Examen Final de Inferencia Estadística para Ciencia de Datos
(Parte Empírica)

Integrantes:
Natali Alba, Tania Díaz, Maria Jose Sierra

Profesor:
Julián Sánchez

Materia:
Inferencia Estadística

Carrera:
Ciencia de Datos

Facultad:
Departamento de Matemáticas

Universidad Externado de Colombia
Bogotá.D.C.
2024-1

Informe Final

En el campo de la estadística, surge la necesidad de utilizar métodos científicos rigurosos para hacer inferencias sobre distintas poblaciones a partir de muestras dadas y poder modelar los comportamientos de la sociedad en varios ámbitos. En el presente informe, se hará un seguimiento de cómo las pruebas de hipótesis son una herramienta poderosa para cumplir dicha función, analizar los puntos débiles y fuertes de estas bajo diferentes contextos y entender a fondo cómo distintos factores influyen finalmente en ciertas conductas de una comunidad mediante un modelo de regresión lineal multivariado.

En un primer momento, es de suma importancia resaltar que tanto para las pruebas de hipótesis como para la regresión lineal, se deben cumplir una serie de supuestos para verificar que los resultados obtenidos sean válidos, confiables y que puedan ser utilizados para la toma de decisiones. A continuación se hará un recorrido de diversos tests usados a lo largo del proyecto para evaluar el cumplimiento o no de estos supuestos:

Prueba de Shapiro-Wilk:
La prueba de Shapiro-Wilk es una prueba de normalidad utilizada para determinar si una muestra de datos proviene de una población con una distribución normal. Es especialmente útil cuando se trabaja con muestras pequeñas.
La prueba de Shapiro-Wilk se basa en la idea de que si una muestra proviene de una población normal, los valores observados deberían estar cerca de la línea diagonal en un gráfico de probabilidad normal (Q-Q plot). La prueba calcula un estadístico W que cuantifica cuánto se desvían los datos de la distribución normal esperada. Bajo la hipótesis nula de que los datos provienen de una población normal, el valor de W se espera que esté cerca de 1.

Si el valor p es menor que un nivel de significancia predefinido (comúnmente 0.05), se rechaza la hipótesis nula y se concluye que los datos no provienen de una distribución normal.

Es importante tener en cuenta que, especialmente con muestras grandes más de 50 observaciones, la prueba de Shapiro-Wilk puede ser sensible a pequeñas desviaciones de la normalidad y, por lo tanto, puede rechazar la hipótesis nula incluso cuando las diferencias con la distribución normal son mínimas.

Prueba de Anderson-Darling

Es una prueba estadística utilizada para evaluar si una muestra de datos proviene de una población con una distribución específica, como la distribución normal. Se basa en la comparación entre los valores observados y los esperados bajo la hipótesis nula de que los datos siguen la distribución especificada.

se calcula utilizando las diferencias entre los valores observados y los valores esperados bajo la distribución específica.

- Hipótesis:

  Hipótesis nula (H0): Los datos provienen de una población con una distribución específica (por ejemplo, una distribución normal).

  Hipótesis alternativa (H1): Los datos no provienen de una población con la distribución especificada.

- Supuestos:

1. Independencia de las observaciones: se supone que las observaciones de la muestra son independientes una de la otra. En otras palabras, el valor de una observación no se ve influido por el valor de otra observación.
2. Datos continuos: la prueba Anderson-Darling es para datos continuos y no es apropiada para datos categóricos u otros.
3. La distribución especificada: la prueba requiere que se especifique la distribución teórica bajo la hipótesis nula. Por ejemplo, si estamos probando la normalidad de los datos, asumimos que los datos están normalmente distribuidos.
4. Tamaño de muestra adecuado: La prueba puede ser sensible al tamaño de la muestra, especialmente a tamaños de muestra pequeños, por lo que el tamaño de muestra debe cumplir con ciertas condiciones.

Prueba de Bartlett:

La prueba de homogeneidad de varianzas de Bartlett se usa para probar que las varianzas son iguales para todas las muestras. Esta prueba proviene de una distribución chi-cuadrado siguiendo la hipótesis que las varianzas de todas las muestras son iguales, es recomendable de usar si vas a hacer alguna prueba que requiera que los datos sigan este supuesto como lo es la prueba ANOVA y la prueba T para dos muestras. Es bastante sensible a desviaciones de normalidad en los datos así que se recomienda usar cuando se está bastante seguro que los datos se distribuyen en forma de montículo.

La prueba de Bartlett proporciona una evaluación de la homogeneidad de las varianzas en varias muestras. Si se rechaza la hipótesis nula, indica que al menos una de las varianzas es diferente, lo que puede tener implicaciones importantes para los análisis estadísticos posteriores, como el análisis de la varianza (ANOVA). Si no se rechaza la

hipótesis nula, indica que no hay suficiente evidencia para concluir que las varianzas son diferentes, lo que sugiere que el supuesto de homogeneidad de varianzas se mantiene.

Prueba de Bondad de Ajuste de Kolmogorov-Smirnov:
Esta prueba es utilizada con el fin de revisar si una muestra dada proviene de una distribución conocida en específico (es decir, comparar una distribución teórica con una distribución empírica) o si dos conjuntos de datos provienen de una misma distribución. Al ser no paramétrica, no tiene supuestos sobre la muestra pero en general es muy utilizada para verificar si un conjunto de datos proviene de una distribución normal. En el contexto de este proyecto, se usó para verificar si las distribuciones de 2 o más grupos tienen una forma similar; en este test se van comparando por parejas de muestras, por lo que su hipótesis nula es que las dos muestras siguen distribuciones similares y su hipótesis alternativa es que las distribuciones de ambas muestras son distintas.

Básicamente, para la realización de esta prueba (en el caso de comparar dos muestras), se calcula la distribución empírica acumulada de ambos grupos y se halla la mayor diferencia vertical entre las dos, este sería el estadístico de prueba (D). Se encuentra un valor crítico de la distribución de Kolmogorov y se compara con el estadístico para saber si rechazar o no la hipótesis nula.

Gráficas y Diagramas usados para visualizar los datos:

Gráfica QQ plot:
El gráfico QQ-plot permite observar cuán cerca está la distribución de un conjunto de datos a alguna distribución ideal en el caso de los datos usados queremos ver que tan cerca se encuentran los datos a una distribución normal ó comparar la distribución de dos conjuntos de datos.

Histogramas:
Un histograma es similar en apariencia a un diagrama de barras, pero en lugar de comparar categorías o buscar tendencias a lo largo del tiempo, cada barra representa cómo se distribuyen los datos en una única categoría. Cada barra representa un rango continuo de datos o el número de frecuencias de un punto de datos específico.

Así mismo para realizar el número de intervalos adecuados para el histograma se tuvo en cuenta la regla Sturges, esta metodología empírica básicamente nos dice el número de clases usadas para el histograma.

Gráficas de Dispersión:
El diagrama de dispersión es una herramienta utilizada cuando se desea realizar un análisis gráfico de datos bivariados, es decir, los que se refieren a dos conjuntos de datos. El resultado del análisis en el diagrama puede mostrar que existe una relación entre una variable y la otra.

Diagrama Boxplot:
Un Boxplot nos proporciona información sobre la centralidad y extensión de los datos mediante la representación de los cuartiles, valores máximos y mínimos y valores atípicos.

Significancia
Asimismo, es necesario hacer la aclaración que la significancia a usar en todos los tests realizados a lo largo del proyecto es de 0.05, pues es un valor estándar que nos da una visión clara sobre desde qué punto es posible rechazar las hipótesis nulas, sin perder confianza en las pruebas o llegar a ser demasiado estrictos en este aspecto.

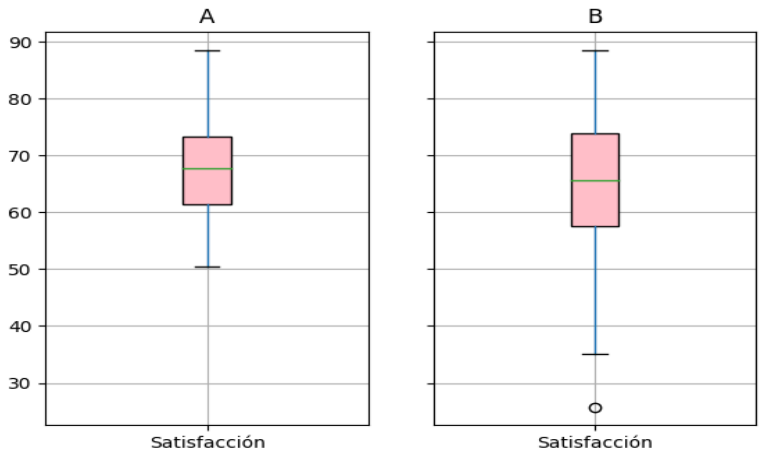
Parte 1: Comparación de Dos Grupos con Prueba T y Mann-Whitney U:

En esta parte nos encontramos con un conjunto de datos en la cual tenemos las puntuaciones de satisfacción de clientes de dos servicios diferentes, la idea de este reto es ver si hay diferencias significativas en la satisfacción de ambos grupos

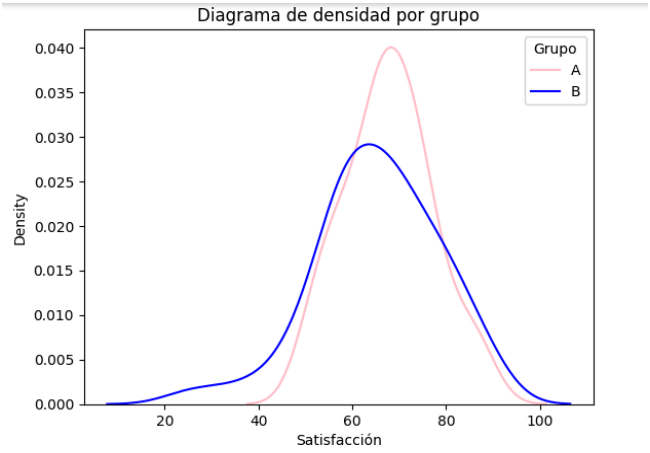
Para mirar esto primero comenzaremos con **el análisis exploratorio** de el conjunto de datos y encontramos que:

	Satisfacción							
	count	mean	std	min	25%	50%	75%	max
Grupo								
A	50.0	67.745261	9.336688	50.403299	61.390209	67.658548	73.362756	88.522782
B	50.0	65.266713	13.114875	25.703823	57.558330	65.691204	73.806111	88.469655

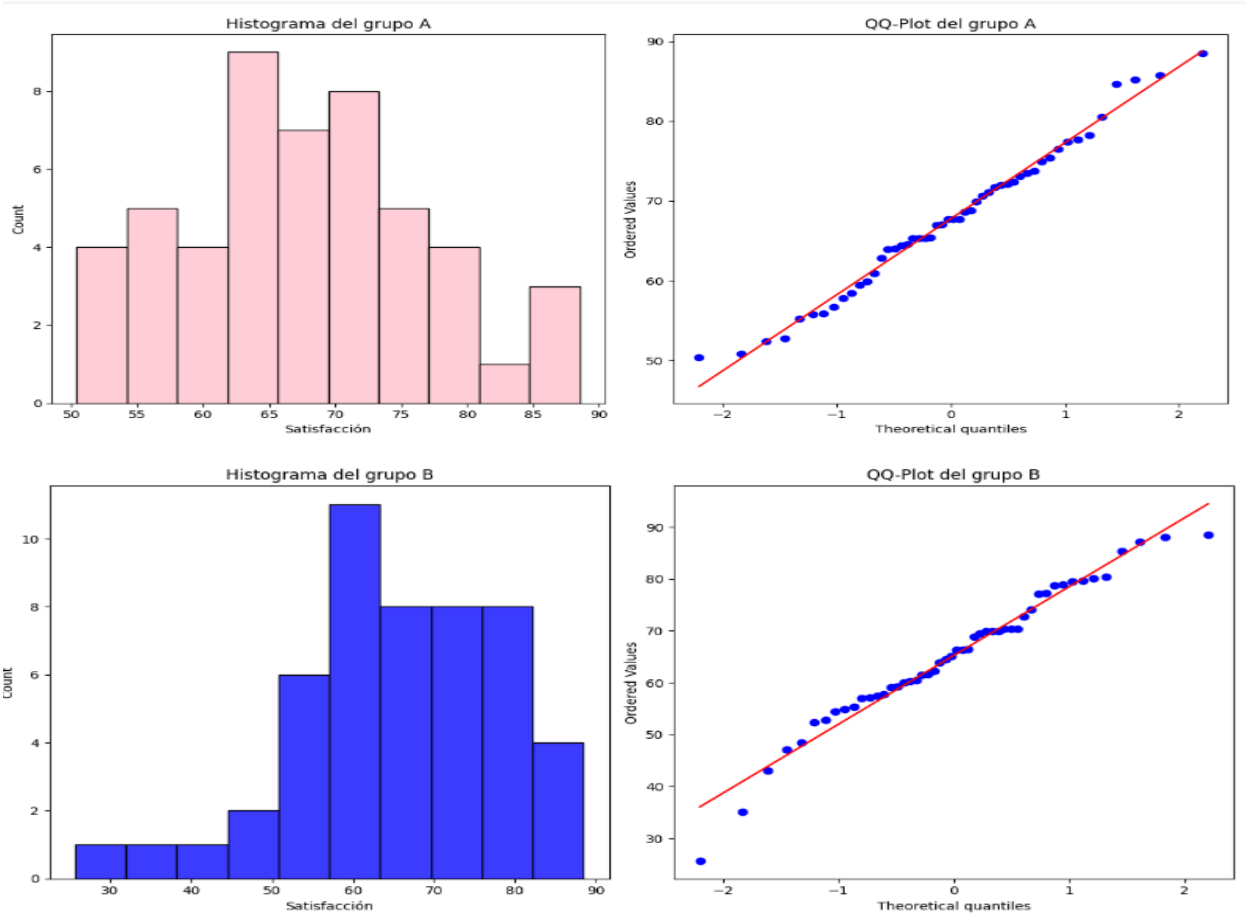
Para observar cómo se distribuyen las puntuaciones de satisfacción en cada grupo hicimos dos box-plot cada uno representativo de un grupo con la intención de comparar la distribución y la tendencia central



de valores numéricos mediante sus cuartiles y además echarle un vistazo superficial a la dependencia que tienen ambos grupos.



Además hicimos una gráfica de distribución de densidad para visualizar la distribución de los datos en un intervalo o período de tiempo continuo.
Luego de ver las gráficas profundizamos un poco más en la visualización de la distribución de las variables, centrándonos esta vez en las distribución normal.



Como se puede ver en los diagramas de QQ ambos grupos en especial el grupo ‘A’ parecen a distribuirse normalmente

Ahora sabiendo que ambas aparentan distribuirse de igual manera, cabe la pregunta del inicio, es decir, ver si hay diferencias significativas en el nivel de satisfacción de ambos grupos.
Para eso usaremos las pruebas Mann-Whitney U y La prueba t de Student para determinar si hay diferencias significativas entre las medias de dos grupos aun con sus diferencias de suposiciones y aplicaciones

Prueba t de Student:
La prueba 't' de student es un estadístico que se usa para verificar la hipótesis de que las media de las muestras son iguales. Para poder realizar nuestra prueba nuestros datos en primera instancia tiene que cumplir con estos dos supuestos principales:

El primer supuesto que nuestras muestras deben cumplir es que tienen una distribución normal o de montículo.

1) El segundo supuesto es que las varianzas de las muestras son iguales, o en otras palabras que la variabilidad de los datos es la misma.

Para comprobar los supuestos se usó:

1) Para comprobar el primer supuesto se realizó a cada una de las muestras la prueba de Shapiro Wilk explicada anteriormente. Bajo la hipótesis nula de que la muestra proviene de una población con una distribución normal, y con una significancia de 5%, estos son los resultados que arrojó para cada muestra.

```
Shapiro-Wilk test para Grupo A: ShapiroResult(statistic=0.9827495217323303, pvalue=0.6722097396850586)
Shapiro-Wilk test para Grupo B: ShapiroResult(statistic=0.9713166952133179, pvalue=0.26161736249923706)
```

   Como podemos ver ambas muestras el p-valor es mayor que la significancia, por lo tanto no hay suficiente evidencia científica para rechazar la hipótesis nula, así que ambas muestras provienen de una distribución normal.
2) Luego para comprobar el supuesto dos usamos la prueba de Bartlett, bajo la hipótesis de que las varianzas de las muestras son iguales, y el estadístico nos arrojó esto.

```
Estadística de prueba de Bartlett: 5.495874887849253
Valor p de la prueba de Bartlett: 0.019061387916391257
```

   Usando una significancia del 5% podemos ver que el p-valor es menor que la significancia. Por lo tanto hay suficiente evidencia científica para rechazar la hipótesis nula, las varianzas de las muestras son diferentes.

Como podemos ver todos los supuestos no se cumplen, por lo tanto vamos a proceder a intentar transformar los datos de tal forma que tratamos que nuestros datos nuevos si cumplan todos los supuestos.

Como pudimos ver en el Análisis exploratorio, se puede observar que el grupo b está sesgado, por lo tanto decidimos por esta razón transformar los datos con el método "Transformación inversa", este método se usa cuando los datos están sesgados. Al momento de volver a hacer las pruebas para comprobar los supuestos nos arroja los siguientes resultados.

```
Prueba de Levene para homogeneidad de varianzas: LeveneResult(statistic=4.030456271317731, pvalue=0.04743905991523835)
```

Con la hipótesis de que las varianzas de los grupo son iguales, y con una significancia de 5% podemos ver que el p-valor el mayor que la significancia por lo tanto existe suficiente evidencia para rechazar la hipótesis inicial. Las varianzas de las muestras no son iguales.

Como observamos a pesar de transformar los datos estos siguen sin cumplir los supuestos. Al momento de realizar la prueba t no dice esto.

```
P-Valor :
probabilidad acumulada =  0.137222564190694
t_valor  =  -1.0996990016239865
p_valor =  0.274445128381388
0.274445128381388 > 0.05
```

Usando la significancia del 5% , el p valor es mayor que la significancia por lo tanto no existe evidencia suficiente para rechazar la hipótesis nula. En ese orden de ideas la prueba t nos acaba de confirmar que las medias de las poblaciones son iguales.

Prueba de Mann-Whitney U:

La prueba de Mann-Whitney U, se suele utilizar cuando los datos no cumplen con las suposiciones de la prueba t de Student, es decir, cuando no se distribuyen normalmente o no tienen homogeneidad de varianza.

Se utiliza para comparar las medianas de dos grupos en lugar de las media

No requiere suposiciones específicas sobre la distribución de los datos y es menos sensible a las desviaciones de la normalidad, por lo que se procede a usar esta fórmula:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

$$U = min(U_1, U_2)$$

Standard error of U

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

z-value

$$z = \frac{U - \mu_U}{\sigma_U}$$

z-distribution

z-Value	p-Value
-0.1825	= 0.855

Al aplicar esta prueba en cuentro conjunto de datos nos arrojó que:

```
U: 2399.0
p_valor: 0.38694307714460696
```

Esto indica que no se rechaza la hipótesis nula por lo que no hay suficiente prueba estadísticas para decir que los dos grupos tienen una diferencia significativa, en otras palabras y en el contexto de la base, la satisfacción de ambos grupos no tiene una diferencia significativa.

Conclusión

En el caso de este conjunto de datos la prueba de Mann-Whitney U es más útil por que los conjuntos de datos no debe cumplir los supuestos que sí debe cumplir la prueba t de Student esto teniendo en cuenta que aun con la transformación de variables los supuestos seguían sin cumplirse, y aunque el resultado de ambas pruebas es la misma es decir que no se rechaza la hipótesis nula la prueba de Mann-Whitney U no puso tanto problema para realizarse.

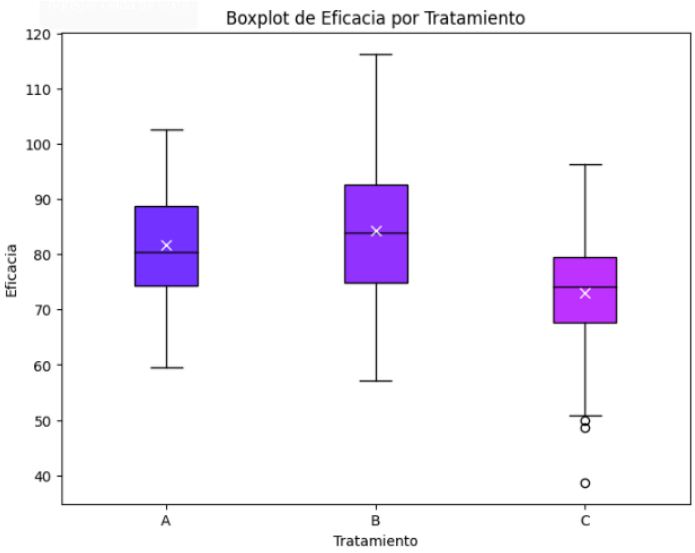
Parte 2: Comparación de Tres Tratamientos Médicos con ANOVA y Prueba de Kruskal-Wallis

Para este problema, se nos fue dada una base de datos para comparar si existen diferencias significativas en la eficacia de tratamientos para la diabetes de 3 grupos. Para hacer este análisis es posible utilizar la prueba ANOVA o Kruskal-Wallis, pues el objetivo de ambas es analizar si hay diferencias significativas entre 3 o más grupos mediante la observación de diferentes medidas de tendencia central, ya sea la media o la mediana.

Entendiendo en un momento inicial las pruebas que se usarán, ahora se hace relevante hablar sobre qué fue encontrado en el dataset. Para cada uno de los grupos de tratamiento, se encontraron estos estadísticos:

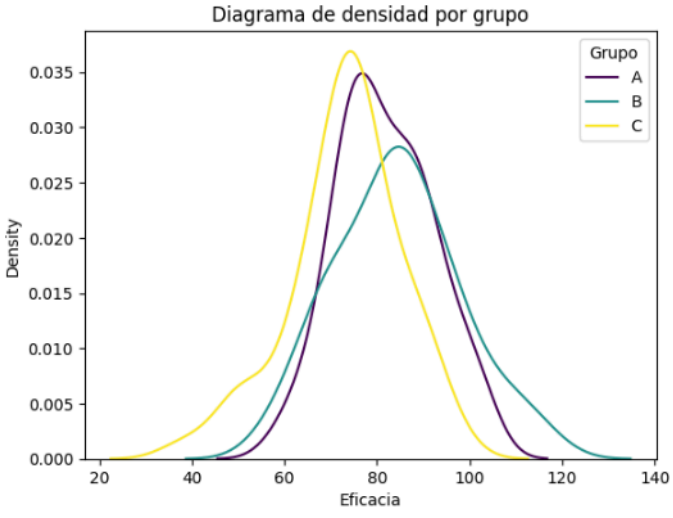
Tratamiento	count	mean	std	min	25%	50%	75%	max
A	50.0	81.78	10.26	59.53	74.30	80.34	88.78	102.66
B	50.0	84.27	13.45	57.12	74.83	83.97	92.68	116.34
C	50.0	73.11	11.90	38.69	67.73	74.13	79.41	96.28

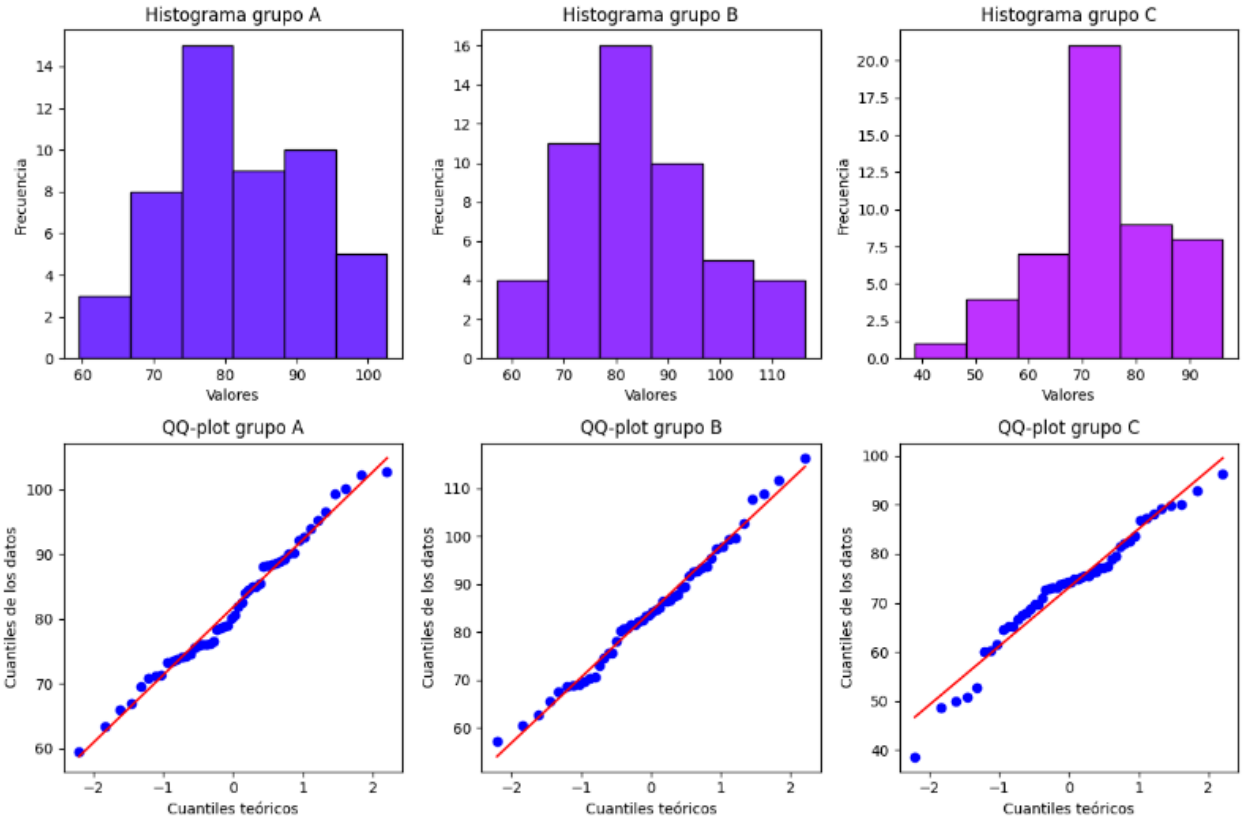
El solo observar esta tabla, nos da indicios iniciales sobre los datos. La cantidad de observaciones en cada grupo está equilibrada, las desviaciones estándar son aproximadamente similares, la media y mediana del grupo C son un poco más bajas que las de los grupos A y B y en general, el grupo del tratamiento C tiene valores menores en los cuartiles en comparación con los otros 2 conjuntos. Estos análisis se pueden realizar más fácilmente al evidenciarse visualmente mediante box plots:



En este mismo sentido, es importante revisar gráficamente si estos grupos tienen una distribución similar y mediante esta comparación de la densidad de los datos, aproximadamente podría decirse que los grupos siguen una distribución normal.

Para hacer más clara esta visualización, se usaron histogramas (siguiendo la regla de Sturges) y gráficas QQ-Plot.





Teniendo esta idea general de los datos, a continuación se presentará todo el procedimiento detrás de las pruebas de hipótesis a realizar bajo esta base de datos.

Prueba de ANOVA de un factor

La prueba ANOVA (Análisis de varianza) es una prueba paramétrica que es usada si tenemos un factor independiente y queremos analizar si al variarla, se producirán cambios significativos en una variable dependiente. En ese sentido, se utiliza para comparar las medias de 3 o más grupos y revisar si existe una diferencia entre estas, esto lo hace al contrastar las varianzas dentro de los grupos con la varianza entre los grupos y así verificar si estos provienen de una misma población. Su hipótesis nula es que las medias de los grupos son iguales, mientras que la hipótesis alternativa es que estas medias son diferentes, pero para usar esta prueba es necesario que se cumplan los siguientes supuestos:

Para poder realizar nuestra prueba nuestros datos en primera instancia tiene que cumplir con estos dos supuestos principales:

1) El primer supuesto que nuestras muestras deben cumplir es que tienen una distribución normal o de montículo.
2) El segundo supuesto es que las varianzas de las muestras son iguales, o en otras palabras que la variabilidad de los datos es la misma.

Para comprobar los supuestos se usó:

1) Para comprobar el primer supuesto se realizó a cada una de las muestras la prueba de Shapiro Wilk explicada anteriormente. Bajo la hipótesis nula de que la muestra proviene de una población con una distribución normal, y con una significancia de 5%, estos son los resultados que arrojó para cada muestra.

```
Prueba de Shapiro-Wilk para Grupo A: ShapiroResult(statistic=0.9823750257492065, pvalue=0.6555035710334778)
Prueba de Shapiro-Wilk para Grupo B: ShapiroResult(statistic=0.9870157837867737, pvalue=0.8535816669464111)
Prueba de Shapiro-Wilk para Grupo C: ShapiroResult(statistic=0.9626771211624146, pvalue=0.11509501188993454)
```

Como podemos ver ambas muestras el p-valor es mayor que la significancia, por lo tanto no hay suficiente evidencia científica para rechazar la hipótesis nula, así que ambas muestras provienen de una distribución normal.

2) Luego para comprobar el supuesto dos usamos la prueba de Bartlett, bajo la hipótesis de que las varianzas de las muestras son iguales, y el estadístico nos arrojó esto.

```
Estadística de prueba de Bartlett: 3.5258958194536407
Valor p de la prueba de Bartlett: 0.17153843791121673
Las varianzas de los grupos son iguales (no rechazamos H0)
```

Como podemos ver ambas muestras el p-valor es mayor que la significancia, por lo tanto no hay suficiente evidencia científica para rechazar la hipótesis nula, las varianzas para las tres muestras es igual.

Como podemos ver los supuestos se cumplen para el ANOVA, ahora podemos realizar la prueba sin ningún problema. Al realizar la prueba ANOVA nos dice esto :

```
Hay suficiente evidencia para rechazar Ho: Las medias poblacionales son diferentes

Tabla ANOVA:
     Fuente Var     Suma Cuad   Gl    Cuadr Med          F   Valor-p
0  Tratamientos   3429.288719    2  1714.644359
1         Error  20958.358238  147   142.573866  12.026358  0.000015
2         Total  24387.646957  149
```

Con una significancia del 5%, el p-valor nos da menor que la significancia por lo tanto existe suficiente evidencia para rechazar la hipótesis nula, eso quiere decir las medias son diferentes, hay un por lo menos un grupo significativamente diferente.

Prueba Kruskal-Wallis

La prueba ANOVA se puede aplicar correctamente bajo contextos muy específicos, por lo que otra opción cuando sus supuestos no se cumplen es su equivalente no paramétrico: Kruskal-Wallis. Esta prueba está basada en los rangos de orden de los datos y no en los valores como tal que toma la variable, por lo que la hace más robusta frente a los atípicos. En esta prueba se busca verificar si las medianas de los grupos estudiados tienen una mediana similar o si estos provienen de la misma distribución. En ese sentido, su hipótesis nula es que la tendencia central de los grupos es similar, mientras que la hipótesis alternativa es que por lo menos la tendencia central de un grupo es diferente a los demás.

En tal sentido, el estadístico de prueba de este test se distribuye Chi Cuadrado con k - 1 grados de libertad (donde k es la cantidad de grupos) y la prueba se realiza a una cola. La cola superior es tomada como la "zona de la hipótesis alternativa", puesto que a partir de cierto valor crítico de acuerdo a la significancia, se considera que el estadístico Chi Cuadrado es "grande", es decir que hay diferencias significativas, en este caso, entre las medianas de los rangos de orden de cada grupo.

Sus supuestos no son tan estrictos pero revisarlos de todas formas es de gran relevancia.

1. **La variable dependiente es ordinal o continua:**
   Este supuesto se cumple dado que la variable 'Eficacia' es continua.

2. **Independencia entre grupos:**
   Esta condición se cumple pues el efecto que tuvo un tratamiento en un grupo de personas, no influye en la eficacia de otro tratamiento en otro conjunto de personas.

3. **Las distribuciones de cada grupo tienen formas similares:**
   Para este supuesto fue utilizado el test de Kolmogorov - Smirnov para dos muestras. Inicialmente se compararon las distribuciones de los grupos de tratamiento A y B, a lo que arrojó estos resultados:

   ```
   Estadística de prueba: 0.18
   Valor p: 0.3959398631708505
   ```

   Por lo que no hay suficiente evidencia estadística para rechazar que los grupos tienen distribuciones diferentes.

   Ahora bien, como los grupos A y B tienen distribuciones similares, solo fue necesario comparar una de ellas (en este caso, A) con la distribución del grupo C. La prueba de Kolmogorov-Smirnov arrojó:

```
Estadística de prueba: 0.34
Valor p: 0.005841778142694731
```

Así que hay suficiente evidencia estadística para rechazar que los grupos tienen distribuciones similares.

En ese sentido, este supuesto no logró ser cumplido por la base de datos, así que se optó por una transformación logarítmica de los datos pues en el histograma fue evidenciada un sesgo hacia la izquierda, pero de todos modos al volver a realizar la prueba de Kolmogorov se obtuvieron exactamente los mismos resultados mencionados previamente. Para realizar la prueba de Kruskal-Wallis se utilizaron tanto los datos en escala logarítmica como los datos sin la transformación pero es necesario resaltar que los resultados obtenidos en el test pueden no llegar a ser del todo confiables por el hecho de que este supuesto no se cumplió.

Luego de haber verificado los supuestos, es posible empezar con el desarrollo de la prueba Kruskal-Wallis. En primer lugar se deben asignar rangos de orden a los datos de la variable 'Eficacia'; como no hay datos repetidos, simplemente se ordenó la base de menor a mayor de acuerdo a los valores de esta variable y se asignaron los números naturales del 1 al 150 (número total de registros).

Posteriormente, se utilizó la siguiente fórmula para obtener el estadístico de prueba H:

$$\left(\frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right)-3(N+1)$$

Donde:

- N es el total de datos en toda la base
- k es el número de grupos
- n_j es el total de datos del grupo j
- R_j es la suma cuadrada de todos los rangos del grupo j

En este caso

```
El estadístico de prueba (Valor H) es 18.45151788079471
```

y para hallar su p-valor, se busca la probabilidad acumulada en la cola superior en una distribución Chi Cuadrado con 2 grados de libertad, a lo que se obtuvo que

```
El p-valor es 9.846996891371695e-05
```

Para finalizar, el valor p obtenido es considerablemente menor a 0.05, por lo que la hipótesis nula es rechazada, lo que significa que por lo menos una de las medianas es significativamente diferente al resto, es decir que por lo menos uno de los grupos de tratamiento contra la diabetes tiene resultados diferentes al resto de grupos.

Comparación y conclusión

Para este problema, mediante ambas pruebas se llegó al mismo resultado de rechazar la hipótesis nula, lo que significa que por lo menos uno de los grupos estudiado es significativamente diferente a los demás. Para el contexto de esta base es más apropiado utilizar la prueba ANOVA, pues se cumplieron todos los supuestos (lo cuál es difícil que ocurra), haciéndola más poderosa, por lo que los resultados obtenidos son más fiables; por el contrario, para la prueba Kruskal-Wallis no se cumplió uno de los supuestos, y en este sentido los resultados arrojados pueden no ser del todo confiables.

Parte 3: Ajuste y Análisis de un Modelo de Regresión Lineal Multivariable

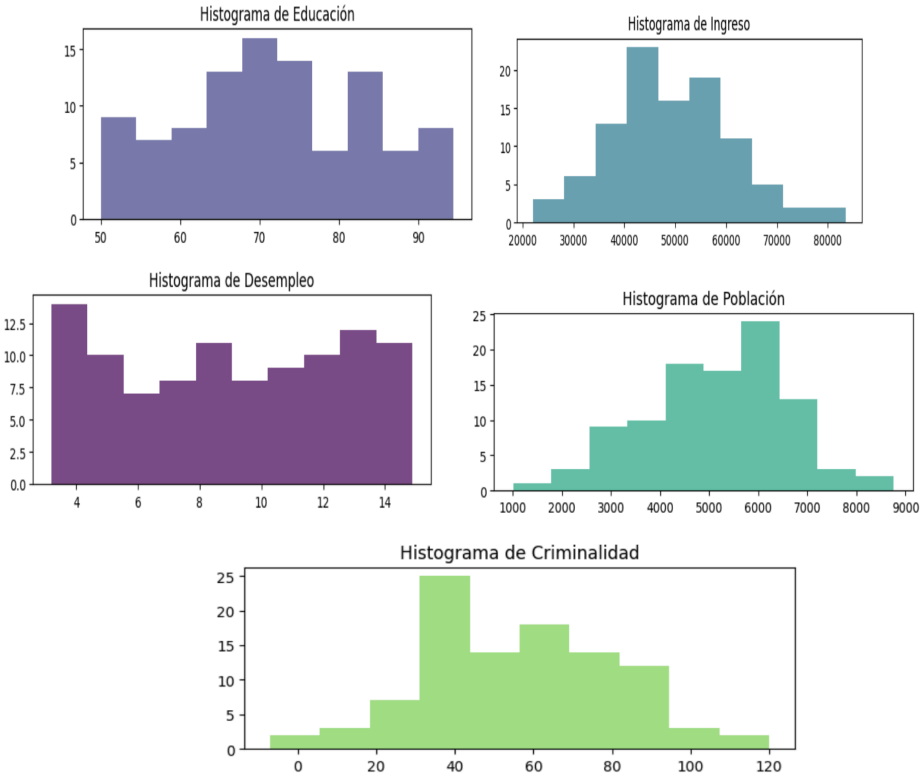
Contexto de la base:

En esta nueva base se nos proporciona información sobre condiciones económicas y demográficas de los diferentes sectores urbanos. Lo que se quiere lograr con el análisis es poder ver qué variables son las que influyen más en la tasa de criminalidad de cada distrito y con esto implementar algún tipo de solución para reducir esta. El modelo que usaremos para predecir será un modelo Regresión Lineal Multivariable ya que es lo más apropiado para predecir la relación entre variables predictoras y objetivo.

-Antes de empezar a formular el modelo, primero debemos ver gráficamente nuestras variables para así poder entenderlas bien y ver si encontramos algún tipo de patrón significante.

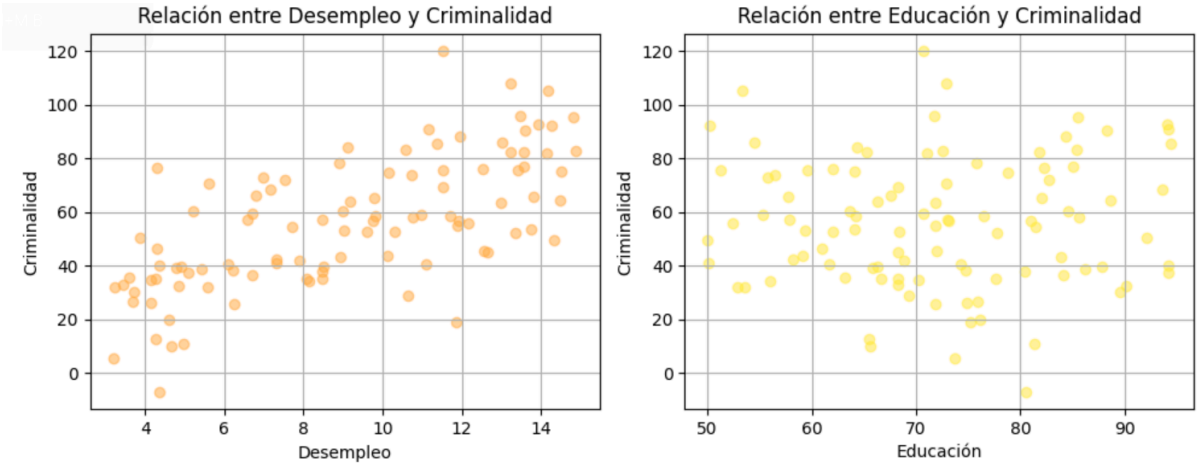
	Desempleo	Educación	Ingreso	Población	Criminalidad
count	100.000000	100.000000	100.000000	100.000000	100.000000
mean	9.074793	71.990294	49457.932857	5170.581667	55.657441
std	3.582284	11.995031	11793.016479	1404.184269	24.329416
min	3.190294	50.016370	21991.174612	1010.121483	-7.157420
25%	5.602564	63.934841	41803.845829	4355.591649	38.285927
50%	9.077954	71.825567	49346.714311	5427.366736	55.403138
75%	11.997881	81.541491	56296.475437	6180.015430	74.956475
max	14.889872	94.393488	83545.613558	8758.263876	120.095150

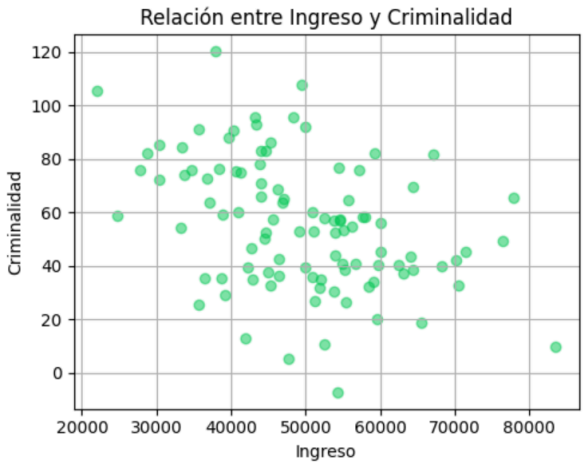
Primero esta tabla nos deja ver como información relevante como la media, mediana,valores mínimos o máximos de las variables, etc.

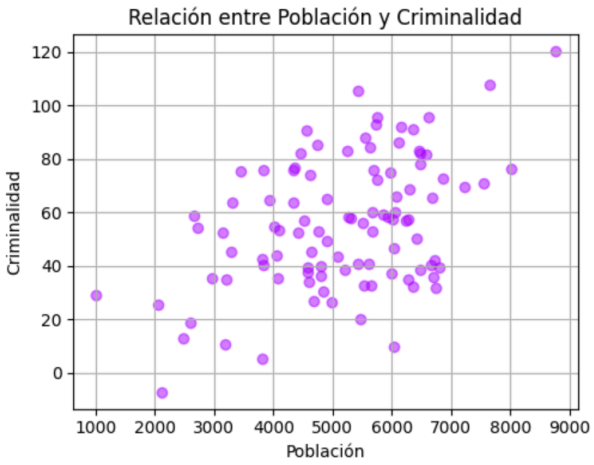


Acá en estos histogramas podemos ver visualmente como se distribuyen nuestros datos. La única gráfica que muestra como una 'anomalía' sería el histograma de 'Desempleo' ya que es la única de las gráficas que parece tener una forma Bimodal esto nos podría dar un indicio de que se están muestreando dos poblaciones,y el histograma de 'Educación' parece que toma una forma concentrada y esto puede ser porque datos provenientes de varias poblaciones con distribución normal, pero por las demás parece tener una forma normal.

-Ahora veremos como se encuentran relacionadas las variables predictoras con la variable objetivo.

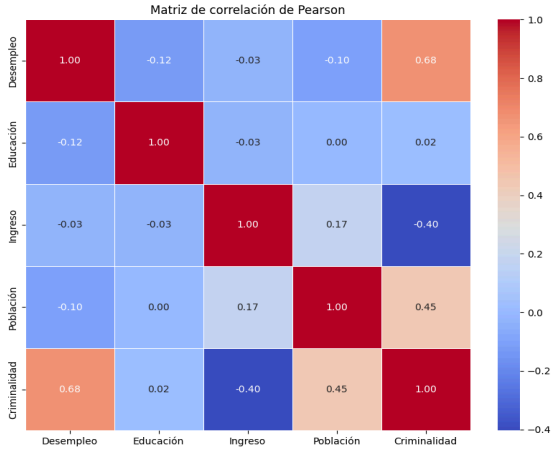


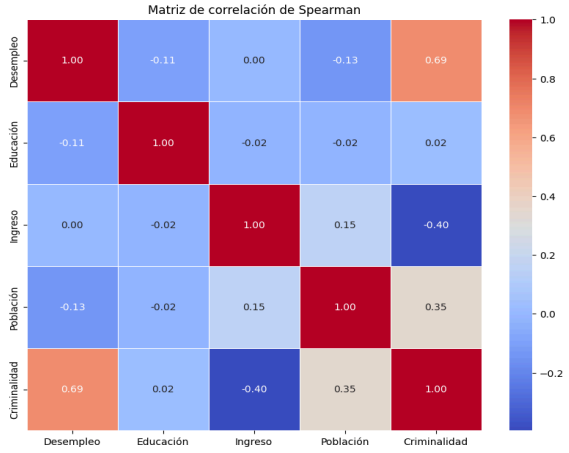




En estas gráficas podemos apreciar varias cosas, las relaciones entre 'Desempleo y Criminalidad' y 'Población y Criminalidad' parecen ser positivas lo que quiere decir que mayor sea el desempleo o la población en el distrito mayoral será la tasa de criminalidad. Por otro lado podemos ver que la relación entre 'Ingreso y Criminalidad' parece ser negativa lo que quiere decir que a mayor ingresos menos será la tasa de criminalidad y por último podemos ver la relación entre 'Educación y Criminalidad' parece ser ''continua'' no se puede ser ningún tipo de tendencia, lo que nos quiere decir que pueden haber criminales con títulos o con ciertos niveles de estudios superiores.

-Por medio de las matrices de correlación vamos a ver que tanto se relacionan las variables y así mismo aplicaremos la prueba VIF para ver la inflación de la varianza entre variables predictoras para así ver la multicolinealidad.





	Variable	VIF
0	const	79.485553
1	Desempleo	1.025018
2	Educación	1.015248
3	Ingreso	1.029761
4	Población	1.038540

Para cada una de las matrices de correlación usamos las correlaciones de Spearman y Pearson para ver la independencia de las variables. Como podemos observar no hay mucha diferencia entre una matriz y otra, ambas nos brindan la misma información varia a variables, lo más destacado de estas matrices sería que nos dejan ver que la variable más fuertemente relacionada con la criminalidad es la variable desempleo, y la más fuerte inversamente relacionada es la variable ingreso.

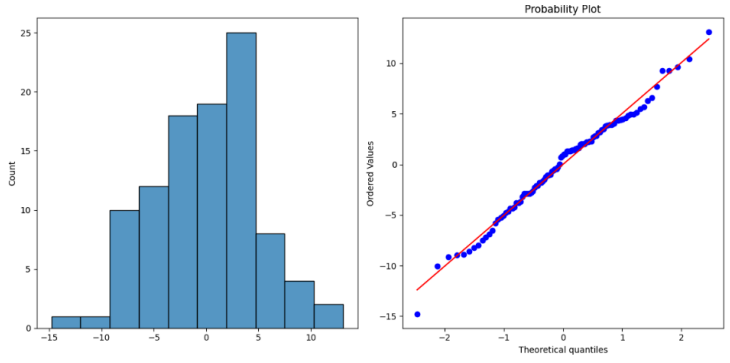
La prueba VIF nos dice que las variables predictoras presentan una fuerte multicolinealidad si muestran resultados mayores a 10, aquí podemos observar que dan valores mucho menores a eso lo que no deja concluir que las varias variables predictoras dandonos la misma información, asegurarnos de este tipo de 'independencia' nos ahorra problemas con sobreajuste más adelante con nuestro modelo.

Verificación de supuestos:

Para este apartado se realizó un modelo inicial de regresión lineal multivariable para hallar los residuos y poder comprobar si cumplían los supuestos requeridos para que el modelo pueda ser interpretado correctamente.

- **Normalidad de los residuos:**

Como la propuesta de esta base involucra una regresión lineal múltiple es importante verificar que los residuos de esta sean normales ya que los modelos de regresión lineal asumen que los residuos tienen una distribución normal con media cero y varianza constante, es por esto que consideramos importante confirmar la normalidad de los residuos y así garantizar que el modelo no tenga problemas de interpretación de los resultados y/o en la utilización del modelo para hacer predicciones o tomar decisiones.



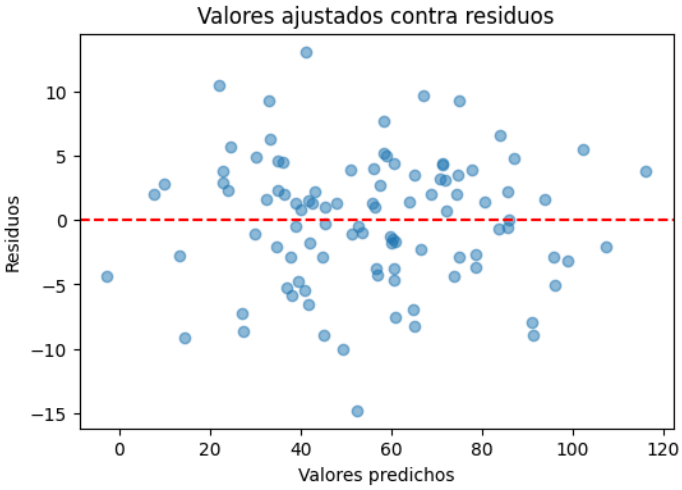
Para visualizar la distribución de los residuos se hizo un histograma junto con un diagrama QQ, la interpretación de este se puede leer en el apartado de pruebas.

Después de realizar esto revisamos que se cumpla lo que se ve en estas gráficas por lo que decidimos aplicarle la prueba de Shapiro-Wilk y de Anderson, los resultados se ven a continuación:

```
Estadística de prueba: 0.9902538657188416
Valor p: 0.6851223111152649
No se puede rechazar la hipótesis nula: los datos parecen ser normales

Estadístico de Anderson: 0.4113467530136319
Con un nivel de significancia de 5.0%, no se rechaza la hipótesis nula (los residuos se distribuyen normalmente)
```

- **Homocedasticidad de los residuos**



En un primer momento, se hizo una visualización de cómo los residuales se comportan respecto a los valores predichos, a lo que se evidenció que no hay un factor temporal en estos. Ahora bien, es necesario hacer una verificación más robusta de este supuesto.

Para este objetivo se utilizó el test de Breusch-Pagan, también conocido como la prueba de heterocedasticidad de Breusch-Pagan. Esta prueba revisa que la varianza de los errores de un modelo de regresión sea constante a lo largo de todas las predicciones, lo cual es de suma importancia dado que si no se cumple este supuesto, los resultados que salgan de la regresión no son fiables. En ese sentido, se tienen las siguientes hipótesis :

- **Hipótesis nula:** los residuos presentan homocedasticidad.
- **Hipótesis alternativa:** en los errores hay heterocedasticidad.

A grandes rasgos se tienen los siguientes supuestos:

- **Linealidad del modelo:** En el modelo de regresión lineal multivariable se tiene que la variable objetivo es una combinación lineal de las variables predictoras, por lo que se cumple el supuesto
- **Los residuos se distribuyen normal con media igual a cero:** La normalidad fue verificada previamente mediante las pruebas Shapiro-Wilk y Anderson. Asimismo, se encontró que la media de los residuos es aproximadamente cero, por lo que este supuesto también se valida.

```
1 residuos.mean()
```

```
-4.031015521377413e-12
```

- **Homocedasticidad de los errores:** este supuesto será validado mediante este test, pero se tiene la idea inicial que el modelo es homocedástico.

Para realizar la prueba de Breusch-Pagan se sigue esta serie de pasos:

1. Entrenar un modelo de regresión. Para este caso se utilizó un modelo de regresión lineal multivariable usando como variables predictoras desempleo, educación, ingreso y población, y como variable objetivo la criminalidad.
2. Calcular el cuadrado de los errores, es decir el cuadrado del valor real menos el valor predicho para cada registro.
3. Entrenar nuevamente un modelo utilizando las mismas variables explicativas, pero cambiando la variable objetivo por el cuadrado de los residuales.

Ahora, para obtener el estadístico de prueba, se usa la siguiente fórmula:

$$LM = nR^2$$

Donde:

- LM es el estadístico de prueba. Se llama así por sus siglas en inglés (Lagrange Multiplicator). Se le da este nombre por la fuerte relación que existe entre este test y la optimización por el método de Multiplicadores de Lagrange.
- n es la cantidad de registros
- R^2 es la métrica R^2 ajustado del nuevo modelo de regresión que fue entrenado con los errores cuadráticos como variable objetivo.

Este estadístico de prueba se distribuye Chi Cuadrado con k grados de libertad (donde k es la cantidad de variables predictoras), por lo que esta distribución es usada ahora para conocer el valor p. A lo que se obtuvieron los siguientes resultados:

```
Estadístico del Multiplicador de Lagrange: 5.2020500108712175
p-valor del estadístico del Multiplicador de Lagrange: 0.2671870038555879
```

Como el p-valor es mayor a la significancia de 0.05, se evidencia que no hay suficiente evidencia estadística para rechazar que los residuos del modelo de regresión lineal no sean homocedásticos.

- **Independencia de los residuos**

La prueba de Durbin-Watson es una prueba estadística utilizada principalmente para identificar la existencia de autocorrelación de primer orden en residuos (indica si existe una relación entre los residuos en un período de tiempo y los residuos en el período de tiempo siguiente).La prueba se calcula a partir de los residuos del modelo de regresión.

La interpretación de la prueba varía en si el valor calculado de DW está cerca de 0 o 4, hay evidencia de autocorrelación positiva o negativa, respectivamente. Si DW está cerca de 2, no hay autocorrelación.

Las hipótesis de las pruebas:

- **Hipótesis:**
  Hipótesis nula (H0) ausencia / no autorización de la autocorrelación de primer orden en los residuos del modelo. En términos más sencillos, los errores no están correlacionados.

  Hipótesis alternativa (H1): existe una relación lineal entre los errores en dos puntos de tiempo sucesivos

- **Supuestos:**
  1. Linealidad: el modelo de regresión es lineal de parámetros.
  2. Que los errores se distribuyen normalmente con una media de 0.
  3.Homocedasticidad: la varianza de los errores es constante en todos los niveles de las variables predictoras.

Explicado lo anterior confirmamos los supuestos:

```
1 residuos.mean()
```

```
-4.031015521377413e-12
```

Como la media es muy cercana a 0 se cumple el supuesto de la media igual a 0

Los demás supuestos fueron comprobados anteriormente, por lo que procedemos a hacer la prueba de Durbin-Watson la cual nos arrojo:

```
Estadístico Durbin-Watson: 2.120728754997438
```

Esta nos dice que no hay autocorrelación en el residuo

Transformacíon de Variables y Modelado:

Cuando realizamos análisis estadísticos no encontramos la necesidad de transformar las variables originales para cumplir con los supuestos de ciertas pruebas. En este contexto, la decisión de no transformar las variables puede ser justificada ya que las pruebas estadísticas anteriores indican que los supuestos necesarios para la regresión se cumplen.

Por lo tanto, al no realizar transformaciones cuando los supuestos de las pruebas estadísticas se cumplen, podemos mantener la simplicidad y la interpretación directa de los resultados de la regresión. Esto nos permite centrarnos en la relación entre las variables originales y comprender mejor cómo se relacionan entre sí sin introducir complicaciones adicionales en el análisis.

En ese sentido, los resultados obtenidos en el modelo son los siguientes:

	Coeficiente	**Error Estándar**	**Valor t**	**Valor p**
Término constante	-7.6867	4.540	-1.693	0.094
Desempleo	4.9806	0.145	34.435	0
Educación	0.1871	0.043	4.351	0
Ingreso	-0.0010	4.4e-05	-22.480	0
Población	0.0104	0	27.931	0

Interpretación y Validación del modelo

- **Interpretación de los coeficientes del modelo:**

Para comprender la relación entre las variables independientes y la variable dependiente en el modelo interpretaremos sus coeficientes:

```
                coef
------------------------
const         -7.6867
Desempleo      4.9806
Educación      0.1871
Ingreso       -0.0010
Población      0.0104
```

Como se mencionó anteriormente una de las variables que más influyen en la predicción de la variable objetivo criminalidad es el desempleo. En nuestro análisis, encontramos que un aumento de un punto en la tasa de desempleo se relaciona con un aumento de 4.9 puntos en la criminalidad. Por otro lado, un aumento de un punto en el ingreso se asocia con una disminución de 0.0010 puntos en la criminalidad.

Con lo anterior se puede generalizar la interpretación de los demás coeficientes de la regresión lineal múltiple.

- **Evaluación de la significancia de cada variable explicativa:**

Las pruebas de hipótesis sobre los coeficientes se hacen para saber si en la población, estos son iguales a cero, tomando esto como hipótesis nula, y que sean diferentes de cero, como hipótesis alternativa. Teniendo en cuenta lo mencionado en la tabla, es necesario resaltar el coeficiente de la constante del modelo. Se obtuvo un valor t de

-1.693 el cual equivale a un p-valor de 0.094. Este valor es mayor a la significancia de 0.05, lo que significa que no hay suficiente evidencia para rechazar que este coeficiente es diferente de cero. Al estar hablando de tasas, esto se interpreta como que el valor mínimo que puede tomar la variable de criminalidad es cero.

Para el caso de las demás variables, todos sus valores p fueron iguales a cero, lo que significa que todos estos los coeficientes en la población son significativamente diferentes de cero, es decir, que todos estos aspectos aportan a la tasa de criminalidad de cada distrito.

- **Calcula y interpreta el R^2**

En pocas palabras el R cuadrado nos indica proporcionalmente que tan bien está siendo prediciendo la variable objetivo con respecto a la variable predictora en un modelo de regresión lineal, por lo tanto entre más alto sea este quiere decir que mayor es la exactitud de nuestro modelo. En ese orden de ideas nuestro modelo calculó un R cuadrado de 0, 958 este resultado nos dice que nuestro modelo tiene una exactitud por así decirlo del 95,8% lo que es un muy buen número por lo tanto podemos concluir que nuestro modelo está bien ejecutado y tendrá muy poco margen de error al predecir.

Conclusión y Recomendaciones

A partir del análisis realizado, se encontró que lo que más aporta a la tasa de criminalidad en los distritos es el desempleo, factor que estaba más correlacionado con la variable objetivo. En general por los resultados obtenidos en las pruebas de significancia de los coeficientes y la métrica del R^2, es notorio que la regresión lineal multivariable es una buena forma de modelar la relación existente entre el desempleo, la educación, el ingreso y la población y cómo estos implican cambios en la criminalidad.

Asimismo, por la poca cantidad de variables explicativas y de observaciones, computacionalmente no hubo mayores dificultades a la hora de entrenar el modelo, ni tampoco interpretativamente a la hora de analizar los coeficientes. Pero de todas formas consideramos que aumentar el número de variables podría mejorar significativamente el modelo, pues al tenerse en cuenta más componentes que influyen en la criminalidad por distrito, esta variable puede ser explicada de una mejor forma y se pueden tomar decisiones interseccionales a la hora de combatirla.

Por lo anterior, a partir de nuestro modelo consideramos que habrían dos caminos por considerar a la hora de crear políticas públicas en materia de reducción de criminalidad:

1. **Línea de educación y empleabilidad para distritos populares:** Se encontró que por lo general, a menor desempleo, la criminalidad tiende a bajar, pero también que la criminalidad puede concentrarse más en distritos con mayor cantidad de habitantes, por lo que una propuesta inicial sería que en estos sectores en los que hay una mayor población se implementen medidas para que las personas tengan más acceso a la una educación de calidad y que se les sea garantizado un empleo estable posteriormente, y que de esta forma la tasa de criminalidad pueda reducirse contundentemente pues las personas ya no tendrán que buscar alternativas ilegales para llenar sus necesidades básicas.
2. **¡Trabajo para todos!** Siguiendo la línea de la propuesta anterior, el Estado debería asegurarse de abrir los suficientes puestos de trabajo, ya sea como servidores públicos o en alianza con grandes empresas, para que las personas al acabar sus estudios tengan distintas opciones de empleabilidad y no tengan que recurrir a la delincuencia. Por otra parte, estos empleos deben contar con salarios justos y dignos, pues los datos mostraron que a mayor ingreso medio por hogar, la tasa de criminalidad es menor, por lo que garantizar un empleo estable con buenos ingresos podría ser un punto importante en la reducción de criminalidad.

Luego de que estas propuestas sean implementadas, se podría usar este modelo para llegar a predecir si los niveles de criminalidad aumentarán o disminuirán a lo largo del tiempo y saber con mayor facilidad si estas políticas públicas lograrán tener el impacto deseado. Las decisiones basadas en modelos y las pruebas de hipótesis harán que la criminalidad disminuya y los ciudadanos sientan que pueden vivir de forma segura en su distrito.

Anexos

- **Cuadernillo en Collab - Punto 1:** Reto1_SatisfaccionClientes.ipynb
- **Cuadernillo en Collab - Punto 2:** Reto2_EficaciaTratamientos.ipynb
- **Cuadernillo en Collab - Punto 3:** Reto3_RegresionEconomica.ipynb

Fuentes:

- https://www.geeksforgeeks.org/kolmogorov-smirnov-test-ks-test/
- https://cienciadedatos.net/documentos/9_homogeneidad_de_varianza_homocedasticidad
- **h**ttps://statologos.com/prueba-de-durbin-watson/
- https://trabajofinal.es/prueba-shapiro-normalidad-ejemplo/
- https://controlautomaticoeducacion.com/control-realimentado/error-en-estado-estacionario/
- https://datatab.es/tutorial/mann-whitney-u-test
- https://statologos.com/como-interpretar-los-coeficientes-de-regresion/
- https://www.statology.org/breusch-pagan-test/
- http://guillermoneria-controlestadistico.weebly.com/164-histograma/december-03rd-2014