SFO Summary Report

The dataset used for this report is adapted from a 2012 passenger survey from San Francisco airport (SFO). The report focuses on fulfilling the following objectives:

1. To report the appropriate descriptive statistics for each variable in the data, such as the mean and standard deviation for continuous variables, and percentages for categorical variables. This will provide an overall understanding of the distribution and patterns of the data.
2. To create a visual representation of the variables 'wait' and 'usa', highlighting observations where the binary outcome was positive. This will help to identify any patterns or trends in the data that may be informative for modelling.
3. To fit a logistic regression model using the predictor variables 'dirty', 'wait', 'last year', and 'usa'. This will provide an estimate of the relationship between these predictor variables and the binary outcome.
4. To choose the 'best' model, which may be different from the one above, based on criteria such as AIC and overall model performance. This will ensure that the model used for prediction is both accurate and parsimonious.
5. To calculate the odds ratio and 95% confidence interval for all predictor variables in the chosen 'best' model. This will provide an estimate of the relative effect of each predictor variable on the outcome and the degree of uncertainty associated with this estimate.
6. To create a classification table (confusion matrix) based on classifying outcomes as "good" if the predicted risk is over 50%, and "bad" otherwise. This will provide a summary of the model's overall performance in terms of its ability to correctly classify outcomes.

1) Descriptive Statistics

1. For the variable **'good'**, the appropriate descriptive statistic would be **percentages**. We can calculate the percentage of observations that have a positive experience (good = 1) and the percentage of observations that have a negative experience (good = 0). The **mean** for the good variable is **0.5848**. This indicates that **58.48%** of the passengers surveyed approved of the airport.

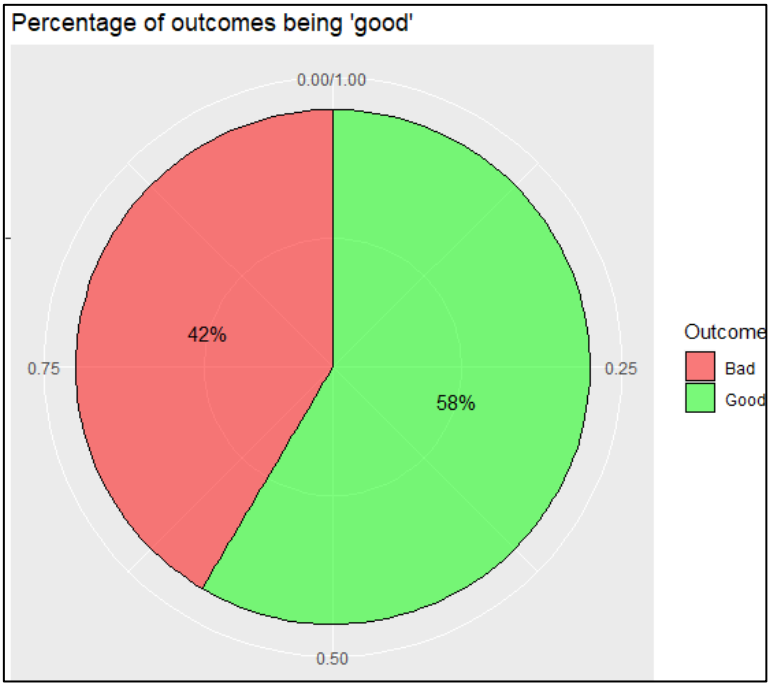

Figure 1: Percentage of outcomes classified as "good".

When visualizing binary data, it can be beneficial to use a pie chart to represent the proportion of observations for each category. In this specific case, the variable 'good' is binary, taking on the values of 1 (good) and 0 (not good). The 58% good and 42% bad indicates that majority of the passengers have a positive opinion of the airport, which is a good sign for the airport management. However, 42% of the passengers have a negative opinion of the airport, this indicates that there is room for improvement, and the airport management should investigate the reasons for dissatisfaction.

2. For the variable **'dirty'**, the appropriate descriptive statistic would be **mean** and **standard deviation** (SD). The **mean** of **0.0928** and the **standard deviation** of **0.3922** for the variable dirty indicate that most of the passengers surveyed felt that a substantially small number of locations were dirty, with a standard deviation of 0.3922 which means that there's a variation in the number of dirty locations reported by the passengers. This could suggest that while most of the passengers felt that the airport was relatively clean, there were a small number of passengers who felt that a larger number of locations were dirty.

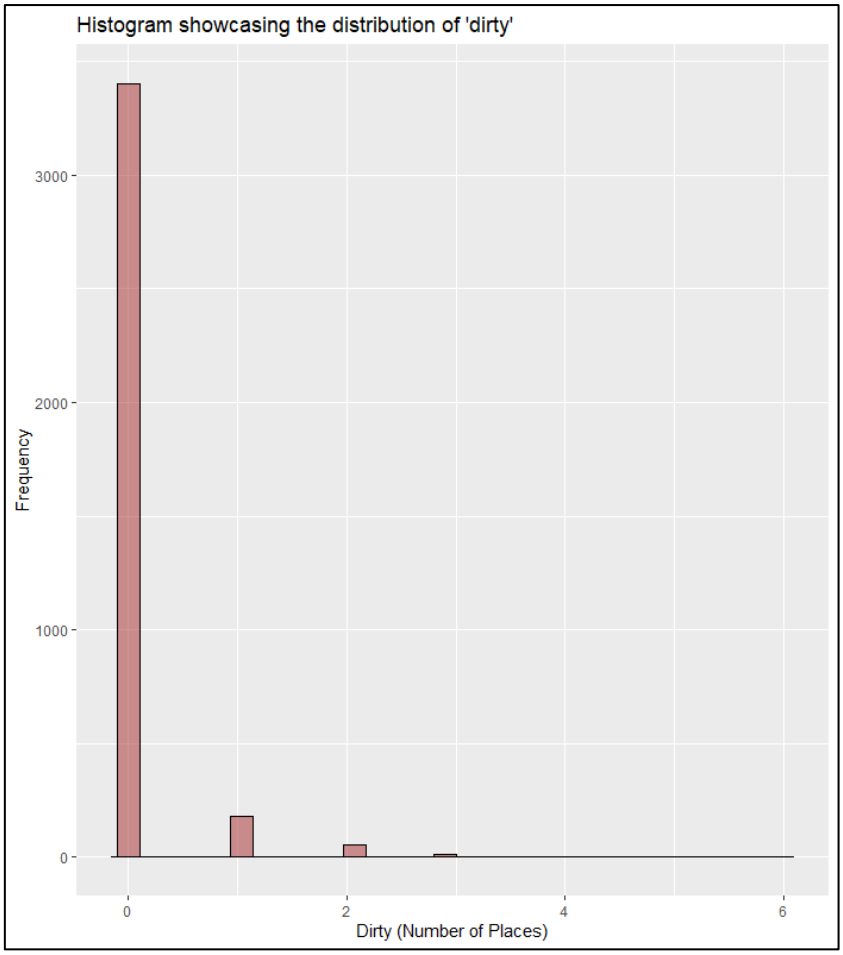

Figure 2: Histogram showcasing the distribution of the predictor variable 'dirty'.

A histogram of the variable 'dirty' allows the airport management to understand the distribution of the dirty count values, including the frequency of different count values. The histogram shows that **most of the passengers did not find any location dirty (dirty = 0) with 3403 observations**, followed by a small number of passengers who found **one location dirty (dirty = 1) with 180 observations**, and a **smaller number of passengers who found two or more locations dirty**. This information can be useful for the airport management to understand the cleanliness of different areas of the airport and to identify areas that need improvement. The mean of 0.09 and the standard deviation of 0.3922 indicate that most of the observations have a dirty count of zero, and **the distribution is skewed towards zero**. The mean is a measure of central tendency, it tells us the average value of the variable, and in this case, it tells us **that the average passenger did not find any location dirty**. The standard deviation is a measure of the spread of the data, it tells us how far the data is spread from the mean, and in this case, it tells us that **most of the observations are close to the mean**, but there are also some observations that are farther away from the mean.

3. For the variable **'wait'**, the appropriate descriptive statistic would be **median** and **quartiles**. The **median** of **1.78, first quartile** of **1.020** and **third quartile** of **4.82** for the wait variable indicate that:

a. The median wait time of 1.78 hours means that half of the passengers spent less than 1.78 hours between arrival and flying, and the other half spent more.
b. The first quartile of 1.02 hours means that 25% of the passengers spent less than 1.02 hours between arrival and flying.
c. The third quartile of 4.82 hours means that 75% of the passengers spent less than 4.82 hours between arrival and flying.

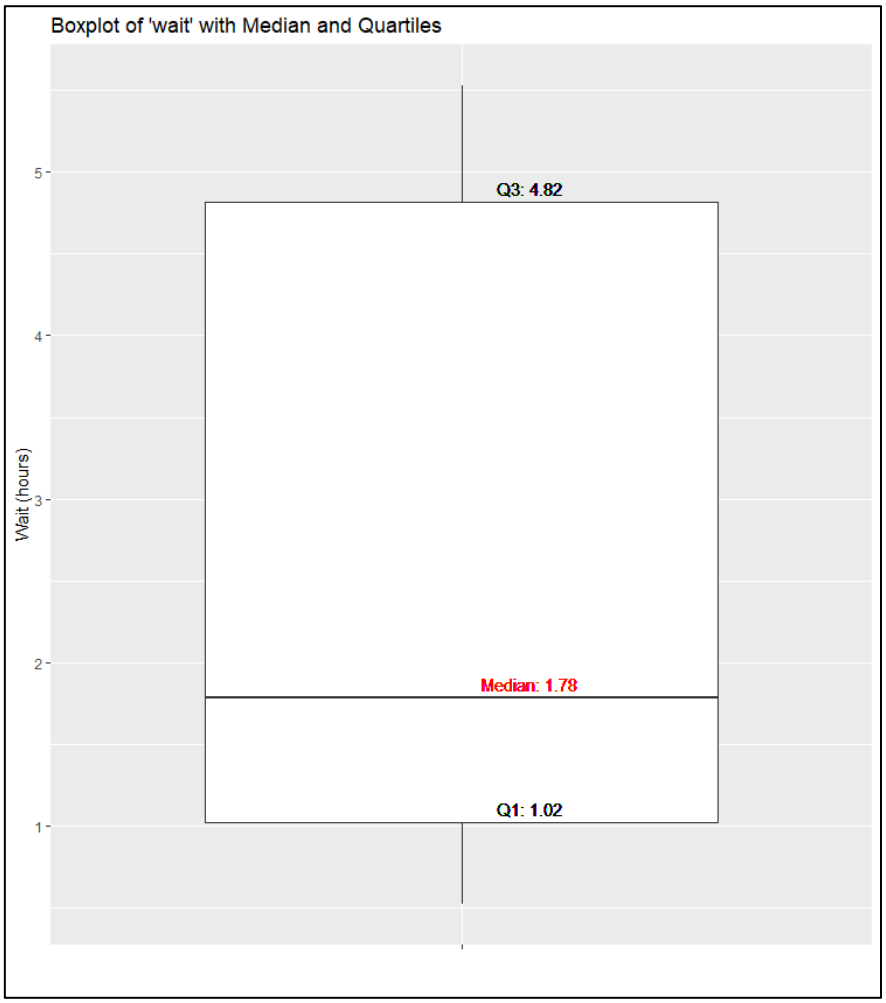

Figure 3: Boxplot of 'wait' with Median and Inter-Quartile Ranges

In the case of the 'wait' variable, a boxplot can be used to understand the distribution of the wait time between arrival and flying, and to identify any outliers that may be affecting the distribution along with the interquartile range (IQR), which represents the middle 50% of the data. Using the median and quartiles for the variable 'wait' and visualising it using a boxplot, we can understand that most of the passengers did not have to wait for a long time between arrival and flying, with half of the passengers waiting for less than 1.78 hours, 25% waiting for less than 1.02 hours and 75% waiting for less than 4.82 hours. This can be considered an important information for the airport management to understand the waiting time of their passengers.

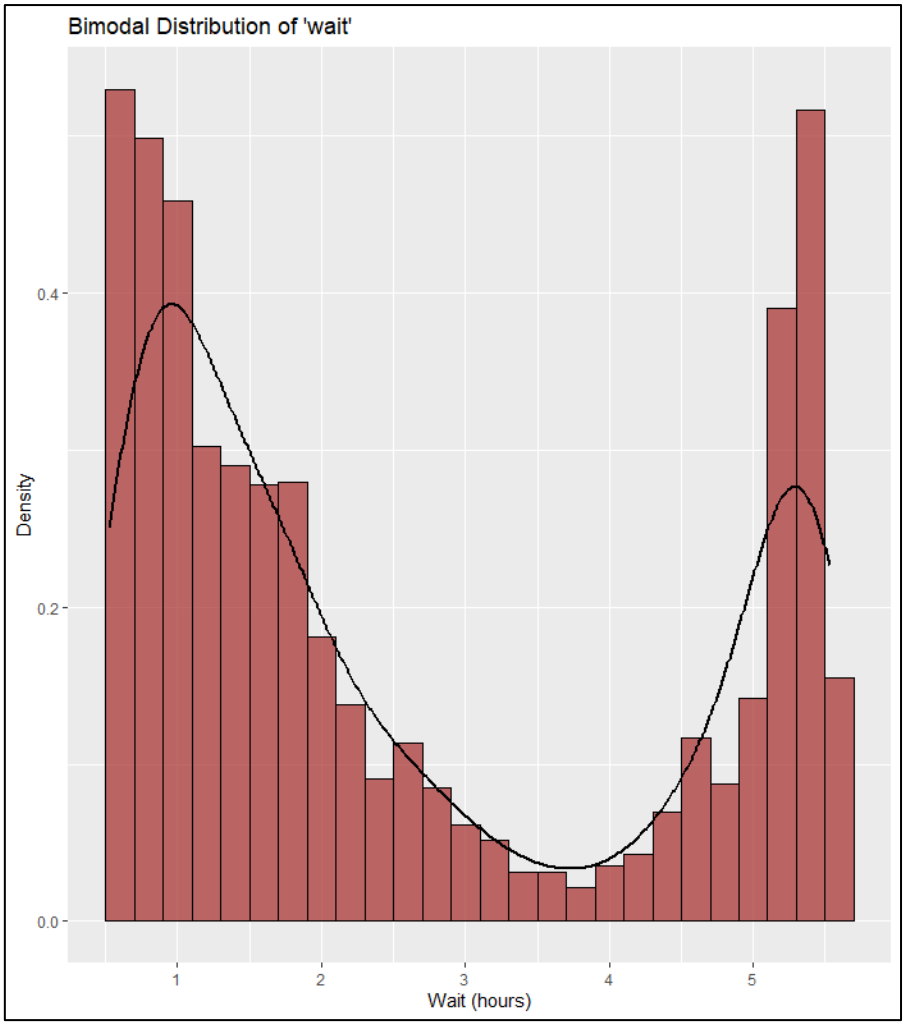

Figure 4: Bimodal Distribution of "wait".

A bimodal distribution refers to a distribution that has two distinct peaks or modes, indicating that there are two distinct groups of data. In the case of the variable 'wait', a bimodal distribution indicates that there are two distinct groups of passengers with different wait times.

For example, **one peak of the distribution represents passengers who had relatively short wait times before flying (<= 3 hours)**, while the **other peak represents passengers who had relatively long wait times before flying (>= 5 hours)**. This suggests that there are t**wo distinct groups of passengers with different wait times.** Also, there can be a strong association between the wait time and whether the passenger has a positive experience at the airport depending on the wait times that are observed in the distribution.

4. For the variable **'lastyear'**, the appropriate descriptive statistic would be **mean** and **standard deviation** (SD). The **mean** of **3.946** and the **standard deviation** of **7.385** for the variable lastyear indicate that most of the passengers surveyed had flown out of SFO an average of 3.946 ≈ 4 times in the previous 12 months, with a standard deviation of 7.385. This indicates that there is a high variation in the

number of times that passengers flew out of SFO in the previous 12 months, with some passengers flying frequently and others flying infrequently. This could suggest that some passengers are regular travellers while others are not.

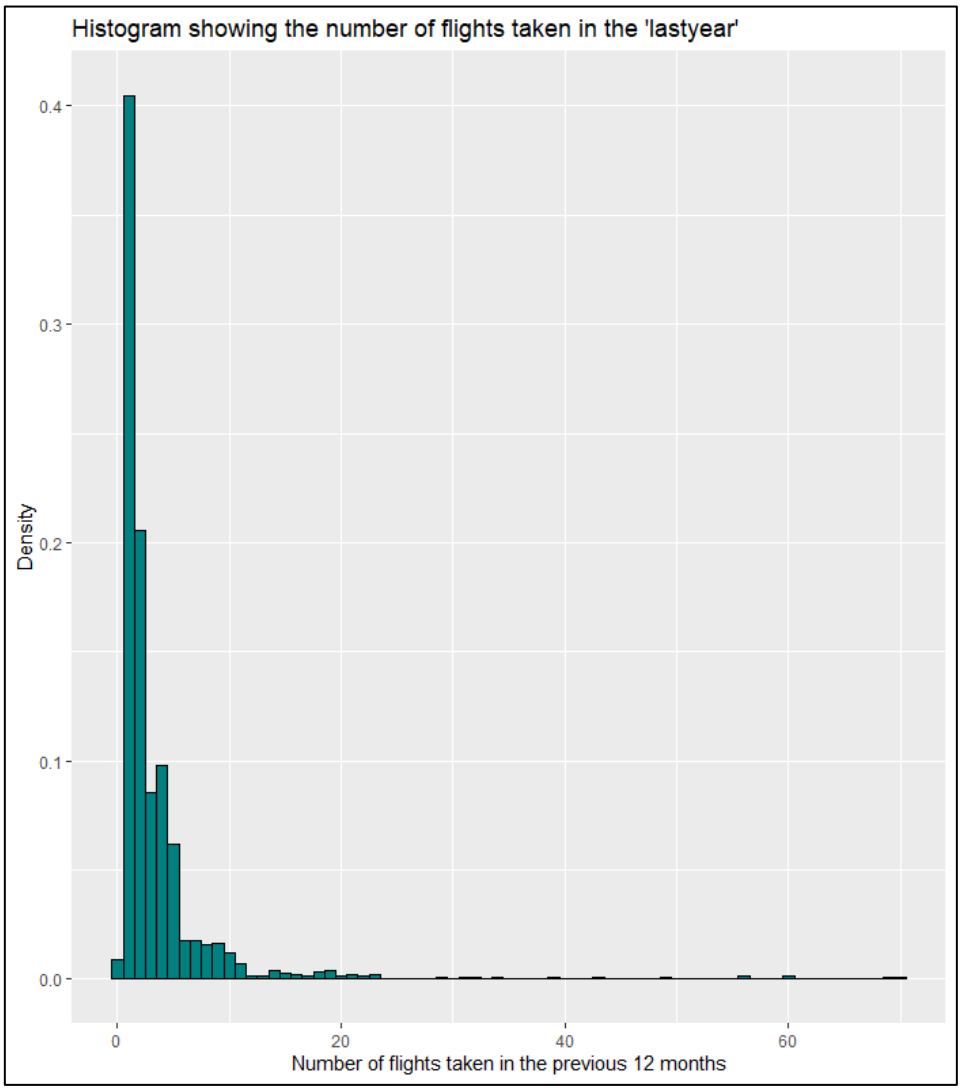

Figure 5: Histogram showing the number of flights taken in the 'lastyear'.

A histogram of the variable 'lastyear' allows the airport management to understand the distribution of the number of flights taken by the passengers in the previous 12 months. The histogram shows the frequency of different count values, and it can help the management to identify patterns and trends in the passenger's travel behaviour. In this case, the mean of 3.946 and the standard deviation of 7.385 indicate that most of the passengers flew out of SFO a few times in the previous 12 months (mean of 3.946) and the data is spread out (standard deviation of 7.385). This **histogram shows a skewed distribution, with a long tail on the right side, indicating that a relatively small number of passengers flew out of SFO many times in the previous 12 months, while most of the passengers flew out of SFO a few times**.

5. For the variable **'usa'**, the appropriate descriptive statistic would be **percentages**. The **average** of **0.7557** for the usa variable indicates that **75.57%** of the

passengers surveyed were flying to a destination **in the USA**. This means that **24.43%** of the passengers were flying to a destination **outside the USA**. This could suggest that most of the passengers surveyed were domestic travellers.

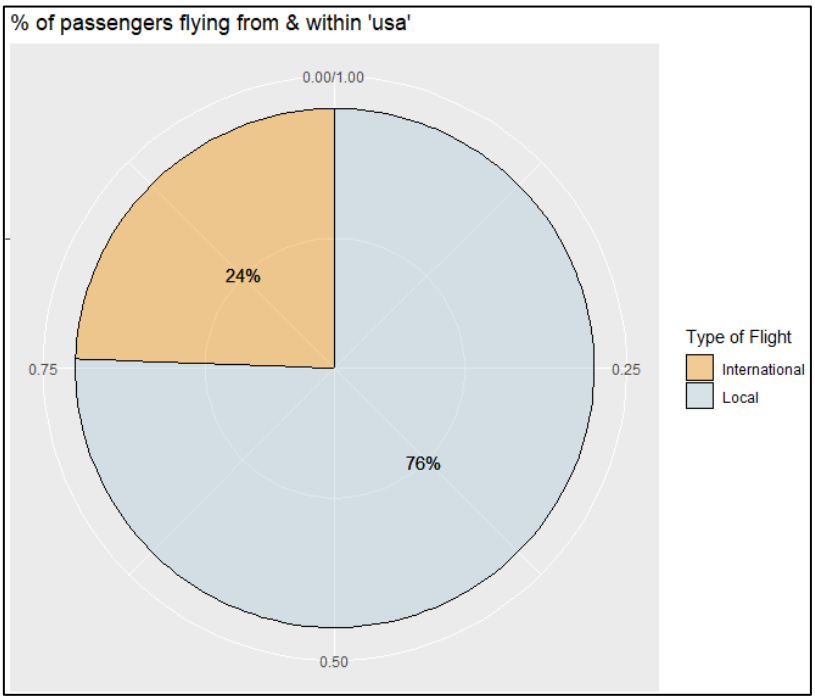

Figure 6: Percentage of passengers flying internationally & locally within 'usa'.

A pie chart is a useful visualization method for categorical data, particularly when the variable has only a few categories. In this case, the variable 'usa' is binary, taking on the values of 1 (USA) and 0 (not USA), and the data shows that a proportion of the observations were flying to USA and the other proportion to other countries. **76% of the observations were flying to USA and 24% were flying to other countries**, this information can be useful for the airport management to understand the main destinations of their passengers, the market share of domestic flights versus international flights, the main destinations of the passengers, and how it's changing over time.

6. Correlation Matrix

Table 1: Correlation Matrix

	good	dirty	wait	lastyear	usa
good	1	-0.13	-0.09	-0.02	-0.0005
dirty	-0.13	1	0.0007	0.082	0.045
wait	-0.09	0.0007	1	0.06	0.099
lastyear	-0.02	0.082	0.06	1	0.082
usa	-0.0005	0.045	0.099	0.082	1

The correlation matrix shows that there is a weak negative correlation between the good variable and the dirty, wait and lastyear variable (-0.13, -0.09, -0.02 respectively) which means that as the number of locations the passenger felt were

dirty, the number of hours the passenger spent at the airport between arrival and flying and the number of times the passenger flew out of SFO in the previous 12 months increases, the likelihood of the passenger approving of the airport decreases slightly.
Also, there is a very weak negative correlation between good and usa (-0.0005) which means that as the passenger is flying to a destination in the USA increases, the likelihood of the passenger approving of the airport decreases slightly.
Additionally, there is a very weak positive correlation between dirty and wait (0.0007) which means that as the number of locations the passenger felt were dirty increases, the number of hours the passenger spent at the airport between arrival and flying increases slightly.
There is a weak positive correlation between dirty and lastyear and dirty and usa (0.082, 0.045 respectively) which means that as the number of locations the passenger felt were dirty increases, the number of times the passenger flew out of SFO in the previous 12 months and the number of passengers flying to a destination in the USA increases slightly.
Furthermore, there is a weak positive correlation between wait and lastyear and wait and usa (0.060, 0.099 respectively) which means that as the number of hours the passenger spent at the airport between arrival and flying increases, the number of times the passenger flew out of SFO in the previous 12 months and the number of passengers flying to a destination in the USA increases slightly.
Finally, there is a weak positive correlation between lastyear and usa (0.082) which means that as the passenger is flying to a destination in the USA increases, the number of times the passenger flew out of SFO in the previous 12 months increases slightly.

2) Visualisation of 'wait' and 'usa'.

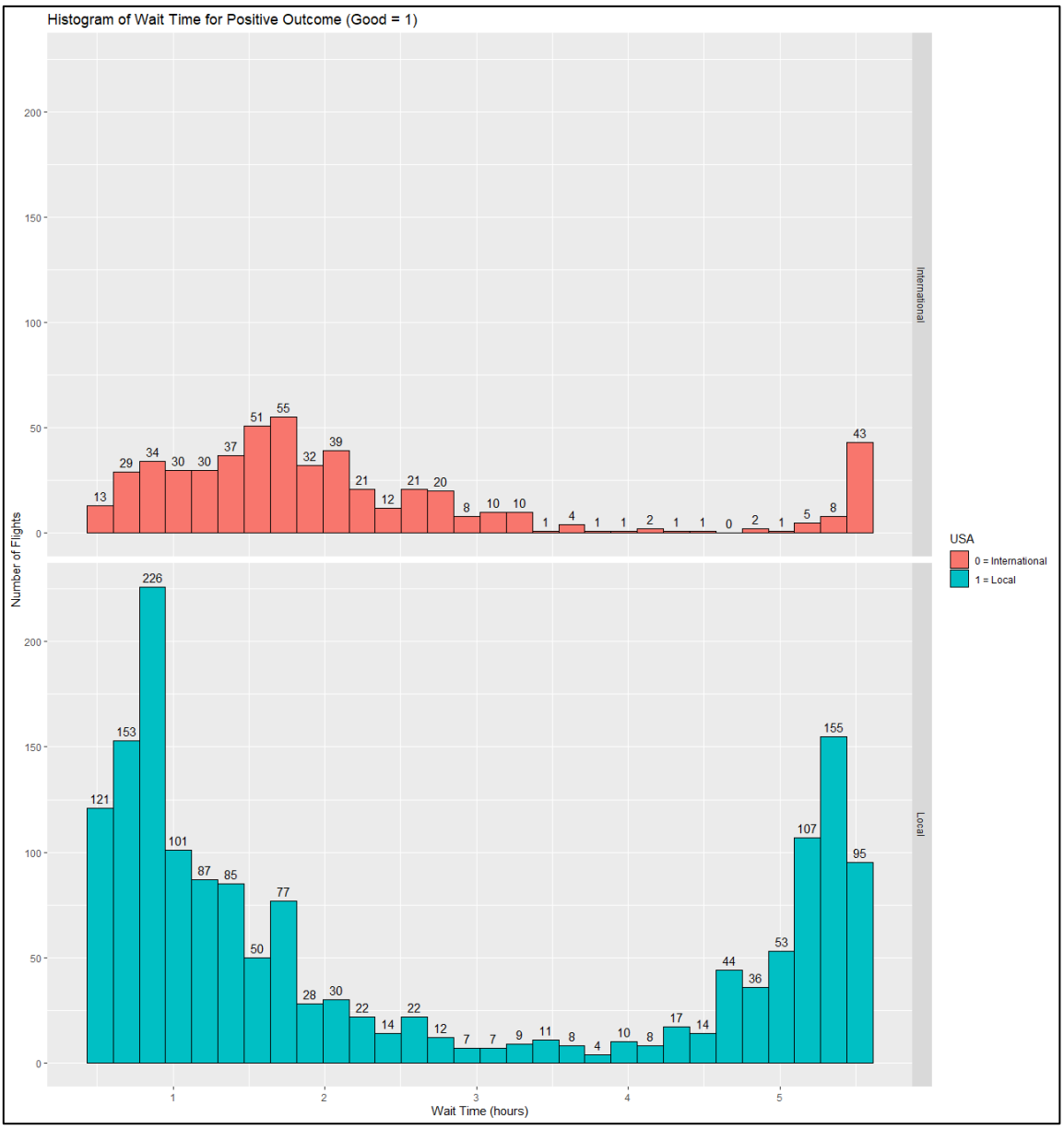

Figure 7: A visualisation of 'wait' and 'usa', highlighting the observations where the binary outcome was positive.

This visualisation shows the number of flights, the wait time, and the type of flight (international or local) for flights that resulted in a positive outcome (good = 1). It shows that **most flights had a wait time of less than 3 hours**, with the highest number of flights (226) having a wait time of 0.86 hours. The visualisation also indicates that **most flights with positive outcomes** had **wait times between 0 and 5 hours**, with **75% of these flights being local flights** within the USA (usa = 1). There were also a significant number of international flights (usa = 0) with wait times between 0 and 5 hours, but the number of flights with positive outcomes decreases as the wait time increases. This suggests that **shorter wait times are associated with a higher likelihood of a positive outcome for international flights**. Strangely **for local flights**, it is noticeable that **even with longer waiting times (>= 5 hours), the**

outcome is positive for more than 400 flights. Overall, the trend suggests that **for international flights, the wait time is relatively short, whereas for local flights, the wait time is relatively long**. In conclusion, the visualisation gives us a good overview of the distribution of wait times for flights that resulted in a positive outcome and how this differs between international and local flights.

3) Logistic Regression Model

A logistic regression model using the predictor variables dirty, wait, lastyear, and usa was created to predict the outcome variable "good".

```
> logregmodel1 <- glm(good ~ dirty + wait + lastyear + usa, data = data, family = "binomial")
> summary(logregmodel1)

Call:
glm(formula = good ~ dirty + wait + lastyear + usa, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4651  -1.2794   0.9276   0.9971   2.1497

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.633590   0.081889   7.737 1.02e-14 ***
dirty       -0.799584   0.106381  -7.516 5.64e-14 ***
wait        -0.103839   0.018675  -5.560 2.69e-08 ***
lastyear    -0.001438   0.004654  -0.309    0.757
usa          0.076705   0.079797   0.961    0.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4955.9  on 3650  degrees of freedom
Residual deviance: 4854.8  on 3646  degrees of freedom
AIC: 4864.8

Number of Fisher Scoring iterations: 4
```

Figure 8: Summary of a logistic regression model using the predictor variables dirty, wait, last year and usa.

The probabilities for "lastyear" and "usa" are considerably higher than "dirty" and "wait" in the model, which suggests that these variables are not as strong predictors of the outcome (good) as "dirty" and "wait" are. This is one way to evaluate the relative importance of predictor variables in a logistic regression model.

It's also important to note that the variables "lastyear" and "usa" may be correlated with other predictor variables, making them less useful as independent predictors. And, in terms of their contribution to the model, it's not substantial enough to justify the complexity they add to the model.

Therefore, it makes sense to consider dropping "lastyear" and "usa" from the model to simplify it and improve its interpretability.

4) Best Model

To evaluate the best model, it was necessary to consider the model which provides the lowest AIC value. AIC can be used with linear regression models, logistic regression models, survival analysis models, and many other models that are based on maximum likelihood estimation (MLE).

Also, AIC is not appropriate for all types of models. For example, AIC cannot be calculated for decision trees, random forests, or support vector machines, because these models are not based on maximum likelihood estimation (MLE), and they do not have a likelihood function.

Therefore, to evaluate the best model we have considered every model that is based on MLE:

1. Linear regression
2. Logistic regression
3. Poisson regression
4. Gaussian mixture models
5. Hidden Markov Models
6. Exponential Family models such as Normal Gaussian and Negative Binomial

We can find the best model by removing lastyear and usa sequentially and then compare the AIC scores of the resulting models.

Table 2: Comparing Logistic Regression models by eliminating variables

Variables	AIC Value
all	4864.8
wait + dirty + usa	4862.9
wait + dirty	**4861.7**

Below is the table of the AIC scores of each of the models with all the variables and after dropping “last year” and “usa”:

Table 3: AIC comparison for viable models

Model Name	**AIC of all variables**	**AIC of “wait” + “dirty”**
Linear Regression	4864.77	4861.75
Logistic Regression	4864.8	**4861.7**
Poisson Regression	6521.96	6518.34
GMM	48805.91	
HMM	8891.089	
Normal Gaussian	5106.44	5103.37
Negative Binomial	6524	6520.4

The logistic regression model mentioned above with just the predictor variables of "wait" and "dirty" is the chosen "best" model as it has the lowest AIC value among all models considered.

The primary reason logistic regression is the preferred model for binary outcome variable is because it allows for easy interpretation of the relationship between the predictor variables (wait + dirty) and the outcome (good), handles both continuous and categorical predictor variables (as wait is continuous and dirty is categorical), and allows for easy estimation of the effect of multiple predictor variables on the outcome.

5) Odds Ratio and 95% Confidence Interval

The odds ratio and 95% confidence interval for the predictor variable "**dirty**" are **0.450 (CI: 0.36 - 0.55)**, indicating that for **every 1 unit increase in the number of locations that the passenger felt were dirty, the odds of the passenger approving of the airport (good = 1) decrease by a factor of 0.450.**

A change in the value of the predictor variable "dirty" impacts the predicted risk by altering the odds of the passenger approving of the airport (good = 1). For example, if a passenger reports 5 dirty locations, the odds of them approving of the airport are 0.45^5 = 0.13 (or 13% of the odds of a passenger with 0 dirty locations).

It's important to note that **when the odds ratio is less than 1**, it means that as the **value of the predictor variable increases, the odds of the outcome being positive decrease**. In other words, a higher number of dirty locations is associated with a lower likelihood of a passenger approving of the airport.

The **95% CI of 0.36 - 0.55** for the predictor variable "dirty" means that there is **a 95% probability that the true odds ratio of the effect of the "dirty" variable on the "good" outcome falls within this range**. In other words, it means that if we were to repeat the study multiple times and calculate the odds ratio for the "dirty" variable each time, we would expect the odds ratio to fall within the range of 0.36 to 0.55 in 95 out of 100 studies. **It's also important to note that a smaller 95% CI means that the results are more precise, and the true odds ratio is more likely to be closer to the point estimate.**

The odds ratio and 95% confidence interval for the predictor variable "**wait**" are **0.90 (CI: 0.87 - 0.93)**, indicating that for **every 1 unit increase in the number of hours that the passenger spent at the airport between arrival and flying, the odds of the passenger approving of the airport (good = 1) decrease by a factor of 0.90**. A change in the value of the predictor variable "wait" impacts the predicted risk by altering the odds of the passenger approving of the airport (good = 1). For example, if a passenger has a wait time of 4 hours, the odds of them approving of the airport are 0.90^4 = 0.65 (or 65% of the odds of a passenger with a wait time of 0 hours).

The **95% confidence interval (CI) of 0.87 - 0.93** indicates that there is a **95% probability that the true odds ratio for the effect of the "wait" variable on the "good" outcome falls within this range**. It indicates that **the results of the study are statistically significant, as the 95% CI does not include 1, which means that there is a statistically significant relationship between the wait time and the approval of the airport** (good = 1).

6) Classification Table (Confusion Matrix)

Based on classifying outcomes as "good" if the predicted risk is over 50%, and "bad" otherwise, the confusion matrix is as follows:

Table 4: Confusion Matrix obtained based on classifying outcomes as "good" if the predicted risk is over 50%, and "bad" otherwise.

Binary Predictions	Bad	Good
Bad	**161**	87
Good	1355	**2048**

In this confusion matrix:

1. **True negatives (TN) = 161**: The model correctly predicted that 161 passengers did not approve of the airport
2. **False negatives (FN) = 87**: The model incorrectly predicted that 87 passengers did not approve of the airport when they did approve
3. **True positives (TP) = 2048**: The model correctly predicted that 2048 passengers approved of the airport
4. **False positives (FP) = 1355**: The model incorrectly predicted that 1355 passengers approved of the airport when they did not approve

The overall accuracy of the model can be calculated by dividing the total number of correct predictions by the total number of observations.

$$\textbf{Accuracy} = (\textbf{TP} + \textbf{TN})/(\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN})$$

In this case, the **accuracy** of the model is (2048+ 161) / (2048+ 161+ 87+ 1355) = **0.60**. It means that the model correctly predicted the outcome 60% of the time.

It is important to note that the confusion matrix can be used to identify areas where the model can be improved. For example, in this case, the model has **a high number of false positives (1355)** which means that it is **incorrectly classifying many passengers as approving of the airport when they do not approve**. This could be an indication that the model needs to be improved or that the data needs to be further analysed to understand why the model is making these incorrect predictions.

Additionally, in this confusion matrix, the number of **false negatives is high (87)** which means that the model is **not identifying a significant number of passengers who approve of the airport as positive**. This could be an indication that the model needs to be refined or that the data needs to be further analysed to understand why the model is missing these observations.

Conclusion

It was found that a substantial proportion of the passengers, 75%, flew within the USA, however, only 58% of all passengers had a positive outcome for the airport. It was observed that shorter wait times are associated with a higher likelihood of a positive outcome for international flights, and for local flights, even with longer waiting times, the outcome is positive for more than 400 flights. The trend shows that for international flights, wait time is short and for local flights, wait time is long. Logistic regression analysis was used to find the best model that can predict the outcome of the passengers' experience at the airport. It was found that the model that uses the predictor variables "dirty" and "wait" as inputs, is the best model as it has the lowest AIC value among all models considered. The odds ratio for "dirty" is 0.450 (CI: 0.36 - 0.55), indicating that as the number of dirty locations increases, the odds of approval decrease. The odds ratio for "wait" is 0.90 (CI: 0.87 - 0.93), indicating that as wait time increases, the odds of approval decrease. Both variables have a statistically significant relationship with the outcome variable and consequently impact the predicted risk of a passenger having a positive experience at the airport. The model is not performing well as it has a high number of false positives and false negatives. This means that the model is incorrectly classifying many passengers as approving of the airport when they do not approve or missing a significant number of passengers who approve of the airport. This could be an indication that the model needs to be improved or that the data needs to be further analysed to understand why the model is making these incorrect predictions.