San Francisco Airport (SFO) Passenger Survey Analysis

CI7330 Data Analytics and Visualisation

Coursework Assignment - 2

Kingston University

Jeevan Mohan Pawar - K2242210

Table of Contents

I. INTRODUCTION

A San Francisco Airport (SFO) dataset adapted from 2012 Passenger Survey needs to be analysed to obtain a suitable predictive model for future use. We need to assess what factors affect the passenger approval.

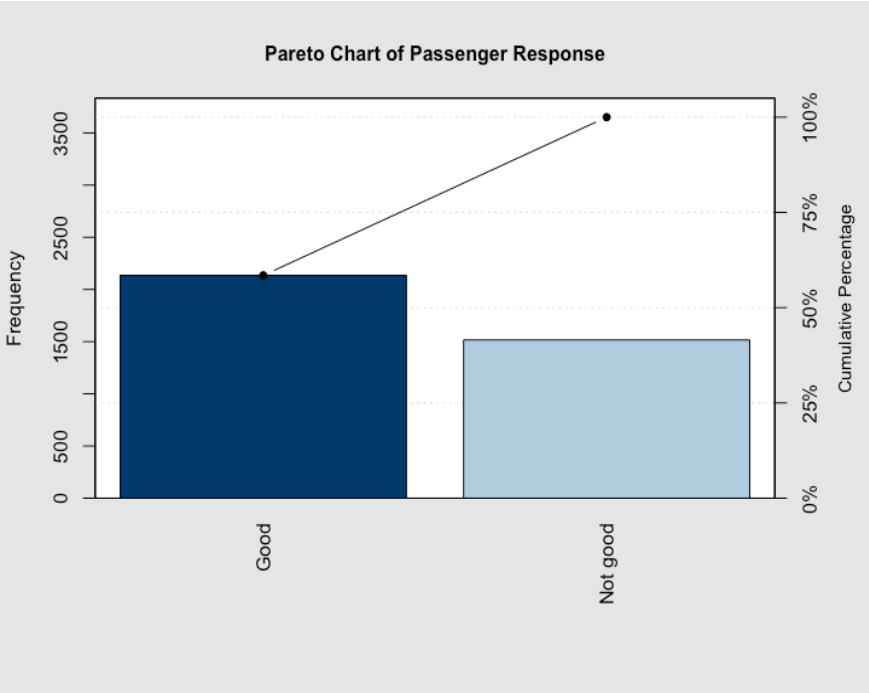
A. Dataset Overview

It consists of **3,651 records**, where each record gives details of a passenger who has flown out from SFO using **5 feature variables**, which have been summarised below:

a. good

It represents a measure of whether a passenger liked or disliked SFO. It is binary in nature, conveying an approval or disapproval by the passenger.

Type : Qualitative/Categorical (Nominal)
Values : 0 represents Disapproval (not good), 1 represents Approval (good)



Pareto chart analysis:

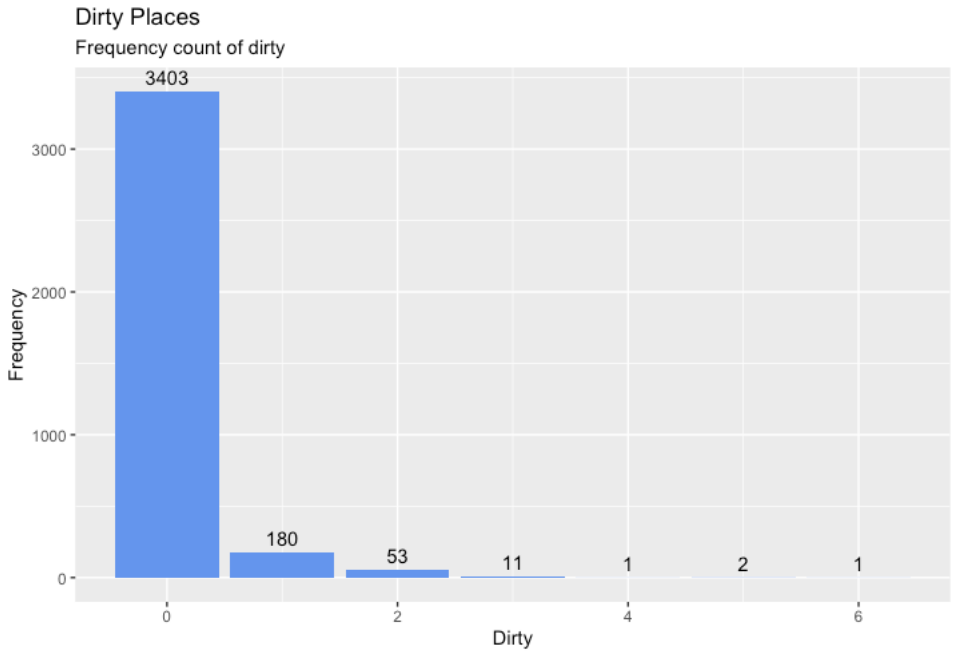
		Frequency	Cum.Freq.	Percentage	Cum.Percent.
Good	(1)	2135	2135	58.477	58.477
Not good	(0)	1516	3651	41.523	100

Around 58.48% of the passengers flying from SFO have approved the airport i.e. consider it 'good'; however, still a significant number of passengers 41.52% approx. have disapproved.

b. dirty

It represents total count of unhygienic or dirty places noticed by a passenger at SFO. These places can refer to areas such as the car parking, restaurants, washrooms, waiting areas, shops etc.

Type : Quantitative/Numerical (Discrete)
Values : 0 represents none of the places were dirty, 1,2,3 etc. represents count of dirty places observed by the passenger



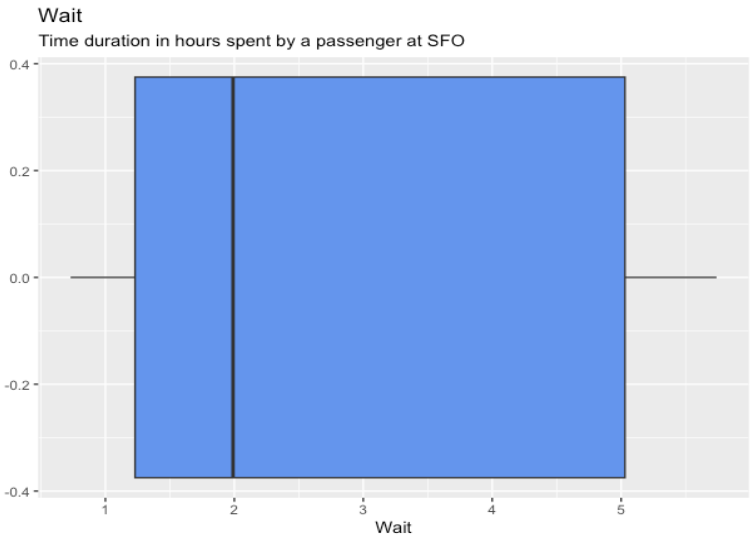
More than 90% of the total passengers responded with 0 count of dirty places, while quite a few responded with counts of 1 or 2, and only a handful number of passengers reported even higher counts as seen above.

Feature	Statistics	Inference
dirty	• Median = 0 • Minimum = 0 • Maximum = 6 • Range = 6 • 1st Quartile (25%) = 0 • 3rd Quartile (75%) = 0	High positive/right skewness, this indicates outliers if present might lie on the right side.

c. wait

It represents the total number of hours spent by a passenger after arriving at SFO and before taking off. This may or may not include any flight delays that could've occurred.

Type : Quantitative/Numerical (Continuous)
Values : 0.73, 3.28, 5.69 etc. represents the wait time in hours

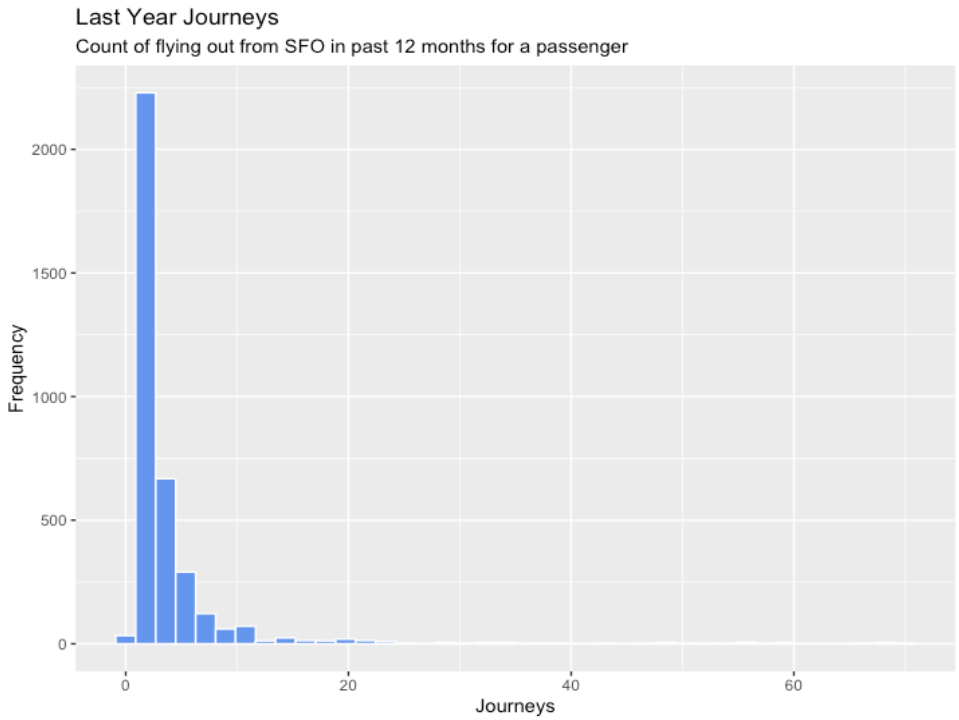


Feature	Statistics	Inference
wait	• Mean = 2.806 • Median = 1.99 • Mode = 1.06 • Minimum = 0.73 • Maximum = 5.74 • Range = 5.01 • 1st Quartile (25%) = 1.23 • 3rd Quartile (75%) = 5.03	Positive/right skewness is observed, no outliers are present as seen from the boxplot.

d. lastyear

It represents the total number of times a passenger has flown out of SFO in the last 12 months or 1 year.

Type : Quantitative/Numerical (Discrete)
Values : 1,3,52 etc. represents total number of times the passenger has travelled from SFO in last 12 months



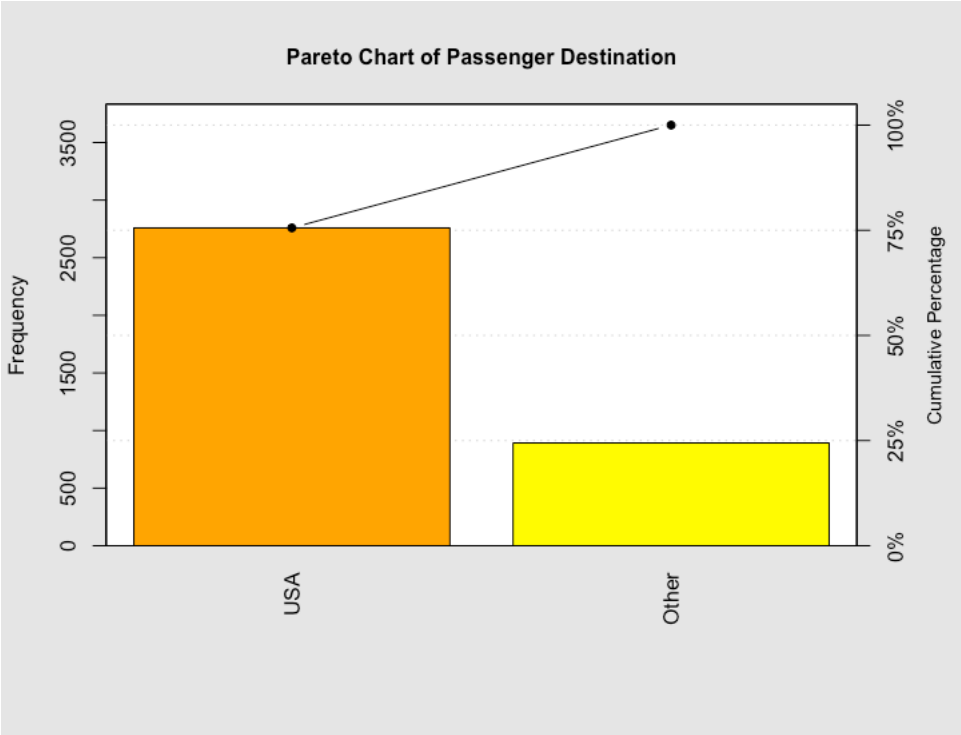
It appears that most of the passengers that flew out from SFO in the last year was between 1 to 5 times, where 1 time being the most common answer followed by 2 times.

Feature	Statistics	Inference
lastyear	• Mean = 3.946 • Median = 2 • Mode = 1 • Minimum = 0 • Maximum = 70 • Range = 70 • 1st Quartile (25%) = 1 • 3rd Quartile (75%) = 4	Positive/right skewness can be observed, outliers if present might lie on the right side.

e. usa

It represents whether the destination of a given passenger was within USA or to some other country.

Type : Qualitative/Categorical (Nominal)
Values : 0 represents some other country, 1 represents USA



Pareto chart analysis:

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
USA (1)	2759	2759	75.568	75.568
Other (0)	892	3651	24.432	100

Approximately 75.57% of the passengers flying from SFO have destination within USA, whereas about 24.43% of the passengers fly to other countries.

Correlation matrix between numerical features in the dataset gives the following matrix:

	dirty	wait	lastyear
dirty	1.000	0.001	0.082
wait	0.001	1.000	0.060
lastyear	0.082	0.060	1.000

There are no null/missing values in the dataset, furthermore, as observed from the correlation matrix there's no multicollinearity between the predictors that will be used for predictive regression models i.e., there is no significant correlation among them.

II. REGRESSION MODELS

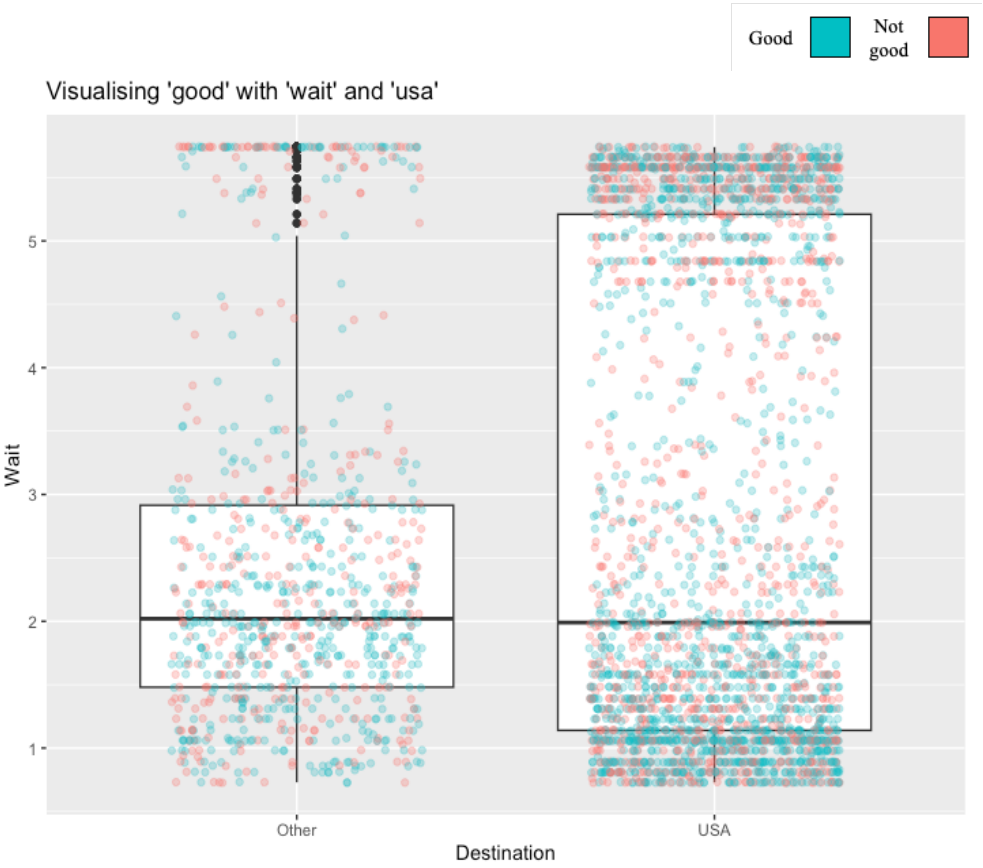
The outcome variable ‘good’ is binary in nature, and there doesn’t appear any linear relationship between the feature variables in the dataset. Logistic regression can be applied to identify the relationship or mapping between the categorical outcome variable (dependent) and many predictor variables (independent) which can be numerical or categorical.

Logistic regression predictive formula is still linear in terms of log-odds:

$$\log w_i = \beta_0 + \beta_i x_i$$

A slope coefficient of β_1 means 1 unit of increment in the predictor will add β_1 to the log-odds. That also translates multiply the odds by $e^{\beta 1}$. Such exponentiated coefficients are called odds ratios. Further, it can convert into probabilities of a sample belonging to one of the categories. An optimal cut-off for this probability can be used to assign sample to certain category. A generally preferred cut-off value is 50%, if a sample scores probability greater than this cut-off, it will be assigned ‘good’ otherwise ‘not good’.

We can analyse how certain predictor variables like ‘wait’, ‘usa’ can influence the outcome variable ‘good’ through visualisations:



It can be observed that the ‘wait’ time for either destination is positive/right skewed. The range of ‘wait’ times as well as a median value of 2hrs is similar in both the cases; however, the spread differs significantly. Almost half of the passengers in ‘other’ have a wait time between 1.5hrs to 2.8hrs, whereas in ‘USA’ it’s

between 1.1hrs to 5.2hrs approximately. Moreover, 'wait' times for 'other' destination passengers are much lesser than passengers with destinations within 'USA'.

Looking at the colour of the points plotted on the above boxplot, it can be observed that most of the passengers with 'wait' time below the median value give a 'good' rating. Further, as the wait time goes beyond median value till the third quartile, there is almost an equal number 'good' and 'not good' responses. Lastly, as the wait time goes beyond the third quartile and touches maximum values above 5hrs or more, most of the passengers responded with 'not good'.

LOGISTIC REGRESSION MODEL – 1

We will use all the possible predictor variables in the dataset, 'good' the outcome variable with predictor variables 'dirty', 'wait', 'lastyear' and 'usa'.

Formula:

good ~ dirty + wait + lastyear + usa

Summary of the model:

Coefficients:

	Estimate	Std. Error	P-value
(Intercept)	**0.655397**	**0.083990**	**6.03e-15**
dirty	**-0.799584**	**0.106381**	**5.64e-14**
wait	**-0.103839**	**0.018675**	**2.69e-08**
lastyear	-0.001438	0.004654	0.757
usa1	0.076705	0.079797	0.336

AIC: 4864.8

P-values of the predictor variables must be less than 0.05 to be statistically significant, observing the p-values in the table above only 'dirty' and 'wait' are significant in predicting the outcome variable 'good'. On the other hand, 'lastyear' and 'usa1' are insignificant, they don't aid much in the predictions. These two can be dropped from the logistic model to avoid overfitting.

LOGISTIC REGRESSION MODEL – 2

We will use only the significant predictor variables as implied from the results of Logistic Regression Model -1, 'good' the outcome variable with predictor variables 'dirty' and 'wait'.

Formula:

good ~ dirty + wait

Summary of the model:

Coefficients:

	Estimate	Std. Error	P-value
(Intercept)	**0.70297**	**0.06362**	**< 2e-16**
dirty	**-0.79709**	**0.10591**	**5.23e-14**
wait	**-0.10227**	**0.01853**	**3.40e-08**

AIC: 4861.7

After comparing the summary of both the models, Logistic Regression Model – 2 appears to be the best model based on the Akaike information criterion (AIC) values. AIC determines a model as best-fit if it explains maximum variance using the least possible independent variables. Models on the same dataset can be compared based on AIC; lower the AIC, better the model.

The following table gives odd ratios and their respective confidence intervals for each of the predictors used in the Logistic Regression Model 2:

Predictor	Log Odds	Odds Ratio	Confidence Interval	
			2.5%	97.5%
dirty	-0.79709	0.451	0.364	0.551
wait	-0.10227	0.903	0.871	0.936

In simple words, these results can be interpreted as follows:

- dirty
  - Provided all other predictors remain constant, an increase in the count of ‘dirty’ places by 1 would result in a decrease of 54.9% in the odds of a passenger approving SFO as ‘good’.
  - Our best estimate of odds ratio for ‘dirty’ is 0.451, with a 95% confidence interval of 0.363 to 0.551. That is, 95% of the times, the true value of this odds ratio lies within this interval.

- wait
  - Provided all other predictors remain constant, an increase in the ‘wait’ time by 1 hr would result in a decrease of 9.7% in the odds off a passenger approving SFO as ‘good’.
  - Our best estimate of odds ratio for ‘wait’ is 0.903, with a 95% confidence interval of 0.871 to 0.936. This is, 95% of the times, the true value of this odds ratio lies within this interval.

It is clearly observed that ‘dirty’ having the least p-value, has a greater influence as compared to wait on the output ‘good’; however, both ‘dirty’ and ‘wait’ influence the approval by a passenger, increments in their values lower the chance of getting a ‘good’ approval by a passenger.

We can look at the marginal effects of the two predictor variables on the output variable for their mean values:

term	dydx	std.error.	p.value	conf.low	conf.high	good	dirty	wait
dirty	**-0.194**	0.026	6.253e-14	-0.244	-0.143	1	0.093	2.80
wait	**-0.025**	0.004	3.355e-08	-0.034	-0.016	1	00.93	2.80

It highlights the change in percentage for probability predicted by the model. For 1 count increase in 'dirty', it will reduce the probability by 19.4%, given the 'wait' time remains constant. Similarly, for every 1hr increase in 'wait' time, the probability reduces by 2.4%, given the 'dirty' count remains constant.

For a passenger with average values of 'dirty' as 0.093 i.e 0 or no dirty places and 'wait' as 2.8 hrs, the predicted risk/probability is 58.47% for a passenger approving the airport as 'good'. If we increase the 'wait' time by 1hr while keeping 'dirty' constant, the predicted risk/probability drops to 55.96% i.e drop of approx. 2.51%. Similarly, keeping 'wait' constant at mean value 2.8hrs and increasing 'dirty' by 1, the predicted risk/probability drops to 38.81% i.e a drop of approx. 19.6%. These numbers match our marginal effects highlighted by the dydx column above.

SFO must try to keep the places as clean as possible and reduce any manner of wait time from their end to increase the probability of a passenger giving an approval as 'good'.

After predicting the outcome variables from this model with a cut-off of 50% for the predicted risk/probability, the confusion matrix obtained is as follows:

Confusion Matrix and Statistics:

	Reference	
Prediction	0	1
0	161	87
1	1355	2048

Accuracy : 0.605
95% CI : (0.589, 0.6209)
Sensitivity : 0.10620
Specificity : 0.95925

The accuracy of the model is 60.5%, with a 95% confidence interval of 58.9% to 62.09%. This model can be used by SFO for future predictions.

III. REFERENCES

[1] "Data visualization with R - GitHub pages." [Online]. Available: https://rkabacoff.github.io/datavis/. [Accessed: 20-Jan-2023].

[2] Datasciencebeginners, "Binary logistic regression with R: R-bloggers," *R*, 27-May-2020. [Online]. Available: https://www.r-bloggers.com/2020/05/binary-logistic-regression-with-r/. [Accessed: 20-Jan-2023].

[3] PETER BRUCE and ANDREW BRUCE, *Practical statistics for data scientists 50+ essential concepts using r and python;50+ essential concepts using r and python*. S.l.: O'REILLY MEDIA, INCORPORA, 2020.

[4] 365 C. Ltd., "Statistics for data science and business analysis," *O'Reilly Online Learning*. [Online]. Available: https://learning.oreilly.com/videos/statistics-for-data/9781789803259/. [Accessed: 20-Jan-2023].