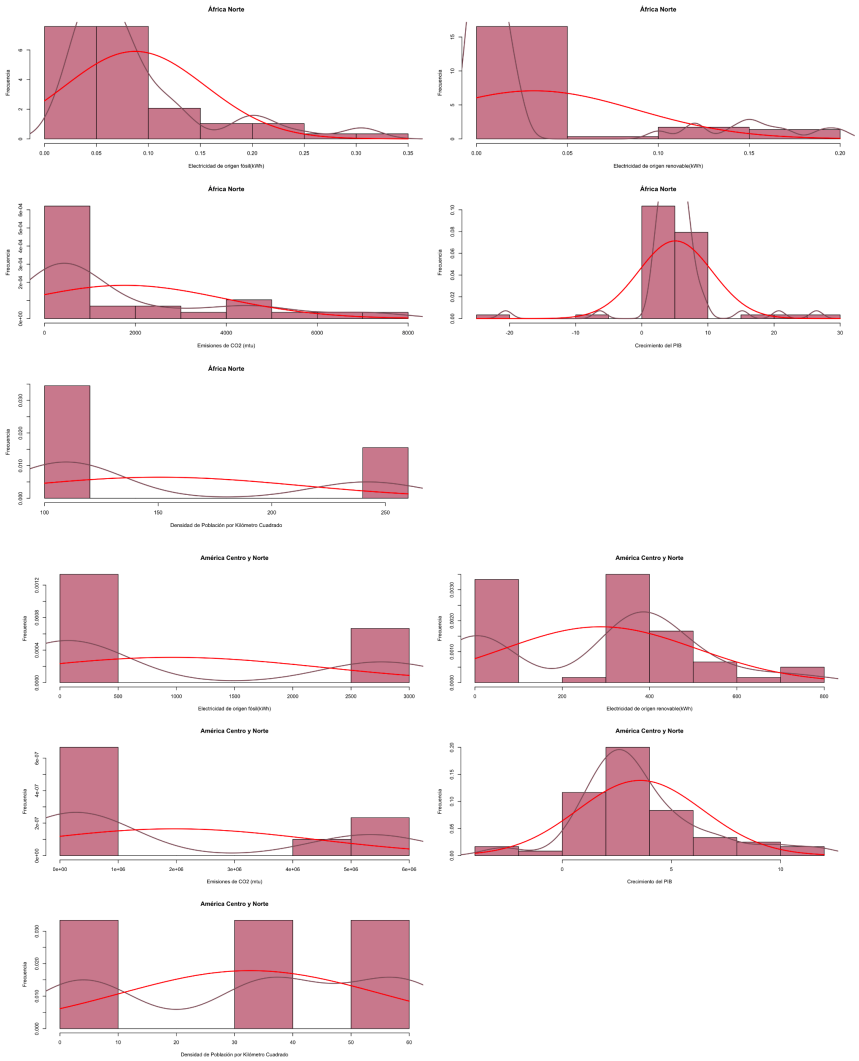
Parte 1
Metodología

Se abrió la base de datos generada en la etapa pasada y se crearon las distintas variables R1 y R2 para África Norte y América Centro y Norte respectivamente para poder analizar la electricidad de origen fósil, de origen renovable, emisiones de dióxido de carbono, y crecimiento anual del PIB.

Lo primero que se hizo fue hacer un histograma de densidad empírica con los datos de cada variable por región, después se encontró la media y desviación estándar para dicha variable y se agregó la curva de densidad empírica y después la curva de densidad normal en el mismo gráfico. Al haber 5 variables por región se crearon 10 histogramas en total con este proceso como se puede apreciar en la siguiente figura:



Histogramas para África Norte y América Centro y Norte

Dentro de los gráficos obtenidos se puede distinguir la función de densidad y los histogramas realizados en la etapa 1. Donde la función de densidad da la probabilidad relativa de cada una de las variables graficadas. Mientras que los histogramas observados sirven para mostrar la distribución y la frecuencia de datos dentro de intervalos específicos.

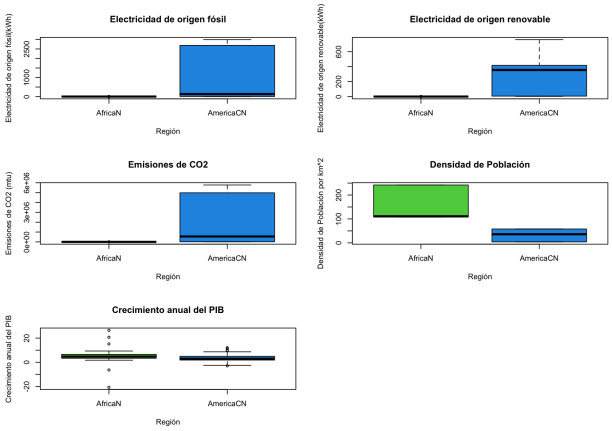
Comparando la función de densidad empírica y la función teórica, podemos observar que las distribuciones obtenidas no podrían ser consideradas como “normales”. Esto, debido a la discrepancia existente entre las curvas ya sea en el centro o con alguna desviación hacia la derecha o hacia la izquierda. Esto es posible atribuirlo a la existencia de datos atípicos o extraños que modifican en gran medida la forma de la curva. Un ejemplo visible para ambas regiones es la gráfica de densidad de población por kilómetro cuadrado. Por otra parte, las gráficas que sí se conforman a una distribución normal son aquellas de Crecimiento del PIB para ambas regiones.

Notablemente, en las gráficas de América del Centro y Norte es sencillo encontrar gráficas más simétricas y con datos con mejor imagen que aquellas de África del Norte**.**

Después se realizó una tabla con las principales medidas de centro por variables para cada región, esto se representa en la siguiente tabla:

	África Norte			**América Centro y Norte**		
	media_R1	**mediana_R1**	**Rm_R1**	**media_R2**	**mediana_R2**	**Rm_R2**
electrd_fosiles	0.08724137931	0.06	0.165	966.2656667	130.69	1494.8
electrd_de_energ_renov	0.03155172414	0.01	0.1	287.4396667	355.05	381.635
emisiones_CO2	1768.103455	664.9999917	3740.000037	1961924.179	549160.0037	2890545
crecimiento_PIB	5.075276339	4.727127094	2.90927294	3.570812885	2.988089907	4.527792692
densidad_pobl_Km2	150.6206897	111	175	32.66666667	36	31

Esto se puede comparar con los histogramas obtenidos en la etapa pasada:



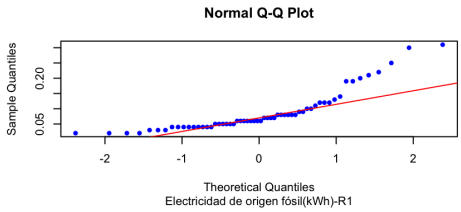
De acuerdo con los diagramas de caja y bigote, el consumo de electricidad fósil y emisiones de CO2 tiene altos máximos con una mediana baja, por lo cual en ambos se presenta una asimetría positiva, esto en el caso de la región de América. Pues la mayoría de los valores se agrupan hacia la parte izquierda ocasionando la disparidad entre el cuartil inferior y el superior.

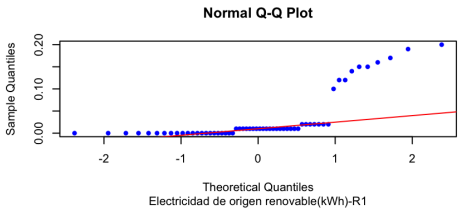
Por el lado contrario, para la electricidad de origen renovable la cual tiene una mediana más alta, se puede observar cómo es que esta presenta una asimetría negativa. Donde la mayoría de los valores se agrupan en la parte derecha, ocasionando que el tamaño del cuartil superior sea más acotado. Además, para la gráfica de la variable del crecimiento del PIB es posible remarcar que entre ambas regiones existe una simetría en la media y entre sus cuartiles. Diferenciándose principalmente por los datos atípicos de la región de África.

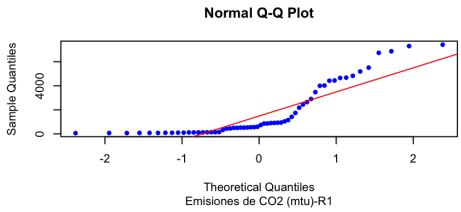
Cabe destacar que debido a la escala de los diagramas de caja y bigote América, aquellos que representan a África, salvo por el de la variable de densidad de población, son difícilmente discernibles.

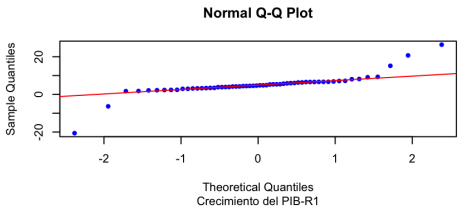
Finalmente, se hace un QQplot. En este gráfico se utilizan los datos de cada variable por región por gráfico y se comparan sus cantidades teóricas con las obtenidas en la muestra y se agrega una línea de tendencia para esta dispersión. Aquellos puntos representados por el color azul que se alejen de la línea roja en donde una gran mayoría la cumple pueden ser considerados valores atípicos ya que no van de acuerdo a una tendencia lineal.

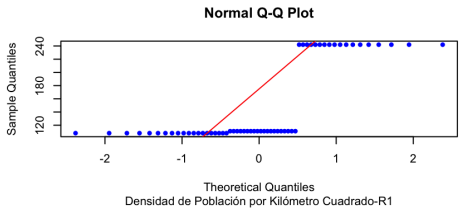
Para África Norte:



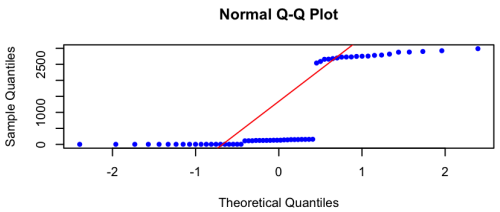


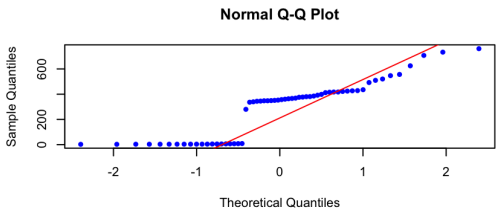


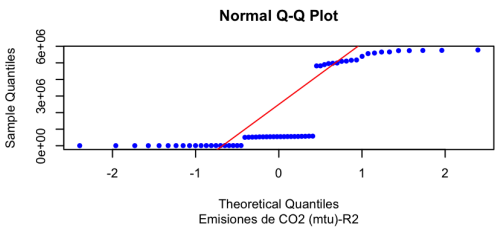


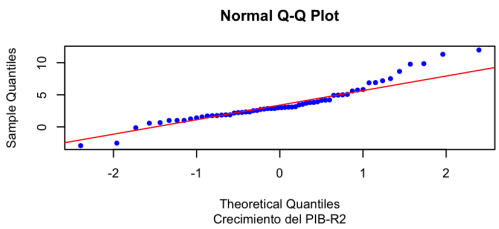


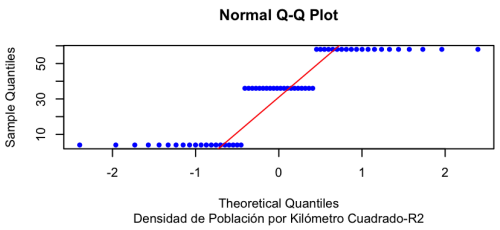
Para América Centro y Norte:











Interpreta cada gráfico: Explica el comportamiento de la gráfica (si hay o no simetría, sí está muy puntiaguda o achatada, si hay datos atípicos, etc). Concluye sobre si los datos se distribuyen aproximadamente normales o no.

Africa Norte:

Electricidad de origen fósil: La gráfica se desvía hacia la izquierda, lo que indica que la distribución no es simétrica, además de que está achatada, lo que significa que existen datos atípicos al momento de modificar los datos.

Electricidad de origen renovable: La distribución de los datos en la gráfica no es simétrica pues está ligeramente achatada. Por otro lado, en la parte superior izquierda hay datos atípicos y parece que se distribuyen como normales.

Emisiones de CO2: La gráfica parece ir en una línea recta; sin embargo, no parece que los datos se distribuyan como normales debido a que no siguen la línea de tendencia

Crecimiento del PIB: En esta gráfica, los datos parecen ir en línea recta, lo que significa que son normales; además de que no hay indicios de que esté achatada o puntiaguda y no hay datos atípicos.

Densidad de Población por Kilómetro Cuadrado: Está última gráfica no parece ser simétrica y está ligeramente achatada y no se distribuyen normalmente.

América centro y norte:

Electricidad de origen fósil: La gráfica no cuenta con una simetría desde la parte superior izquierda hacia la inferior derecha. Por otro lado, este gráfico cuenta con datos atípicos y no están distribuidos normalmente, esto se demuestra en la punta superior de la distribución.

Electricidad de origen renovable: Al igual que la gráfica anterior, el origen renovable cuenta con valores atípicos en la cola inferior de la distribución, esto no permite que haya simetría; y esto se denota en la distribución de datos, pues la cola inferior de los datos está ligeramente achatada y los datos no se distribuyen normales.

Emisiones de CO2: De igual manera, los datos no cuentan con una distribución normal, pues está ligeramente achatada; sin embargo, no hay datos que parezcan atípicos.

Crecimiento del PIB: A diferencia de las otras gráficas, a simple vista parece que los datos son aproximadamente normales, pues son más apegados a la línea de tendencia mostrada; sin embargo, está ligeramente achatada.

Densidad de Población por Kilómetro Cuadrado: Por último, la densidad de población igualmente no está distribuida normalmente y no hay simetría; además de que está achatada; sin embargo, no aparece tener datos atípicos.

Discusión y conclusiones

Las variables con más simetría son las que están relacionadas con el PIB. Esto es porque el PIB de los lugares se conectan con todas las otras variables que tenemos, siendo la cantidad de energía que cada fuente de energía produce.

Es por esto que tiene sentido que las variables del PIB sean simétricas, ya que su conexión con las otras variables resultaría en una tendencia y por consecuencia una simetría no vista en las otras variables.

En general la mayoría de estas gráficas carecen de una simetría definida, algunas cuentan con una simetría parcial más sin embargo por lo mismo no pueden ser consideradas como gráficas simétricas, y es por esto mismo que las gráficas no pueden contar con una normalidad perfecta, además de la simetría en estas gráficas también se encuentran achatamientos y valores atípicos.

Durante la primera fase analizamos opciones para las variables respuesta y las variables regresoras. Llegamos a la conclusión que teníamos dos opciones posibles.

Una de las opciones era escoger la variable de electricidad de combustibles fósiles, ya que esta parecía tener una buena correlación con las otras variables. Ahora que terminamos esta fase, vemos que la variable no tiene una correlación tan grande como esperábamos por lo que decidimos cambiar el método de encontrar dichas variables.

La segunda opción, mientras tanto, resultó ser muy similar a las tendencias que vemos ahora. Escribimos que podíamos escoger la variable del PIB para encontrar una correlación con las otras variables, y así argumentar los beneficios de usar otras fuentes de energía. Como vemos ahora, las variables relacionadas con el PIB tienen una correlación muy alta con las otras variables, mostrando que es una muy buena variable dependiente.

Parte 2
Introducción

De acuerdo con Good New Energy (2022), las fuentes de energía sostenibles son aquellas que utilizan recursos naturales sin poner en riesgo su existencia para el uso de generaciones futuras. Este término es utilizado para englobar varios tipos de energías, entre las más famosas se encuentran las energías renovables. La energía renovable se refiere a los tipos de energía que pueden reponerse más rápidamente de lo que son producidas, en la mayoría de los casos esto se refiere a recursos naturales como la energía solar o eólica (del viento) que son vastos y se pueden utilizar para generar energía sin dañar al planeta.

La necesidad de priorizar la producción de energía sostenible viene de que generan una cantidad considerablemente menor de emisiones dañinas al planeta que su contraparte: la quema de combustibles fósiles (Naciones Unidas, 2023). Estas emisiones son dañinas tanto para las personas con enfermedades causadas por una mala calidad de aire como para la naturaleza del planeta, lo que hace que la quema de combustibles fósiles no solo esté afectando en el presente, sino que también tendrá un efecto irreversible en futuras generaciones (Greenpeace México, 2021),

La energía es algo que se ha vuelto indispensable para la vida diaria de toda la población a la vez que un elemento clave del desarrollo que ésta ha estado teniendo en los últimos años. Los beneficios de este constante desarrollo están siendo opacados por los efectos negativos que causa una generación eléctrica no sustentable. Ya que la energía se consume a diario en aparatos del hogar, en casi la mitad de la industria y en transporte tanto público como privado (Junta de Castilla y León, s.f). Es por esto que durante este trabajo se pretende encontrar relación entre elementos como la producción de energía, consumo de energía, densidad de la población, crecimiento del producto interno bruto, etc para poder modelar una regresión que infiera las estadísticas de algunos de estos elementos en un futuro.

Metodología
El equipo tuvo una parte de una base de datos con elementos relacionados con la energía de 6 entidades diferentes con datos tomados entre los años 2000 y 2020. Estas entidades son parte de la región de África del Norte (“Benin”, “Sao Tome and Principe” y “Sierra Leone”) y de América Centro y Norte (“Panama”, “Canada” y “United States”).

En la primera etapa de esta situación problema se pudo notar que el coeficiente de correlación más fuerte entre las variables analizadas fue de aquellas relaciones con las emisiones de dióxido de carbono, por lo que la variable respuesta elegida es la de “emisiones_CO2”, en donde su más fuerte correlación es aquella con la variable de generación de electricidad por quema de fósiles “electrd_fosiles” con un coeficiente de 0.996 para América Centro y Norte y uno de 0.791 para África del Norte. Para cada región se hace una regresión lineal y se obtienen los siguientes datos:

América Centro y Norte					**África del Norte**				
	Estimate	Std. Error	t value	Pr(>|t|)		Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	144961	33182	4.369	0.00005236	(Intercept)	-452	289	-1.564	0.1234
x	1880	20.77	90.52	3.807e-64	x	25448	2627	9.686	0.000000000001434

Regresión de América Centro y Norte: $y = 1880x + 144961$
Regresión de África del Norte: $y = 25448 - 452$
(Donde x son las emisiones de CO2 y “y” la generación de energía fósil)

América Centro y Norte				**África del Norte**			
Observations	Residual Std. Error	R^2	Adjusted	Observations	Residual Std. Error	R^2	Adjusted
60	204665	0.993	0.9928	58	1341	0.6262	0.6195

Se puede interpretar por el coeficiente de determinación de cada región que la relación entre las emisiones de dióxido de carbono en América Centro y Norte tiene una correlación alta y en África del Norte hay una correlación moderada. Esto significa que muchas más emisiones de dióxido de carbono provienen de la generación de energía a partir de combustibles fósiles. Después, se genera un gráfico con la correlación y los residuales:

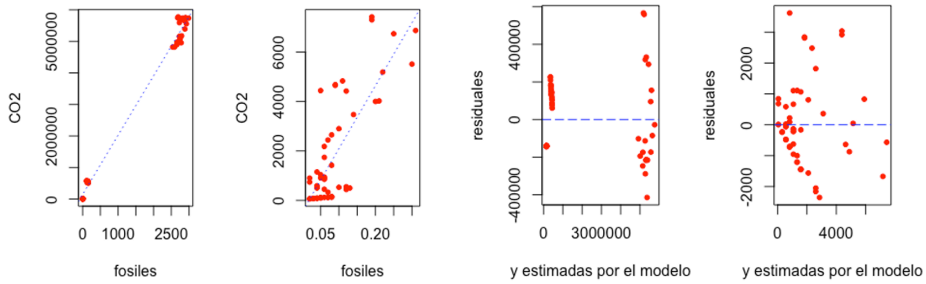
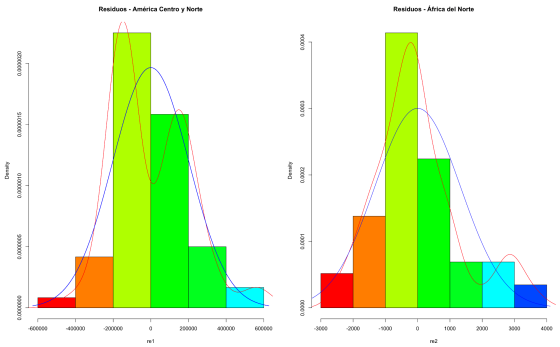
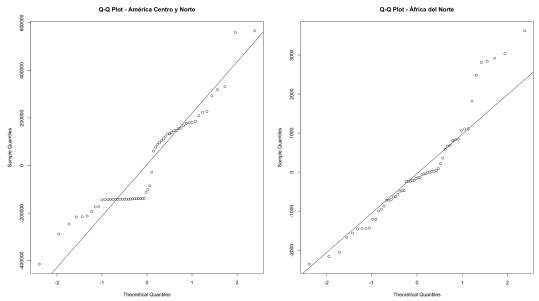


Gráfico de Correlación y de Residuales para América Centro y Norte y África del Norte

Luego, se plantea la hipótesis nula de que el promedio de los residuales sea igual a 0 y la hipótesis alternativa de que el promedio de los residuales sea diferente a 0. Para esto se utilizó una representación visual de un histograma con los datos y después una función de distribución normal con la media y desviación estándar de cada región para poder ver si el valor de 0 se encuentra cerca de la media.



Como se puede ver por la distribución normal de la línea azúl, la media de los residuales de ambas regiones se encuentra en el 0, aunque en la tabla anterior se pudo notar que el error estándar de los residuales es considerablemente alto. Para esto se utilizaron las pruebas de Shapiro-Wilk para determinar si el modelo se adapta a las cualidades de una distribución normal y de Breuch-Pagan para determinar si existe homocedasticidad en la varianza residual:



América Centro y Norte		**África del Norte**	
W	0.90556	W	0.93076
p-value	0.000211	p-value	0.00261

Se rechaza la hipótesis nula de que los datos provienen de una población normal con significancia de 5% y se acepta la hipótesis alternativa de que no provienen de una población normal con la prueba de Shapiro-Wilk para ambas regiones.

América Centro y Norte		**África del Norte**	
BP	11.234	BP	3.5817
df	1	df	1
p-value	0.000803	p-value	0.05842

Se rechaza la hipótesis nula de que hay homocedasticidad en la varianza residual para América Centro y Norte pero se acepta para África del Norte con significancia de 5% y se acepta la hipótesis alternativa de que no hay homocedasticidad en la varianza residual con la prueba de Breuch-Pagan para América Centro y norte.

Conclusiones

Mientras que se cumplió con un modelo de regresión lineal para poder representar los datos, no sería correcto decir que es un buen modelo solamente con ver coeficiente de determinación de los datos porque aunque se trate de una moderada y alta para cada región, no se puede decir que los residuales tienen una media de 0, lo cual sería ideal para un modelo lineal correcto. Esto debido a que aunque los datos de los residuales tienen visualmente una media de 0 al ser representados en un histograma con una función que represente su normalidad, estos datos no pasan la prueba para saber si siguen un comportamiento normal y solo África del Norte pasa la prueba que dicta que existe homocedasticidad en la variancia residual de la correlación con un valor “p” muy cercano al de una significancia del 5%.

Al no pasar la prueba de Shapiro-Wilk, se puede concluir que los residuos del modelo lineal para ambas regiones no tiene un error estándar que se pueda estimar correctamente (University of Wisconsin Madison, 2021). Al no pasar la prueba de Breuch-Pagan, se conoce que el modelo lineal para América Centro y Norte no tiene una variación constante en sus datos y ésta va cambiando según el valor de las emisiones de dióxido de carbono. Mientras que la región de África del Norte sí pasó está prueba y se puede decir que la varianza es constante a un nivel de significancia de 5% (XLSTAT, 2023).

En conclusión, el modelo lineal que infiere la cantidad de electricidad generada a partir de fuentes fósiles basándose en la cantidad de emisiones de dióxido de carbono en la región no es óptimo. Aunque se podría decir que el modelo de África del Norte es mejor al tener homocedasticidad en su varianza residual, su determinación es moderada a comparación de la determinación alta de América Centro y Norte.

Autoevaluación

- **Propósito:** ¿Cuál cree que es el propósito de aprendizaje de la actividad que realizó? (mencione qué es lo que piensa que su profesor quería que aprendiera cuando realizó esta actividad)

  El propósito de esta actividad es poder aplicar los contenidos vistos en clase junto con una herramienta de programación y análisis estadístico como lo es R en un caso del mundo real en donde se pudo juntar en análisis matemático, la modelación gráfica y las pruebas establecidas para comprobar una hipótesis para llegar a una conclusión sobre la relación entre los datos y la capacidad de representarla utilizando un modelo lineal.

- **Estrategia:** ¿Qué plan siguió para lograr estos objetivos? ¿Cambió de plan más de una vez? Si lo hizo, ¿qué fue lo que le hizo cambiarlo? Coméntelo. ¿Fue claro lo que tenía que hacer? ¿En qué momento le quedó claro cuál era el objetivo de aprendizaje?

  Para lograr estos objetivos tuve que familiarizarme con el lenguaje de programación R y entender cómo transferir la teoría de lo que vimos en clase a los distintos comandos para poder utilizarlos durante esta situación problema. No hubo cambios de planes pero si muchos obstáculos al momento de utilizar la programación ya que en teoría los procedimientos y el análisis estaba bien pero hubo veces en donde se graficó una región dos veces o bajo el nombre de la otra región por lo que ayudó mucho estar verificando si hacía sentido lo que se obtuvo visualmente con los datos en la base de datos. Siempre fue claro lo que tenía que hacer, desde el principio me di cuenta de que el objetivo de aprendizaje era poder aplicar los conocimientos vistos en clase a una situación con datos reales.

- **Resultado:** ¿Logró el propósito de aprendizaje que enunció arriba? ¿Qué fue lo más importante que aprendió de esta actividad? Enuncia al menos las tres cosas más importantes que haya aprendido de ella.

  Sí, logré mi propósito de aprendizaje ya que pude familiarizarme con el lenguaje de programación R, las distintas formas de poder interpretar datos de una base de datos y las maneras en las que temas como distribución normal o residuales son importantes para poder probar una hipótesis ya que no basta con una sola característica como tener un alto coeficiente de determinación. Lo más importante que aprendí fueron las diversas pruebas para determinar si un modelo lineal es apropiado para una muestra de datos.

Referencias

4 Normality | Regression Diagnostics with R. (2021). https://sscc.wisc.edu/sscc/pubs/RegDiag-R/normality.html

Collection, D. G. N. G. I. (2022, 30 mayo). Para luchar contra el cambio climático son necesarias desesperadamente fuentes de energía más limpias, pero algunos expertos dicen que hay que hacer más para mantener intactos los ecosistemas desérticos. *National Geographic*. https://www.nationalgeographic.es/medio-ambiente/2022/05/por-que-las-energias-renovables-pueden-suponer-una-amenaza-para-la-biodiversidad

Greenpeace México. (s. f.). *¿Cómo afectan los combustibles fósiles a la salud humana? - Greenpeace México*. https://www.greenpeace.org/mexico/blog/9853/como-afectan-los-combustibles-fosiles-a-la-salud-humana/

Junta de Castilla y León. (s. f.). *APLICACIONES DE LA ELECTRICIDAD*. https://energia.jcyl.es/web/es/biblioteca/suministro-electricidad-aplicaciones.html

Pezzi, S. (2022, 19 enero). *La energía sostenible y el valor de la eficiencia*. Good New Energy. https://goodnewenergy.enagas.es/sostenibles/la-energia-sostenible-y-el-valor-de-la-eficiencia/

United Nations. (2023). *¿Qué son las energías renovables? | Naciones Unidas*. https://www.un.org/es/climatechange/what-is-renewable-energy

XLSTAT. (2023). *Tests heterocedasticidad Breusch-Pagan & White*. https://help.xlstat.com/es/6656-breusch-pagan-white-heteroscedasticity-tests-excel