Quiz 1 Mineria de datos.

David Moreno

2024-01-30

**Punto 1**

• Explique cada una de las variables que están incluidas en el dataset. ¿Cuántas variables y cuántas

observaciones tiene el dataset? ¿Qué tipo de datos tiene? Resuma la información del dataset y muestre

una figura utilizando las primeras 100 observaciones. ¿Ve alguna tendencia o correlación relevante?

Confirme este resultado usando una función de correlación de Pearson.

**Descripción del Dataset**

• year, month, day: Fecha del vuelo.

• dep\_time, arr\_time: Hora de salida y llegada.

• sched\_dep\_time, sched\_arr\_time: Hora programada de salida y llegada.

• dep\_delay, arr\_delay: Retraso en salida y llegada.

• carrier: Código de la aerolínea.

• flight: Número de vuelo.

• tailnum: Número de registro del avión.

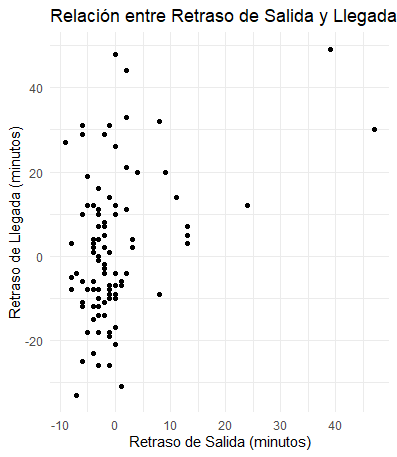
• origin, dest: Aeropuertos de origen y destino.

• air\_time: Tiempo en el aire.

• distance: Distancia del vuelo.

• hour, minute: Hora y minuto de salida programada.

• time\_hour: Fecha y hora de salida programada



Para crear el primer gráfico, me parece interesante explorar la relación que podría existir entre los retrasos en la salida de los vuelos y los retrasos en su llegada. Esto se debe a que si un vuelo experimenta un retraso en su salida y, al mismo tiempo, en su llegada, podría haber una influencia mutua. En este caso, observamos una correlación de 0.37, lo cual es un valor significativo y sugiere una relación positiva entre ambas variables.

**Punto 2**

2. Realice un histograma para examinar la distribución de los retrasos en las salidas de todos

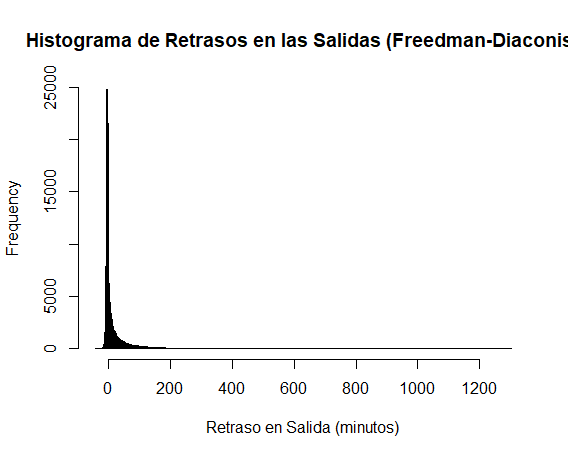
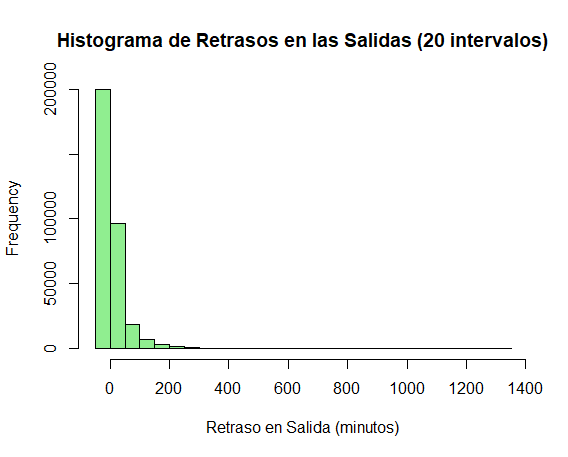
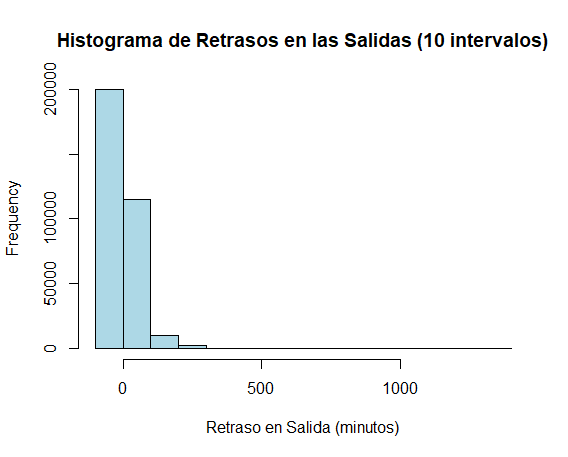
los vuelos. ¿Qué puede mencionar sobre esta distribución? Muestre el histograma con un

número de intervalos de 10, 20 y un ancho de banda de 2 ∗ 𝐼𝑄𝑅 ∗ 𝑛ି con 𝐼𝑄𝑅 la

distancia intercuarƟl de los datos y 𝑛 el tamaño de observaciones (Regla Freedman y

Diaconis), ¿cuál considera mejor?

Primero, vamos a crear el gráfico utilizando 10 intervalos, luego 20 intervalos, y finalmente aplicaremos la fórmula Greedman-Daconis.



Los resultados sugieren que la distribución podría ajustarse a una variable exponencial. Para confirmarlo, realizamos la prueba de Kolmogorov-Smirnov, la cual nos permite evaluar si los datos se ajustan a esta distribución específica. Al aplicar la prueba con un nivel de confianza del 95%, concluimos que los datos no siguen una distribución exponencial.

En cuanto a determinar cuál es el mejor histograma, es importante considerar cómo se dividen los intervalos. Utilizando la fórmula de Freedman-Diaconis, dividimos los datos en intervalos más pequeños, lo que nos permite obtener un histograma que se ajusta más estrechamente a los datos reales.

**Punto 3**

a. **La distribución es unimodal.**

La distribución muestra unimodalidad, evidenciada por un pico prominente en los valores entre -25 y 0, donde se observa la mayor frecuencia.

b. **La distribución está sesgada a la derecha.**

La media de los retrasos es -9.143, superando a la mediana de -13. Esto indica una sesgo hacia la derecha en la distribución de los datos.

(x)c. **Ningún vuelo se retrasa más de 2 horas.**

Esta afirmación no es correcta. El retraso máximo registrado en los datos es de 205 minutos, lo cual excede las 2 horas (120 minutos).

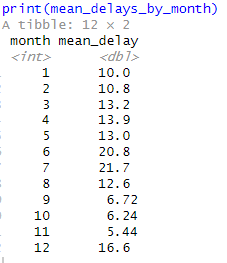
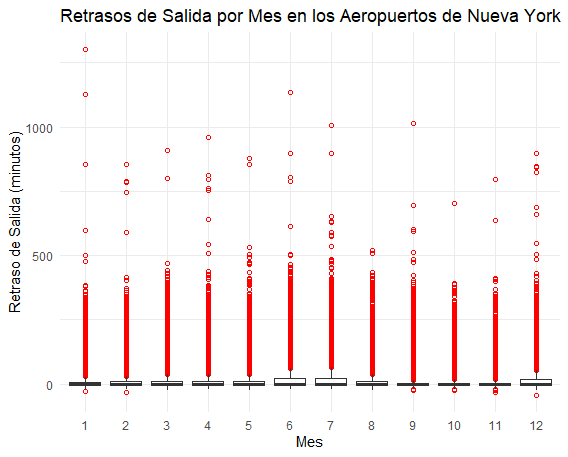
d. **La distribución tiene varios valores extremos en el lado derecho.**

La distribución muestra la presencia de valores atípicos en su extremo derecho, particularmente alrededor de 200 minutos y en valores superiores.

e. **Más del 50% de los vuelos llegan a tiempo o antes de lo programado.**

Esta afirmación es acertada. Aproximadamente el 73.88% de los vuelos llegaron a tiempo o incluso antes de la hora programada.

**Punto 4**



Se observa que los retrasos más significativos ocurren en el mes de julio. Al analizar la gráfica, se destaca que junio y julio son los meses con mayor incidencia de retrasos en los aeropuertos de Nueva York. Esto podría estar relacionado con el aumento de viajeros durante la temporada de verano, una época caracterizada por ser alta en turismo.