ENTREGA FINAL PROYECTO: UBER VS LYFT

FECHA DE ENTREGA: 21 de Nov de 2022

PROFESORA: Ana María Beltrán Cortés

ASIGNATURA: Probabilidad y Estadística

INTEGRANTES:

- David Santiago Buitrago Prada
- Danna Sofia Marin Guacaneme
- Julián David Tovar Gaitán
- Santiago Uribe Luna

DESARROLLO:

1. **Descripción del problema que se ha identificado**
   En los últimos años, las plataformas que prestan el servicio de alquiler de automóvil con conductor han incrementado su uso a nivel mundial, lo anterior por la facilidad que representa para el usuario solicitar un automóvil a cualquier hora del día desde una aplicación en su celular y así mismo ha sido bien acogida por los consumidores pues se ha convertido en una forma segura de conseguir transporte debido a que se tiene los datos claros del conductor, placas del vehículo y hasta se permite rastrear el viaje.
   Aterrizando este contexto a nuestro proyecto, hemos decidido escoger la base de datos "Uber and Lyft Dataset Boston" la cual nos facilita información sobre el tema en la ciudad de Boston, Estados Unidos. Aquí encontramos las dos plataformas más utilizadas en la zona, las cuales son Uber y Lytf. El banco de datos nos suministra cifras que nos permiten poner a competir las ventajas de ambas empresas, ahora, reconociendo la gran oportunidad que puede significar para muchas personas trabajar con este tipo de plataformas, puesto que puede ser un método de conseguir ingresos adicionales y más cuando la tarifa estándar de estas aplicaciones se multiplica, esto también puede ocasionar disgusto en los usuarios ya que genera un incremento en los costos de los viajes.
   Por lo anterior se han presentado varias novedades en noticieros o diarios locales que expresan molestia por parte de los pasajeros pues según el diario El Planeta (2018) se han encontrado con la incómoda situación de pagar tarifas inusualmente altas en comparación con otros tiempos, entonces teniendo en cuenta lo anterior, el objetivo principal de este proyecto es demostrar y analizar ¿Cuáles factores afectan el costo del servicio en las plataformas de movilidad en la ciudad de Boston, EE.UU, durante los meses de noviembre y diciembre?

2. **Acerca de la base de datos**
   Nuestra base fue extraída de la plataforma Kaggle y a priori se obtuvieron cerca de 693.070 filas y 57 columnas. Al empezar el análisis se realizó una limpieza de filas repetidas o vacías y datos poco útiles para tener una base más exacta, finalmente se consiguieron 637.976 filas y 27 columnas para trabajar.
   Decidimos no tener en cuenta las variables que mencionaremos en la seccion del anexo, expuestas en el apartado "variables eliminadas"

Todas las relacionadas con la temperatura se descartaron pues en una sola variable ya se almacena la temperatura puntual que se registró al pedir el servicio, así mismo con todos los relacionados con "Time", son un serial que hace constancia del registro de la información.

3. **Diccionario de variables:** Revisar anexo, apartado "Diccionario de variables"

4. **4.1. Análisis descriptivo de datos**

Para el desarrollo del análisis descriptivo se identificaron dos variables cualitativas y tres cuantitativas principales. En primer lugar, los datos registrados nos permitieron observar la preferencia de los usuarios por una de las dos plataformas, Uber con un 51% de viajes del total consignado en la base, es la plataforma de preferencia

cab_type	count	%
<chr>	<int>	<dbl>
Lyft	307408	48.18488
Uber	330568	51.81512

para continuar se analizan las siguientes variables cualitativas: tiempo meteorológico y tipo de servicio por plataforma, de acuerdo con lo anterior, se muestran las tablas de frecuencia absoluta y relativa.

short_summary	count	%
<chr>	<int>	<dbl>
Clear	80256	12.579784
Drizzle	6725	1.054115
Foggy	8292	1.299735
Light Rain	50488	7.913777
Mostly Cloudy	134603	21.098443
Overcast	201429	31.573131
Partly Cloudy	117226	18.374672
Possible Drizzle	17176	2.692264
Rain	21781	3.414078

Viendo la tabla nos podemos percatar que los días que más se utilizó este tipo de servicio son aquellos que estaba nublado el cielo ya sea muy nublado o parcialmente nublado; esta condición concentra aproximadamente el 70% de los viajes realizados en las aplicaciones. Mientras que los días con lluvia, neblina, rayos, etc. aproximadamente conforman un 16% del total de viajes y el restante fueron días despejados. Por ende, podríamos sospechar que el clima si tiene una afectación para este tipo de compañías.

En cuanto a los distintos tipos de servicio por parte de Uber y Lyft tenemos:

name	count	%
<chr>	<int>	<dbl>
Lux	51235	16.66678
Lux Black	51235	16.66678
Lux Black XL	51235	16.66678
Lyft	51235	16.66678
Lyft XL	51235	16.66678
Shared	51233	16.66612

name	count	%
<chr>	<int>	<dbl>
Black	55095	16.66677
Black SUV	55096	16.66707
UberPool	55091	16.66556
UberX	55094	16.66646
UberXL	55096	16.66707
WAV	55096	16.66707

Tanto la plataforma Uber como Lyft ofrecen seis tipos de servicio para los cuales en la muestra se tomaron en torno a 50000 registros con proporciones iguales.
En segundo lugar, sobre las variables cuantitativas: distancia (dada en kilómetros) y precio (dado en dólares) el análisis se divide en tres partes: medidas de tendencia central, medidas dispersión y medidas de forma.

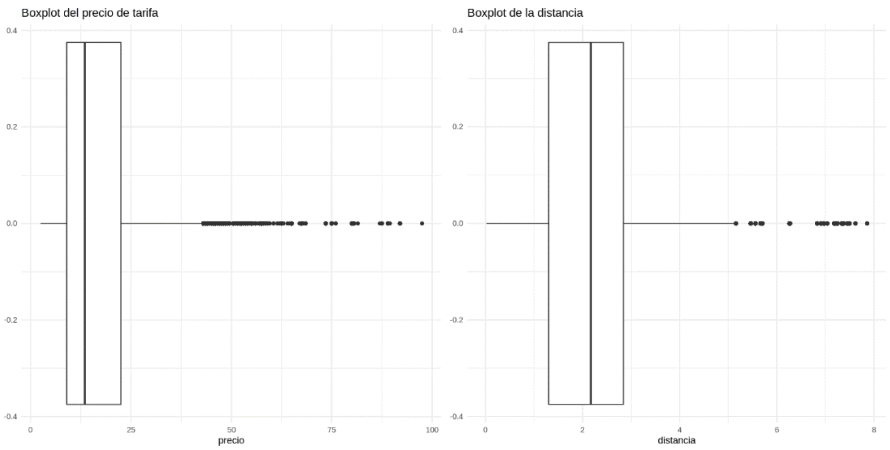
Para el precio y la distancia se emplean los siguientes resúmenes.

```
     price          distance
Min.   : 2.50   Min.   :0.020
1st Qu.: 9.00   1st Qu.:1.270
Median :13.50   Median :2.160
Mean   :16.55   Mean   :2.189
3rd Qu.:22.50   3rd Qu.:2.930
Max.   :97.50   Max.   :7.860
```

```
[ ]  mfv(newrideshare$price)

     7
```

Se concluye que el precio promedio de los viajes es de 16.55 dólares. Además, hay tantos viajes que superan 13.5 dólares como viajes con un precio menor; el precio más frecuente es de 7 dólares. Por parte de la distancia, en promedio se realizaron viajes de 2.189 kilómetros con un 50% por debajo de 2.16 kilómetros y otro 50% superior a este valor.

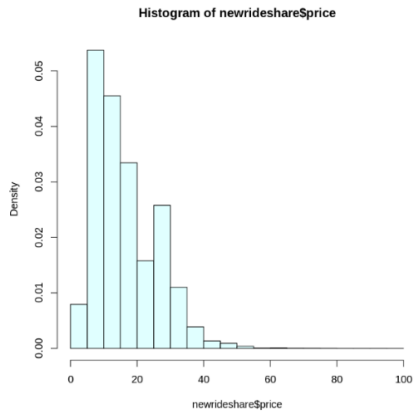
Para las medidas de posición se construyen los diagramas de caja con los cuartiles dados en las tablas anteriores.



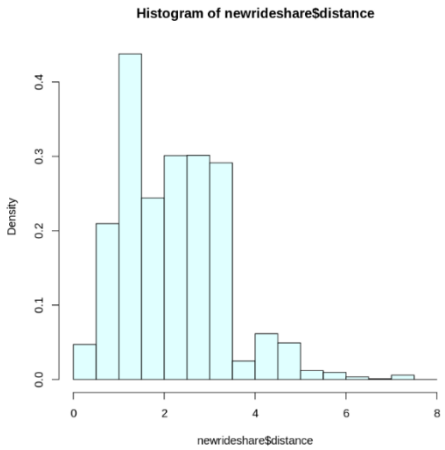
De acuerdo con las gráficas, el precio toma valores desde los 2.5 hasta los 97.5 dólares con un rango de 95. Es claro que, el 25% de las tarifas más baratas pagadas por los usuarios de estas aplicaciones no supera los 9 dólares. También, las tarifas por encima de los 22,5 dólares conforman el 25% más costoso de la base; el rango intercuartílico es de 13.5. En cuanto a la distancia, toma valores desde los 0.02 hasta los 7.86 kilómetros con un rango de 7.08598. Los viajes entre el 25% más cortos son menores a 1.27 kilómetros y el 25% de los viajes más largos son mayores a 2.93 kilómetros; el rango intercuartílico es de 1.66.

Para el análisis de dispersión se utilizan las siguientes herramientas: varianza y desviación muestral. El precio presenta una varianza muestral de 86.94 y una desviación estándar de 9.32 lo cual indica que, en promedio las tarifas se encuentran a 9.32 dólares de la media. Repitiendo el proceso para la distancia, varianza y desviación muestrales son respectivamente 1.29 y 1.14.

Finalmente, sobre el comportamiento visual de las gráficas y gracias a los coeficientes de Asimetría y curtosis se puede evidenciar en el precio una asimetría positiva con coeficiente CAp=1.05>0 y de tipo leptocúrtica o de colas pesadas con CKp=1.22>0.



Para la distancia se tiene una curva asimétrica positiva con coeficiente CAp=0.82>0 y colas pesadas puesto que CKp =1.15 > 0.

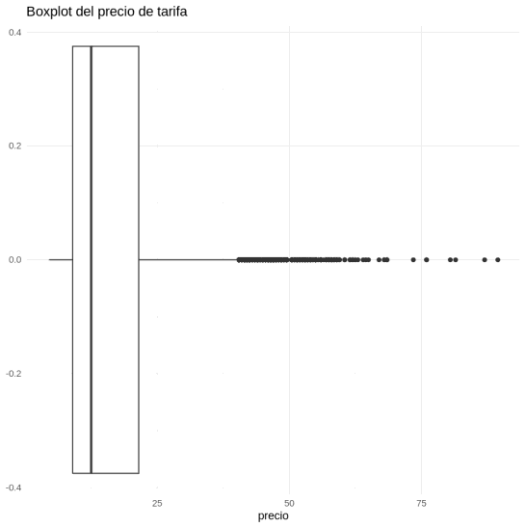


4.2. Análisis descriptivo de datos por plataforma

4.2.1 Uber: Para el precio y la distancia se aplica el siguiente análisis:

price		distance	
Min.	: 4.5	Min.	:0.020
1st Qu.	: 9.0	1st Qu.	:1.300
Median	:12.5	Median	:2.170
Mean	:15.8	Mean	:2.191
3rd Qu.	:21.5	3rd Qu.	:2.840
Max.	:89.5	Max.	:7.860

Partiendo de la gráfica anterior, podemos ver un precio mínimo de 4.5 USD y un máximo de 89.6 USD, el valor en promedio de la muestra de los viajes es de 15.8 USD.

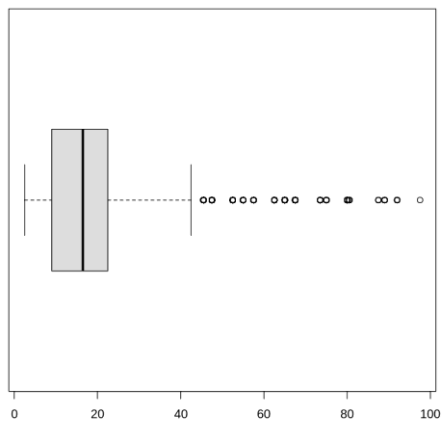


Evidenciamos que, el 25% de las tarifas más baratas pagadas por los usuarios no superan los 9 dólares. Así mismo, las tarifas por encima de los 21,5 dólares conforman el 25% más costoso de la base; el rango intercuartílico es de 12.5. Para el análisis de dispersión, observando la varianza y desviación muestral, notamos que el precio presenta una varianza muestral de 73.27 y una desviación estándar de 8.56 lo cual indica que, en promedio las tarifas de Uber se encuentran a 8.56 dólares de la media.

4.2.2 Lyft: Para el precio y la distancia se aplica el siguiente análisis:

price		distance	
Min.	: 2.50	Min.	:0.390
1st Qu.	: 9.00	1st Qu.	:1.270
Median	:16.50	Median	:2.140
Mean	:17.35	Mean	:2.187
3rd Qu.	:22.50	3rd Qu.	:2.970
Max.	:97.50	Max.	:6.330

Partiendo de la gráfica anterior, podemos ver un precio mínimo de 2.5 USD y un máximo de 97.5 USD, el valor en promedio de la muestra de los viajes es de 16.8 USD.



Notamos en la caja de bigotes, que el 25% de las tarifas más baratas pagadas por los usuarios no supera los 9 dólares. Las tarifas por encima de los 22,5 dólares conforman el 25% más costoso de los datos; el rango intercuartílico es de 13.5. Para el análisis de dispersión, consideramos la varianza y desviación muestral. El precio presenta una varianza muestral de 100.38 y una desviación estándar de 10.01 lo cual indica que, en promedio las tarifas de Lyft se encuentran a 10.1 dólares de la media. Repitiendo el proceso para la distancia, varianza y desviación muestrales son respectivamente 1.28 y 1.13.

4.3 Analisis Comparativo: Uber vs Lyft

Después del análisis descriptivo, podemos evidenciar datos importantes sobre el precio de estas plataformas, aunque Lyft registra la tarifa más baja entre ambas, es esta la aplicación más costosa, pues tanto su promedio como su máximo registro superan por unos dólares a los valores de Uber. Así mismo si nos fijamos en la desviación, Lyft tiene una desviación mayor que Uber (en aproximadamente 2 USD), lo cual nos dice que la tarifa de Lyft esta más alejada de su media que la de Uber.

5. **Anexo notebook de Colab con las salidas descriptivas y análisis asociados al dataset**
   https://colab.research.google.com/drive/1aZR1uoNgptz5P_XnkCIAZBcLbkJ23BDu?usp=sharing

6. **Propuesta de posibles hipótesis**

$$U = Uber \quad L = Lyft \quad P = Precio$$

Como se presenta en el análisis descriptivo, la evidencia muestral sugiere una preferencia entra las plataformas donde Uber es más usada que Lyft, por lo tanto, cabe preguntar si:

- ¿La evidencia muestral es suficiente para argumentar que la proporción de viajes en la plataforma Uber es mayor que la de Lyft?
  H0: πU ≤ πL
  Ha: πU > πL

Por otro lado, teniendo en cuenta las medias muestrales del precio para Uber y Lyft expuestas en el análisis descriptivo, resulta de interés saber si:

- ¿La evidencia muestral comprueba que el precio medio de Lyft es mayor al precio medio de Uber?
  H0: μpU ≥ μpL
  Ha: μpU < μpL

7. **Estimadores puntuales y por intervalo**
   Como se analizó anteriormente, el precio tiene una gráfica asimétrica de comportamiento no normal por lo que se realiza la transformación ln (U), ln (L) y se calculan los estimadores sobre estas variables
   Uber:
   7.1. Estimador puntual para precio promedio:

```
[ ]  mean(log(newrideshare_uber$price))
     2.62609351987929
```

Estimador por intervalos de confianza para precio: con una confianza del 97.5% se construye el intervalo que contiene el precio promedio de la muestra el cual es: (2.624;2.628)

Lyft:

7.2: Estimador puntual para precio promedio:

```
[ ]  mean(log(newrideshare_lyft$price))

     2.67520103940255
```

Estimador por intervalo de confianza para precio: con una confianza del 97.5% se construye el intervalo que contiene el precio promedio de la muestra el cual es: (2.672;2.677)

8. **Pruebas de hipótesis**

- ¿La evidencia muestral es suficiente para argumentar que la proporción de viajes en la plataforma Uber es mayor que en Lyft?

Por p-valeu se calcula sobre la evidencia muestral y suponiendo H0 cierta

$$Pu - Pl \sim N(0, 0.5006(1-0.5006)((\frac{1}{330568}) + (\frac{1}{307408})))$$

```
prop.test(x=c(330568,307408), n=c(637976, 637976),alternative='greater', conf.level=0.025)

        2-sample test for equality of proportions with continuity correction

data:  c(330568, 307408) out of c(637976, 637976)
X-squared = 1681.4, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
2.5 percent confidence interval:
 0.03803472 1.00000000
sample estimates:
   prop 1    prop 2
0.5181512 0.4818488
```

Análisis: Con una significancia del 0.025, se rechaza Ho y se concluye que, si hay una diferencia en la proporción de viajes registrados en la base de datos por parte de Uber y Lyft, esto quiere decir que los usuarios tienen mayor preferencia por la plataforma Uber ya que el IC para la diferencia de proporciones no contiene a 0 y es positivo.

- ¿La evidencia muestral comprueba medias distintas para el precio entre los viajes por cada plataforma?

Usando la evidencia muestra y bajo H0 cierta $Ln(U) - Ln(L) \sim N(0; 0,288)$

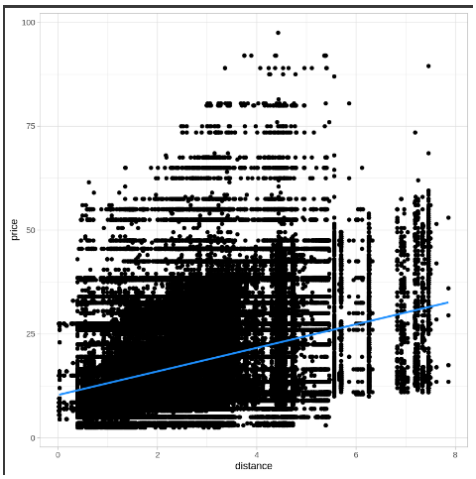
por el metodo de p-valeu:

```
[ ]  t.test(x=log(newrideshare_uber$price), y=log(newrideshare_lyft$price), alternative="less",conf.level=0.025)

            Welch Two Sample t-test

data:  log(newrideshare_uber$price) and log(newrideshare_lyft$price)
t = -34.236, df = 592959, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
2.5 percent confidence interval:
       -Inf -0.0519189
sample estimates:
mean of x mean of y
 2.626094  2.675201

t.test(x=log(newrideshare_uber$price), y=log(newrideshare_lyft$price), alternative="two.sided",conf.level=0.025)$conf.int
-0.0491524708068701 -0.0490625682396557
```

Análisis: Con una significancia del 0.25 se rechaza Ho y teniendo en cuenta el IC para la diferencia de medias (-0.04915; -0.04906) que es negativo y no contiene a 0 se concluye que el precio promedio de Lyft es mayor que el de Uber.

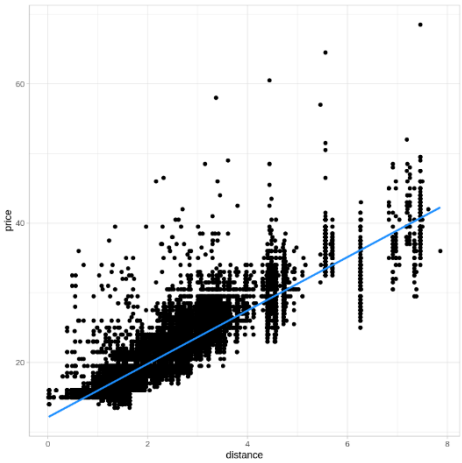
9. **Regresión lineal**

- Análisis de diagramas de dispersión:

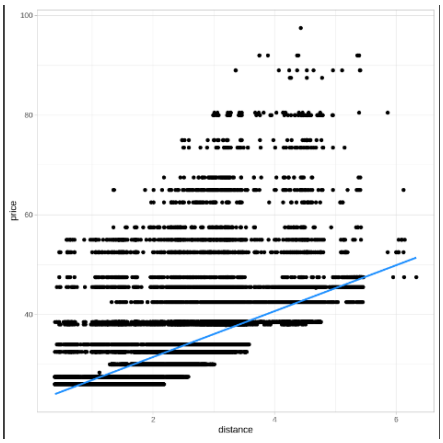


Es de interés utilizar las variables de distancia y precio con el objetivo de determinar si existe alguna relación. Determinando el ρ para las plataformas, se calcula que ρ=0.34 y también teniendo en cuenta la gráfica se puede suponer la inexistencia de asociación lineal entre las variables.

Ahora analizando la distancia con los tipos de servicios preferidos por los usuarios, encontramos resultados interesantes.



Filtrando el análisis por los tres tipos de servicios más utilizados por los usuarios de Uber con respecto a la distancia y determinando el coeficiente de relación lineal (ρ). Podemos concluir que Uber Black tiene una fuerte asociación lineal con respecto a la distancia ya que al calcular el ρ de Uber Black nos arroja un valor de 0.913, el cual es > 0.8.



En Lyft, al igual que en Uber tomamos los tres tipos de servicios que más frecuentan solicitar los usuarios y determinado el coeficiente de correlación lineal (ρ). Podemos concluir que Lyft XL tiene una fuerte asociación lineal con respecto a la distancia ya que al calcular el ρ de Lyft nos da un valor de 0.815, el cual es > 0.8

Una vez realizada la regresión para el servicio Uber black y Lyft XL, se busca determinar el grado de ajuste del modelo con respecto al CME. Para esto se desarrollan las pruebas de hipótesis con una significancia del 2,5%. Por supuestos de normalidad en las distribuciones de los residuales se sabe que CMR/CME se distribuyen F en parámetros 1 y n-2.
Para las pruebas de hipótesis de cola derecha, se consideran Ho ciertas.

$$Ho: \frac{CMR}{CME} = 1 \text{ vs } H1: \frac{CMR}{CME} > 1.$$

- Tabla ANOVA: Uber

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor critico de F
Regresión	1	1126705.4	CMR=224342.7	Fp=276690.8867	p-value
Residuos	55093	224342.7	CME=55093		2.2e-16
Total	55094	1351048.1			

Como el P−valor < 0,025 se rechaza Ho y se concluye que el modelo: Y=12.1182 + 3.83 X + ε describe el comportamiento del precio respecto a la distancia.

-Tabla ANOVA: LYFT

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor critico de F
Regresión	1	705355	CMR=357889	Fp=100973.9129	p-value
Residuos	51233	357889	CME=51233		2.2e-16
Total	51234	1063224			

Como el P−valor < 0,025 se rechaza Ho y se concluye que el modelo: Y=7.84 + 3.41 X + ε describe el comportamiento del precio respecto a la distancia

Resultado de la regresión:

Uber:

	Coeficientes	Error Típico	Estadístico t	Probabilidad	Intervalo de confianza	
Intercepción	β_o=12.1182	2.018	Tp_o=6.0050	2.2e-16	β_o=12.1175	β_o=12.1187
Pérdida	β_1=3.8357		Tp_1=1.900	2.2e-16	β_1=3.8855	β_1=3.8859

En Uber como β_1=3.8357 entonces el precio se incrementa 3.8 dólares por cada unidad de distancia recorrida. Por otra parte, en promedio el valor inicial para los viajes en la plataforma de Uber es β_o=12.1182.

Lyft:

	Coeficientes	Error Típico	Estadístico t	Probabilidad	Intervalo de confianza	
Intercepción	β_o=7.84169	2.643	Tp_o=2.9669	2.2e-16	β_o=7.8408	β_o=7.8425
Pérdida	β_1=3.41462		Tp_1=1.2919	2.2e-16	β_1=3.4142	β_1=3.4149

En Lyft como β_1=3.41462 entonces el valor aumenta 3.4 dólares por cada unidad de distancia recorrida. Por otro lado, en promedio el costo inicial para los viajes en la aplicación de Lyft es β_o=7.84169.

De acuerdo con lo anterior, la distancia impacta más la tarifa al servicio de Uber (Uber Black) que al servicio de Lyft (Lyft XL) y además la tarifa mínima para un viaje es mas costosa en Uber que el Lyft

10. **Prueba de hipótesis sobre bondad de ajuste (Para variable continúa discretizada):**
Se utiliza la regla de Sturges para definir el número de intervalos con los cuales se dividen los datos, se utiliza el estadístico de prueba X_p^2 y se concluye por medio del método de valor crítico; se busca demostrar los supuestos:

- **La distribución normal de los residuales del modelo para Uber Black**:

$$Ho: \epsilon \sim N(\mu = 0, \sigma^2 = 4.0721) \text{ Vs } H1 = \neg Ho$$

Se halla que se deben definir 14 clases, sin embargo, el valor esperado para las ultimas 6 clases es menor a 5; luego se fusionan y quedan 5 clases.

Clase	Valor	Ob	Pi	Esperado(Pi)
1	[-11.241,-8.622)	24	0.0167035	920.278213
2	[-8.622,-6.0028)	73	0.0535200	2948.682986
3	[-6.0028,-3.3836)	1183	0.1327852	7315.801594
4	[-3.3836,-0.76442)	18874	0.2225389	12260.779858
5	[-0.76442,1.8548)	26948	0.2500740	13777.826463
6	[1.8548,4.4739)	6907	0.1884228	10381.151562
7	[4.4739,7.0931)	842	0.0951908	5244.535652
8	[7.0931,9.7123)	129	0.0322273	1775.563818
9	[9.7123,12.331)	34	0.0073074	402.601475
10	[12.331,14.951)	33	0.0011097	61.137637
11	[14.951,33.285)	48	0.0001125	6.200531

Con una significancia de 2.5%, puesto que $X_c^2 = 54439.33 > X_p^2 = 31992.3676$ no rechaza Ho y se concluye que los residuales se distribuyen normal de parámetros μ = 0 y σ^2 = 4.072, esto soporta el modelo de regresión ajustado a los precios de Uber Black con respecto a la distancia.

La distribución normal de los residuales del modelo para Lyft XL:

$$Ho: \epsilon \sim N(\mu = 0, \sigma^2 = 6.9855) \text{ Vs } H1 = \neg Ho$$

Se halla que se deben definir 17 clases, sin embargo, el valor esperado para las ultimas 6 clases es menor a 5; luego se fusionan y quedan 11 clases.

Clase	Valor	Ob	Pi	Esperado(Pi)
1	[-4.4688,-1.7007)	9754	0.4038240	20689.92112
2	[-1.7007,1.0675)	31859	0.1569046	8039.00493
3	[1.0675,3.8356)	7125	0.1477963	7572.34242
4	[3.8356,6.6038)	1107	0.1192372	6109.11726
5	[6.6038,9.372)	567	0.0823810	4220.79170
6	[9.372,12.14)	390	0.0487415	2497.27069
7	[12.14,14.908)	176	0.0246996	1265.48498
8	[14.908,17.676)	118	0.0107190	549.18716
9	[17.676,20.445)	61	0.0039844	204.14121
10	[20.445,23.213)	30	0.0012672	64.92703
11	[23.213,42.59)	48	0.0004452	22.81149

Con una significancia de 2.5%, que $X_c^2 = 50602.53 < X_p^2 = 86846.9693$ rechaza Ho y se concluye que los residuales no se distribuyen normal de parámetros μ = 0 y σ^2 = 6.985, esto significa que el modelo lineal no es el que mejor describe el comportamiento de Lyft XL con respecto a la distancia.

11. Resumen ejecutivo del problema:

Teniendo en cuenta que en la actualidad el uso de plataformas de transporte es tan común, quisimos conocer un poco más sobre cómo funcionan sus tarifas, en especial saber las razones por las cuales pueden varias sus precios. La meta principal de nuestro análisis era conocer los factores que generaban un aumento en el valor de los viajes de las plataformas de transporte.

Las herramientas principales para desarrollar nuestro trabajo comenzaron por la base de datos conseguida en Kaggle, este banco almacena miles de datos acerca de las dos plataformas principales de transporte por alquiler en la ciudad de Boston, Estados Unidos, nuestra mano derecha para conseguir de forma más precisa los datos que le darían rumbo a la investigación fue la implementación de R, esta nos facilitó gráficos exactos y el cálculo de coeficientes con los cuales se desarrolló un análisis descriptivo. Se llevaron a cabo dos modelos de regresión lineal y se comprobó la capacidad para describir los datos recogidos usando análisis de varianzas y pruebas de hipótesis de bondad de ajuste para los residuales.

Partiendo del análisis descriptivo general, encontramos que el valor en promedio que pagan los Bostonianos por este tipo de servicio ronda entre los 16 USD, así mismo los usuarios registrados tienen mayor preferencia por la aplicación de Uber y esto podría deberse a otro dato importante que descubrimos y es que Uber en promedio tiene una tarifa más económica que Lyft. Quisimos analizar si existía algún tipo de relación entre la distancia del viaje con respecto al costo en la totalidad de los registros, pero este fue un parámetro que pudimos descartar con rapidez pues evidenciamos que no tienen algún tipo de relación. Se procede a filtrar la base de datos con respecto a los 3 servicios mas frecuentes de cada plataforma; así encontramos que los servicios Uber Black y Lyft XL tiene una fuerte correlación lineal entre la distancia y el precio. Para las pruebas de hipótesis de análisis de varianza se encontró que por el método del p-valor se concluye con gran contundencia la capacidad para describir el precio con respecto a la distancia, sin embargo, con las pruebas de bondad de ajuste, los residuales del modelo para Lyft XL no son normales.

Finalizando el análisis de nuestra base datos pudimos encontrar dos principales variables que impactan el costo de la tarifa, estas son la distancia y el tipo de servicio que solicita el usuario. En cuanto al tipo de servicio, se encontró que todos comparten una correlación lineal entre el precio y la distancia, pero se estudiaron las 2 que tenían mayor coeficiente de correlación (una por plataforma), lo anterior nos permitió descubrir que por cada kilómetro recorrido en los servicios Uber Black y Lyft XL, el costo aumenta entre (3.4USD;3.9USD). El proyecto ha sido limitado porque se ejecutó el estudio de regresión solo para los dos servicios principales de mejor ajuste, además, se rechazó el supuesto de normalidad para los residuales de Lyft por lo que las conclusiones de este modelo se ven limitadas. Si se busca solicitar un servicio en alguna de estas plataformas de transporte a un valor económico, se debe considerar la distancia del viaje y tipo de servicio solicitado, ya que el precio mínimo que se puede pagar por un servicio de corta distancia en el servicio Uber Black es mayor al cobro que se puede generar en Lyft. Aunque la media de la tarifa de los viajes realizados en Lyft es mayor que la de Uber, esto se puede deber a que los usuarios realizan viajes mas largos en Lyft que en Uber ya que tanto para viajes largos como para viajes cortos, resulta menos costoso contratar el servicio Lyft XL.

12. Anexos:

Diccionario de variables:

- **Hora**
  Nombre: Hour
  Tipo: Cuantitativa Discreta
  Definición: Registra el numero de la hora que se solicitó el servicio
  Unidades: Int
- **Dia**
  Nombre: Day
  Tipo: Cuantitativa Discreta
  Definición: Registra el numero del dia que se solicito el servicio
  Unidades: Int
- **Mes**
  Nombre: Month
  Tipo: Cuantitativa Discreta
  Definicion: Registra el numero del mes que se solicito el servicio
  Unidades: Int
- **Dato del Tiempo**
  Nombre: DataTime
  Tipo: Cuantitativa Continua
  Definición: Registra la fecha y hora exacta en la que se solicitó el servicio
  Unidades: Chr
- **Origen**
  Nombre: Source
  Tipo: Cualitativa Nominal
  Definición: Lugar desde donde se solicitó el servicio
  Unidades: Chr
- **Destino**
  Nombre: Destination
  Tipo: Cualitativa Nominal
  Definición: Lugar donde finaliza el servicio
  Unidades: Chr
- **Aplicación empleada**
  Nombre: Cab_Type
  Tipo: Cualitativa Nominal
  Definición: Plataforma que escogió el usuario para realizar el viaje
  Categorías: Uber, Lyft
  Unidades: Chr

- **Tipo de servicio solicitado**
  Nombre: Product_id
  Tipo: Cualitativa Nominal
  Definición: Registra el id del servicio (para Lyft existen categorías, por el contrario, para Uber es un serial que no brinda mayor información)
  Categorías para Lyft: lyft_line, lyft_lux, lyft_luxsuv, lyft_plus, lyft_premier
  Unidades: Chr
- **Nombre del servicio**
  Nombre: Name
  Tipo: Cualitativo Nominal
  Definición: Registra el nombre del servicio que el cliente quiere pagar, (estos servicios se diferencian porque tienen más o menos beneficios que otros)
  Categorías para Uber: Uber Pool, Uber X, Uber XL, WAV, Black SUV
  Categorías para Lytf: Lux, Lux Black, Lux Black XL, Lyft, Lyft XL, Shared.
  Unidades: Chr
- **Precio**
  Nombre: Price
  Tipo: Cuantitativo Continuo
  Definición: Registra el valor del viaje
  Unidades: Num
- **Distancia recorrida**
  Nombre: Distance
  Tipo: Cuantitativo Continua
  Definicion: Marca la distancia que recorrió el usuario
  Unidades: Num
- **Multiplicación del precio**
  Nombre: Surge_multiplier
  Tipo: Cuantitativo Discreto
  Definición: Registra el % de aumento que tiene el valor del servicio respecto a la tarifa estándar.
  Categorías: 1, 1.25, 1.5, 1.75, 2, 2.5
  Unidades: Num
- **Temperatura**
  Nombre: Temperature
  Tipo: Cuantitativo Continua
  Definición: Registra la temperatura que se presentó al momento de solicitar el servicio.
  Unidades: Num

- **Corto resumen**
  Nombre: Short _Summary
  Tipo: Cualitativo Nominal
  Definición: Registra en un corto texto, información sobre el clima que se presentó al momento de solicitar el servicio
  Categorías: Despejado, llovizna, nublado, lluvia ligera, mayormente nublado, parcialmente nublado, posible llovizna, lluvia
  Unidades: Chr
- **Largo resumen**
  Nombre: Long_Summary
  Tipo: Cualitativo Nominal
  Definición: Registra información más detallada sobre el clima que se presentó al momento de solicitar el servicio
  Categorías: Niebla por la mañana, lluvia ligera por la mañana, lluvia ligera por la mañana y durante la noche, lluvia ligera hasta la noche, mayormente nublado durante todo el día, nublado durante todo el día, parcialmente nublado durante todo el día, posible llovizna por la mañana, lluvia por la mañana y por la tarde, lluvia durante todo el día, lluvia hasta la mañana, comenzando de nuevo por la tarde.
  Unidades: Chr
- **Intensidad de precipitación**
  Nombre: PrecipIntensity
  Tipo: Cuantitativo Continua
  Definición: Guarda el registro en mililitros de la precipitación que se presentó en el momento que se solicitó el servicio
  Unidades: Num
- **Probabilidad de precipitación**
  Nombre: PrecipProbability
  Tipo: Cuantitativo Continua
  Definición: Registra la probabilidad de lluvia que se presento en el momento que se solicito el servicio
  Unidades: Num
- **Humedad**
  Nombre: Humidity
  Tipo: Cuantitativo Continua
  Definición: Registra el % de humedad que se present al momento de solicitor el servicio
  Unidades: Num
- **Velocidad del viento**
  Nombre: WindSpeed
  Tipo: Cuantitativa Continua

Definición: Guarda el valor de la velocidad del viento en kilómetros por hora. Unidades: Num

- **Clima**
  Nombre:Icon
  Tipo: Cualitativa Nominal
  Definición: Registra en texto el clima que se present al momento de solicitor el servicio
  Unidades: Chr
- **Nublar**
  Nombre:CloudCover
  Tipo: Cualitativa Continua
  Definición: Registra el % de nubes que se presentó al solicitar el servicio
  Unidades: Num
- **Incite de rayos UV**
  Nombre: UVIndex
  Tipo: Cualitativa Discreta
  Definición: Registra el índice de rayos UV
  Categorías: 0, 1, 2 (todos los registros indican bajo riesgo)
  Unidades: Int

13. Variables eliminadas:

Decidimos no tener en cuenta las siguientes variables:

- ”Timestamp”
- ”Latitude”
- ”Longitude”
- ”WindGustTime”
- ”ApparentTemperature”
- ”ApparentTemperatureHigh”
- ”ApparentTemperatureHighTime”
- ”ApparentTemperatureLow”
- ”ApparentTemperatureLowTime”
- ”TemperatureHighTime”
- ”TemperatureHigh”
- ”TemperatureLowTime”
- ”TemperatureLow”
- ”TemperatureMin”
- ”TemperatureMax”
- ”TemperatureMinTime”
- ”TemperatureMaxTime”
- ”ApparentTemperatureMin”
- ”ApparentTemperatureMinTime”
- ”ApparentTemperatureMax”
- ”ApparentTemperatureMaxTime”
- ”Pressure”
- ”UvIndexTime”
- ”Visibility.1”
- ”precipIntensityMax”
- ”Ozone”
- ”SunriseTime”
- ”SunsetTime”
- “Ozone”
- ”MoonPhase”

Verificación de normalidad para los residuales

Modelo: Uber Black

Prueba KS

```
Ksn<- ks.test(modelo2$residuals, "pnorm", mean =Ajusten$estimate[1], sd= Ajusten$estimate[2])
Ksn
```

```
Warning message in ks.test.default(modelo2$residuals, "pnorm", mean = Ajusten$estimate[1], :
“ties should not be present for the Kolmogorov-Smirnov test”

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  modelo2$residuals
D = 0.040237, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Prueba Anderson Darling

```
Adn<- ad.test(modelo2$residuals, "pnorm", mean =Ajusten$estimate[1], sd= Ajusten$estimate[2])
Adn
```

```
        Anderson-Darling test of goodness-of-fit
        Null hypothesis: Normal distribution
        with parameters mean = 7.08876726634716e-17, sd = 2.01790121879715
        Parameters assumed to be fixed

data:  modelo2$residuals
An = Inf, p-value = 1.089e-08
```

Prueba Shapiro-Wilk

```
Swn<-shapiro.test(modelo2$residuals[1:5000])#como Shapiro-Wilk trabaja hasta con 5000 datos
Swn
```

```
        Shapiro-Wilk normality test

data:  modelo2$residuals[1:5000]
W = 0.93858, p-value < 2.2e-16
```

Lyft XL

Prueba KS

```
Ksn<- ks.test(modeloB$residuals, "pnorm", mean =Ajustenl$estimate[1], sd= Ajustenl$estimate[2])
Ksn
```

```
Warning message in ks.test.default(modeloB$residuals, "pnorm", mean = Ajustenl$estimate[1], :
“ties should not be present for the Kolmogorov-Smirnov test”

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  modeloB$residuals
D = 0.16102, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Prueba Anderson-Darling

```
Adnl<- ad.test(modeloB$residuals, "pnorm", mean =Ajustenl$estimate[1], sd= Ajustenl$estimate[2])
Adnl
```

```
        Anderson-Darling test of goodness-of-fit
        Null hypothesis: Normal distribution
        with parameters mean = -4.169325853087e-17, sd = 2.64296093447202
        Parameters assumed to be fixed

data:  modeloB$residuals
An = Inf, p-value = 1.171e-08
```

Prueba Shapiro-Wilk

```
SwnL<-shapiro.test(modeloB$residuals[1:5000])#como Shapiro-Wilk trabaja hasta con 5000 datos
SwnL
```

```
        Shapiro-Wilk normality test

data:  modeloB$residuals[1:5000]
W = 0.72781, p-value < 2.2e-16
```

Verificación de normalidad para la transformación del precio ln(U), ln(L)

Modelo Uber:

Prueba KS

```
Ksn<- ks.test(log(newrideshare_uber$price), "pnorm",mean=Ajusten$estimate[1],sd=Ajusten$estimate[2])
Ksn
```

```
Warning message in ks.test.default(log(newrideshare_uber$price), "pnorm", mean = Ajusten$estimate[1], :
“ties should not be present for the Kolmogorov-Smirnov test”

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  log(newrideshare_uber$price)
D = 0.11941, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Prueba Anderson-Darling

```
#require(goftest)
Adn<- ad.test(log(newrideshare_uber$price), "pnorm",mean=Ajusten$estimate[1],sd=Ajusten$estimate[2])
Adn
```

```
        Anderson-Darling test of goodness-of-fit
        Null hypothesis: Normal distribution
        with parameters mean = 2.62609351987929, sd = 0.508704737096341
        Parameters assumed to be fixed

data:  log(newrideshare_uber$price)
An = 5674.2, p-value = 1.815e-09
```

Prueba Shapiro-Wilk

```
sw <- log(newrideshare_uber$price)
Swn<-shapiro.test(sw[1:5000])#como Shapiro-Wilk trabaja hasta con 5000 datos
Swn
```

```
        Shapiro-Wilk normality test

data:  sw[1:5000]
W = 0.95034, p-value < 2.2e-16
```

Modelo Lyft XL

Prueba KS

```
Ksn<- ks.test(log(newrideshare_lyft$price), "pnorm", mean =Ajustenl$estimate[1], sd= Ajustenl$estimate[2])
Ksn
```

```
Warning message in ks.test.default(log(newrideshare_lyft$price), "pnorm", mean = Ajustenl$estimate[1], :
“ties should not be present for the Kolmogorov-Smirnov test”

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  log(newrideshare_lyft$price)
D = 0.10865, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Prueba Anderson-Darling

```
Adnl<- ad.test(log(newrideshare_lyft$price), "pnorm", mean =Ajustenl$estimate[1], sd= Ajustenl$estimate[2])
Adnl
```

```
        Anderson-Darling test of goodness-of-fit
        Null hypothesis: Normal distribution
        with parameters mean = 2.67520103940255, sd = 0.625973434090293
        Parameters assumed to be fixed

data:  log(newrideshare_lyft$price)
An = 2735.3, p-value = 1.952e-09
```

Prueba Shapiro-Wilk

```
SwnL<-shapiro.test(log(newrideshare_lyft$price)[1:5000])#como Shapiro-Wilk trabaja hasta con 5000 datos
SwnL
```

```
        Shapiro-Wilk normality test

data:  log(newrideshare_lyft$price)[1:5000]
W = 0.96997, p-value < 2.2e-16
```