Universidad Tecnológica de Panamá

Facultad en de Ingeniería De Sistemas Computacionales Carrera Licenciatura en Ingeniería de Software

Estadística con apoyo informático

Proyecto semestral

Estudiante:

|  |  |
| --- | --- |
| Donovan Garcia | 6-725-700 |

Profesor: Juan Castillo, PhD

Semestre 1, 2023

26/7/2023

# Introducción

La estadística es una herramienta fundamental en la investigación científica que nos permite explorar, comprender y extraer conocimientos significativos a partir de datos. En este proyecto, hemos llevado a cabo un estudio exhaustivo en el que analizamos 10 bases de datos de distintos dominios con el objetivo de llevar de lo teórico dado en clase a lo práctico.

A lo largo de este informe, presentaremos los resultados obtenidos en cada una de las bases de datos desde sus graficas para entender los datos hasta los porcentajes de proporcionalidad de variables en la regresión lineal.

Índice

[Introducción 2](#_Toc141227593)

[Análisis del movimiento de Reexportación por Vía, identificando el mes de mayor actividad 3](#_Toc141227594)

[Estudio de accidentes vehiculares según rango de edades 8](#_Toc141227595)

[Estudio de importaciones a Panamá según el valor cif y peso neto 14](#_Toc141227596)

[Estudio de la tabla pescado 20](#_Toc141227597)

[Estudio de movimiento de página web 24](#_Toc141227598)

[Estudio de nacimientos 50+ 30](#_Toc141227599)

[Estudio de tabla de accidentes vehiculares 36](#_Toc141227600)

[Estudio de table de tren (fallido) 42](#_Toc141227601)

[Estudio del oro precio de entrada y salida del oro 45](#_Toc141227602)

[Consumo de combustible en New York 49](#_Toc141227603)

[Conclusión 54](#_Toc141227604)

# Análisis del movimiento de Reexportación por Vía, identificando el mes de mayor actividad

Tipo de estudio: Curiosidad

Tipos de variables: categóricas



En base a los datos analizados, se puede concluir que los meses de septiembre y octubre son de gran relevancia en términos de movimiento de reexportación. Estos meses registran un notable aumento en el volumen de movimientos en comparación con otros periodos del año. Además, se destaca que la vía terrestre es la opción preferida para el transporte de mercancías en el contexto de reexportación.

Este análisis proporciona una base sólida para comprender mejor los patrones de movimiento de reexportación y puede servir como punto de partida para futuras investigaciones sobre factores específicos que impulsan el aumento de la actividad durante estos meses. Además, estos resultados pueden ser de utilidad para las autoridades y las empresas que participan en el comercio internacional, al brindarles información sobre los periodos más activos y las preferencias de transporte en el ámbito de la reexportación.

**Análisis del mayor movimiento según el mes**

|  |  |
| --- | --- |
| Media | 3073.454545 |
| Mediana | 3104 |
| Varianza | 126288.4727 |
| Desv. Estand | 355.3708946 |





Se creó esta tabla mediante la media, desviación estándar y la fórmula de distribución normal.

|  |  |
| --- | --- |
| x | f(x) |
| 2007.341862 | 1.24711E-05 |
| 2362.712756 | 0.000151928 |
| 2718.083651 | 0.000680896 |
| 3073.454545 | 0.001122608 |
| 3428.82544 | 0.000680896 |
| 3784.196335 | 0.000151928 |
| 4139.567229 | 1.24711E-05 |

Una distribución normal que cumple con la curiosidad realizada.



para un análisis más profundo y para hacer predicciones precisas, sería necesario considerar otras variables y técnicas de modelado más adecuadas. Debido a la baja relación entre variables

# Estudio de accidentes vehiculares según rango de edades

Este estudio es por curiosidad para relacionar los rangos de edades con los accidentes vehiculares sin efectos del alcohol.

Los datos son categóricos.







Análisis de distribución normal



Análisis de regresión lineal



Una regresión altamente predecible debido a los coeficientes de determinación. El modelo es altamente ajustado y preciso, con una relación extremadamente fuerte entre las variables.

# Estudio de importaciones a Panamá según el valor cif y peso neto

Estudio por curiosidad de la tabla de importación en Panamá

Los valores de peso y cif son numéricos.

El valor de aranceles es categórico.

Utilizando una muestra de 200 datos esta es la gráfica que obtenemos que en relaciona el valor cif de cada arancel con su peso neto.

Nuevamente hice un muestreo para tener una cantidad más manipúlale de 20 aranceles.=



Con la ayuda de esta tabla podemos realizar:

|  |  |  |
| --- | --- | --- |
|  | CIF | Peso |
| Media | 2865.5 | 149.15 |
| Mediana | 914.5 | 49.5 |
| Varianza | 20542244 | 43738.45 |
| Desv. Estandar | 4532.355 | 209.1374 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rango | Frecuencia | Frecuencia Relativa | Frecuencia acumulativa | Frecuencia acumulativa relativa |
| 0-500 | 10 | 0.5 | 10 | 0.5 |
| 500-1000 | 5 | 0.25 | 15 | 0.75 |
| 1000-1500 | 3 | 0.15 | 18 | 0.9 |
| 1500-2000 | 2 | 0.1 | 20 | 1 |
|  | 20 | 1 |  |  |

Distribución normal:



Análisis de la regresión lineal



Esta regresión lineal no es tan predictiva debido a que la correlación moderada entre las variables. La proporción de variabilidad es de 29%

# Estudio de la tabla pescado

En este estudio veremos la correlación entre el peso con la altura del pescado a ver si nos proporciona algún conocimiento general.

Los datos de altura y peso serán numéricos.







Análisis de distribución normal



Análisis de la regresión lineal



Tenemos una proporción de variabilidad de 63% por lo que estas 2 variables no tienen una correlación para poder hacer una predicción.

# Estudio de movimiento de página web

Este estudio se va a revisar la relación entre las veces q que la pagina se carga en relación con la primera visita de usuarios.

Los datos de cantidad cargada y primera visita es numérica

El dato de días es categórico

|  |  |  |
| --- | --- | --- |
| **Dia** | **Catidad de carga** | **Primera visita** |
| Sunday | 44943 | 28374 |
| Monday | 66135 | 41481 |
| Tuesday | 67213 | 41949 |
| Wednesday | 64562 | 40561 |
| Thursday | 59785 | 37472 |
| Friday | 47200 | 28441 |
| Saturday | 30776 | 18694 |
| **Grand Total** | **380614** | **236972** |

Lo separamos para tener una mejor vista de los datos

Con esta grafica observamos que solo un 37.8% de las veces que se cargan esta pagina es la primera visita.



Estudio de frecuencias de las veces cargadas por día



Estudio de la distribución normal



Estudio de la regresión lineal



Podemos observar que el coeficiente re correlación es de 0.99 lo que se aproxima mucho a 1 que significa que es predecible y la proporción de variabilidad es de 99% por lo que es un buen caso de estudio ya que es un modelo altamente ajustado y preciso, con un coeficiente de determinación muy cercano a 1.

# Estudio de nacimientos 50+

Tome la tabla de Nacimientos Vivos en la República de Panamá del INEC dashboard porque me interesa saber el rango de edad más frecuente de nacimientos.

Los datos son categóricos a excepción de la variable nacimientos que es numérica.

Una vez obtenida la tabla de conteo

Pude notar que el rango más frecuente es de los 20-24 pero haciendo este estudio me sorprendió el rango de 50 y más. Ya Hablando en términos de números; tenemos que, A partir de los 40 años, la posibilidad de lograr un embarazo es de apenas un 5% por ciclo, y desde los 45, el riesgo de un aborto espontáneo llega al 95%. Post 50 la fertilidad es casi nula y lograr un embarazo de manera natural es posible, pero extremadamente inusual.

procedí a seguir investigando sobre el tema y obtuve esta gráfica:

Esta tabla nos proporciona que en Panamá tienen una media de **5 nacimientos exitosos** de una mujer con 50 años o más. Esto me parece una locura ya que la fertilidad de una mujer disminuye significativamente a medida que envejece debido a la disminución de la reserva ovárica y la calidad de los óvulos. Además, el riesgo de complicaciones durante el embarazo, como la hipertensión gestacional, la diabetes gestacional y el parto prematuro, también aumenta con la edad materna avanzada. 





Estudio de distribución normal



Estudio de regresión lineal



Podemos observar que tenemos un coeficiente de determinación de 0.86 por lo que el modelo de tiene una buena capacidad para explicar la variabilidad en la variable dependiente y proporciona una correlación fuerte con la variable independiente.

# Estudio de tabla de accidentes vehiculares

Este estudio es por curiosidad para ver que sexo es más afectado por accidentes automovilísticos.

Todos los datos son categóricos.

|  |  |  |
| --- | --- | --- |
| **Cuenta de Rango edad** | **Column Labels** |  |
| **Row Labels** | **Hombre** | **Mujer** |
| 2010 | 72,060 | 14,197 |
| 2011 | 56,168 | 9,662 |
| 2012 | 65,500 | 11,568 |
| 2013 | 70,516 | 13,736 |
| 2014 | 69,188 | 13,497 |
| 2015 | 80,133 | 12,837 |
| 2016 | 92,457 | 14,858 |
| 2017 | 92,854 | 16,943 |
| 2018 | 88,734 | 17,682 |
| 2019 | 82,668 | 17,165 |
| **Grand Total** | **770,278** | **142,145** |

Esta grafica nos proporciona una vista de diferencia de los más afectados.





Análisis de distribución normal



Análisis de regresión lineal



el modelo de regresión lineal tiene una buena capacidad para explicar la variabilidad en la variable dependiente y proporciona una correlación fuerte con las variables independientes utilizadas.

# Estudio de table de tren (fallido)

Este estudio era para ver si las distancias recorridas en un tren son proporcional mente variable del costo y tiempo

Los datos son numéricos







Debido a la mala toma de las variables la regresión lineal en este caso tiene un ajuste deficiente a los datos, y el modelo no es muy útil para explicar o predecir la variable dependiente en función de la variable independiente.

# Estudio del oro precio de entrada y salida del oro

Este estudio trata de ver la correlación entre el valor del oro

Los datos de entrada y salida son numéricos

El único dato categórico seria la fecha

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Open** | **Sum of Close** |
| Jan | 96540.46 | 96637.72 |
| Feb | 87272.36 | 87273.39 |
| Mar | 104285.63 | 104446.26 |
| Apr | 98493.19 | 98443.74 |
| May | 103814.35 | 103719.73 |
| Jun | 98404.93 | 98319.81 |
| Jul | 94375.23 | 94360.84 |
| Aug | 95856 | 95806.2 |
| Sep | 90083.87 | 89968.74 |
| Oct | 89462.31 | 89464.06 |
| Nov | 90332.81 | 90466.38 |
| Dec | 91497.42 | 91587.82 |
| **Grand Total** | **1140418.56** | **1140494.69** |







Modelo altamente ajustado y preciso, con una relación extremadamente fuerte entre las variables. Debido al alto porcentaje de proporcionalidad de variables

# Consumo de combustible en New York

Estudio por curiosidad

El consumo de combustibles será proporcional al cilindraje.

mpg: Cantidad de millas recorridas por galón (numérico)

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

Histogramas

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a number of blue lines

Description automatically generated

A graph of a number of blue lines

Description automatically generated

Diagrama de Pareto

A graph with a red line

Description automatically generated

Media: 23.5145

Varianza: 61.08961

Desviación Estándar: 7.8156

Coeficiente de correlación entre mpg y displacement: -0.80420

Análisis de Regresión Lineal

A graph with blue dots and red line

Description automatically generated

A graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated

A blue background with white numbers

Description automatically generated

R cuadrado: 0.6467421834257859

A blue screen with white text

Description automatically generated



regresión lineal tiene una buena capacidad para explicar la variabilidad en la variable dependiente y proporciona una correlación fuerte con las variables independientes utilizadas. Pero no es tan precisa ni predecible.

# Conclusión

Con la realización de este proyecto he podido entender las gran herramienta de la estadística ya que es una nos ayuda a realizar predicciones desde algo mínimo como en el caso de los nacimientos hasta ya algo industrial y complicado como son las importaciones.

Este proyecto de estadística ha sido una exploración enriquecedora que nos ha permitido descubrir patrones y relaciones entre variables en 10 bases de datos distintas. La correlación es una herramienta poderosa para desentrañar conexiones entre fenómenos aparentemente independientes, y sus hallazgos contribuyen al conocimiento en cada uno de los campos de estudio considerados.