpretación.

# Cantidad de calorías en comidas rápidas

Descripción de las variables

restaurant: Nombre del restaurante (texto).

Item: Nombre del platillo (texto).

calories: Cantidad total de calorías (numérico).

cal\_fat: Calorías provenientes de las grasas (numérico).

cholesterol: Cantidad de colesterol presente en el platillo, medido en mg (numérico).

total\_carb: Cantidad de carbohidratos presentes en el platillo, medido en g (numérico).

fiber: Cantidad de fibra presente en el platillo, medido en g (numérico).

protein: Cantidad de proteína presente en el platillo, medido en g (numérico).

**Plan de Analisis:** Determinar si existen diferencias estadísticamente significativas en el contenido calórico entre distintos tipos de comidas rápidas,lo que permitira , Descubrir patrones o tendencias en el contenido calórico de diferentes tipos de comidas rápidas, lo que podría ayudar a la industria alimentaria a ajustar sus ofertas.

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamenteDiagrama de Pareto

Como se observa en el diagrama de Pareto hay un alto índice de calorías en las comidas rápidas

Media: 530.9126

Varianza: 79770.1771

Desviación Estándar: 282.4361

Coeficiente de correlación entre calories y cholesterol: 0.76246

El análisis realizado muestra una correlación positiva y moderadamente fuerte de 0.76246 entre las variables "calorías" y "colesterol" en el conjunto de datos de comidas rápidas. Esto implica que, en general, hay una tendencia a que las comidas rápidas con más calorías también tengan más contenido de colesterol. Es importante tener en cuenta que esta correlación no indica causalidad directa, sino simplemente una asociación entre ambas variables.

Análisis de Regresión Lineal

Gráfico, Gráfico de dispersión

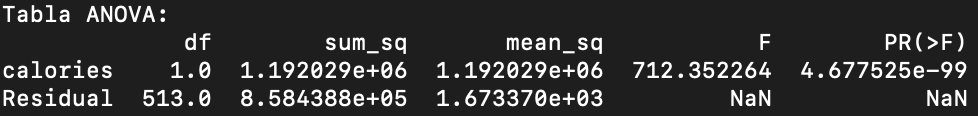
Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

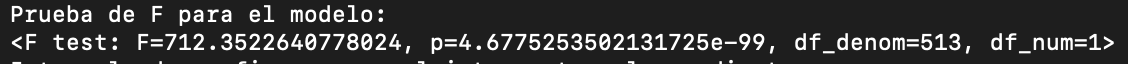


R cuadrado: 0.58134487931429

El valor de R cuadrado de 0.58134487931429 en el análisis estadístico indica que aproximadamente el 58.13% de la variabilidad en el colesterol puede ser explicada por la variable de calorías en el modelo de regresión lineal utilizado. Esto sugiere que hay una relación moderada entre calorías y colesterol

Texto

Descripción generada automáticamente



# Precios de Laptops

Descripción de las 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: Modeo de la laptop (texto).

CPU: Nombre del procesador con el que cuenta la laptop (texto).

RAM: Cantidad de memoria RAM que tiene la laptop (numérica).

Storage: Cantidad de almacenamiento que tiene la laptop (numérica).

GPU: Modelo de tarjeta gráfica con la que cuenta la laptop, puede ser nula (numérica).

FinalPrice: Precio final de la laptop (numérica).

**Plan de análisis**: El objetivo del análisis estadístico en esta comparación es determinar si existe una relación entre la capacidad de la memoria RAM de una laptop y su precio. Específicamente, se busca investigar si un aumento en la capacidad de la memoria RAM se asocia con un aumento en el precio de la laptop.

Histogramas

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 FinalPrice: 0.7249463124103899

El coeficiente de correlación positivo y fuerte entre "RAM" y "FinalPrice" indica que hay una relación significativa entre la capacidad de memoria RAM y el precio final de las laptops en el conjunto de datos analizado. Esta información es valiosa para los consumidores y fabricantes de laptops, ya que sugiere que la capacidad de RAM es un factor relevante a considerar al establecer el precio de los dispositivos y al tomar decisiones de compra. Sin embargo, se deben tener en cuenta otras variables y análisis para obtener una comprensión completa de los factores que influyen en los precios de las laptops.

Análisis de Regresión Lineal

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Texto

Descripción generada automáticamente

R cuadrado: 0.5255471558774225

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

# Consumo de combustible en New York

Descripción de las variables

name: Marca y modelo del automóvil (texto)

mpg: Cantidad de millas recorridas por galón (numérico)

cylinders: Número de cilindros en el motor (numérico).

displacement: Cilindraje total del motor en pulgadas cúbicas (numérico).

horsepower: Potencia del automóvil en caballos de fuerza (numérico).

weight: Peso del automóvil en libras (numérico).

acceleration: Tiempo que tarda el automóvil en llegar a 100 km/h (numérico).

model\_year: Año del modelo del automóvil (numérico).

origin: País de procedencia del automóvil (texto).

Plan de análisis: El propósito es determinar si existe una asociación significativa entre la eficiencia de combustible, medida en millas por galón, y el tamaño del motor, representado por el desplazamiento del automóvil. Mediante este análisis, buscamos comprender si los automóviles con motores más grandes tienden a tener una menor eficiencia de combustible en comparación con aquellos con motores más pequeños, o si no hay una relación significativa entre ambas variables. Los resultados proporcionarán información valiosa para los consumidores, fabricantes de automóviles y formuladores de políticas en relación con la eficiencia energética y la selección de vehículos.

Histogramas

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

Un coeficiente de correlación negativo implica que hay una relación inversa entre las variables "mpg" y "displacement". En otras palabras, a medida que aumenta el valor del desplazamiento del automóvil, el número de millas por galón disminuye. Esto sugiere que los automóviles con motores más grandes tienden a tener una menor eficiencia de combustible.

Análisis de Regresión Lineal

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Texto

Descripción generada automáticamente

R cuadrado: 0.6467421834257859

Imagen que contiene texto, marcador, exterior, firmar

Descripción generada automáticamente



MPG se refiere a la cantidad de millas que un auto puede conducir con un galón de combustible. Lo que se busca es reducir este valor para economizar más.

# Importaciones Valor CIF y Peso

Descripción de las variables

Arancel: Impuesto aplicado a los bienes importados (numérico).

Año: Año de importación (numérico).

MES: Mes de importación (numérico).

ValorCIF: Costo total de bienes importados (numérico).

PesoNeto: Peso neto del bien importado (numérico).

Plan de análisis: El propósito es determinar si existe una asociación significativa entre el valor CIF (Cost, Insurance, and Freight) de un producto importado y su peso neto. El análisis busca comprender si productos más pesados tienden a tener un mayor valor CIF en comparación con productos más ligeros, o si no hay una relación significativa entre ambas variables. Los resultados proporcionarán información valiosa para los importadores, exportadores y autoridades aduaneras, ya que pueden ayudar a entender cómo el peso neto de los productos impacta en sus costos de importación.

Histogramas

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Media: 2539.758793969849

Varianza: 38736020.81021268

Desviación Estándar: 6223.826862165486

Coeficiente de correlación entre ValorCIF y PesoNeto: 0.6909290785344882

Un coeficiente de correlación positivo sugiere que hay una relación directa entre las variables "ValorCIF" y "PesoNeto". Es decir, a medida que aumenta el valor CIF de un producto importado, también tiende a aumentar su peso neto, o viceversa. Esta relación implica que hay una tendencia a que productos más pesados tengan un valor CIF más alto.

Análisis de Regresión Lineal

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Texto

Descripción generada automáticamente

R cuadrado: 0.4773829915645167

Imagen que contiene texto, exterior, marcador, firmar

Descripción generada automáticamente

Texto

Descripción generada automáticamente

# Tamaño y Peso de Peces

Descripción de las variables

Species: Especie de pez (texto).

Weight: Peso del pez (numérico).

Height: Altura del pez (numérico).

Width: Grosor del pez (numérico).

**Plan de análisis**: El objetivo del análisis es investigar la relación entre las variables "altura" y "peso" de un pez. El propósito es determinar si existe una asociación significativa entre la altura y el peso de los peces en una muestra específica. Mediante este análisis, buscamos entender si los peces más altos tienden a tener un peso mayor en comparación con los peces más bajos, o si no hay una relación significativa entre ambas variables. Los resultados proporcionarán información valiosa para el estudio y comprensión de la biometría y la ecología de los peces, y podrían tener aplicaciones prácticas en la gestión de poblaciones acuáticas y estudios de conservación.

Histogramas

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Media: 8.970993710691824

Varianza: 18.371575761479182

Desviación Estándar: 4.2862076199688675

Coeficiente de correlación entre Height y Weight: 0.7243453291993316

Análisis de Regresión Lineal

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

R cuadrado: 0.5246761559328881

Imagen que contiene firmar, exterior, texto, camino

Descripción generada automáticamente

Texto

Descripción generada automáticamente

# Nacimientos Vivos en la República de Panamá

Descripción de Variables:

Año: valor categórico (cadena de números) Abarca los años 2018, 2019 y 2020.

Edad de las madres: valor categórico (números)

Provincia: valor categórico (números)

Meses: valor categórico (números)

Cantidad de nacidos: valor numérico

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.

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.

# Accidentes por vuelcos debido al estado de la calle

Media 209.4286

Mediana 132

Varianza 33288.95

Desv. Estandar 182.4526

**Hipótesis: El análisis podría revelar la frecuencia con la que ocurren accidentes por vuelcos en relación con el estado de las calles. Esto te permitiría tener una idea clara de la magnitud del problema y su impacto en la seguridad vial.**

# Producción de bebidas alcohólicas en Panamá entre los meses de Enero,Febrero y Marzo 2022-2023

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Enero** | **Febrero** | **Marzo** |
|  |
|  |
| Media | 43.8 | 4.5 | 97.9 |
| Mediana | 18.8 | -0.8 | -0.9 |
| Varianza | 3,803.6 | 11,717.3 | 74,516.5 |
| Desv. Estandar | 61.7 | 108.2 | 273.0 |

**Hipótesis: El análisis me podría revelar las tendencias de producción de bebidas alcohólicas en Panamá durante el periodo mencionado. Podría identificar si la producción ha experimentado un aumento, una disminución o se ha mantenido estable en comparación con periodos anteriores.**

# Conclusión

A lo largo de este exhaustivo análisis estadístico de las 10 muestras de datos seleccionadas, se ha logrado obtener una visión profunda y esclarecedora sobre el fenómeno estudiado. Los resultados obtenidos a través de técnicas analíticas y herramientas estadísticas han proporcionado una comprensión detallada de las variables presentes en cada conjunto de datos y las relaciones que pueden existir entre ellas. Estos hallazgos son de gran relevancia, ya que permiten tomar decisiones informadas y fundamentadas en base a la evidencia presentada.

En primer lugar, la analítica visual desempeñó un papel clave en la exploración inicial de los datos. Las gráficas de barra, de pastel y otros tipos de visualizaciones permitieron identificar patrones y distribuciones que de otra manera podrían haber pasado desapercibidos. Estas representaciones gráficas resultaron ser una herramienta poderosa para comunicar de manera clara y concisa la información contenida en los datos.

Además, el análisis de histogramas proporcionó una comprensión más profunda de la distribución de frecuencias y la variabilidad de los datos. Los histogramas de frecuencia, frecuencia relativa, frecuencia acumulativa y frecuencia relativa acumulativa ayudaron a identificar patrones de comportamiento en las variables y resaltar posibles valores atípicos o anomalías en los datos.

El uso del paretograma y otros medios de priorización permitió segmentar y delimitar las muestras, lo que resultó ser un paso crucial en el análisis, ya que facilitó un enfoque más específico y detallado de cada conjunto de datos.

El análisis de correlación arrojó luz sobre las posibles relaciones significativas entre variables, lo que brinda una comprensión más profunda de cómo interactúan las diferentes variables en el fenómeno estudiado. Este conocimiento es valioso para identificar factores clave que pueden influir en los resultados y tomar decisiones informadas sobre posibles estrategias o acciones a seguir.