Taller #1 – Electiva I

Jhon Sebastián Aparicio Mesa - 201721324, Carlos Alberto Salamanca Sanchez - 201711906, Erika Valentina Tinjacá Cely – 201711674

Hipótesis

Los climas de lluvia provocan un aumento en las malas entregas. Esto obedece a la idea de que los climas de lluvia, considerados dentro de la escala brindada por la alcaldía como clima desfavorable, puede llegar a aumentar la cantidad de bicicletas entregadas tarde o en malas condiciones.

¿Por qué le podría servir a la alcaldía?

El demostrar esta hipótesis puede alertar a la alcaldía que en estos días puede generarse una mayor cantidad de entregas malas. Para subsanar esta falla la alcaldía puede generar estrategias ya sea ampliar el tiempo de préstamo o imponer una pequeña multa o penalización.

Análisis de calidad de los datos

1. Cargue del conjunto de datos distinguiendo NA.

```
library(readr)
datosalcaldia <- read_delim("C:/Users/carlo/Documents/semestre10/electiva i/clases/ejemplosclase10/dataset_c
                            , delim=",", na="")
View(datosalcaldia)
```

Imagen 1. Uso de función read_delim para cargar el conjunto de datos

El cargue del dataset se hace por medio de la librería '*readr*' y la función *'read_delim'*. Para el caso los Na se distinguieron como un especio vacio.

2. Identificación de variables necesarias para la validación de la hipótesis, de modo que se descarten las variables que no son necesarias.

```
View(datosalcaldia)
```

Imagen 2. Uso de función View para explorar el conjunto de datos

Por medio del comando *'View'* se visualiza a completitud el dataset. Con esto logramos identificar con que variables se cuenta en la información. En este paso se identifica que con las variables *'categoria_clima'* y *'mala_entrega'* se puede obtener información relevante para la alcaldia.

3. Verificación del tipo de dato que estas variables contienen, si son categóricas o no. Haciendo uso de la función *'class'* se obtiene que las variables son de tipo numéricas.

```
class(datosalcaldia$categoria_clima)
class(datosalcaldia$mala_entrega)
```

Imagen 3. Uso de función class para determinar el tipo de dato

4. Conteo de la cantidad de valores NA, para así generar una medida de acción.

```
sum(is.na(datosalcaldia$categoria_clima))
sum(sum(is.na(datosalcaldia$mala_entrega)))
```

Imagen 4. Uso de función sum en el conteo de datos faltantes

Por medio de la función *'sum'* se obtiene la cantidad de registros que son Na de cada una de las variables identificadas. El resultado de esta exploración es que en ninguna de estas variables se encuentran valores nulos.

5. Se hace también una validación de las variables restantes.
6. Modificación de variables:
   6.1. **'año'**: Se elimina debido a que esta información está reflejada en la variable fecha.
   6.2. **'mes'**: Se elimina debido a que esta información está reflejada en la variable fecha.
   6.3. **'dia_laboral'**: Se elimina debido a que implícitamente esta información está contenida en 'dia_semana.'
   6.4. **'temperatura'**: Se elimina ya que esta información no es relevante para corroborar la hipótesis.
   6.5. **'Humedad'**: Se elimina ya que esta información no es relevante para corroborar la hipótesis.
   6.6. **'Velocidad'**: Se elimina ya que esta información no es relevante para corroborar la hipótesis

```
datosalcaldia<-select(datosalcaldia,-'a-o',-mes,-dia_laboral,-temperatura,-velocidad,-humedad)
```

Imagen 5. Uso de función select para filtrar las columnas deseadas en el conjunto de datos

Por medio de la aplicación de la función '*Select'* se eliminan las variables mencionadas en los ítems 6.1, 6.2, 6.3, 6.4, 6.5 y 6.6.

6.7. **'temporada'**: Conversión de variable numérica a categórica con el fin de evitar la operación aritmética de estos datos. Se cambian los números a su correspondiente estación.

```
datosalcaldia$temporada<- cut(datosalcaldia$temporada, breaks=c(0,1,2,3,4),
                              labels=c("Primavera","Verano","Otoño","Invierno"))
```

Imagen 6. Uso de función cut para convertir la variable 'temporada'

La función '*Cut'* permite convertir una variable numérica a categórica. Se transforma la información original la cual esta presentada en valores del 1 a 4 a su estación correspondiente.

6.8. Cambio de nombre variable **'temporada'** a **'estación'**, para brindar más claridad a la variable. Por medio de la función '*Rename'* se aplica el cambio de nombre de la variable.

```
datosalcaldia<-rename(datosalcaldia, estacion=temporada)
```

Imagen 7. Uso de función rename para renombrar la variable 'temporada'

6.9. Cambio de nombre variable **'categoría_clima'** a **'estado_clima',** para brindar más claridad a la variable. Por medio de la función '*Rename'* se aplica el cambio de nombre de la variable.

```
datosalcaldia<-rename(datosalcaldia, estado_clima=categoria_clima)
```

Imagen 8. Uso de función rename para renombrar la variable 'categoria_clima'

6.10. Cambio de nombre variable '**instante'** a **'Nro_registro'**, para brindar más claridad a la variable. Por medio de la función '*Rename'* se aplica el cambio de nombre de la variable.

```
datosalcaldia<-rename(datosalcaldia, Nro_registro=instante)
```

Imagen 9. Uso de función rename para renombrar la variable 'instante'

7. Validación de la hipótesis

7.1. Verificar si las variables estado_clima y mala_entrega tienen una distribución normal y aplicar el test de normalidad en cada una de las variables.

En primer lugar, se representa la información en forma de histograma con la finalidad de ver si los datos de cada variable podrían presentar una distribución normal. A continuación, en la figura 11. se presenta el histograma de la variable estado_clima y en la figura 12. se presenta el histograma de la variable mala_entrega.

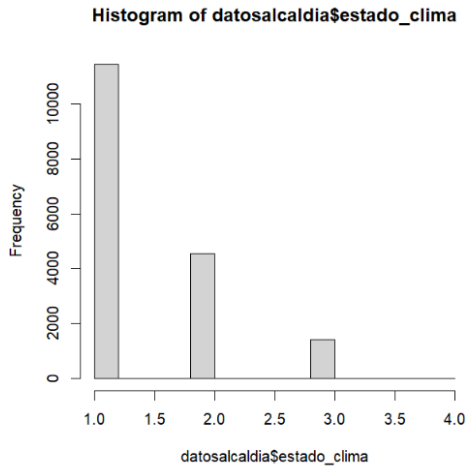

Imagen 11. Histograma estado_clima

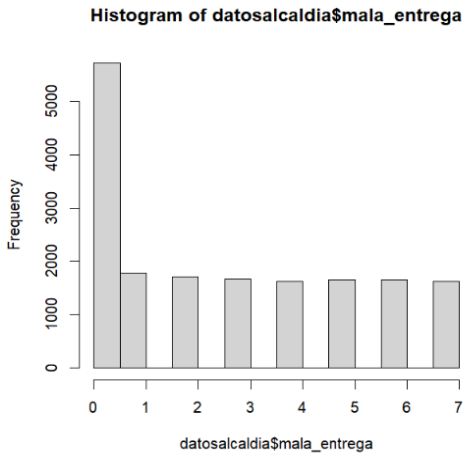

Imagen 12. Histograma mala_entrega

Para validar la normalidad de los datos se generan las hipótesis respectivas para cada variable y se hace la aplicación del método de kolmogorov, ya que en el conjunto de datos se cuenta con más de 5000 registros. La imagen 13 refleja el resultado del test de normalidad para la variable estado_clima:

Hipótesis estado_clima

Ho: El estado del clima tiene una distribución normal
H1: El estado del clima no tiene una distribución normal

Hipótesis mala_entrega
Ho: El estado del clima tiene una distribución normal
H1: El estado del clima no tiene una distribución normal

```
data:  datosalcaldia$estado_clima
D = 0.40375, p-value < 2.2e-16
```

Imagen 13. Resultado de test de kolmogorov para la variable estado_clima

Sabiendo que el valor del test de normalidad es menor que el nivel de significancia (2.2e-16<0.05) se puede concluir que la variable estado_clima no cuenta con una distribución normal. Por otro lado, al aplicar el test de normalidad a la variable mala_entrega se obtiene el resultado reflejado en la imagen 14.

```
data:  datosalcaldia$mala_entrega
D = 0.18586, p-value < 2.2e-16
```

Imagen 14. Resultado de test de kolmogorov para la variable mala_entrega

Ya que en este caso se obtiene el mismo valor también se pude decir que el valor del test de normalidad es menor que el nivel de significancia, en consecuencia, la variable mala_entrega no cuenta con una distribución normal. Una vez se determino que los datos a analizar no cuentan con distribución normal se procede a aplicar la prueba no paramétrica de Kendall, en la cual se obtiene el siguiente resultado:

```
> cor(x=datosalcaldia$estado_clima, y= datosalcaldia$mala_entrega, method="kendall")
[1] -0.005180569
>
```

Imagen 15. Prueba no paramétrica con el método de Kendall sobre las variables estado_clima y mala_entrega

Al comparar este valor con las escalas de correlación presentada en la imagen 16. se concluye que la correlación de estas variables es NULA.

±0.96 , ±1.0	PERFECTA
±0.85 , ±0.95	FUERTE
±0.70 , ±0.84	SIGNIFICATIVA
±0.50 , ±0.69	MODERADA
±0.20 , ±0.49	DÉBIL
±0.10 , ±0.19	MUY DÉBIL
±0.09 , ±0.0	NULA

Imagen 16. Escala de correlación

Por ultimo se representa la información de las variables ya mencionadas en un grafico con el fin de verificar el resultado de manera grafica. Esta grafica se presenta en la imagen siguiente y corrobora la información presentada anteriormente.

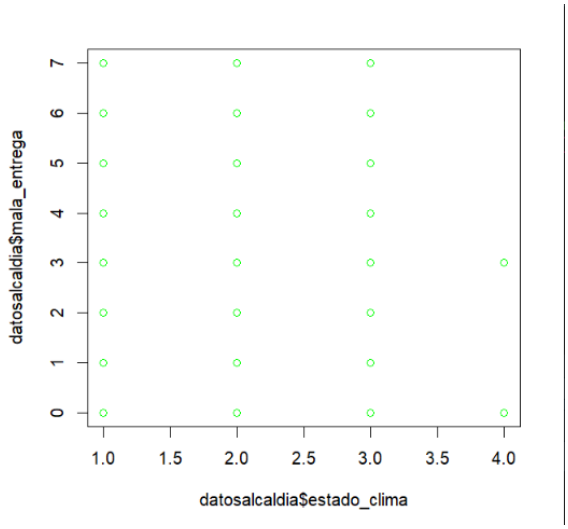

Imagen 17. Grafico de correlación por el método Kendall

Conclusión

Con base en los resultados obtenidos se concluye que no existe una relación entre la cantidad de malas entregas y el estado del clima, lo que desmiente la hipótesis planteada inicialmente. Por tanto, no es necesario realizar sugerencia alguna para la alcaldía de Madrid en lo que respecto a la hipótesis planteada.