**Predict Flights Delay**

Please use the dataset(“Delay\_v3.csv”) to explore the following questions.

Departure Time: DEP\_TIME

Flight Date: FL\_DATE

Flight Number: FL\_NUM

Day of the Week: DAY\_WEEK

Flight Status (where it’s ontime or delayed)

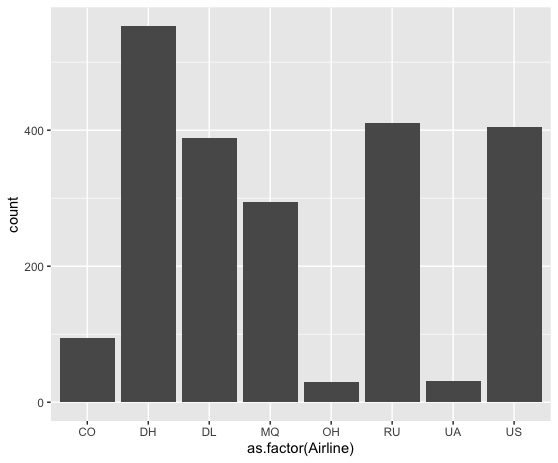
Q1. Understand the Data (Range, Data type, Visualize the variables, etc.)

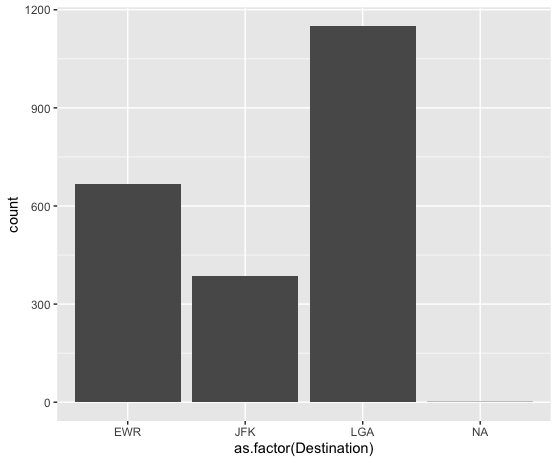
A screenshot of a cell phone

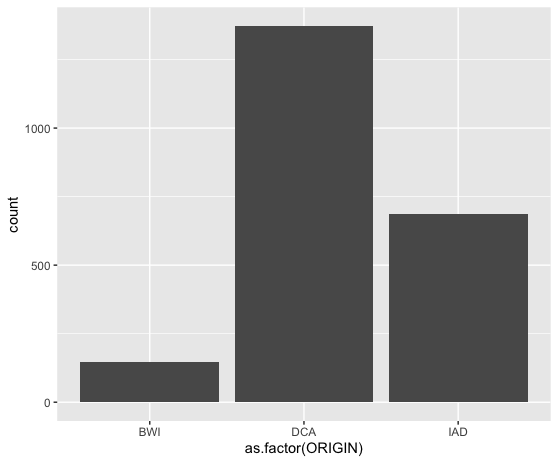
Description automatically generated

A screenshot of a cell phone

Description automatically generated







A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Q2. Remove missing values

There’s one missing value in the ‘Destination’ column.

Q3. Remove duplicated in the dataset

Use *unique()* function to remove duplicates.

Q4. Convert the data type

1. Convert to factor: Weather, Day\_week

Use *factor()* function.

1. Convert Scheduled “DEP\_TIME” to Date data type

Use *as.time()* function in the *datetime* library.

Q5. Insights

5.1 Which day of the week sees the most delays?

Day 1

A screenshot of a cell phone

Description automatically generated

5.2 Which Destination sees the most delay?

LGA

A screenshot of a cell phone

Description automatically generated

Q6. Predict flights delay by using logistic regression and evaluate model performance

A close up of text on a white background

Description automatically generated

pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE)

2.298644e-42

|  |  |  |
| --- | --- | --- |
| predict |  |  |
|  | ontime | delay |
| ontime | 1156 | 617 |
| delay | 139 | 289 |

Accuracy = (1156 + 289) / (1156 + 617 + 139 + 289) = 65.65%