FALCON AIRLINES
NONSTOP YOU

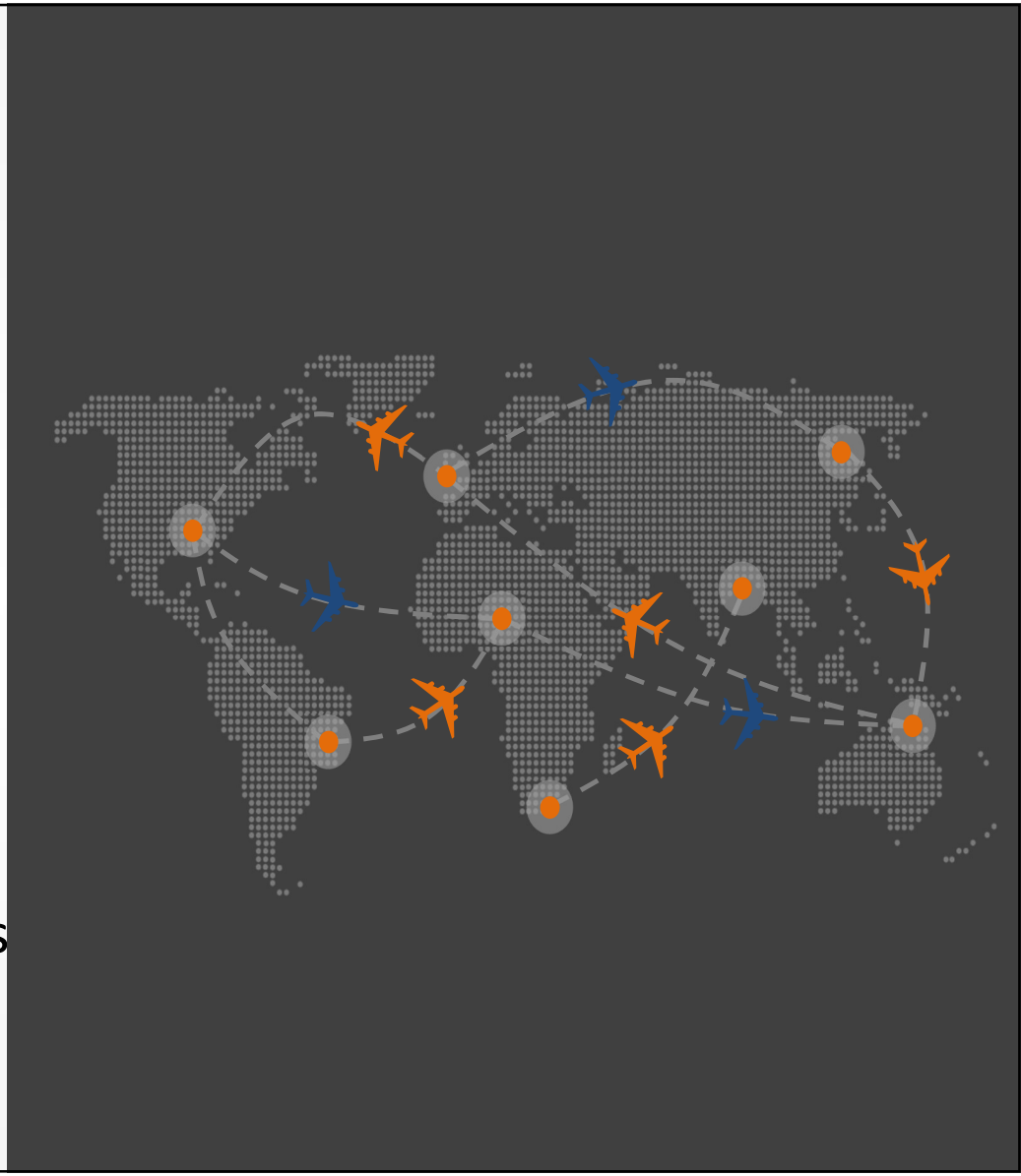
Falcon Airlines

Final Report
Submitted by: Mohanned T. Gomaa
Date: 06 July 2020

FALCON AIRLINES
NONSTOP YOU

Table of Contents

(*) Including Model Evaluation and Selection Criteria.



FALCON AIRLINES
NONSTOP YOU

Executive summary

We Understand that…

…Falcon Airlines has been recently **losing passengers** to the competition, due to overall **drop** in Passengers' **satisfaction** of their flights' services. The Airline has been working to resolve this issue by understand and map out the **key factors affecting** overall levels of **satisfaction** and the **best means** to **predict it.**

This recent customers' churn has been affecting Falcon Airlines on **three main fronts**;

Market share loss.

~45% of passenger sample are not satisfied.

Reputational damage.

High levels of dissatisfaction will have severe damage to reputation.

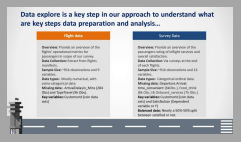
Financial loss.

Can reached up to ~60% revenue loss for ECO & ECO plus classes, and ~29% for Business class.

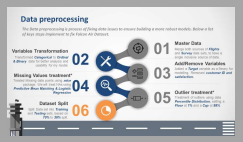
As a response to this issue we have designed a **six-step approach** that utilizes advances analytics including machine learning to **map out the key factors affecting satisfaction and supports future prediction.**

1. Understanding Business Need,
Were we defined the problem statement and its implications.

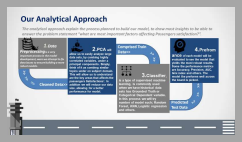
2. Data collection and exploration,
covered the initial exploration of data and mapping data issues.



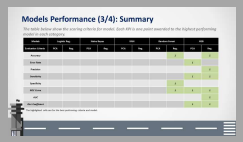
3. Understanding my data,
included a step by step approach for data processing.



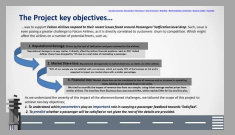
4. Modelling
This step covered the steps followed in building our classifiers models.



5. Model Eval. & selection,
We've built performance criteria to compare select the best model.



6. Business value,
through the understanding of the key factors affecting satisfaction and being able to predict it



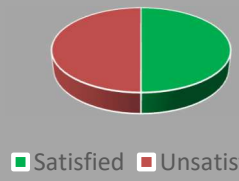
Our Key findings Mapped to Our Approach Stages (1/6)

1 Understanding Business Need | 2 Data collection and processing. | 3 Understand my data. | 4 Modelling | 5 Model Evaluation and Selection. | 6 Business value

The **need** for This Project was triggered by a noticeable **drop in customers' satisfaction** drop for Falcon Airlines services, this is **leading to customers' churn** in favor of competition. If this churn fully took effect, it would hit the business on three keys fronts. Below a glimpse of the impact quantified to show case the magnitude of the issue.

Market share loss.

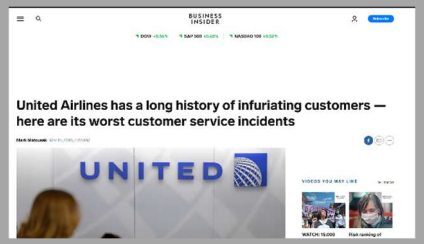
Based on the outcome of a satisfaction survey that captured the opinion of 91k passengers, ~45% of sample were not satisfied. This can be translated that the Airlines is at a risk to losing nearly Half of their business, if customer continue leaving and services were not improved. This will directly translate to market share loss with a similar percentage.



Most of the impact will hit "Loyal customers", which results showed that ~38% (29k passenger) are not satisfied and ~76% (13k passengers) for "Disloyal customers".

Reputational damage.

The high levels of dissatisfaction will have severe impact to Falcon Airlines brand and reputation. This might more difficult to recover from compared to financial or market share loss, due to the severity if its impact and the long-lasting impact this stays with our customers. For instance Untied Airlines Stocks has a severely been hit in 2017, by ~4% due to reputational damage. This to show case that reputation damage can directly lead to a financial loss if not mitigated.



Financial loss.

The severity of the financial is very evident when we take a closer look at revenue by class, then we will be able to see that two most money generating class ECO and ECO plus

that contributes to at least ~70% of plans seating capacity. Based on the graph presented above we can see that Risk of revenue loss for ECO & ECO plus classes is ~60% (48k pass.), and ~29% for Business class (43k pass.). To put this a monetary terms Business class loss might reach U$63m and ECO & ECO plus might be U$34m, based on avg. ticket prices.

Our Key findings Mapped to Our Approach Stages (2/6)

1 Understanding Business Need | **2 Data collection and processing.** | 3 Understand my data. | 4 Modelling | 5 Model Evaluation and Selection. | 6 Business value

We've received **two datasets** from Falcon Airlines management that covered **~91k passengers** to measure their overall satisfaction of their inflight services, along with flights operational details. Based on our initial exploration of the datasets we noted the following:

Flight data

This dataset provided an overview of the flights' operational details for 91k passengers surveyed, such as departure and arrival delays, flights durations and passengers' basic info. We understand that these dataset was extracted from the Airlines systems and was not collected; hence we have not validated the accuracy of the information.

Data Set Observations	
No. of instances	91k
No. of variables	9
Data type	Mostly Numerical data, continues.
Data Issues	Missing Values, outliers and variable transformations.

Survey Data

This dataset provided was in a survey format that covered 16 key questions for a sample of 91k passengers; the survey was designed to be a multiple choice that consisting of a rating between 1-5, where 1 is the lowest and it represents highly dissatisfied. In addition to an explicit question to rate their overall Satisfaction of the Airlines.

Data Set Observations	
No. of instances	91k
No. of variables	16
Data type	Mostly Categorical data, Ordinal.
Data Issues	Missing Values and variable transformations.

The main key takeaways from this stage was these **data sets can't be looked at separately**, and by combining them we can **understand further the impact of passengers' demographics' and flight operation impact of Satisfaction**. Also, we were able to map out the **data key issues** with data to be treated in the following stage.

Our Key findings Mapped to Our Approach Stages (3/6)

1 Understanding Business Need | 2 Data collection and processing | 3 **Understand my data.** | 4 Modelling | 5 Model Evaluation and Selection. | 6 Business value

In this step of our approach we mainly focus **preparing our data sets for the modelling** stage, mainly focusing in three main areas;

Merge and split data

In this area we focuses into two main things;
1- Creating a Master dataset by merging both flight and survey data received from the Airlines. This will allow us to combine both survey results, passengers' demographics and flights details, to have a better understanding of what is affecting Satfication.

Master dataset with 100% of passengers sampled (91k).

2- Split data sets once it is cleaned into Train and Test datasets, to be used in our modelling and validation process.

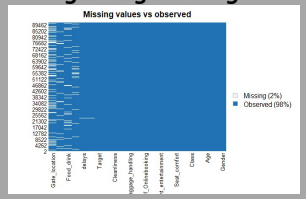
Train dataset with 70% of sample.

Test dataset with 30% of sample.

This use the train data to train the model and evaluate performance on test data, to avoid model overfitting.

Fixing data issues

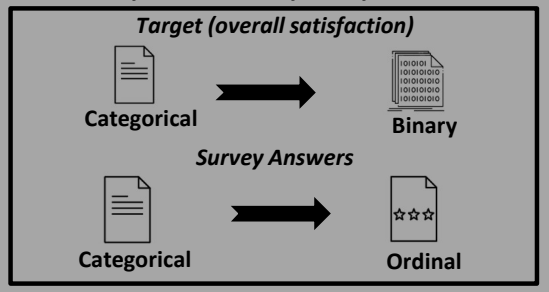
In this area we mainly focus into two main issues we mapped out in the previous stage;
1- Missing values treatment, that either can due to mistake in data extraction or simply questions that have not been answered by passenger. Leaving missing data untreated will impact the model bias and precision, affecting overall misleading results or can simply stop the model from running. We treated missing data using ML package called "MICE" that uses Predictive Mean Matching & Logistic Regression



2- Outliers can affect the overall model accuracy and cause the model to overfit, to preform good on test data and badly on test data. Treatment of outliers using data Percentile Distribution, setting a Floor at 1% and a Cap at 99%.

Variables additions and transformations

In this last area we focused on data engineering including;
1- Variables transformations, were we transformed survey Categorical to Ordinal & Binary data for better analysis and usability for my model.



2- Variables addition mainly focused on adding a Target column to be used in training and testing for overall Satfication and removed unnecessary columns such customer ID.

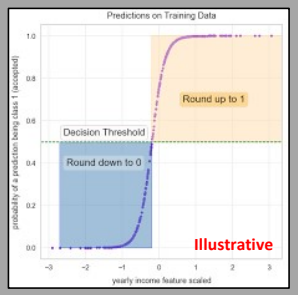
Our Key findings Mapped to Our Approach Stages (4/6)

1 Understanding Business Need | 2 Data collection and processing | 3 Understand my data. | 4 **Modelling** | 5 Model Evaluation and Selection. | 6 Business value

In this step of our approach we mainly focus **the modelling** process, were we select the relevant model to solve the problem at hand. So, we understand that Falcon Airlines wants to **predict** and be able to **class** passenger into two buckets either **Satisfied or Not**. Consequently, the best model for the issue at hand would be a **classifier models,** which class instances to the two-classes based on each instances probabilities' outcome. We have selected **five different algorithms** for this project.

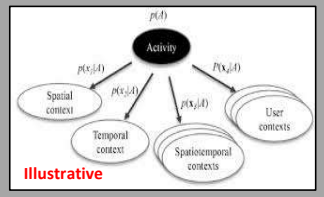
Logistic Regression

Is a machine learning model, is considered as a special case of linear regression model that is used to predict categorical variables. It simply class Target based by modelling probabilities, if higher than a certain cut-off its 1 and lower would be 0.



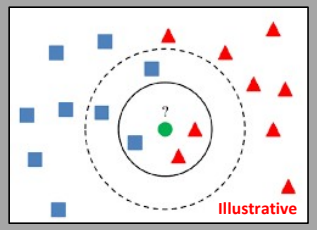
Naïve Bayes

Naïve Bayes is a classification model that is uses Bayes Theorem. It mainly assumes each feature is independent of each other and uses conditional probabilities between the Target and the features. It classes instance based by figuring out the probability of different variables being associated with a certain class.



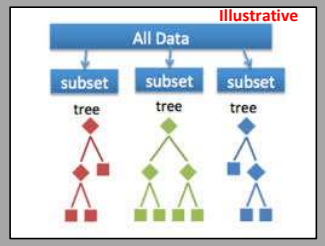
K Nearest Neighbor

K nearest neighbor "KNN" is a classification model, that groups instances based on their proximity to neighbor that belong to certain class. Proximity is measured by a measure of distance, most used is Euclidean distance.



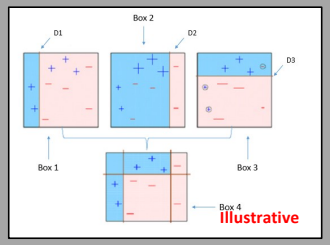
Random Forest

Random Forest is a type of ensemble modelling that uses bootstrap. Which build number of independent trees using random subsets of data, then uses mode of results to classify to the Target. It provides better predictive ability compared to bagging mainly due to random features selection .



X Gradient Boost

Extreme Gradient Boosting "XGB" is another ensemble modelling, but unlike RF, it is sequential leaner. So, deploy number models one after the other and which each one learns from its precedent, they learn from each others' mistakes, through penalization.

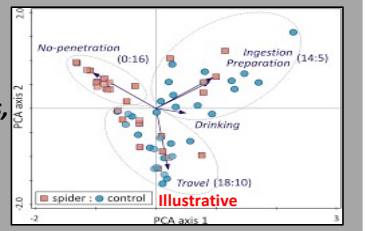


Our Key findings Mapped to Our Approach Stages (4/6)

1 Understanding Business Need | 2 Data collection and processing | 3 Understand my data. | 4 **Modelling** | 5 Model Evaluation and Selection. | 6 Business value

In our model built we have used two types of data, regular data and comprised data using **Dimensionality reduction**, which is a type of unsupervised machine learning that helps in reducing high dimensional data into a smaller version that still remains meaningful and retains the properties of the bigger set of data. This is usually helpful when dealing a large set of data, or when you want to understand the key theme that affects the overall data set.

Principal component analysis (PCA) *is one of the famous methods used in dimensionality reduction that increase our understanding of the data and minimize information loss due to reduction. PCA will allow us to easily analyze large data sets, by combing highly correlated variables, under a principal components. Simply, think of it as combing similar topics under on subject domain. This will allow us to understand are the key areas that affects the passengers Satisfactions'. In addition we will reduce our data size, allowing for a better performance for model.* **PCA resulted has in reducing the data set from 24 to 10 dimensions. These ten dimensions captures c.86% of overall variance of passengers'.** **These are the most important variable to predict Satisfaction**



PCA 1: ***Online Support*** *This PCA groups all online services offered by Falcon Airlines.*	**PCA2:** ***Boarding ease*** *This relates to boarding the plane, including gate location and other boarding services*	**PCA3:** ***Services & Facilities*** *This group planes cleanness and facilities*	**PCA4:** ***Timeliness*** *This groups departure and arrival delays and how timeliness affects Satisfaction.*	**PCA5:** ***Entrainment*** *Onboard Entrainment*
PCA6: ***Check-ins*** *Counter check-ins*	**PCA7:** ***Age*** *Age of the passenger*	**PCA 8:** ***Distance*** *Flight distance*	**PCA 9:** ***Seating*** *Seating and leg room*	**PCA 10:** ***Onboard service*** *On board of the plane*

Our Key findings Mapped to Our Approach Stages (5/6)

1 Understanding Business Need | 2 Data collection and processing | 3 Understand my data. | 4 Modelling | 5 **Model Evaluation and Selection.** | 6 Business value

We have defined a criteria to evaluate and select the best preforming model. Our criteria covers **four areas performance measures**, to make sure we are selecting the best preforming model. We have listed below top preforming models in order.

Confusion Matrix

The Confusion matrix show case the model prediction power, using number of KPIs. The overall model accuracy and Precision ability to predict +ve instances correctly.

Below a list of the models sorted based on their performance based in this area.

1. ***XGBoost***
2. ***Random Forest***
3. ***Logistic Regression***
4. ***Naïve Bayes***
5. ***KNN***

ROC

Plots the model performance by plotting true positive rate compared to false positive rate.

Below a list of the models sorted based on their performance based in this area.

1. ***XGBoost***
2. ***Random Forest***
3. ***Logistic Regression***
4. ***Naïve Bayes***
5. ***KNN***

Ks

A higher the KS Statistics is desired, which implies which is the better model, and it means separation between class-1(i.e. Satisfied) and class-2(i.e.Not).

Below a list of the models sorted based on their performance based in this area.

1. ***Random Forest***
2. ***XGBoost***
3. ***Logistic Regression***
4. ***Naïve Bayes***
5. ***KNN***

Concordance Ratios

The Concordance ratios below compare probabilities for each instance to assess the model predictive power.

Below a list of the models sorted based on their performance based in this area.

1. ***Random Forest***
2. ***XGBoost***
3. ***Logistic Regression***
4. ***Naïve Bayes***
5. ***KNN***

Based on the results shown we recommend using **ensemble model either XGBoost or Random Forest**, which more favor towards ***Random forest due it preforming better on the Ks area***. These models can provide prediction to Falcon Airlines with **~95% plus accuracy**.

Our Key findings Mapped to Our Approach Stages (6/6)

1 Understanding Business Need | 2 Data collection and processing | 3 Understand my data. | 4 Modelling | 5 Model Evaluation and Selection. | 6 **Business value**

We started with project with two main objectives to attempt to provide a solution for the ongoing challenge Falcon Airlines is facing with Satfication. Our Objectives were ;

1. To understand which parameters play an important role in swaying a passenger feedback towards 'Satisfied'.

2. To predict whether a passenger will be satisfied or not given the rest of the details are provided.

Project Objective no. one outcomes

- **We were able to sort variables based on their importance and below is a list of the top 6 variables in predicting satisfaction are:**

1- Seat comfort **2- *Entrainment*** ***3- Checking-services***
4- Customer type ***5- Travel type*** ***6- Online support***

- ***If Falcon airlines works to improve these top services, we believe they might see a noticeable improvement in Passengers' overall*** **Satisfaction.**
- **Also, based on our PCA results we were able to reduce our variables from 24 to 10 only. These ten dimensions captures c.86% of overall variance of passengers' satisfaction.**

Project Objective no. two outcomes

- **The Two recommended models for predicting Satfication were;**
  1. **Random Forest**
  2. **XGBoost**
- **Both are ensemble method, which was reflected in their performance.**
- **We recommend using either of these two model as they provide impressive performance with both high accuracy *95%+* and precision *95%+*. We would recommend using either RF or XGBoost in predicting passengers' satisfaction.**

Our Key findings Mapped to Our Approach Stages (6/6)

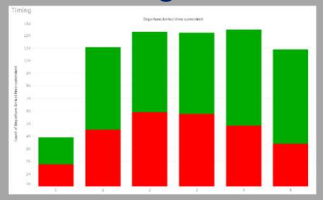
1 Understanding Business Need | 2 Data collection and processing | 3 Understand my data. | 4 Modelling | 5 Model Evaluation and Selection. | 6 **Business value**

In attempt to create value to Falcon Airlines we extracted key business insights from the data to create value to the business, by improving the decisions making process regrading corrective actions to improve overall customer Satisfaction. We recommend the following;

Age Groups

To focus efforts on improving the experience of the riskiest Age groups are between 0-30 years and 60+ years old.

Timing



Business travelers are one of Falcon strong customers, improving flights timing will help in improving the overall experience.

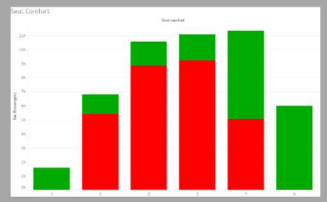
Type of Passengers

We have a noticeable issues with Disloyal customers, Falcon airlines need to introduce some perks that improves their experience.

By Class

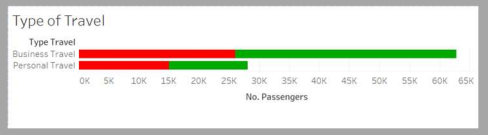
Falcon Should consider reviewing the perks of ECO class to improve passenger experience

Seat Comfort



This a very serious issue for Eco and Eco plus classes, which affecting overall experience of the passengers.

Type of Travel



Falcon Should consider improving the experience of personal travelers, not to be branded solely as a business airlines.

FALCON AIRLINES
NONSTOP YOU

Section 1:

Introduction

Introduction (1/2) – Project Overview

This project is centered around the airlines industry, focusing on one of the reputable US airline carrier '**Falcon Airlines**'. They have been recently **losing passengers** to competition, due to overall **drop** in clients' **satisfaction** of their flights' services. The Airlines are working to understand and map out the **key factors affecting** overall **satisfaction**.

In attempt to understand the root cause of the issue Falcon Airlines management has **survey ~91k passengers** to measure their overall satisfaction of their inflight services.

Data was collected from two sources, **Flight data** and **Satisfaction surveys**. Flights data provides overview of the flights' operational matrices, including distance, duration, delays and etc. While, survey data capture the passengers' satisfaction rating for inflight services, such meals, entertainment and etc.

In the survey, the passengers were explicitly asked whether they were satisfied with their overall flight experience and that is captured in the data of survey report under the variable labelled 'Satisfaction'.

The **overall objects** of this projects are;

1. To understand which **parameters** play an **important** role in swaying a passenger feedback towards 'satisfied'.
2. To **predict** whether a passenger will be satisfied or not given the rest of the details are provided.

Introduction (1/2) - Project Approach

FALCON AIRLINES
NONSTOP YOU

We have designed a six-steps approach that explain the key steps that we will undertake in our attempt to solve Falcon Airlines' problem statement and create value.
Our approach can be a reiterative process, as in this cycle might be repeated to based on our key insights discovered at step six.

6. Business Value:
- Extracting insights
- Business Decision

5. Model Evaluation:
- Model Evaluation and Tuning;
- Selecting the best preforming model;
- Model Deployment *(Out of scope)*.

4. Modelling:
- Select the appropriate ML Algorithm(s); &
- Train and Test Model(s).

3. Understand my data:
- Exploratory Data Analysis; &
- Feature Engineering.
- Data Split

2. Data:
- Data Collection *(Out of scope)*;
- Initial Data Exploration; &
- Data Processing and Preparation.

1. Business Need:
- Understand business context; &
- Define into a problem statement.

FALCON AIRLINES
NONSTOP YOU

Section 2:

Business Problem

Our approach starts with an understanding of Falcon Airlines' Business Need…

Project Background

Falcon Airlines is a US reputed airlines that has receiving several ***mixed reviews*** recently from their passengers regrading overall satisfaction. Consequently, number of ***passengers*** have ***leaving Falcon*** to other airlines.
Recently, Falcon has ***surveyed ~91k passengers*** to measure their overall satisfaction and understand what are the key factors affecting it.

What is the problem statement?

We attempt to understand what are ***they key factors*** that are ***affecting overall satisfaction***, leading to ***passengers' churn***.
The problem statement defined above is a very common for a highly competitive and dynamic industries such as airlines.

Why do we need to tackle this Issue?

Once we this mapping of key factors, we will be able to ***measure and predict passengers' satisfaction level***. This will unable Falcon Airlines to device corrective actions to tackle key issues to ***improve*** passengers ***experience***.

What is the potential impact?

Improving passengers' experience will lead to a ***higher retention rates***. Higher passengers' retention rates can be translated into a ***bigger market share*** and ***higher profits*** for the airlines.

FALCON AIRLINES
NONSTOP YOU

Section 3:

Dataset Overview

Data explore is a key step in our approach to understand what are key steps data preparation and analysis…

Flight data

Overview: Provide an overview of the flights' operational metrics for passengers in scope of our survey.
Data Collection: Extract from flights manifests.
Sample Size: ~91k observations and 9 variables.
Data types: Mostly numerical, with some categorical data
Missing data:. ArrivalDelayin_Mins (284 Obs) and TypeTravel (9k Obs).
Key variables: CustomerId (Join data sets)

Survey Data

Overview: Provide an overview of the passengers rating of inflight services and overall satisfaction.
Data Collection: Via surveys at the end of each flights.
Sample Size: ~91k observations and 16 variables.
Data types: Categorical ordinal data
Missing data: Departure.Arrival. time_convenient (8kObs.), Food_drink (8k Obs.) & Onboard_services (7k Obs.)
Key variables: CustomerId (Join data sets) and Satisfaction (Dependent variable or Y)
Balanced data: Nearly a 50%-50% split between satisfied or not.

Detailed look at data sets (1/4) – Flight Data…

The table below provide a descriptive statistics for flight data set…

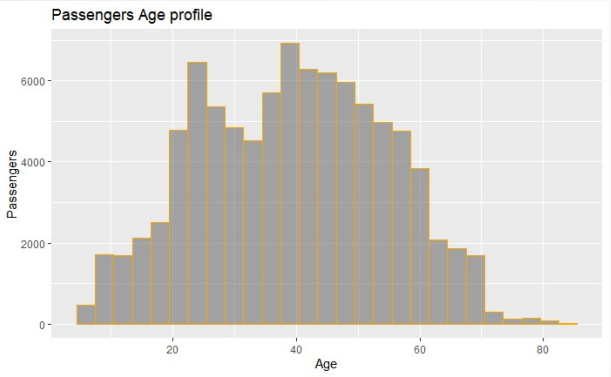
Varibale		Data type	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID	**1**	Numerical	90917	195423	26245.62	195423	195423	33698.02	149965	240881	90916	0	-1.20004	87.04309
Gender	**2**	Categorical	90917	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CustomerType	**3**	Categorical	90917	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Age	**4**	Numerical	90917	39.44717	15.12979	40	39.46007	17.7912	7	85	78	-0.00065	-0.71861	0.050178
TypeTravel	**5**	Categorical	90917	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Class	**6**	Categorical	90917	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Flight_Distance	**7**	Numerical	90917	1981.629	1026.78	1927	1938.535	879.1818	50	6950	6900	0.460165	0.350692	3.405296
DepartureDelayin_Mins	**8**	Numerical	90917	14.68659	38.66926	0	5.76831	0	0	1592	1592	7.364971	118.1917	0.128246
ArrivalDelayin_Mins	**9**	Numerical	90633	15.05893	39.03852	0	6.064945	0	0	1584	1584	7.202062	111.988	0.129673

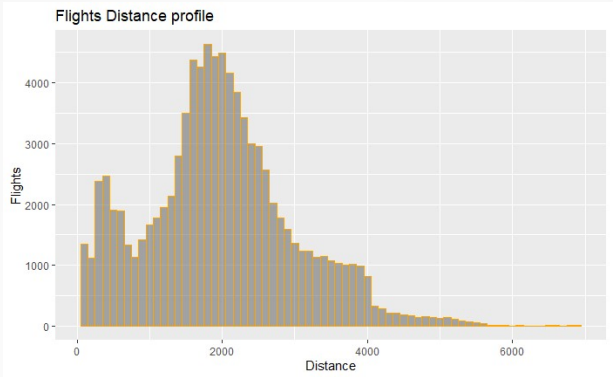
Key insights:

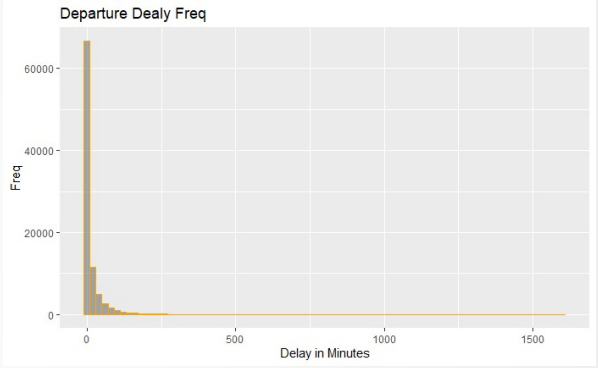
1. **CustomerID:** is a variable this gives passenger ID onboard of our flights, although it is numerical it is commonly treated as string. It can be used to join data.
2. **Gender:** is a categorical variable that indicates the gender of my passengers.
3. **CustomerType:** is a categorical variable that indicates the type of my passengers, either frequent flyer or not.
4. **Age:** Passengers age range between 7-85 years, with average age of 39 years. Age is skewed to the left indicating older passengers profile. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
5. **TypeTravel:** is a categorical variable that indicates the type of travel for each passengers, either business or pleasure.
6. **Class:** is a categorical variable that indicates the type of ticket class, Business vs economy.
7. **Flight_Distance:** Distance ranges between 50-6950 miles, with average age of 1981 miles. Distance is skewed slightly to the right indicating more frequent shorter flights. leptokurtic Kurtosis (+ve) indicates more outliers in my data.
8. **DepartureDelayin_Mins:** Delay ranges between 0-1592 miles, with average age of 15 mins. Distance is skewed to the right indicating long delays are less frequent. leptokurtic Kurtosis (+ve) indicates high number outliers in my data.
9. **ArrivalDelayin_Mins:** Delay ranges between 0-1584 miles, with average age of 15 mins. Distance is skewed to the right indicating long delays are less frequent. leptokurtic Kurtosis (+ve) indicates high number outliers in my data.

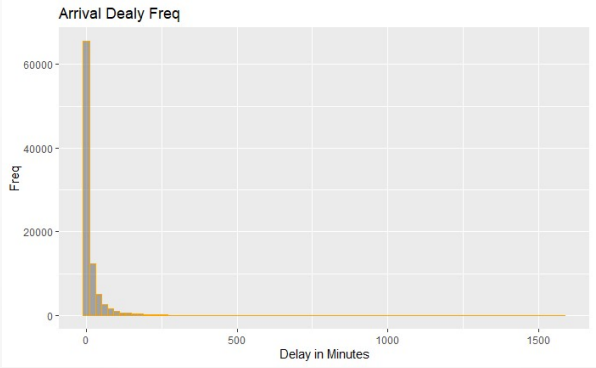
Visual inspection of data sets – Flight Data…

Passengers Age profile seem to around the mean of 40 years, who fly with around an average of 2000 miles with low or insignificant delays.









Detailed look at data sets (2/4) – Survey Data…

The table below provide a descriptive statistics for Survey data set, we had to transform data ordinal data into its original rating as a numerical variables to get insightful stats. Generally data seems to be skewed to the left (i.e –ve) meaning more higher ratings, with –ve kurtosis meaning lower number of outliers …

Variables		Data Type	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerId	*1*	Numerical	90917	195423	26245.62	195423	195423	33698.02	149965	240881	90916	0	-1.20004	87.04309
Satisfaction	*2*	Categorical	90917	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Seat_comfort	*3*	Numerical	90917	2.838831	1.393582	3	2.844848	1.4826	0	5	5	-0.09242	-0.94214	0.004622
Departure.Arrival.time_convenient	*4*	Numerical	82673	2.993251	1.525231	3	3.055051	1.4826	0	5	5	-0.25304	-1.08662	0.005305
Food_drink	*5*	Numerical	82736	2.850102	1.443017	3	2.86995	1.4826	0	5	5	-0.11434	-0.98417	0.005017
Gate_location	*6*	Numerical	67532	2.276373	0.818698	3	2.345482	0	0	3	3	-0.5477	-1.29483	0.00315
Inflightwifi_service	*7*	Numerical	90917	3.251559	1.320115	3	3.315763	1.4826	0	5	5	-0.19492	-1.12241	0.004378
Inflight_entertainment	*8*	Numerical	90917	3.383955	1.342158	4	3.507954	1.4826	0	5	5	-0.60173	-0.53234	0.004451
Online_support	*9*	Numerical	90917	3.519133	1.307794	4	3.648917	1.4826	0	5	5	-0.57558	-0.81262	0.004337
Ease_of_Onlinebooking	*10*	Numerical	90917	3.47561	1.304658	4	3.594666	1.4826	0	5	5	-0.49575	-0.90488	0.004327
Onboard_service	*11*	Numerical	83738	3.466503	1.269375	4	3.583159	1.4826	0	5	5	-0.50904	-0.77792	0.004387
Leg_room_service	*12*	Numerical	90917	3.486994	1.291758	4	3.599546	1.4826	0	5	5	-0.49926	-0.83423	0.004284
Baggage_handling	*13*	Numerical	90917	3.697416	1.154341	4	3.823194	1.4826	1	5	4	-0.74542	-0.22573	0.003828
Checkin_service	*14*	Numerical	90917	3.340761	1.260548	3	3.425957	1.4826	0	5	5	-0.39198	-0.79283	0.004181
Cleanliness	*15*	Numerical	90917	3.707887	1.148017	4	3.833863	1.4826	0	5	5	-0.75767	-0.19387	0.003807
Online_boarding	*16*	Numerical	90917	3.352475	1.299698	4	3.440709	1.4826	0	5	5	-0.36698	-0.94103	0.00431

Key insights:

1. **CustomerID:** is a variable this gives passenger ID onboard of our flights, although it is numerical it is commonly treated as string. It can be used to join data.
2. **Satisfaction:** is a categorical variable and it indicates the overall Satfication of the passengers. It is our dependent variable Y, that we aim to predict.
3. **Seat_comfort:** is a ordinal variable that indicates passengers' seat comfort level. Average rating is ~3 or "acceptable". Data is slightly skewed to the left, meaning that some passengers have rated seats to be more comfortable. Platykurtic Kurtosis (-ve) indicates less outliers in my data.

Detailed look at data sets (3/4) – Survey Data (cont.)…

The table below provide a descriptive statistics for Survey data set, we had to transform data ordinal data into its original rating as a numerical variables to get insightful stats. Generally data seems to be skewed to the left (i.e –ve) meaning more higher ratings, with –ve kurtosis meaning lower number of outliers …

Key insights:

4. **Departure.Arrival.time_convenient:** is an ordinal variable that indicates passengers' Satisfaction level with flights timing. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated timing Satfication on the higher end. Platykurtic Kurtosis (-ve) indicates less data outliers.
5. **Food_drink:** is an ordinal variable that indicates passengers' If they enjoy onboard meals. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated meals Satfication on the higher end. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
6. **Gate_location:** is an ordinal variable that indicates passengers' feel the gates are far or not. Average rating is ~2 or "need improvement". Data is skewed to the left, meaning that more passengers have rated gates higher than average. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
7. **Inflightwifi_service:** is an ordinal variable that indicates passengers' happens with onboard Wifi. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated Wifi higher than average. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
8. **Inflight_entertainment:** is an ordinal variable that indicates passengers' Satfication level with onboard entertainment. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated ent. higher than average. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
9. **Online_support:** is an ordinal variable that indicates passengers' Satfication level with online support. Average rating is ~4 or "good". Data is skewed to the left, meaning that more passengers have rated support Satfication higher than average. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
10. **Ease_of_Onlinebooking:** is an ordinal variable that indicates passengers' Satfication level with online booking. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated online booking higher than avg. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
11. **Onboard_service:** is an ordinal variable that indicates passengers' Satfication level with onboard services. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated services on the higher end. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
12. **Leg_room_service:** is an ordinal variable that indicates passengers' Satfication level with leg room. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated leg room Satfication higher than avg. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
13. **Baggage_handling:** is an ordinal variable that indicates passengers' Satfication level with baggage handle. Average rating is ~4 or "good". Data is skewed to the left, meaning that more passengers have rated handling higher than avg. Platykurtic Kurtosis (-ve) indicates less outliers in my data.

Detailed look at data sets (4/4) – Survey Data (cont.)…

The table below provide a descriptive statistics for Survey data set, we had to transform data ordinal data into its original rating as a numerical variables to get insightful stats …

Key insights:

14. **Checkin_service:** is an ordinal variable that indicates passengers' Satfication level with check-in. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated check-in higher than avg. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
15. **Cleanliness:** is an ordinal variable that indicates passengers' Satfication level with Cleanliness. Average rating is ~4 or "good". Data is skewed to the left, meaning that more passengers have rated Cleanliness higher than avg. Platykurtic Kurtosis (-ve) indicates less outliers in my data.
16. **Online_boarding:** is an ordinal variable that indicates passengers' Satfication level with Online_boarding. Average rating is ~3 or "acceptable". Data is skewed to the left, meaning that more passengers have rated Online_boarding higher than avg. Platykurtic Kurtosis (-ve) indicates less outliers in my data.

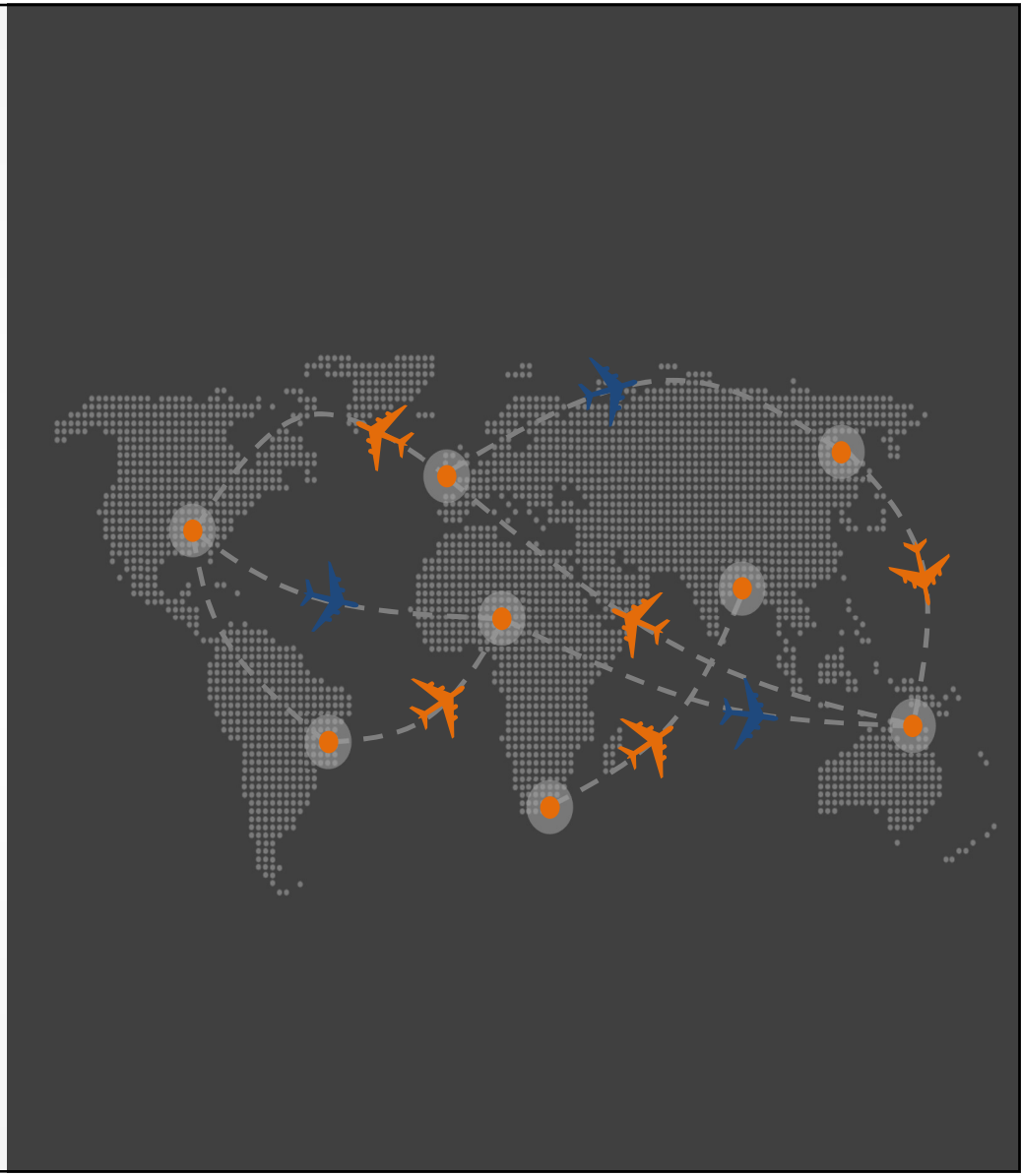
FALCON AIRLINES
NONSTOP YOU

Section 4: Initial Data exploration

Project Submission Notes I

FALCON AIRLINES
NONSTOP YOU

Table of Contents



Data Relationships (1/2) – Correlation Matrix

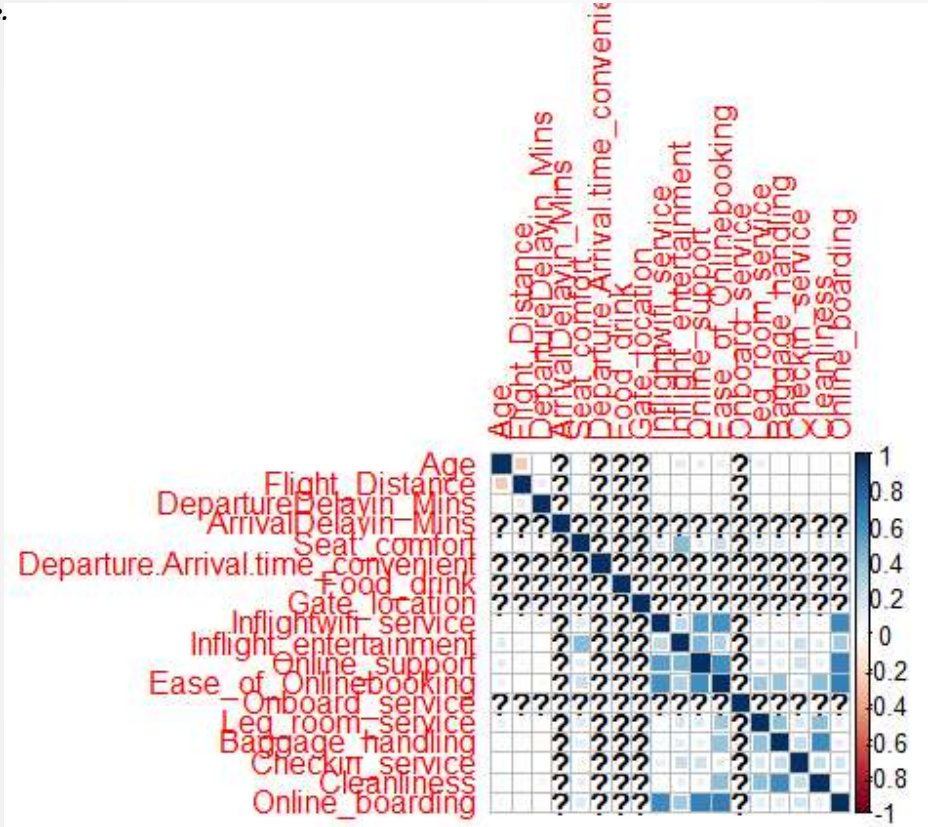
The table below provide an overview of the relation between by data points and how variable react to each other, mostly their positive relationship highlighted in Green

	Age	Flight_Distance	DepartureDelayin_Mins	ArrivalDelayin_Mins	Seat_comfort	Departure.Arrival.time_convenient	Food_drink	Gate_location	Inflightwifi_service	Inflight_entertainment	Online_support	Ease_of_Onlinebooking	Onboard_service	Leg_room_service	Baggage_handling	Checkin_service	Cleanliness	Online_boarding
Age	1.0	-0.3	0.0		0.0				0.0	0.1	0.1	0.1		0.1	0.0	0.0	0.0	0.0
Flight_Distance	-0.3	1.0	0.1		0.0				0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0
DepartureDelayin_Mins	0.0	0.1	1.0		0.0				0.0	0.0	0.0	0.0		0.0	0.0	0.0	-0.1	0.0
ArrivalDelayin_Mins				1.0														
Seat_comfort	0.0	0.0	0.0		1.0				0.1	0.4	0.1	0.2		0.1	0.1	0.0	0.1	0.1
Departure.Arrival.time_convenient						1.0												
Food_drink							1.0											
Gate_location								1.0										
Inflightwifi_service	0.0	0.0	0.0		0.1				1.0	0.3	0.6	0.6		0.0	0.0	0.1	0.0	0.6
Inflight_entertainment	0.1	0.0	0.0		0.4				0.3	1.0	0.4	0.3		0.2	0.1	0.2	0.1	0.4
Online_support	0.1	0.0	0.0		0.1				0.6	0.4	1.0	0.6		0.1	0.1	0.2	0.1	0.7
Ease_of_Onlinebooking	0.1	0.0	0.0		0.2				0.6	0.3	0.6	1.0		0.4	0.4	0.1	0.4	0.7
Onboard_service													1.0					
Leg_room_service	0.1	0.0	0.0		0.1				0.0	0.2	0.1	0.4		1.0	0.4	0.2	0.4	0.1
Baggage_handling	0.0	0.0	0.0		0.1				0.0	0.1	0.1	0.4		0.4	1.0	0.2	0.6	0.1
Checkin_service	0.0	0.0	0.0		0.0				0.1	0.2	0.2	0.1		0.2	0.2	1.0	0.2	0.2
Cleanliness	0.0	0.0	-0.1		0.1				0.0	0.1	0.1	0.4		0.4	0.6	0.2	1.0	0.1
Online_boarding	0.0	0.0	0.0		0.1				0.6	0.4	0.7	0.7		0.1	0.1	0.2	0.1	1.0

The underlying issues with the data can be seen in this table above, where some of the data are not in the correct type or missing this affected the completeness of the above table as you see.

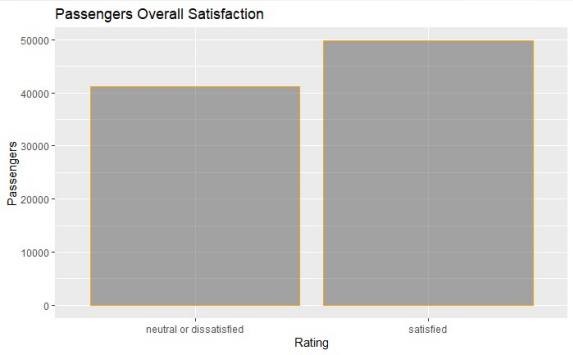
Data Relationships (2/2) – Correlation Matrix

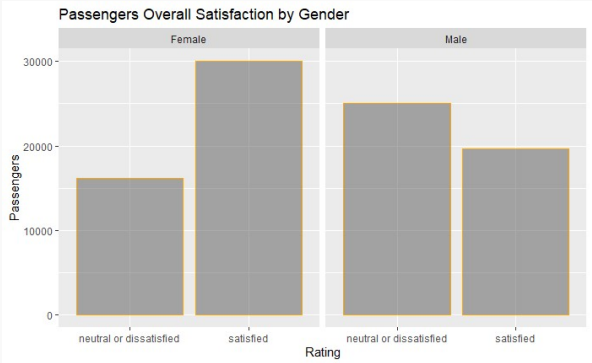
The underlying issues with the data can be seen in this graph below, where some of the data are not in the correct type or missing this affected the completeness of the above graph as you see.

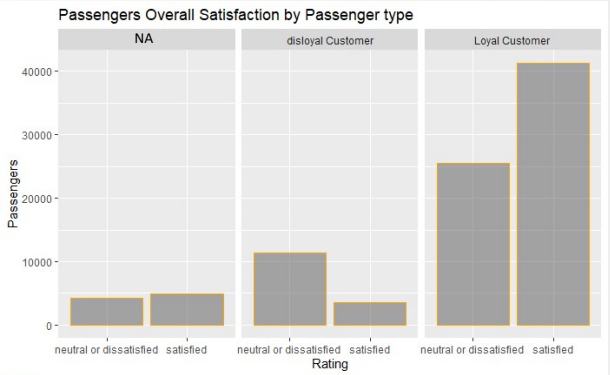


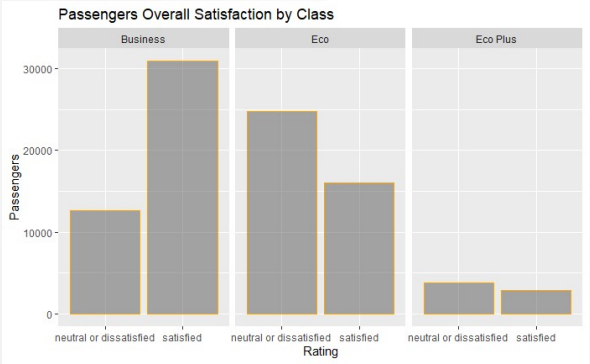
Visual inspection of survey results (1/6) – Satisfaction …

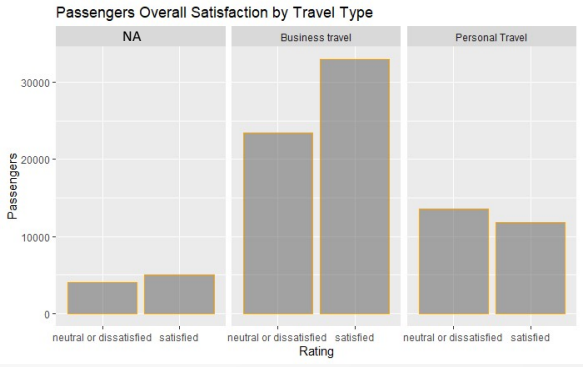
~54% of overall population are satisfied, mostly are loyal passengers from business class that are on a business trip…





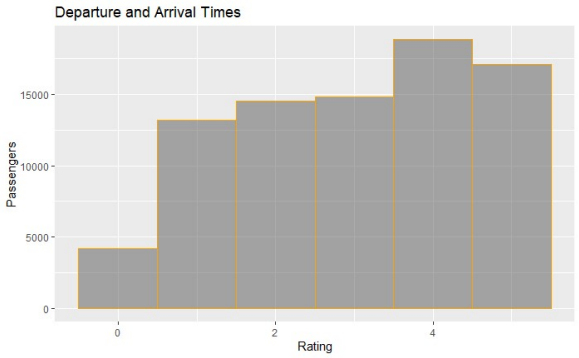


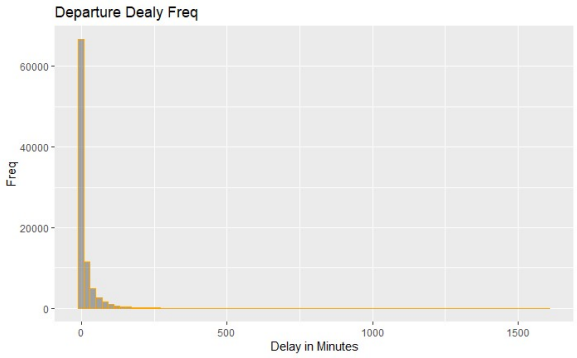


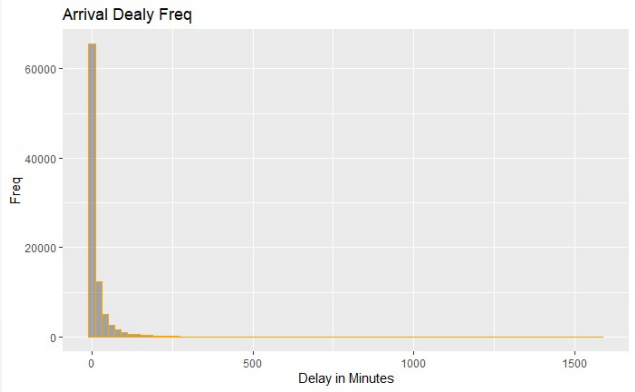


Visual inspection of survey results (2/6) – Timeliness…

Overall delays seems to not as frequent with our Airlines. While, flights timing Satfication seems to be fragmented showing a uniform distribution.

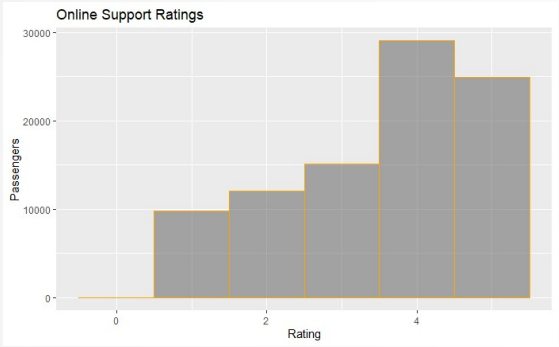


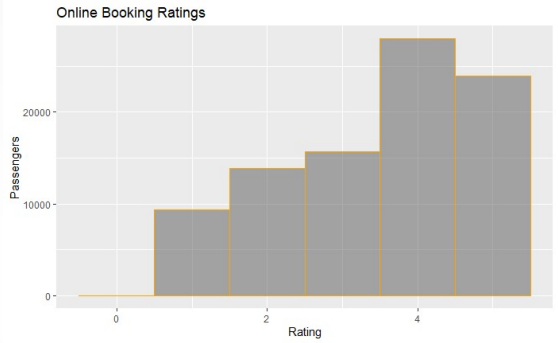


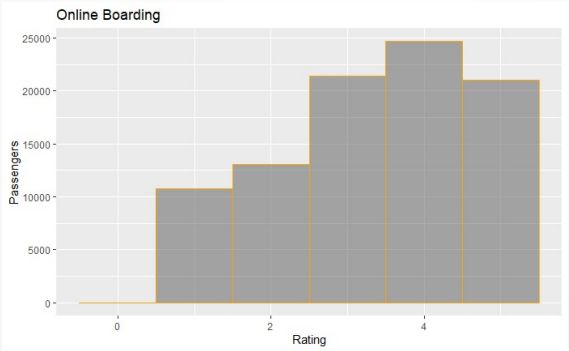


Visual inspection of survey results (3/6) – Online Services…

Overall passengers seem to be happy with our online services, majority rated 4 or good and above…

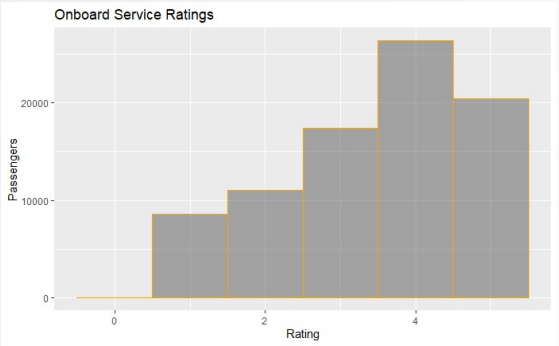


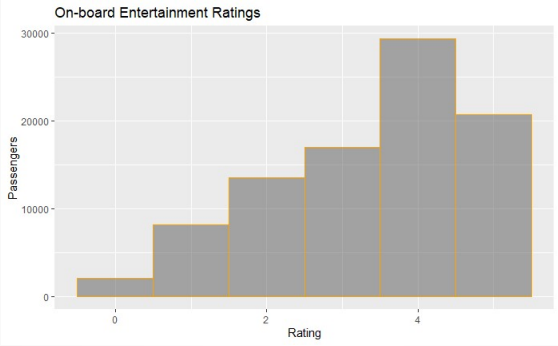


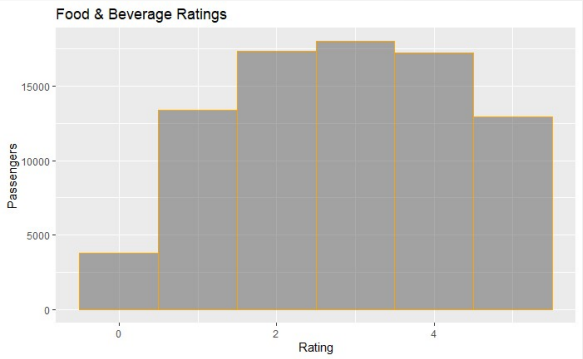


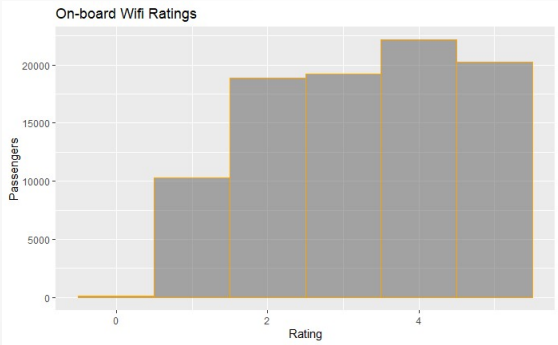
Visual inspection of survey results (4/6) – Onboard Exp…

Overall passengers are mostly happy with our onboard services, maybe F&B and WIFI would need some improvements…



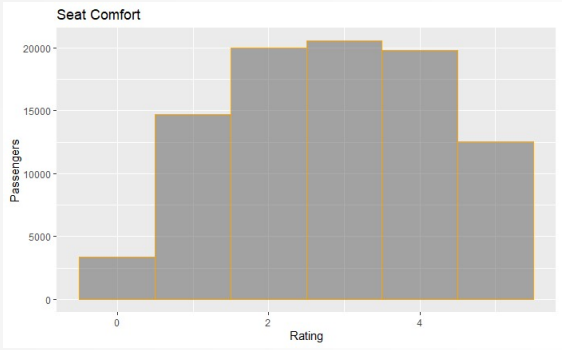


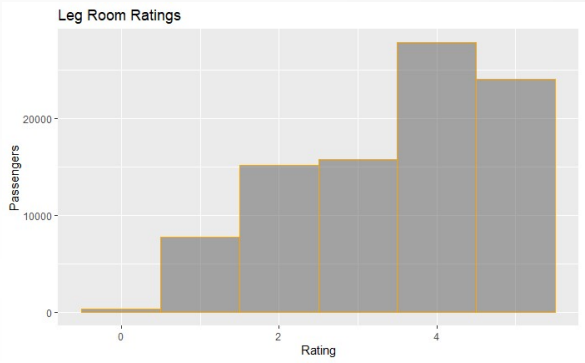


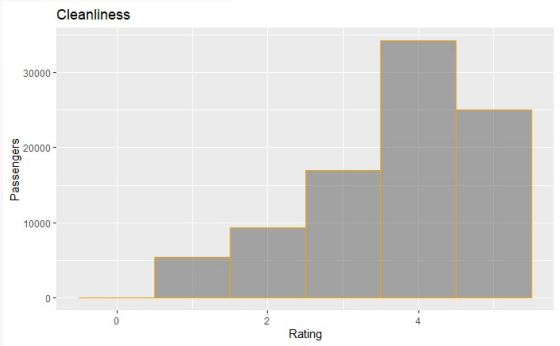


Visual inspection of survey results (5/6) – Planes…

Overall passengers seems to be happy with leg room and plan cleanliness, while seat comfort might need some improvement, being concentrated around the average.

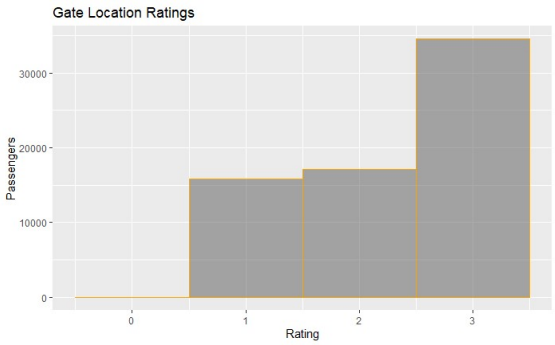


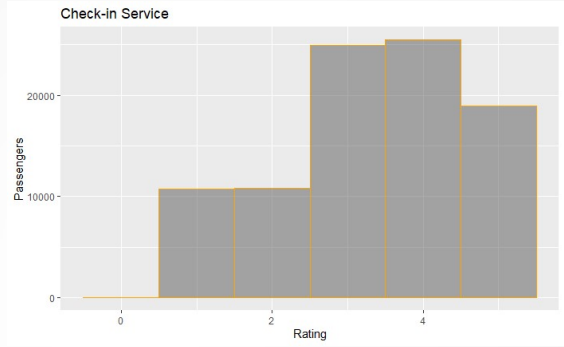


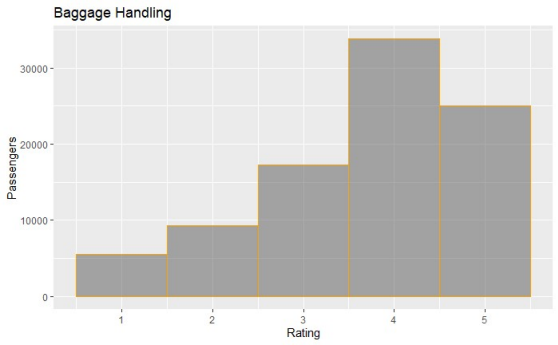


Visual inspection of data sets (6/6) – Before and After…

Overall passengers seems to find our before and after services acceptable, good at most. Gates location is an area worth looking into for further improvements…



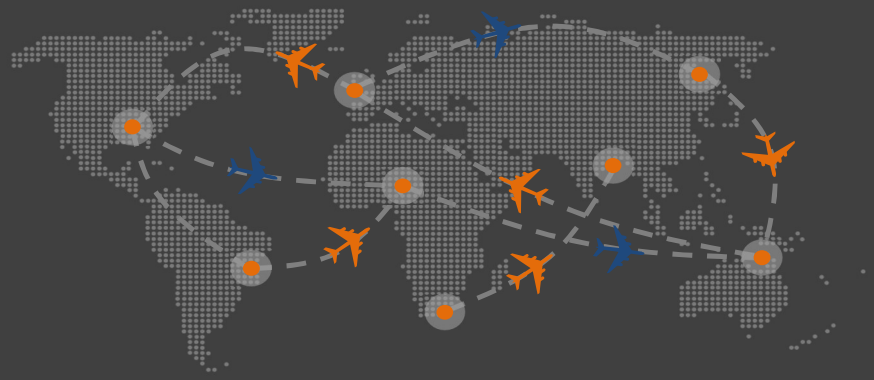




FALCON AIRLINES
NONSTOP YOU

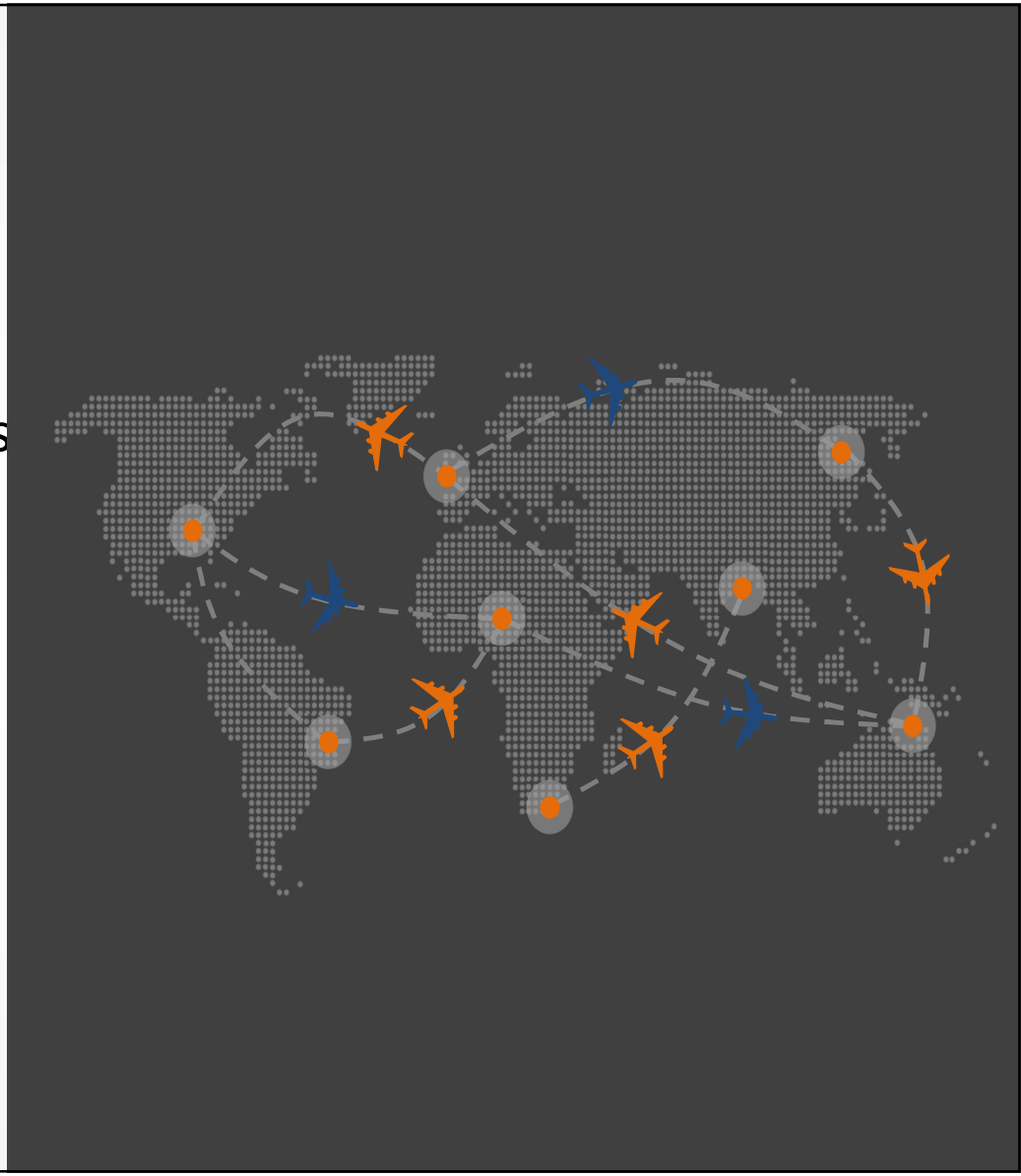
Section 5: Data Preprocessing

Project Submission Notes II



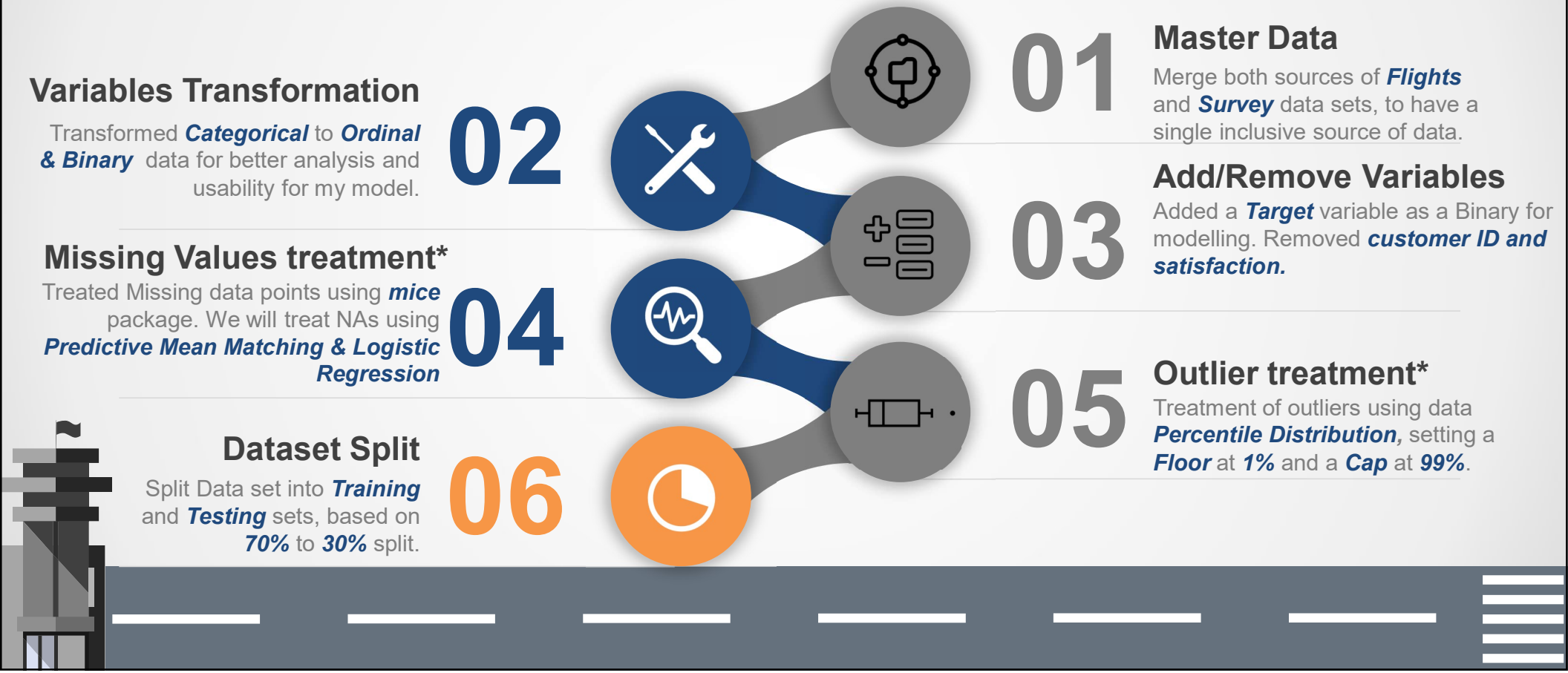
FALCON AIRLINES
NONSTOP YOU

Table of Contents



Data preprocessing

The Data preprocessing is process of fixing data issues to ensure building a more robust models. Below a list of keys steps implement to fix Falcon Air Dataset.



01 Master Data

Merge both sources of ***Flights*** and ***Survey*** data sets, to have a single inclusive source of data.

02 Variables Transformation

Transformed ***Categorical*** to ***Ordinal & Binary*** data for better analysis and usability for my model.

03 Add/Remove Variables

Added a ***Target*** variable as a Binary for modelling. Removed ***customer ID and satisfaction.***

04 Missing Values treatment*

Treated Missing data points using ***mice*** package. We will treat NAs using ***Predictive Mean Matching & Logistic Regression***

05 Outlier treatment*

Treatment of outliers using data ***Percentile Distribution,*** setting a ***Floor*** at ***1%*** and a ***Cap*** at ***99%***.

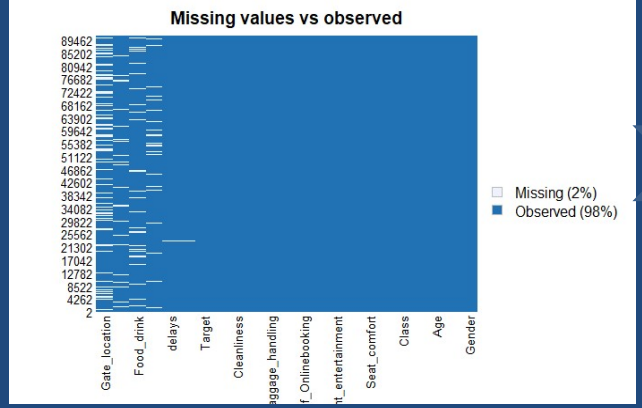
06 Dataset Split

Split Data set into ***Training*** and ***Testing*** sets, based on ***70%*** to ***30%*** split.

Data preprocessing: Step 04

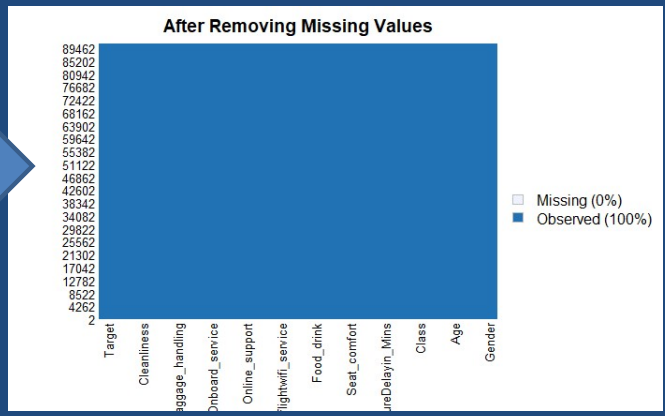
Visual depiction of missing values for variables shown below.

Before



We were able to get 0% missing values using MICE package

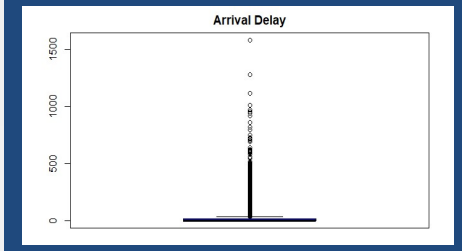
After

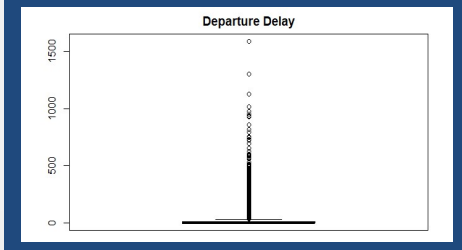


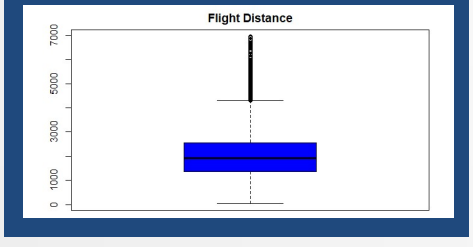
Data preprocessing: Step 05

Visual depiction of outliers for three main variables shown below.

Before

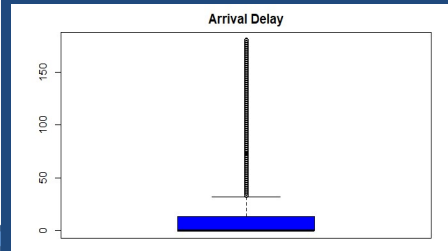


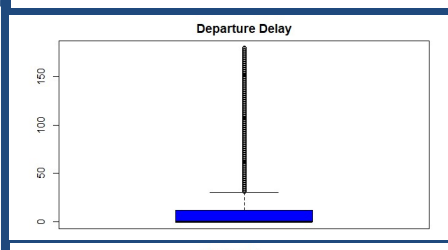


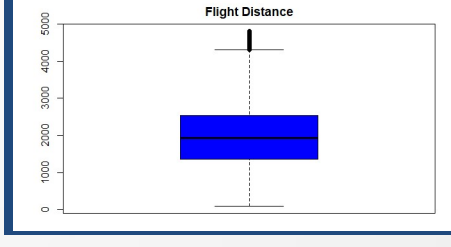


We did not get rid of outliers, we removed everything beyond the 99% percentile and below the 1% percentile

After

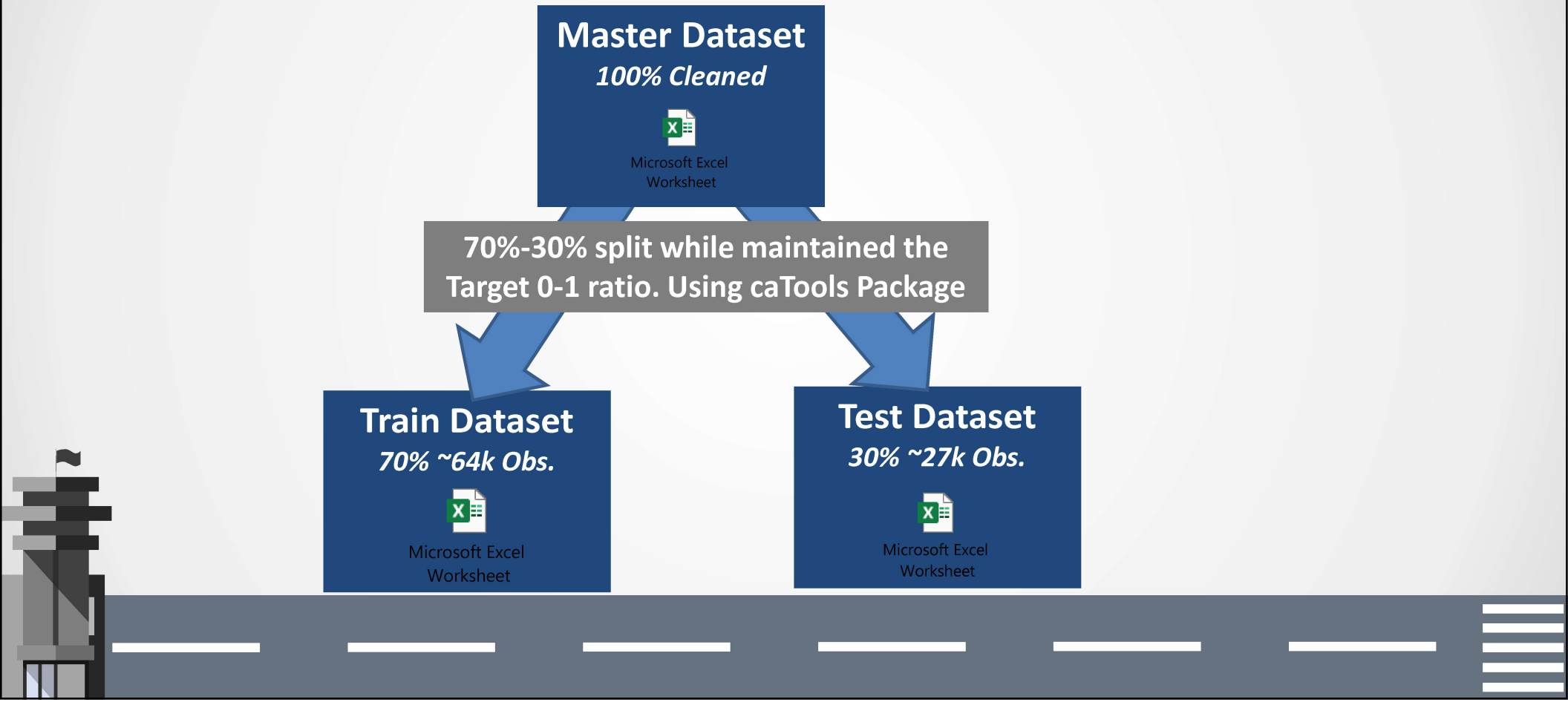






Data preprocessing: Step 06

Below Attached copies for Train and Test sets…



Exploratory Data Analysis: Relational Analysis

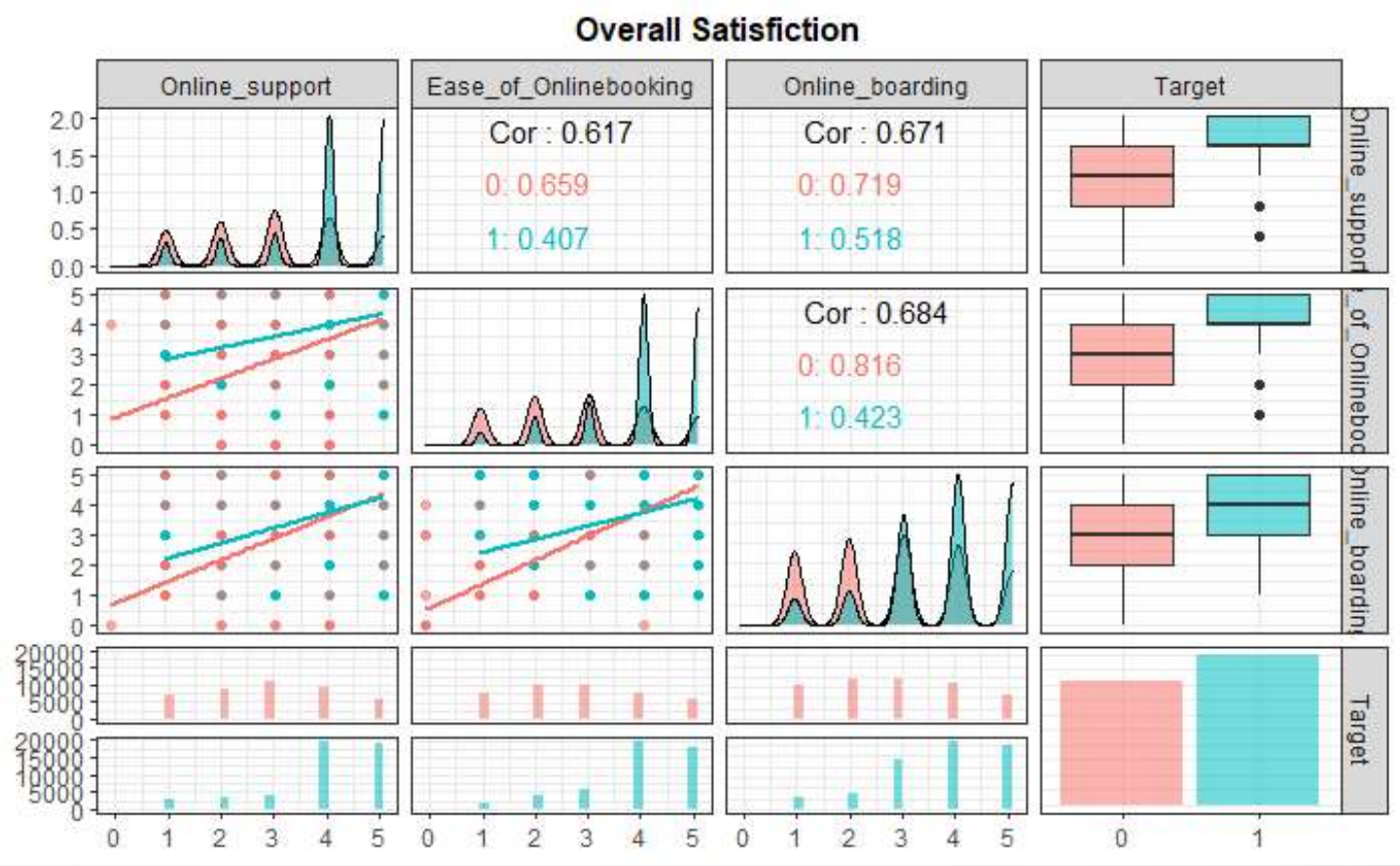
Relationship among variables can be presented by a Correlation plot and a matrix, as shown below

	Age	Flight_Distance	Departure Delay	Arrival Delay	Seat Comfort	Travel time Convenient	Food Drink	Gate Location	Inflight WIFI	Inflight Entertain.	Online Support	Ease of Onlinebooking	Onboard Service	Leg Room	Baggage handling	Checkin Service	Cleanliness	Online Boarding
Age	1.00	-0.25	-0.01	-0.01	0.01	0.04	0.02	0.00	0.01	0.13	0.12	0.07	0.07	0.09	-0.01	0.03	-0.02	0.04
Flight_Distance	-0.25	1.00	0.10	0.10	-0.05	0.00	-0.01	0.00	0.01	-0.03	-0.03	-0.02	-0.03	-0.03	0.02	0.00	0.01	0.01
Departure Delay	-0.01	0.10	1.00	0.94	-0.02	0.00	-0.01	0.01	-0.03	-0.03	-0.04	-0.04	-0.04	0.00	-0.01	-0.02	-0.06	-0.02
Arrival Delay	-0.01	0.10	0.94	1.00	-0.03	0.00	-0.02	0.01	-0.04	-0.04	-0.04	-0.04	-0.05	0.00	-0.02	-0.03	-0.07	-0.03
Seat Comfort	0.01	-0.05	-0.02	-0.03	1.00	0.44	0.72	0.41	0.13	0.42	0.12	0.21	0.12	0.14	0.12	0.05	0.11	0.13
Travel time Convenient	0.04	0.00	0.00	0.00	0.44	1.00	0.53	0.55	-0.01	0.08	0.00	0.00	0.06	0.03	0.07	0.06	0.06	0.00
Food Drink	0.02	-0.01	-0.01	-0.02	0.72	0.53	1.00	0.53	0.03	0.37	0.03	0.04	0.04	0.08	0.04	0.02	0.03	0.02
Gate Location	0.00	0.00	0.01	0.01	0.41	0.55	0.53	1.00	0.00	0.00	0.00	0.00	-0.02	-0.01	0.00	-0.03	0.00	0.00
Inflight WIFI	0.01	0.01	-0.03	-0.04	0.13	-0.01	0.03	0.00	1.00	0.25	0.56	0.60	0.06	0.04	0.04	0.09	0.04	0.63
Inflight Entertain.	0.13	-0.03	-0.03	-0.04	0.42	0.08	0.37	0.00	0.25	1.00	0.44	0.32	0.18	0.16	0.12	0.23	0.11	0.35
Online Support	0.12	-0.03	-0.04	-0.04	0.12	0.00	0.03	0.00	0.56	0.44	1.00	0.62	0.16	0.14	0.10	0.20	0.10	0.67
Ease of Onlinebooking	0.07	-0.02	-0.04	-0.04	0.21	0.00	0.04	0.00	0.60	0.32	0.62	1.00	0.43	0.36	0.40	0.14	0.42	0.68
Onboard Service	0.07	-0.03	-0.04	-0.05	0.12	0.06	0.04	-0.02	0.06	0.18	0.16	0.43	1.00	0.41	0.53	0.25	0.55	0.14
Leg Room	0.09	-0.03	0.00	0.00	0.14	0.03	0.08	-0.01	0.04	0.16	0.14	0.36	0.41	1.00	0.41	0.17	0.41	0.11
Baggage handling	-0.01	0.02	-0.01	-0.02	0.12	0.07	0.04	0.00	0.04	0.12	0.10	0.40	0.53	0.41	1.00	0.24	0.63	0.11
Checkin Service	0.03	0.00	-0.02	-0.03	0.05	0.06	0.02	-0.03	0.09	0.23	0.20	0.14	0.25	0.17	0.24	1.00	0.24	0.18
Cleanliness	-0.02	0.01	-0.06	-0.07	0.11	0.06	0.03	0.00	0.04	0.11	0.10	0.42	0.55	0.41	0.63	0.24	1.00	0.11
Online Boarding	0.04	0.01	-0.02	-0.03	0.13	0.00	0.02	0.00	0.63	0.35	0.67	0.68	0.14	0.11	0.11	0.18	0.11	1.00

As, we may see that most of the data set is positivity correlated, with less significant negative relations. We have noticed a strong relation between variables that are of the same domain, for example all online support, online boarding, ease of online boarding and inflight Wi-Fi are correlated. This even reinforce the option to apply a PCA to group this items

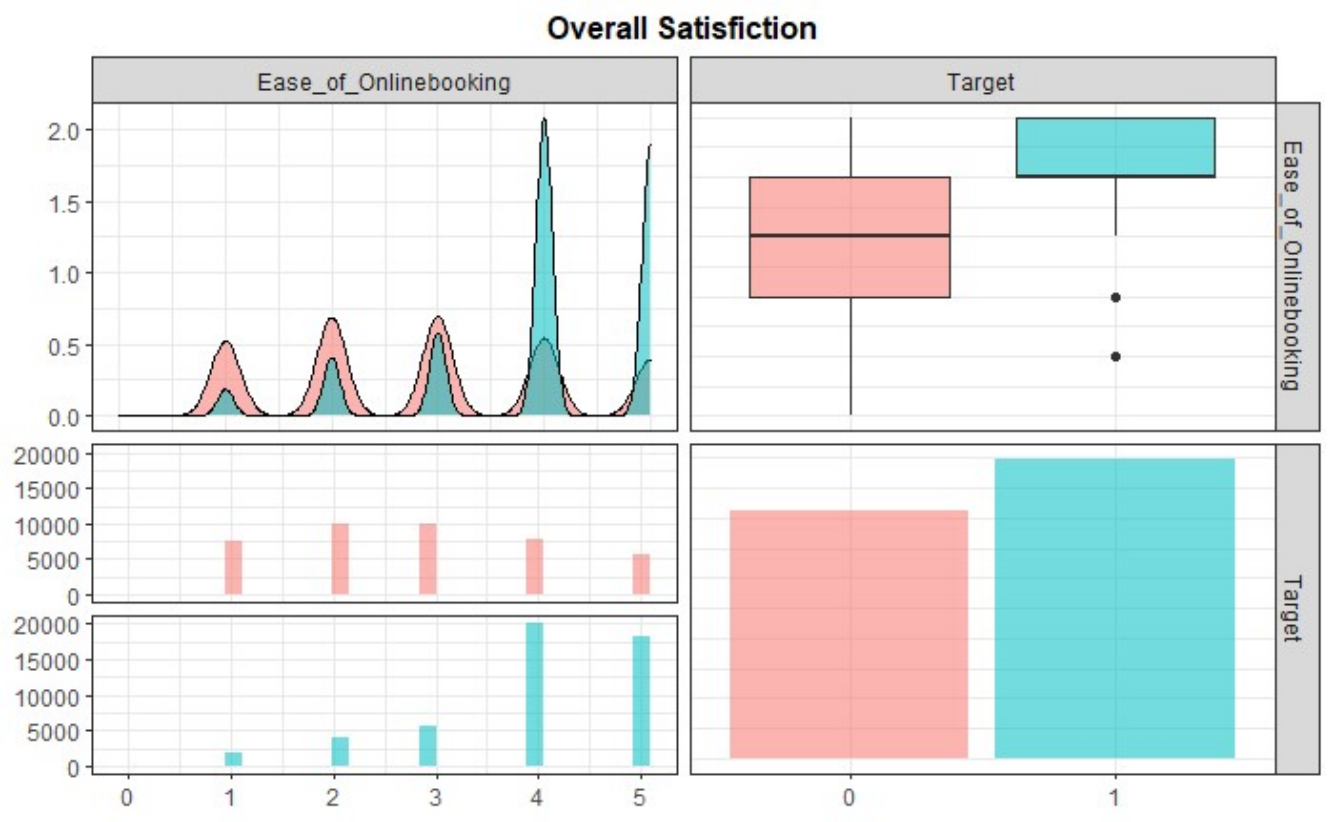
Exploratory Data Analysis: Online Services (1/4)

We have created number of dashboard to show the relation between variables tied to the Target variable overall Satisfaction...



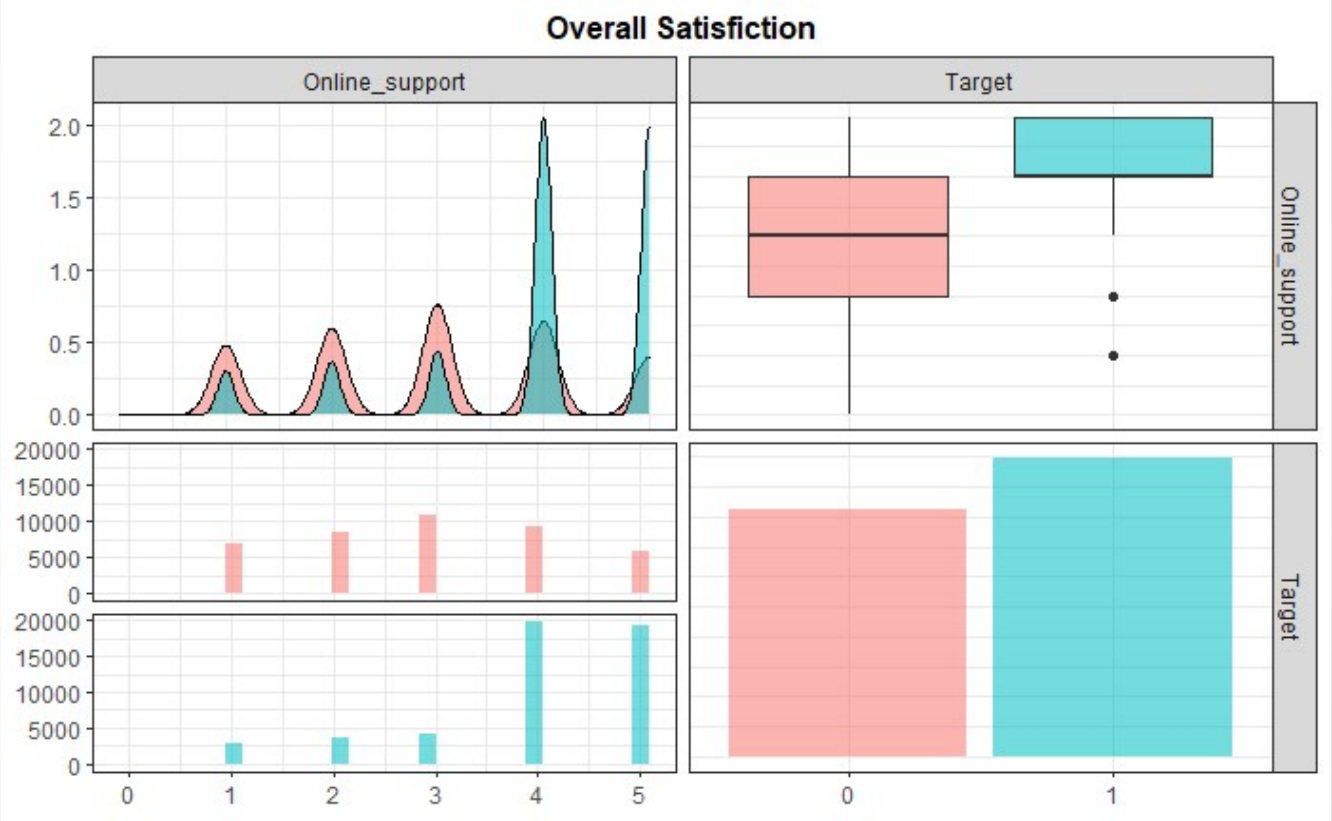
Exploratory Data Analysis: Online Services (2/4)

This Dashboard shows the relationship between Ease of online booking and overall Satfication.



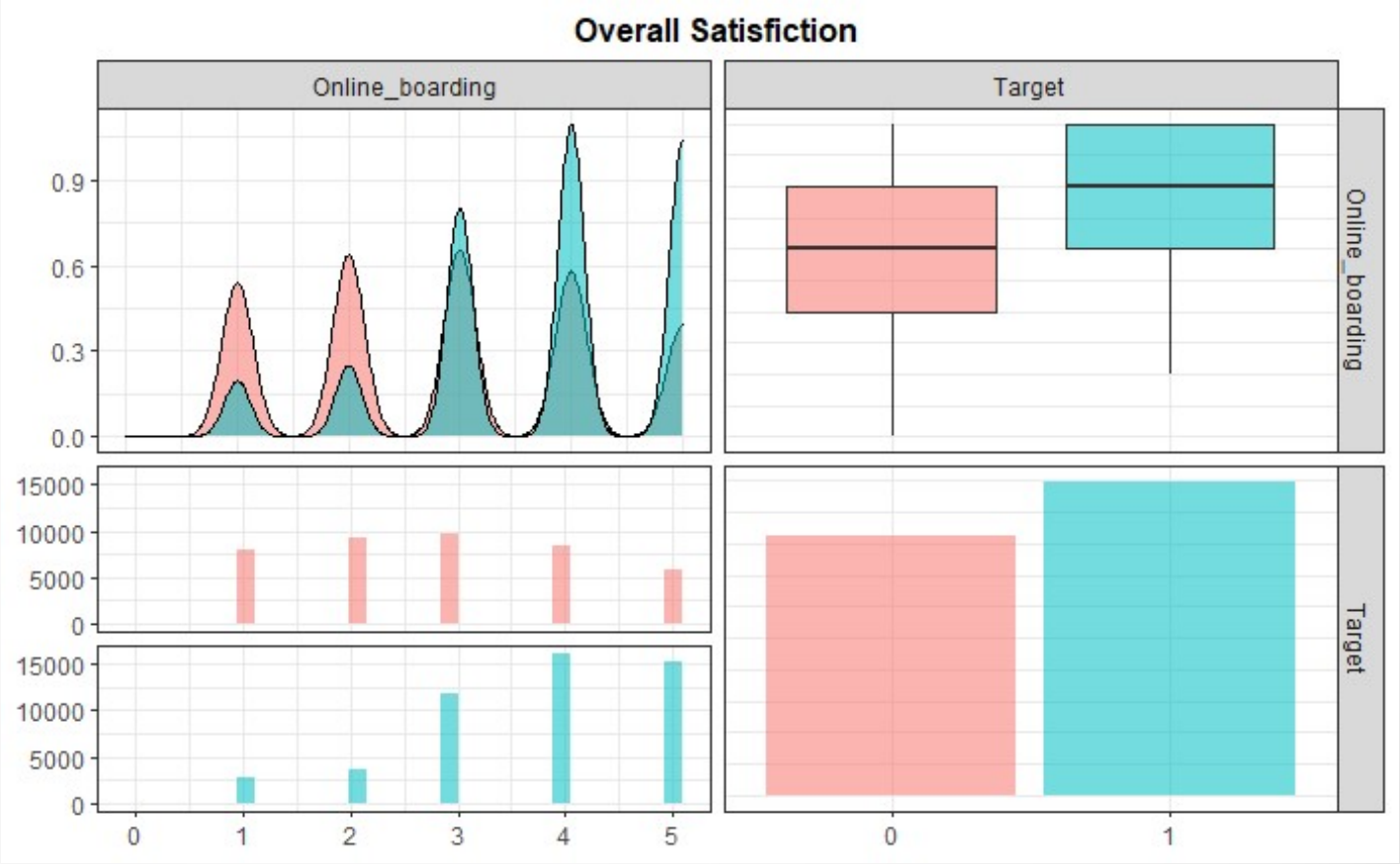
Exploratory Data Analysis: Online Services (3/4)

This Dashboard shows the relationship between online support and overall Satfication.



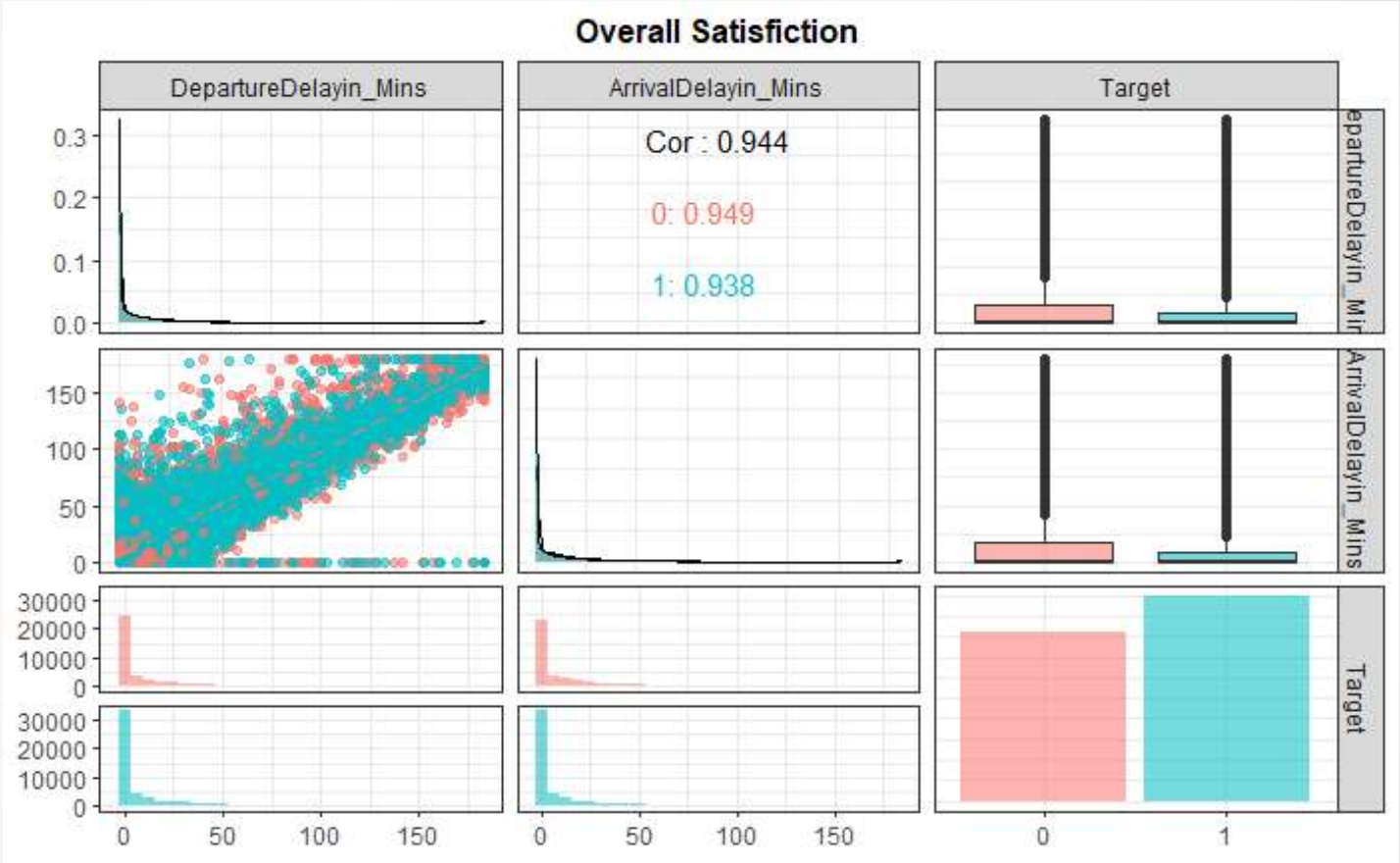
Exploratory Data Analysis: Online Services (4/4)

This Dashboard shows the relationship between Ease of online boarding and overall Satfication.



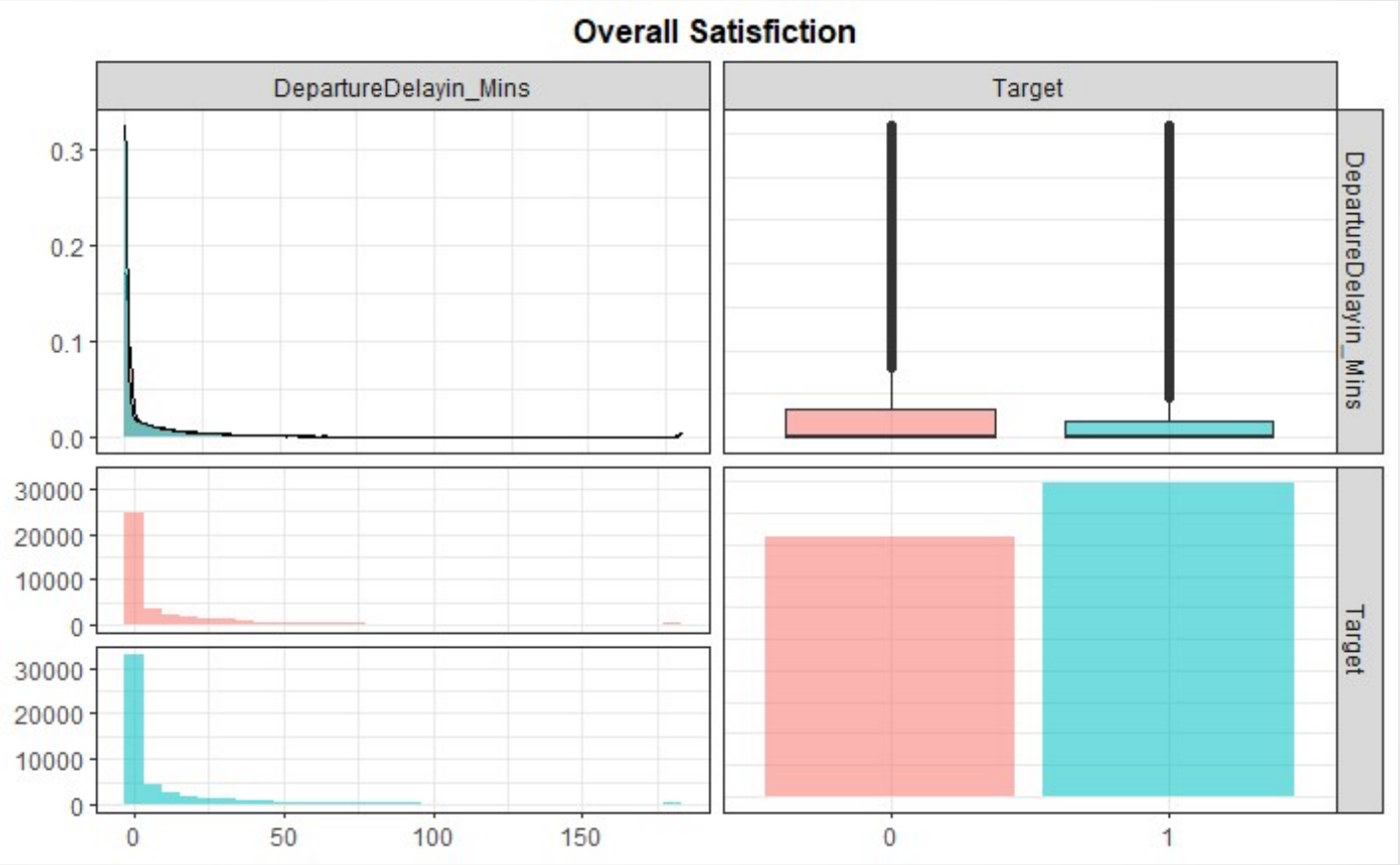
Exploratory Data Analysis: Delays (1/3)

We have created number of dashboard to show the relation between variables tied to the Target variable overall Satisfaction...



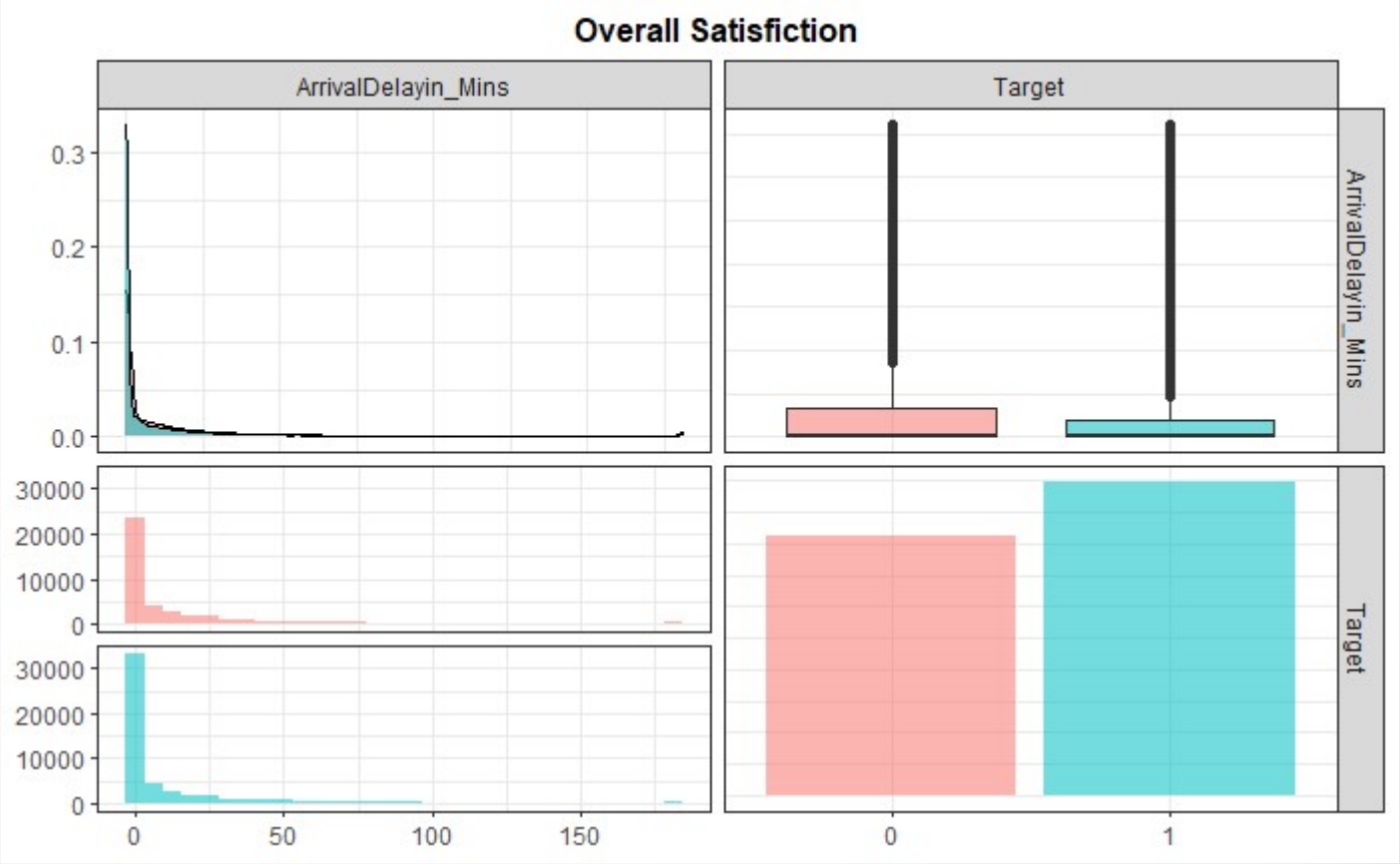
Exploratory Data Analysis: Delays (2/3)

This Dashboard shows the relationship between Ease of departure delays and overall Satfication.



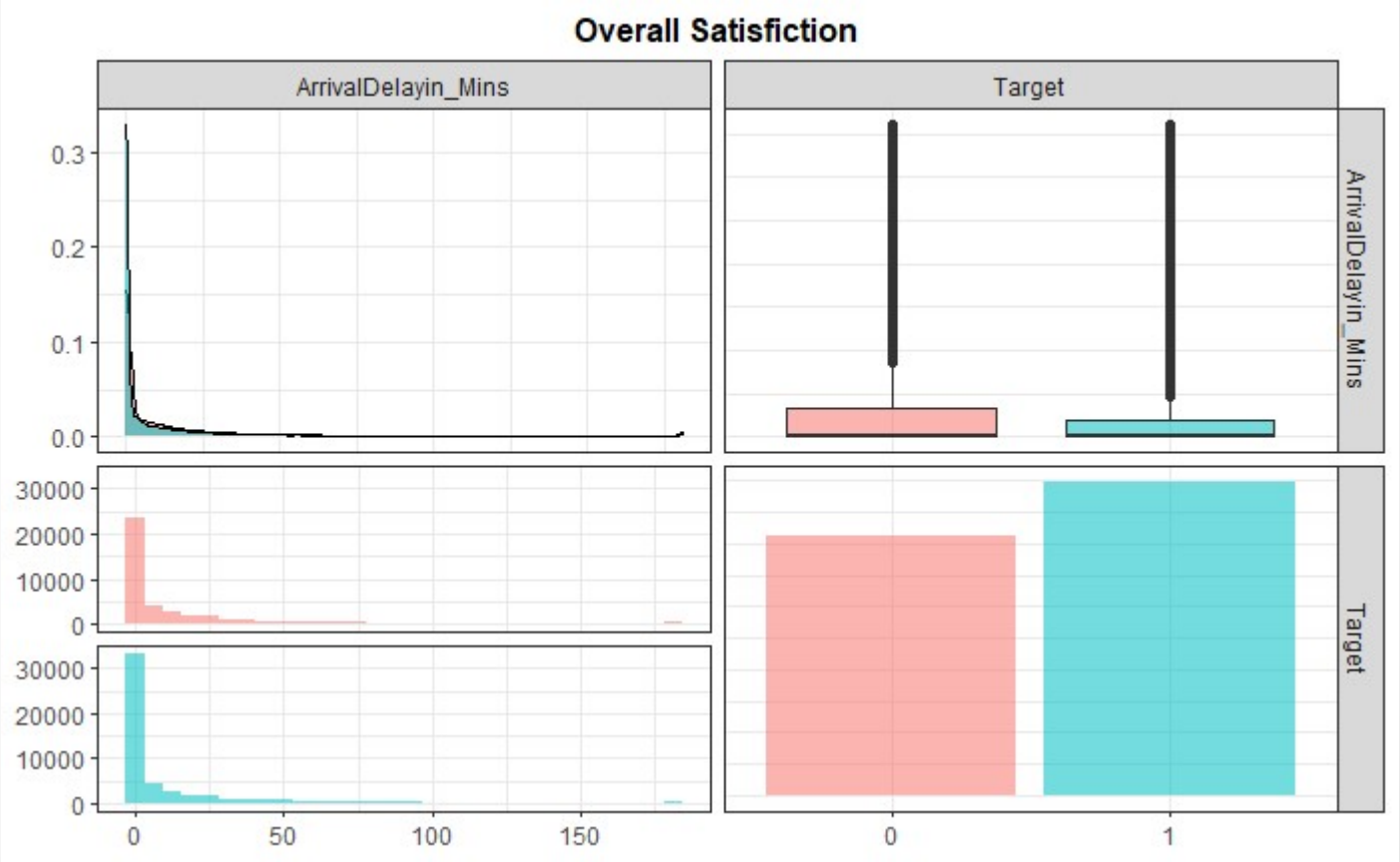
Exploratory Data Analysis: Delays (3/3)

This Dashboard shows the relationship between arrival delays and overall Satfication.



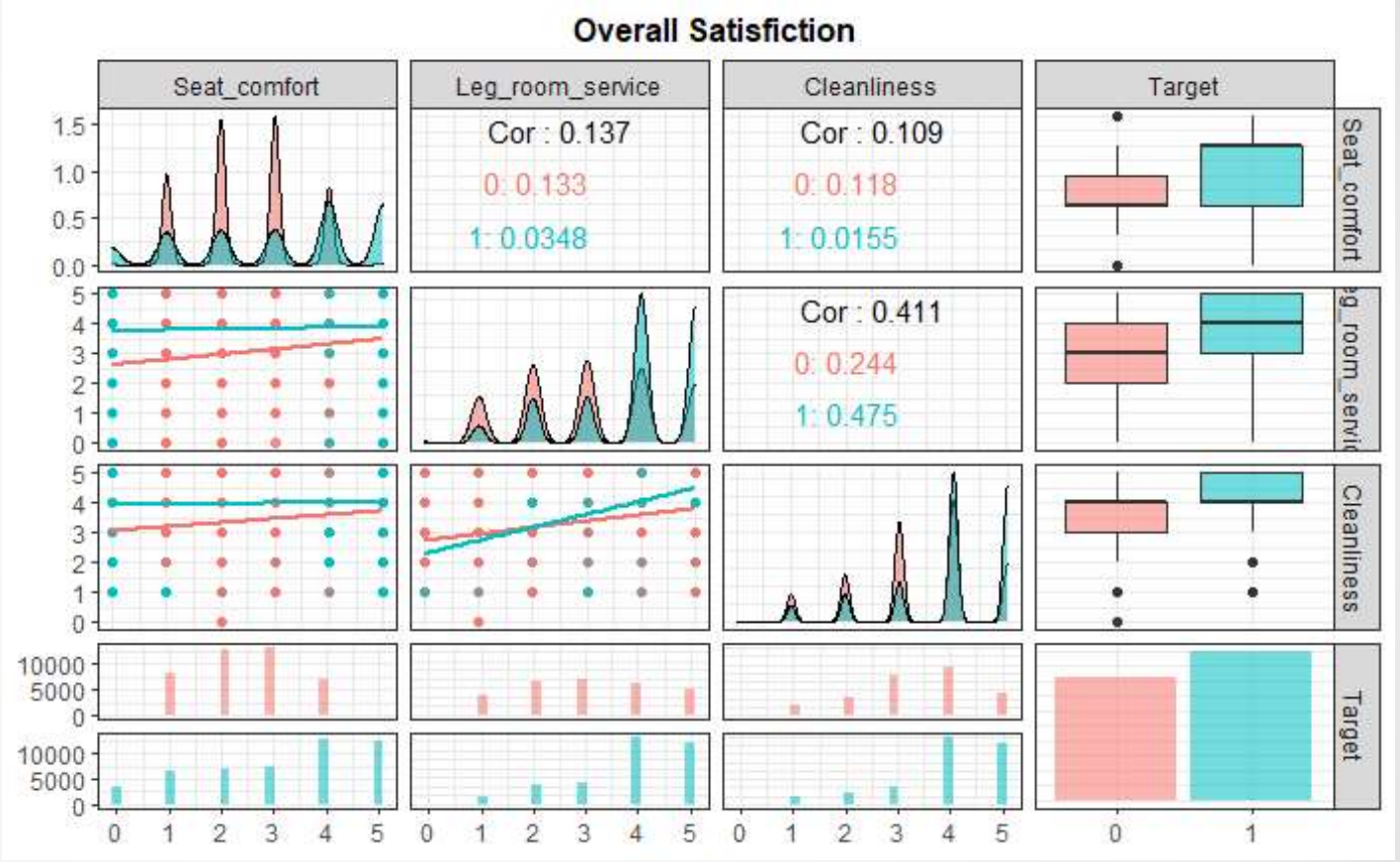
Exploratory Data Analysis: Plans (1/4)

This Dashboard shows the relationship between planes as a facility and overall Satfication.



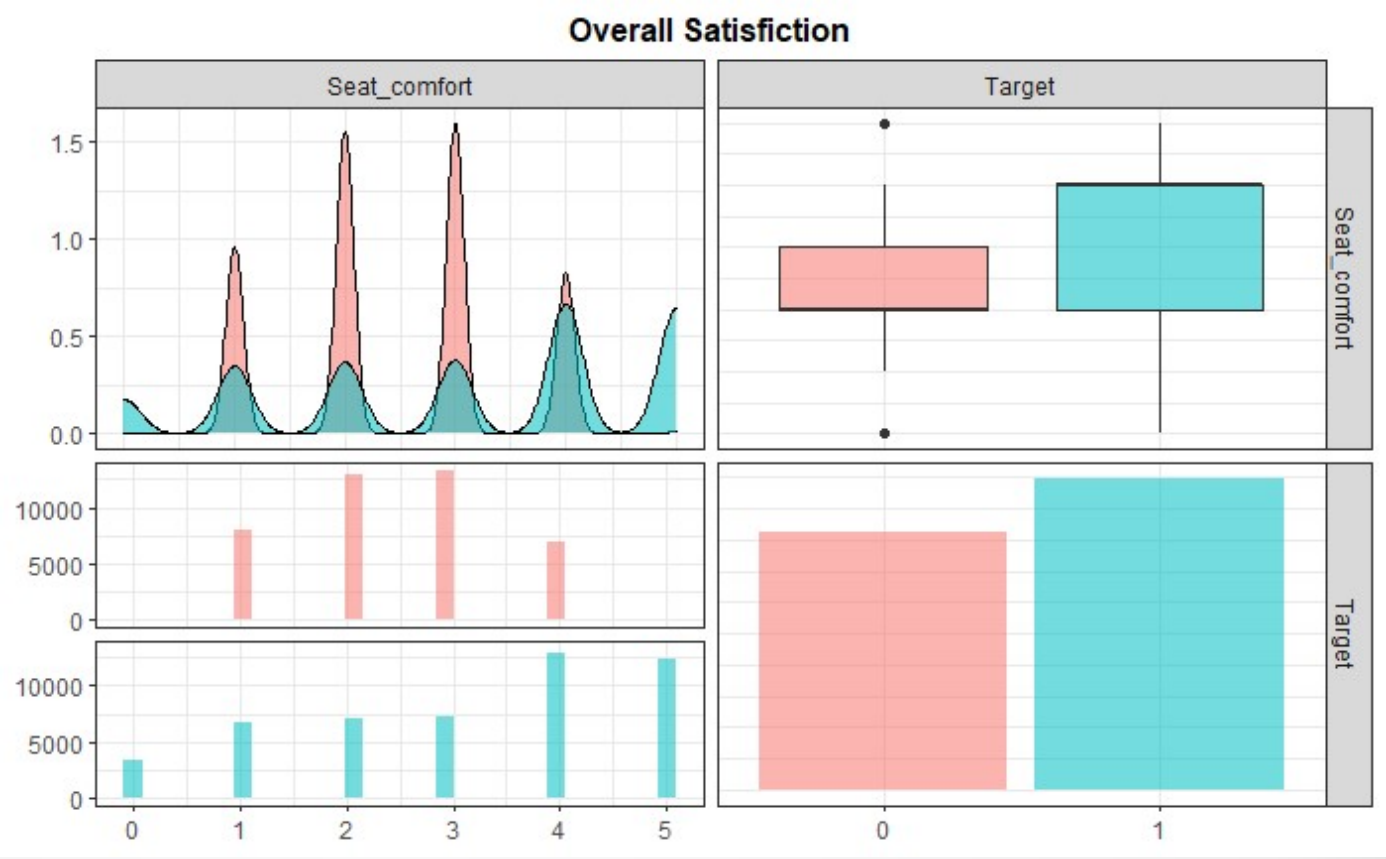
Exploratory Data Analysis: Plans (1/4)

This Dashboard shows the relationship between planes as a facility and overall Satfication.



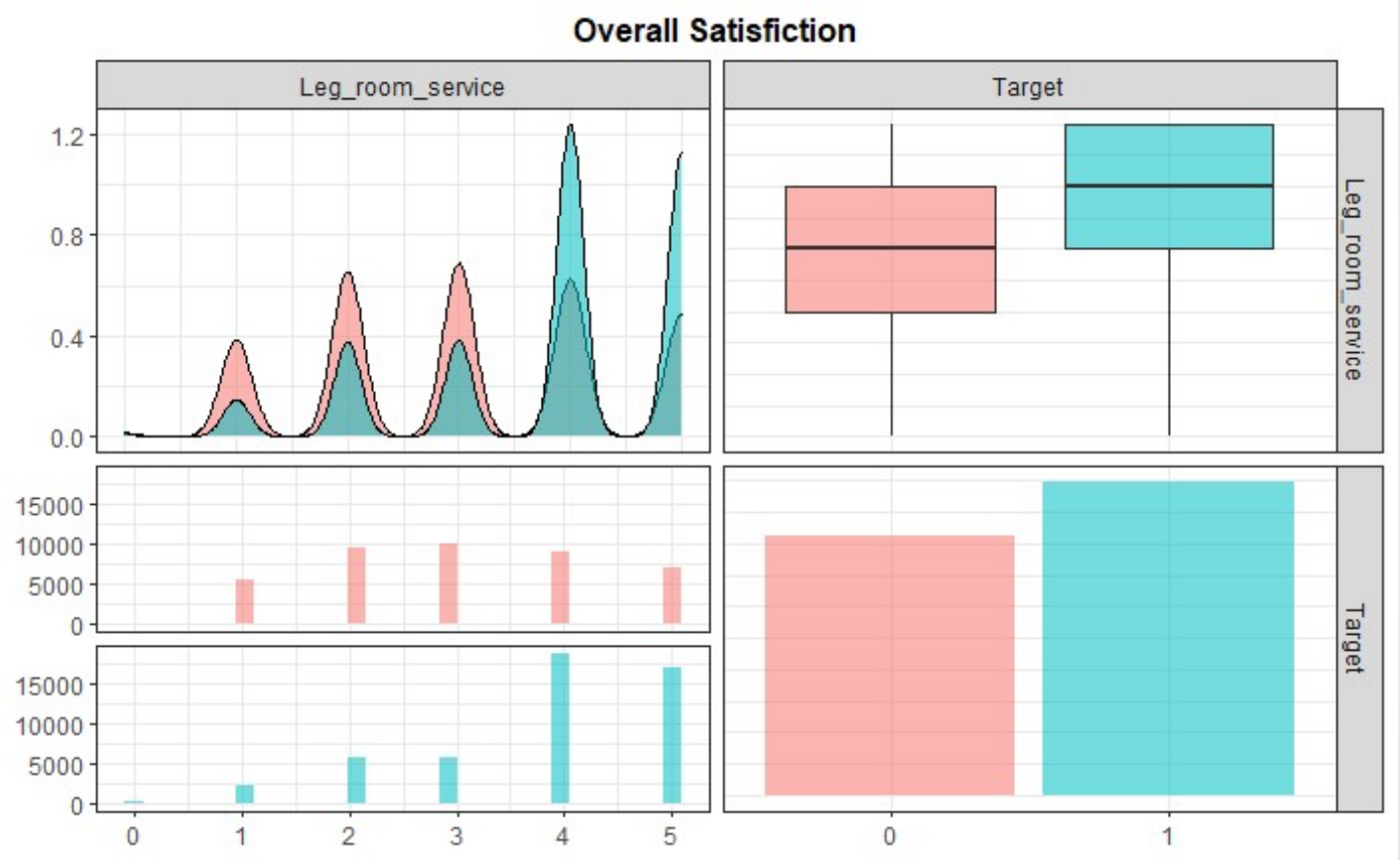
Exploratory Data Analysis: Plans (2/4)

This Dashboard shows the relationship between seat comfort and overall Satfication.



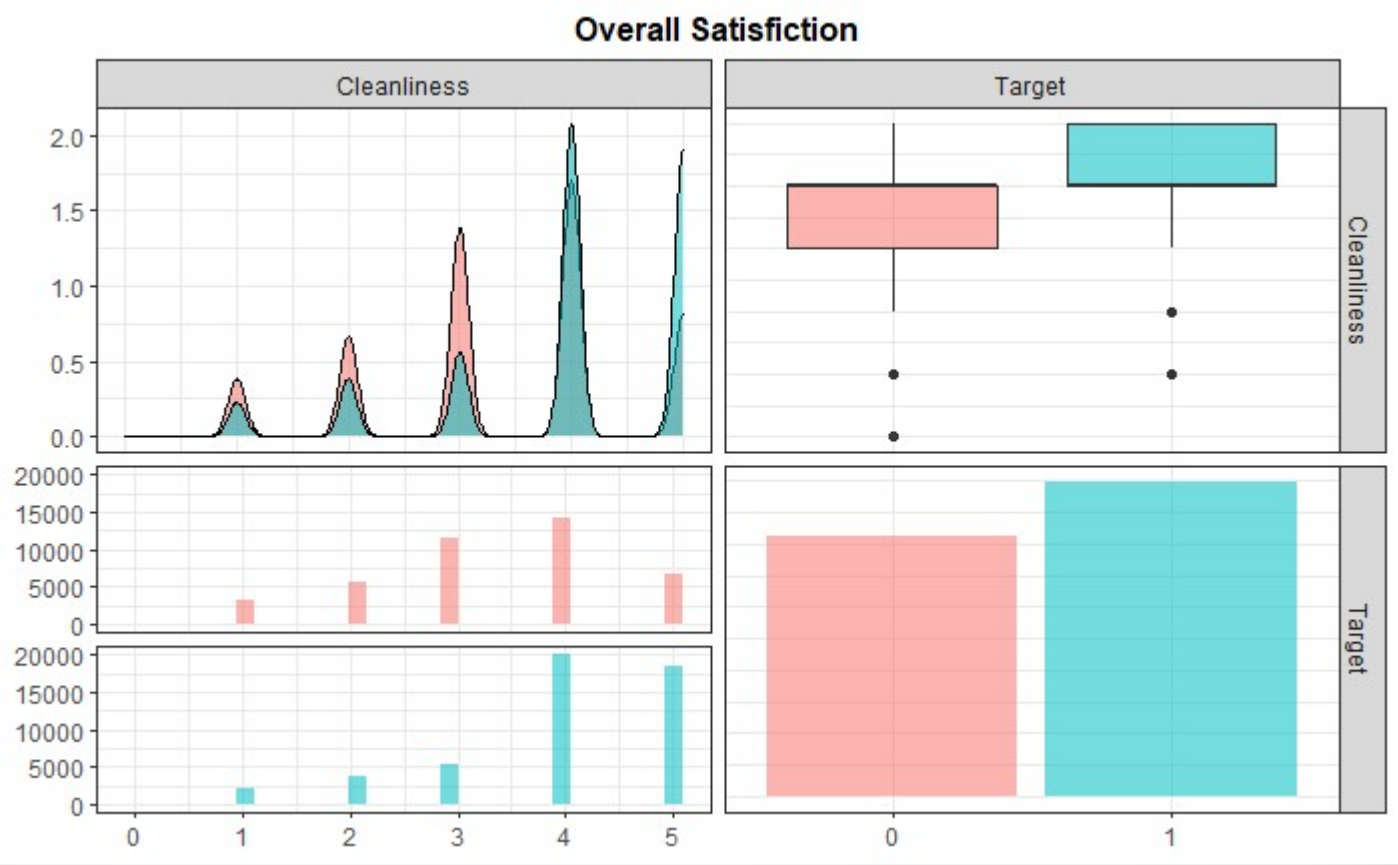
Exploratory Data Analysis: Plans (3/4)

This Dashboard shows the relationship between seats leg room and overall Satfication.



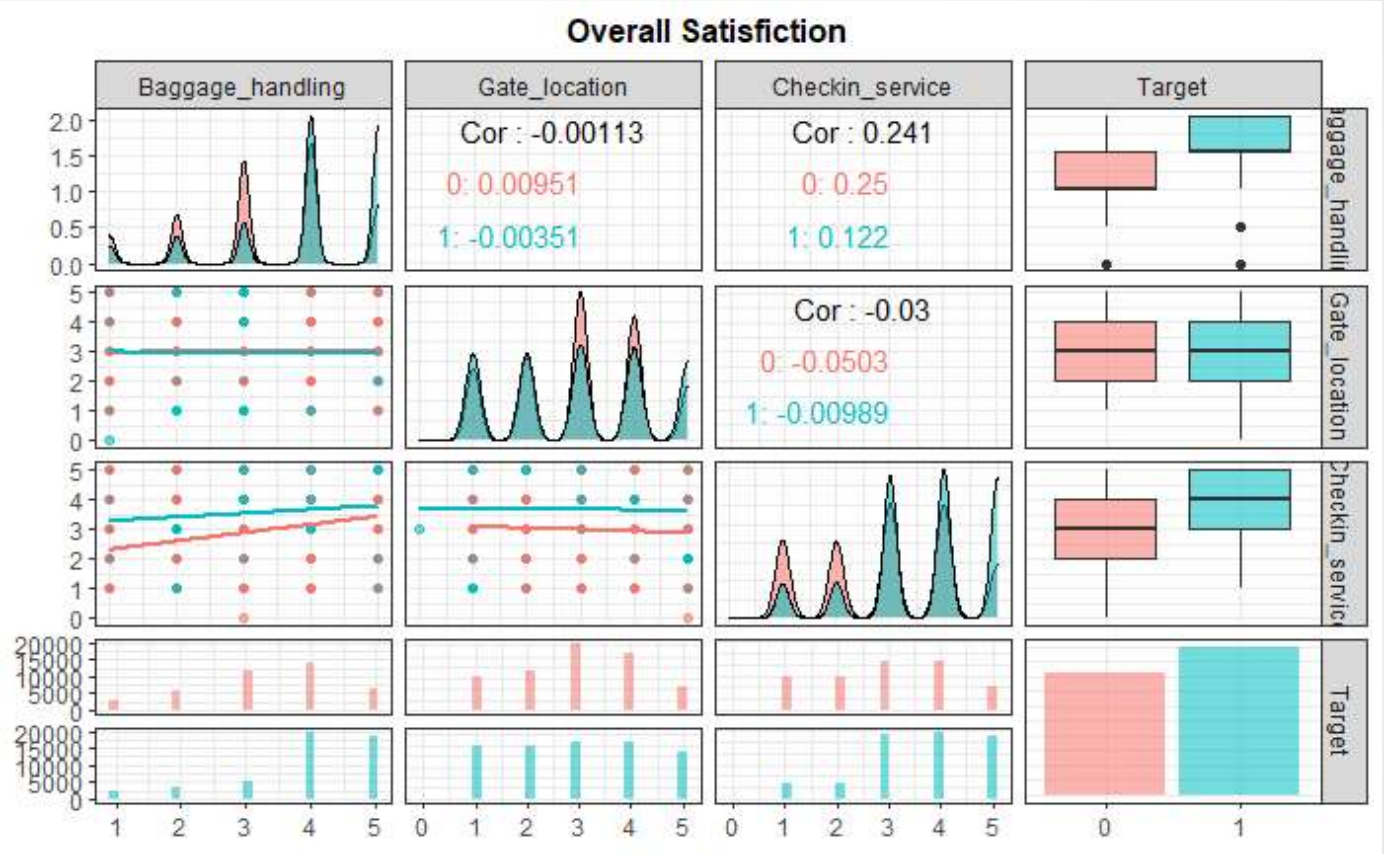
Exploratory Data Analysis: Plans (4/4)

This Dashboard shows the relationship between how clean my planes and overall Satfication.



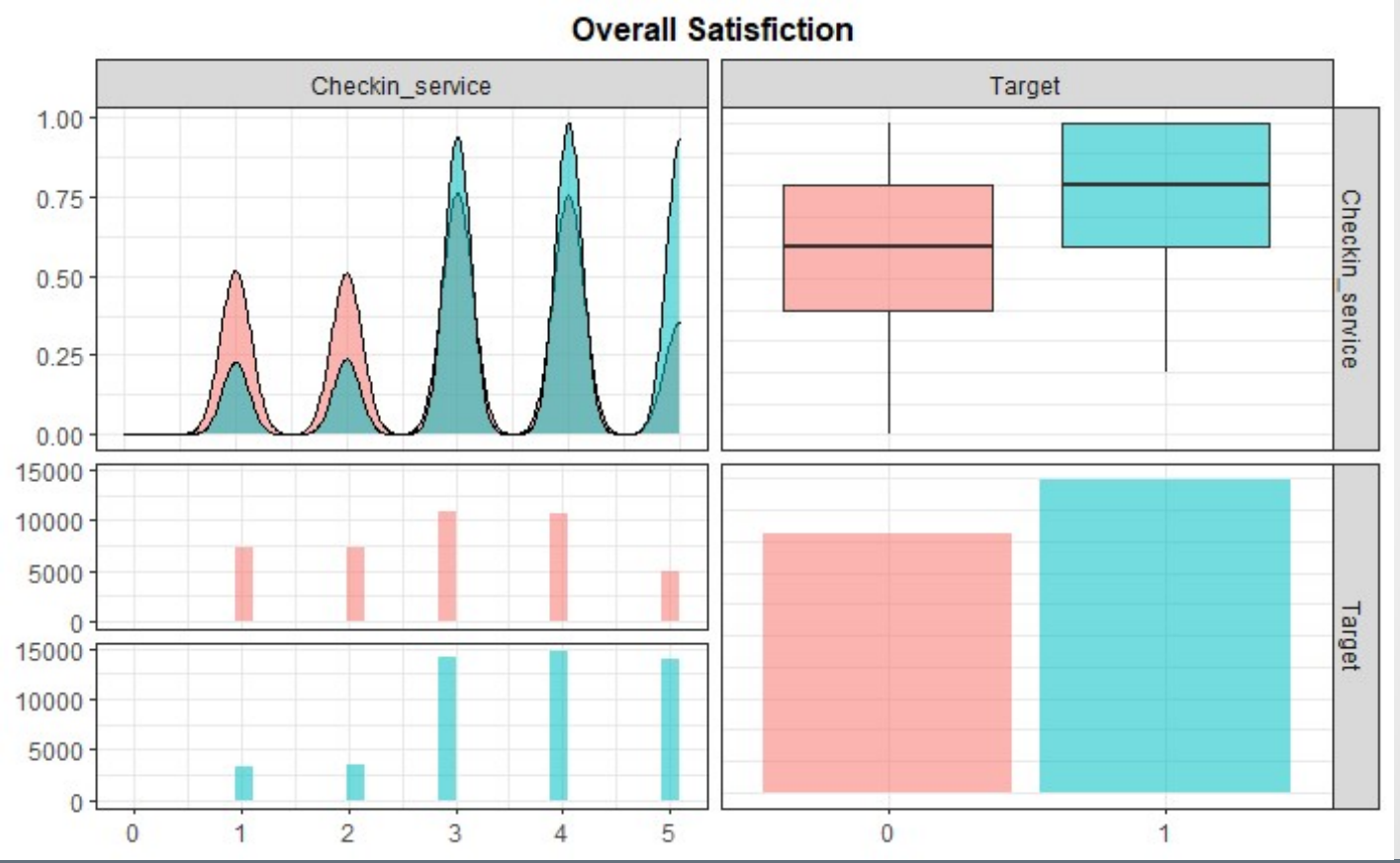
Exploratory Data Analysis: Before and After (1/4)

This Dashboard shows the relationship between my before & after flight services and overall Satfication.



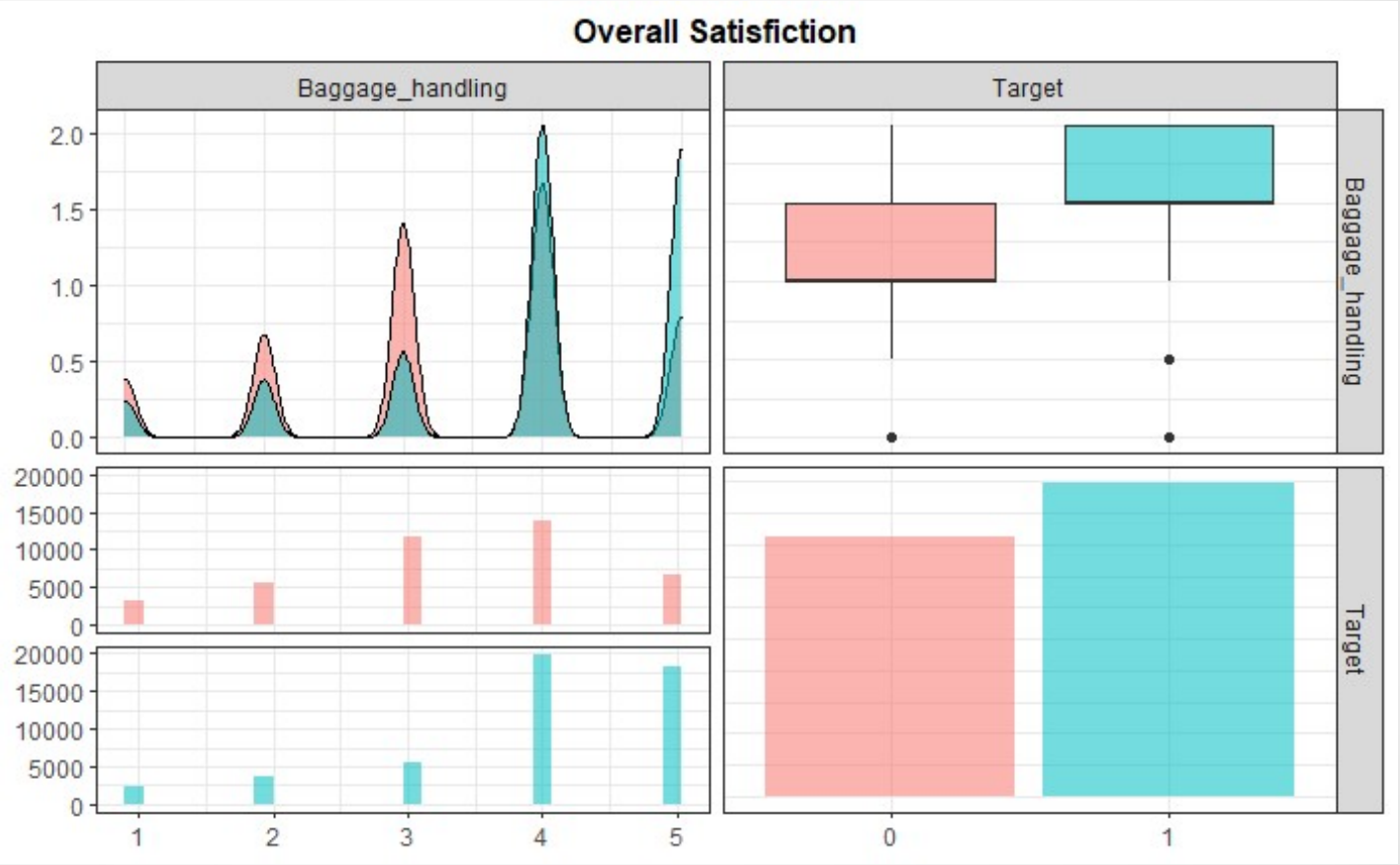
Exploratory Data Analysis: Before and After (2/4)

This Dashboard shows the relationship between check-in services and overall Satfication.



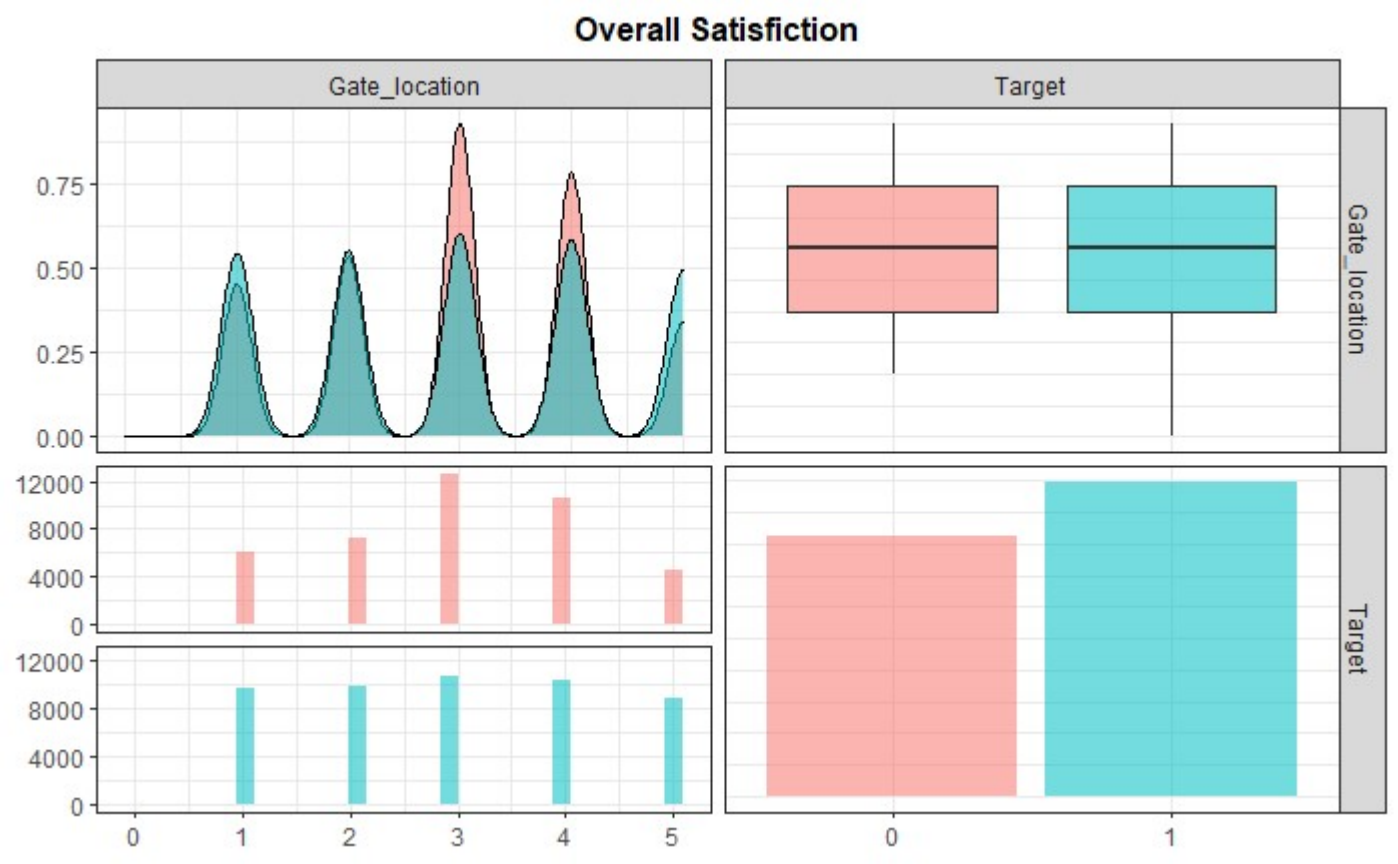
Exploratory Data Analysis: Before and After (3/4)

This Dashboard shows the relationship between baggage handling and overall Satfication.



Exploratory Data Analysis: Before and After (4/4)

This Dashboard shows the relationship between gate location and overall Satfication.

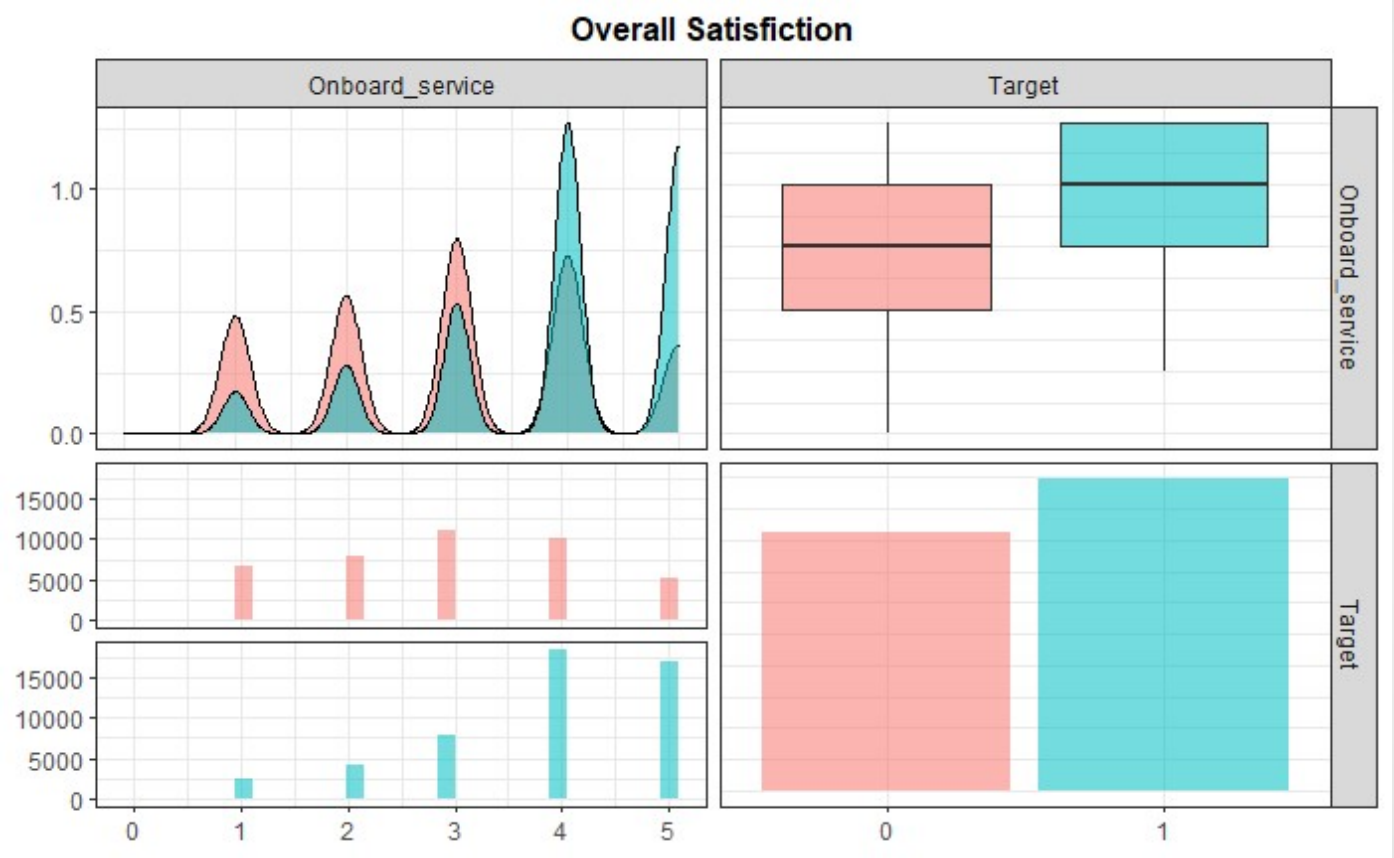


Exploratory Data Analysis: Onboard Exp (1/5)

This Dashboard shows the relationship between Onboard services and overall Satfication.

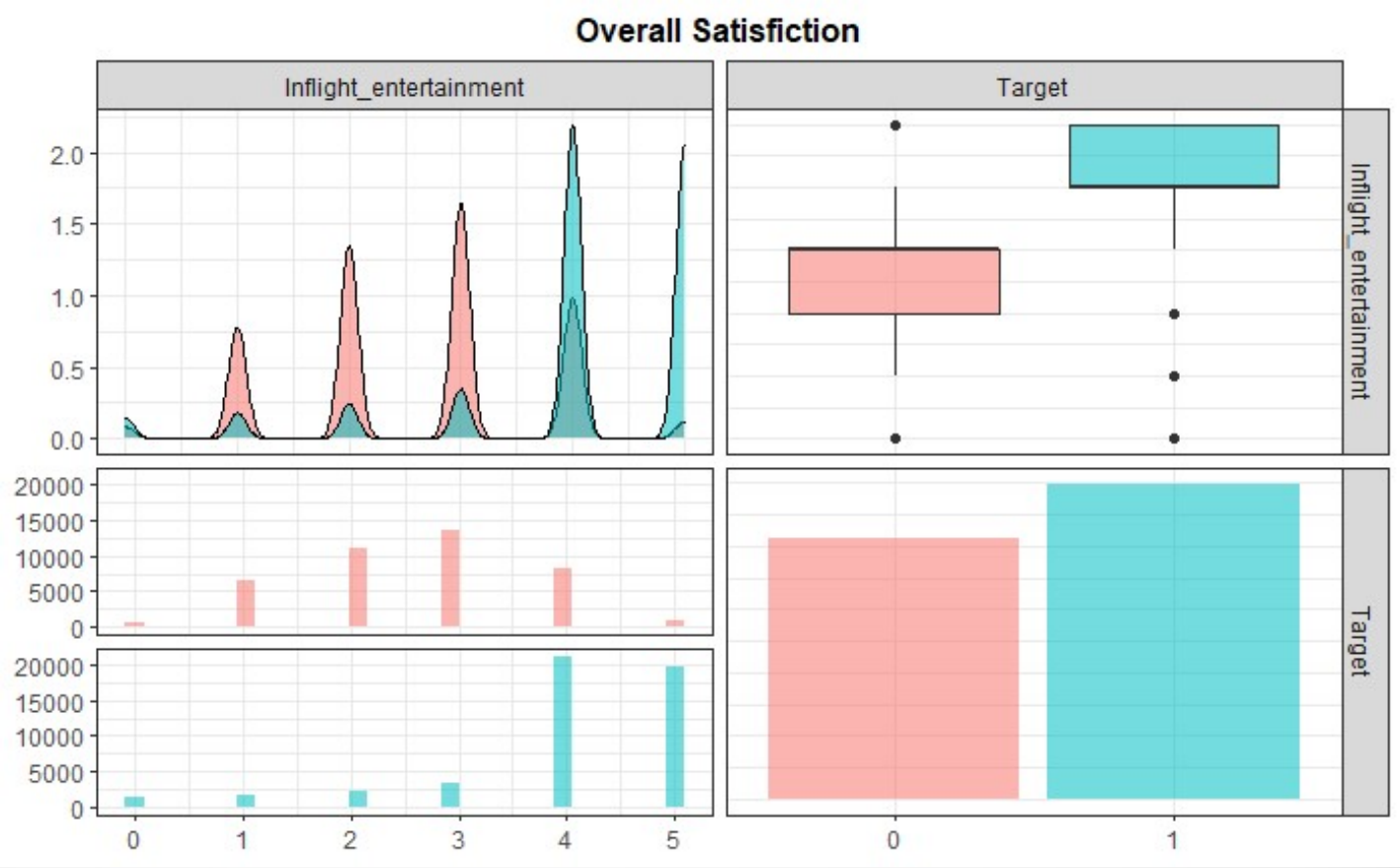
Exploratory Data Analysis: Onboard Exp (2/5)

This Dashboard shows the relationship between Onboard Services and overall Satfication.



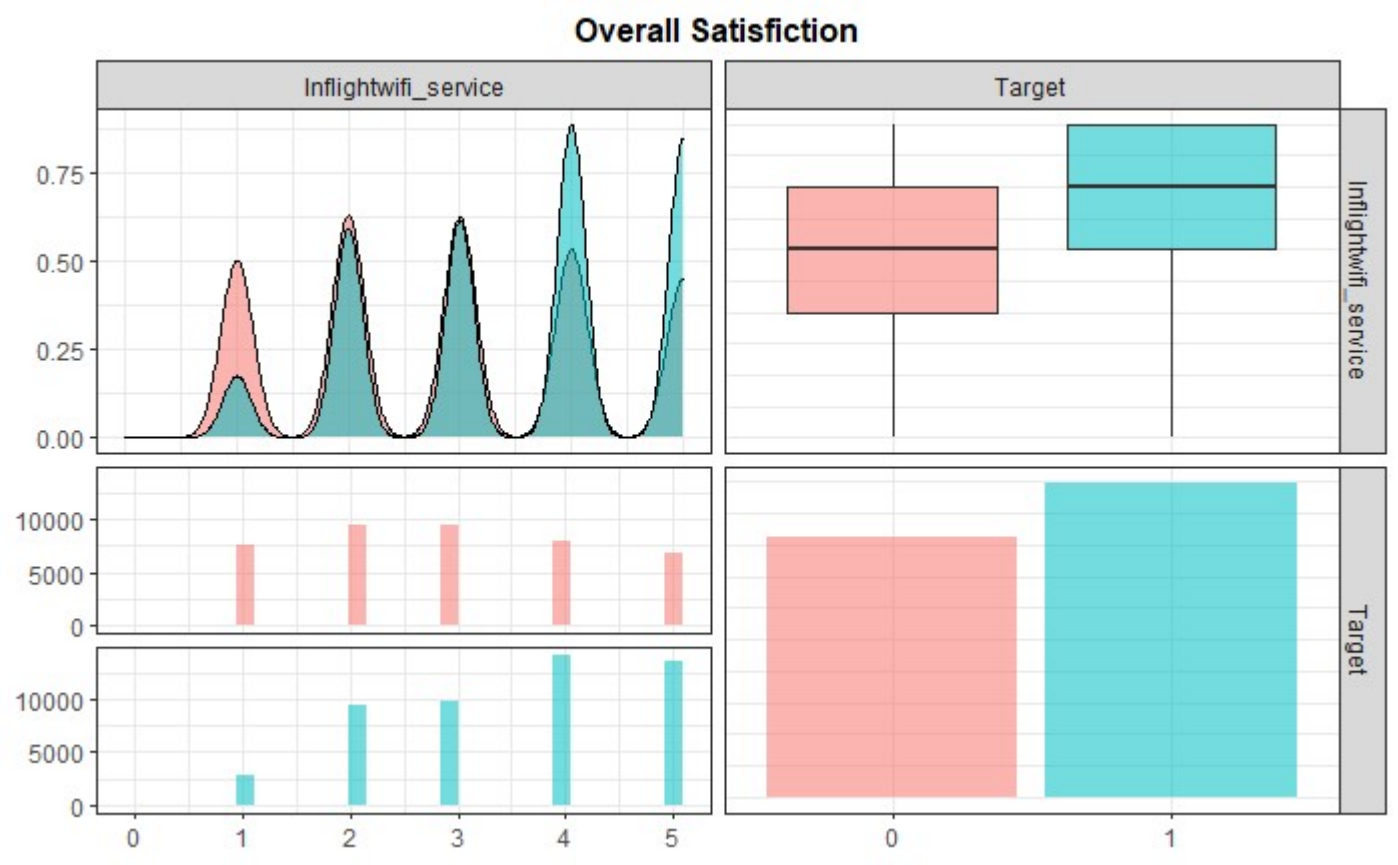
Exploratory Data Analysis: Onboard Exp (3/5)

This Dashboard shows the relationship between onboard entertainment and overall Satfication.



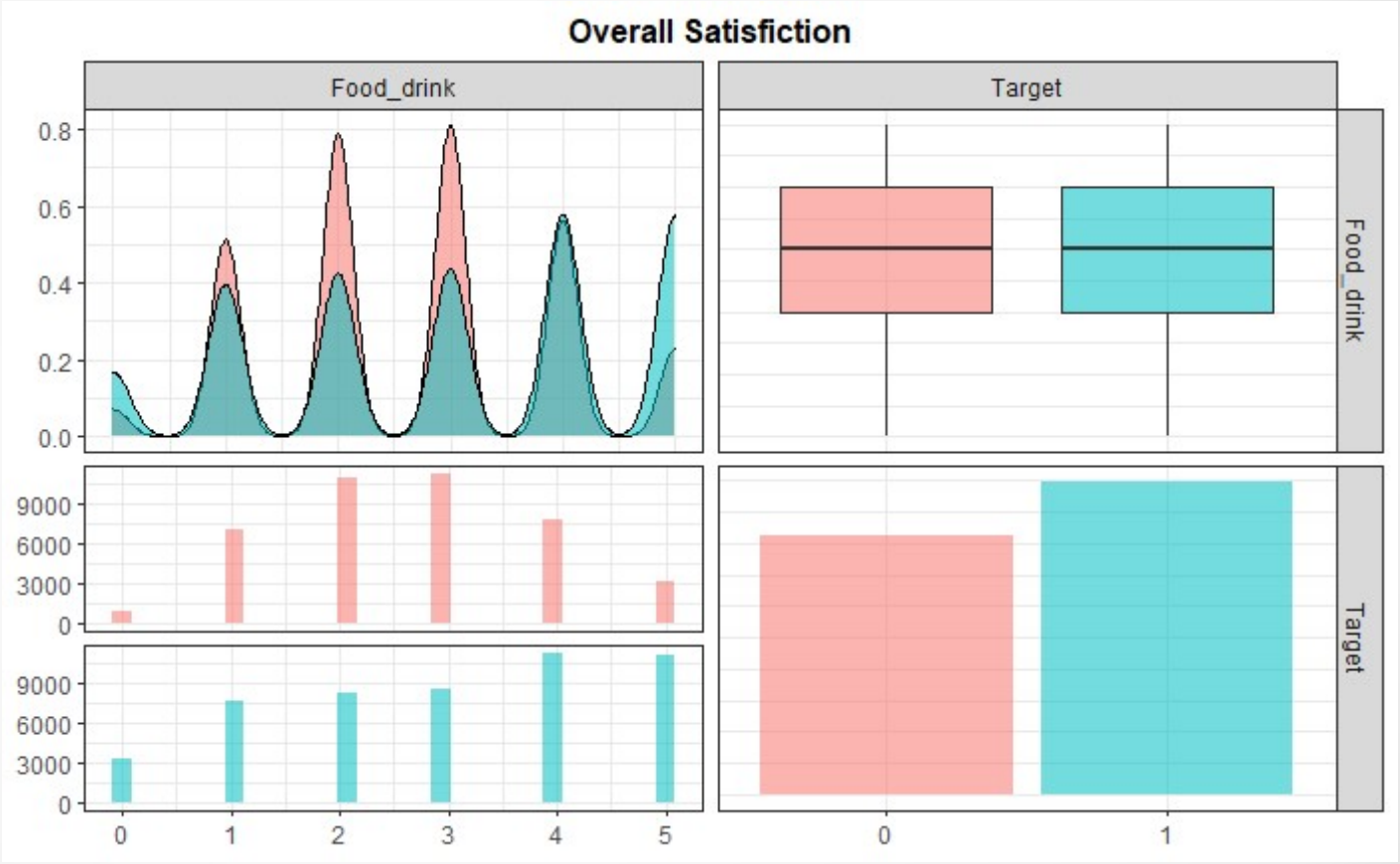
Exploratory Data Analysis: Onboard Exp (4/5)

This Dashboard shows the relationship between onboard WIFI and overall Satfication.



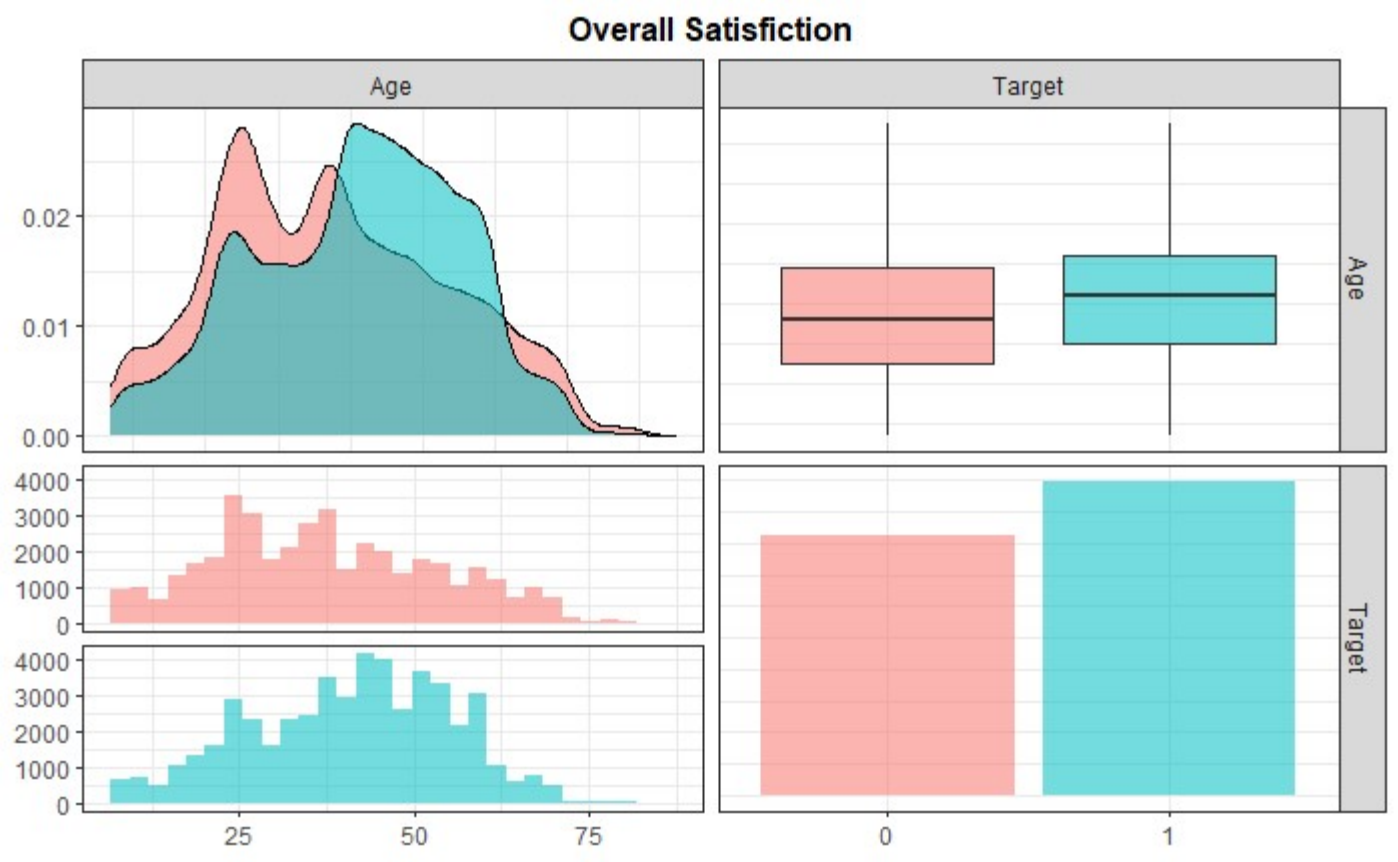
Exploratory Data Analysis: Onboard Exp (5/5)

This Dashboard shows the relationship between food & Beverage on board and overall Satfication.



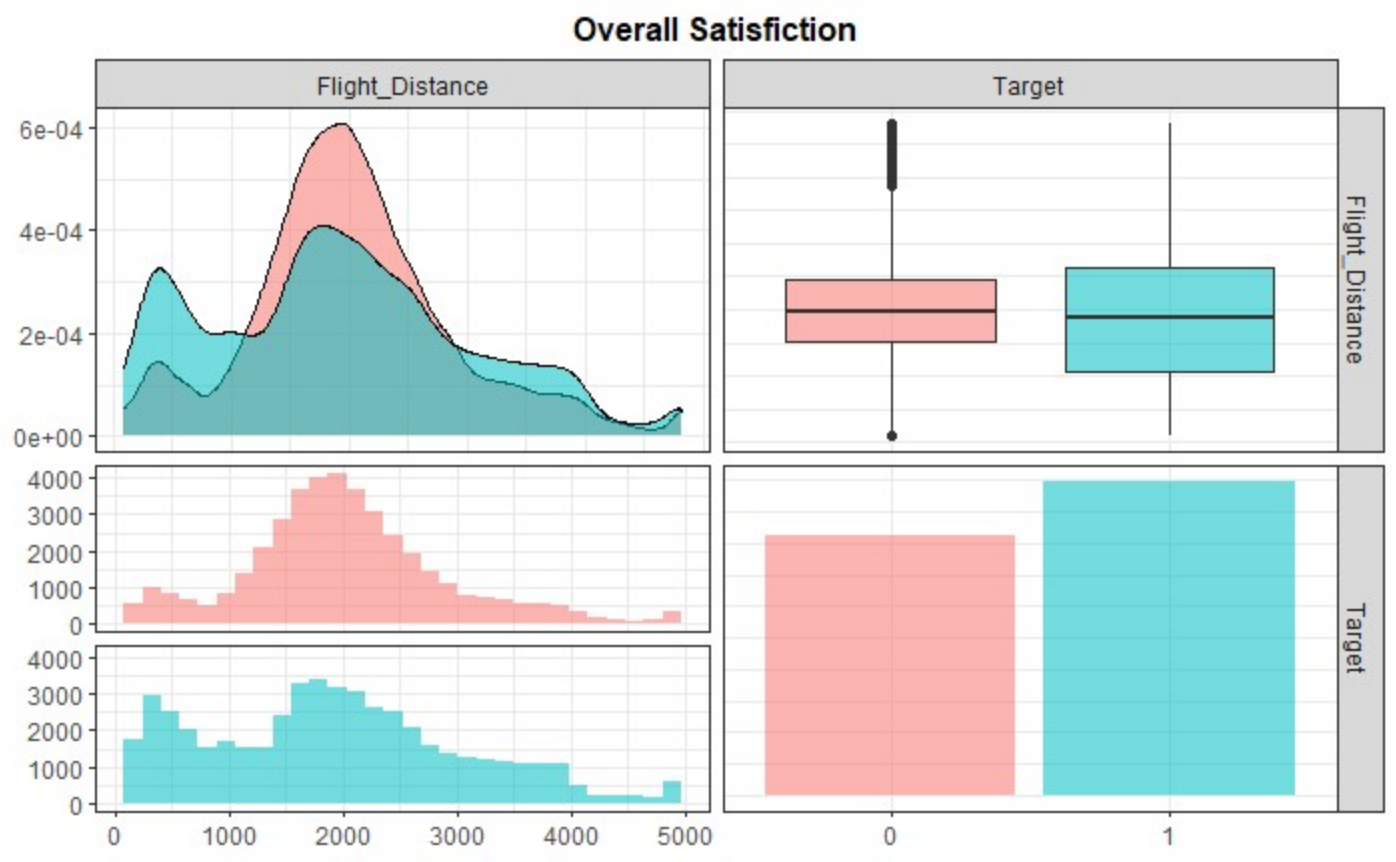
Exploratory Data Analysis: Age Profile

This Dashboard shows the relationship between age profile and overall Satfication.



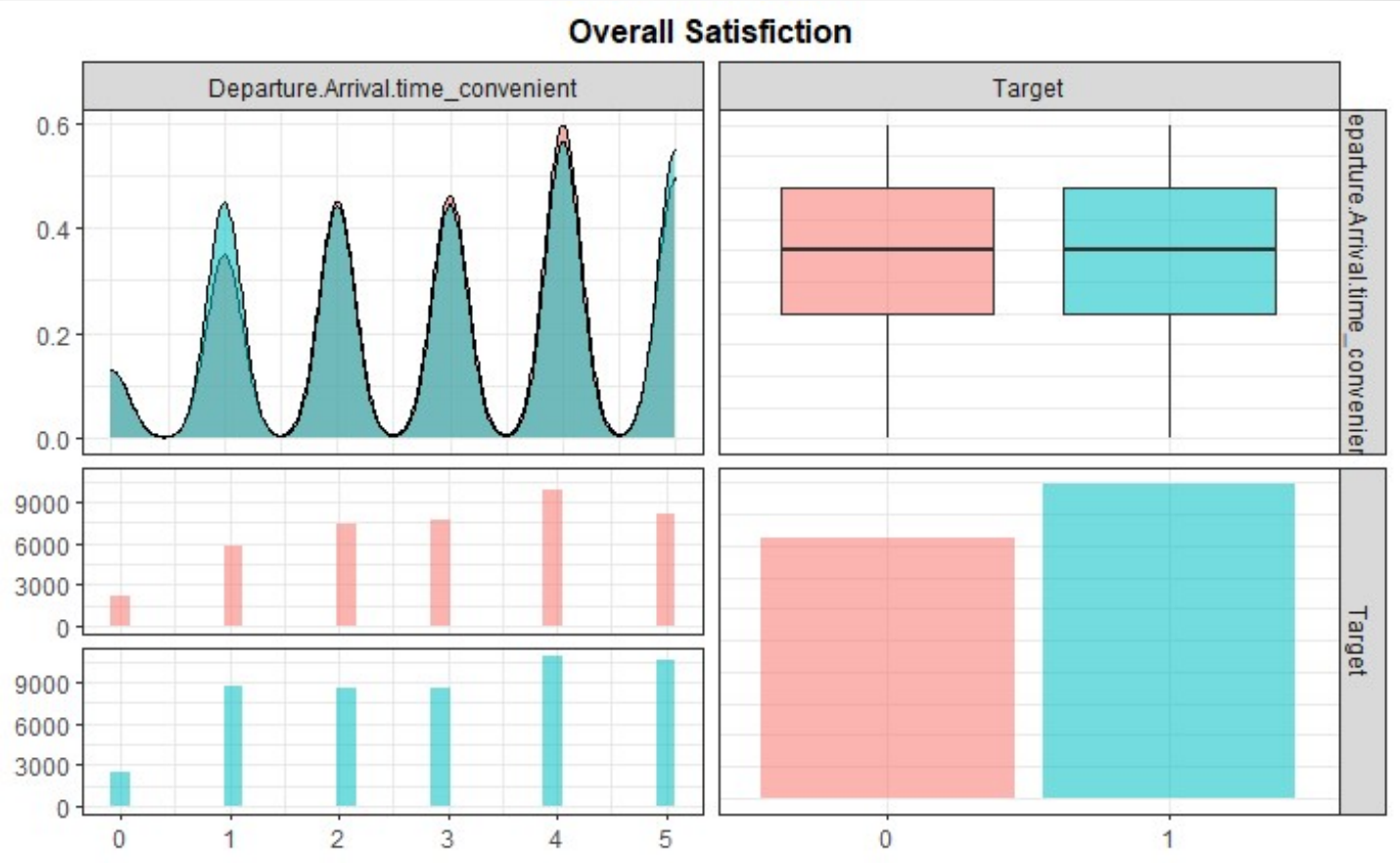
Exploratory Data Analysis: Age Profile

This Dashboard shows the relationship between flight distance and overall Satfication.



Exploratory Data Analysis: Flight Timing

This Dashboard shows the relationship between Flight Timing and overall Satfication.



Analytical Approach

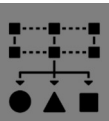
The analytical approach explain the process planned to build our model, to draw most insights to be able to answer the problem statement "what are most important factors affecting Passengers satisfaction?".

1.Data Preprocessing *is a very important process in the model development, were we attempt to fix data issues to ensure building a more robust models.*

Cleaned Data>>

2.PCA will allow us to easily analyze large data sets, by combing highly correlated variables, under a principal components. Simply, think of it as combing similar topics under on subject domain. This will allow us to understand are the key areas that affects the passengers SatisfactionS. In addition we will reduce our data size, allowing for a better performance for model.

Comprised Train Data>>



3.Classifier, is a type of supervised machine learning, is commonly used when we have historical data sets has Grounded Truth or Categorical Dependent variable. In this process we will try number of model such; Random Forest, KNN, Logistic regression and others.

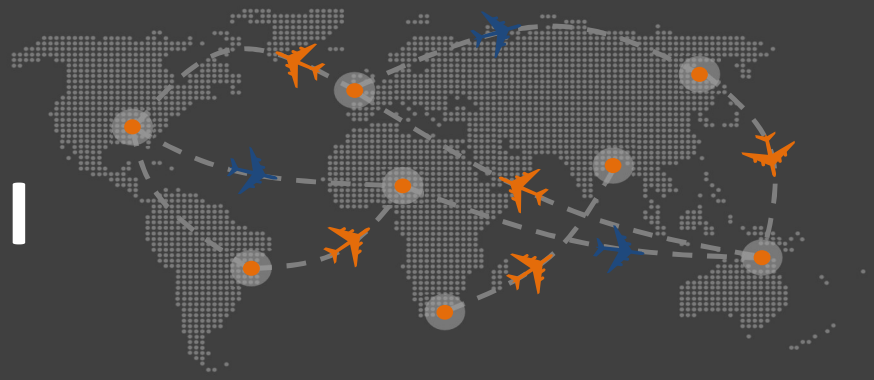
4.Prefromance of each model will be evaluated to see the model that yields the most robust results. Some the performance metrics are Accuracy, Precision, AUC, Gini index and others. The model the preforms well across the board is picked.

Predicted Test Data

FALCON AIRLINES
NONSTOP YOU

Section 6: Modelling*

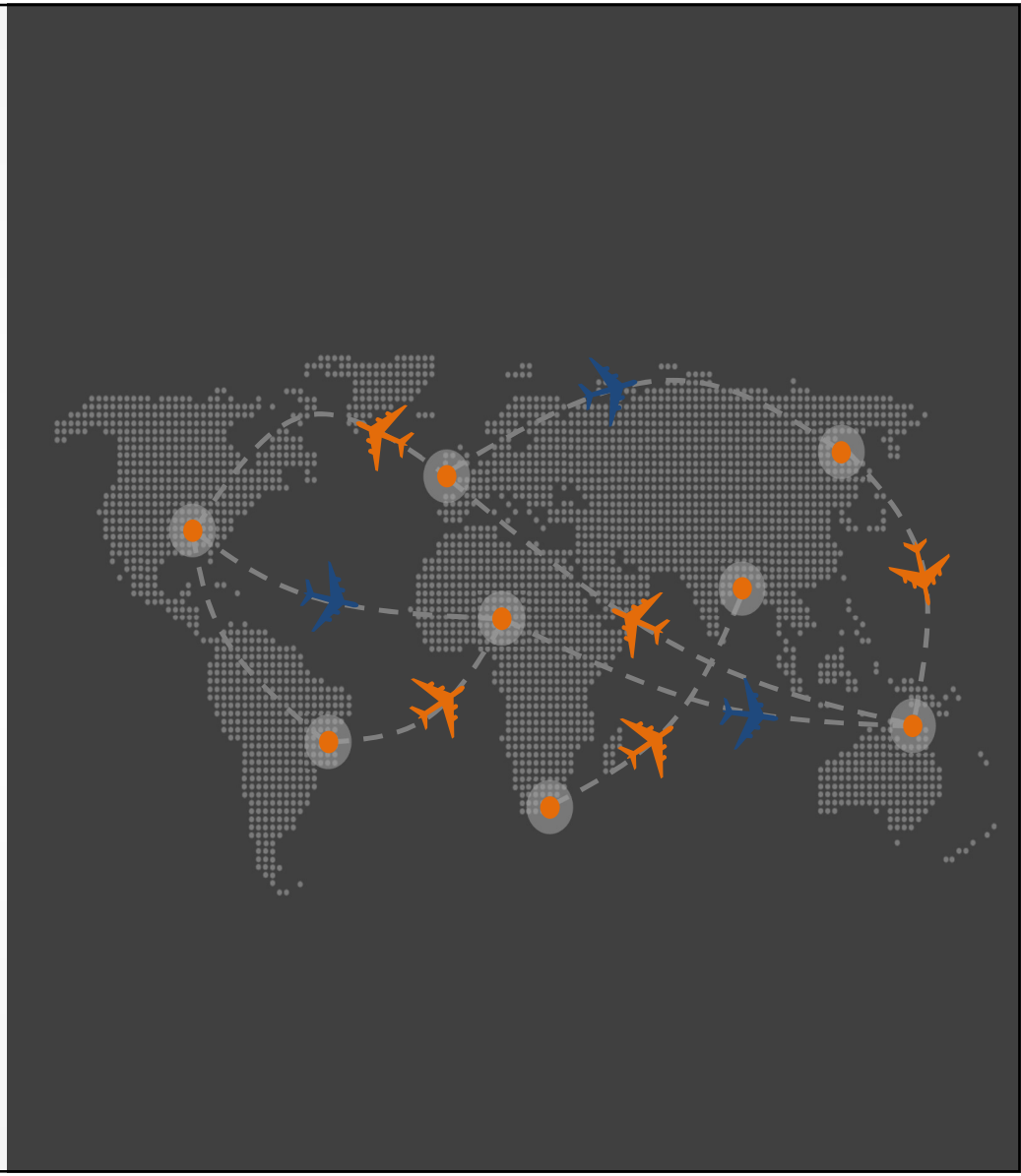
Project Submission Notes III



(*) Including Model Evaluation and Selection Criteria.

FALCON AIRLINES
NONSTOP YOU

Table of Contents



Our Analytical Approach

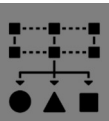
The analytical approach explain the process planned to build our model, to draw most insights to be able to answer the problem statement "what are most important factors affecting Passengers satisfaction?".

1.Data Preprocessing *is a very important process in the model development, were we attempt to fix data issues to ensure building a more robust models.*

Cleaned Data>>

2.PCA will allow us to easily analyze large data sets, by combing highly correlated variables, under a principal components. Simply, think of it as combing similar topics under on subject domain. This will allow us to understand are the key areas that affects the passengers Satisfactions'. In addition we will reduce our data size, allowing for a better performance for model.

Comprised Train Data>>



3.Classifier, is a type of supervised machine learning, is commonly used when we have historical data sets has Grounded Truth or Categorical Dependent variable. In this process we will try number of model such; Random Forest, KNN, Logistic regression and others.

4.Prefromance of each model will be evaluated to see the model that yields the most robust results. Some the performance metrics are Accuracy, Precision, AUC, Gini index and others. The model the preforms well across the board is picked.

Predicted Test Data

Models Overview

The table below gives a bird's eye view of the models used in our exercise and a brief comparison of their pros and cons.

Models	Logistic Reg.	Naïve Bayes	KNN	Random Forest	XGB
Overview	*The logistic regression is a special case of linear regression that is used to predict categorical variables. It simply class Target based by modelling probabilities, if higher than a certain cut-off its 1 and lower would be 0.*	*Naïve Bayes "NB" is a classification model that is uses Bayes Theorem. It mainly assumes each feature is independent of each other and uses conditional probabilities between the Target and the feature.*	*K nearest neighbor "KNN" is a classification and a regression model, that groups instances based on proximity. Proximity is measured by a measure of distance, most used is Euclidean distance.*	*Random Forest "RF" is a type of ensemble modelling that uses bootstrap. Which build number of independent trees using random subsets of data, then uses mode of results to classify the Target. It provides better predictive ability compared to bagging mainly random features selection .*	*Extreme Gradient Boosting "XGB" is another ensemble modelling, but unlike RF, it is sequential leaner. So, deploy models in order and they leaner from each others mistakes, through penalization.*
Pros	• *Easy and Efficient* • *Low computational resources* • *Easily tuned* • *Less prune to overfitting*	• *Simplification assumptions* • *Can predict continuous data* • *Can run without Xs'.* • *Easy and fast*	• *Simple and easy* • *No Training needed* • *Works for both classification & regression*	• *Works for both classification & regression* • *Strong predictive ability* • *Robust to outliers* • *Stable*	• *Fast and efficient* • *Don't require feature engineering*
Cons	• *The assumption of linearity between X&Y.* • *Needs a cleaned data* • *Can predict continuous data*	• *Assumption of independent is not very realistic.*	• *Data should naturally clustered* • *Slow* • *Don't work well large no of variables* • *Outlier sensitive*	• *Inherently less interpretable than decision tress.* • *High computational need*	• *Only numeric data* • *High computational need*

Models Performance (1/4): Summary

The table below provides summary to the key KPIS used to select the model and how the impact the model performance.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**
Accuracy	*Is defined as the number of correctly predicted data points out of all the data points.*									
Error Rate	*Is defined as the number of wrongly predicted data points out of all the data points. It equals to 1-Accuarcy.*									
Precision	*Is defined as the number of relevant data that is correctly predicted. In our case this is passengers Satisfaction.*									
Sensitivity	*It's the proposition of the total desired outcome that is predicated correctly by the model. For example, how many 1 or satisfied is correctly classed.*									
Specificity	*It's proportion of total negative that are correctly identified. measures the proportion of actual negatives that are correctly identified*									
ROC Curve	*The ROC curve Plots the model true positive rate (Sensitivity) and false positive rate (1-Specificity) for every possible decision rule cutoff between 0& 1 for a model. The bigger the curve the better the model, it means higher true positive rate and lower false positive rate.*									
AUC	*The area under the curve (AUC) is area under the ROC curve and it measure how good is the model. The bigger it gets the better is the model.*									
Gini Coefficient	*The Gini is 2*AUC-1, and it helps in normalizing AUC so the model classifies low scores as zero and higher score as one. The higher the Gini the better the model performance*									

Models Performance (2/4): Summary

The Confusion matrix below show case the model prediction power, using number of KPIs. The overall model accuracy and Precision ability to predict +ve instances correctly.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**
Ks	*The Ks measures how good the model's classification. It measure the degree of separation between –ve and +ve distribution. The bigger it gets the better the model.*									
Concordance	*It compares instances probabilities of being true or being false, if being true is higher then the pair are in concordance. Simply measures the predictive power of the model*									
Discordance	*It compares instances probabilities of being true or being false, if being false is higher then the pair are in disconcordance. Simply measures the predictive power of the model*									
Tie	*It compares instances probabilities of being true or being false, if two are tied then it's a tie. Simply measures the predictive power of the model*									

Models Performance (3/4): Summary

The table below show the scoring criteria for model. Each KPI is one point awarded to the highest performing model in each category.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.
Accuracy								*1*		*1*
Error Rate									*1*	
Precision										*1*
Sensitivity									*1*	*1*
Specificity								*1*		
ROC Curve								*1*	*1*	*1*
AUC										*1*
Gini Coefficient									*1*	*1*

The highlighted cells are for the best preforming criteria and model.

Models Performance (4/4): Summary

The table below show the scoring criteria for model. Each KPI is one point awarded to the highest performing model in each category.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.
Ks								*1*		
Concordance								*1*		
Discordance								*1*		
Tie								*1*		
Total								***7***	***4***	***6***

Overall results are very close, nearly a tie between Random Forest "RF" and XGBoost "XGB", both are ensemble leaners. Their performance can be attributed to fact that these algorithms develop a huge number of weaker models to learn and improve performance so they can finally build a stronger one. Results are very close in every aspect, especially models ran on regular data both would be good model to predict overall Satfication for Falcon Air.

The highlighted cells are for the best preforming criteria and model.

Models Performance (1/5): Confusion matrix

The Confusion matrix below show case the model prediction power, using number of KPIs. The overall model accuracy and Precision ability to predict +ve instances correctly.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**
Confusion matrix										
Accuracy	*45%*	*84%*	*52%*	*81%*	*52%*	*68%*	*54%*	*95%*	*94%*	*95%*
Error Rate	*55%*	*16%*	*48%*	*19%*	*48%*	*32%*	*46%*	*5%*	*6%*	*5%*
Precision	*50%*	*86%*	*57%*	*84%*	*57%*	*72%*	*55%*	*94%*	*95%*	*96%*
Sensitivity	*45%*	*84%*	*52%*	*80%*	*52%*	*68%*	*50%*	*93%*	*94%*	*94%*
Specificity	*45%*	*83%*	*52%*	*81%*	*52%*	*68%*	*59%*	*96%*	*95%*	*95%*

The highlighted cells are for the best preforming criteria and model.

Models Performance (2/5): ROC

The table below shows the two main KPIS AUC or area under the curve and Gini Coefficient.

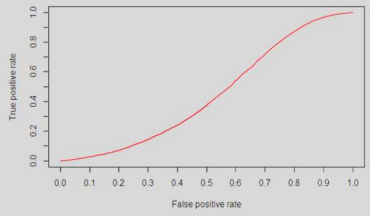
Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.
ROC										
AUC	*45%*	*91%*	*53%*	*89%*	*43%*	*75%*	*55%*	*99%*	*99%*	*99.1%*
Gini Coefficient	*33%*	*37%*	*34%*	*43%*	*41%*	*24%*	*38%*	*44%*	*45%*	*45%*

The highlighted cells are for the best preforming criteria and model.

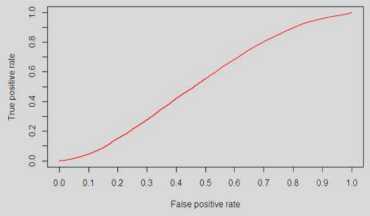
Models Performance (3/5): ROC Plots

The highlighted charts are for the top preforming models showing a higher curve for ROC, which indicates lower FPR and higher TPR.

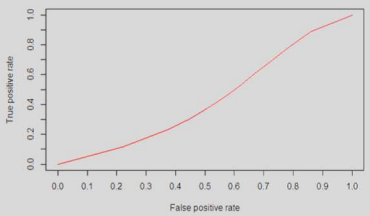
Logistic Regression – PCA Data



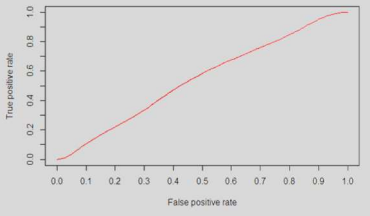
Naïve Bayes – PCA Data



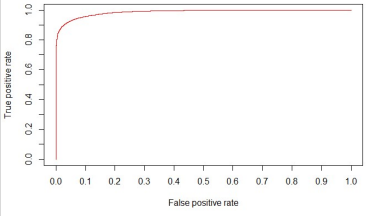
KNN – PCA Data



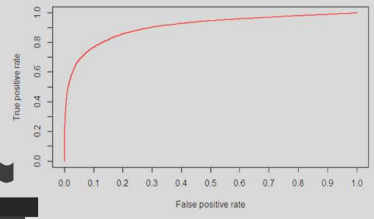
Random Forest – PCA Data



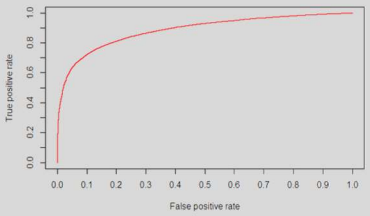
XGB – PCA Data



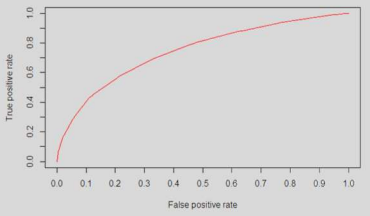
Logistic Regression – Reg. Data



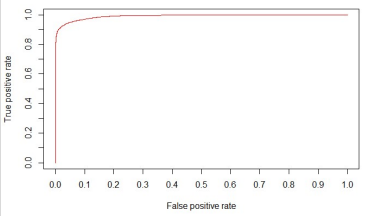
Naïve Bayes – Reg. Data



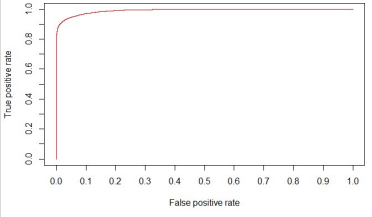
KNN – Reg. Data



Random Forest – Reg. Data



XGB – Reg. Data



The highlighted cells are for the best preforming criteria and model.

Models Performance (4/5): Ks Statistics

A higher the KS Statistics is desired, which implies which is the better model, and it means separation between class-1(i.e. Satisfied) and class-2(i.e.Not). The bigger the Ks the more accurate the model.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.	PCA	Reg.
Ks Statistics										
Ks	*8%*	*67%*	*11%*	*63%*	*3%*	*36%*	*8.5%*	*90%*	*89%*	*89%*

The highlighted cells are for the best preforming criteria and model.

Models Performance (5/5): Concordance ratios

The Concordance ratios below compare probabilities for each instance to assess the model predictive power.

Models	Logistic Reg.		Naïve Bayes		KNN		Random Forest		XGB	
Evaluation Criteria	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**	**PCA**	**Reg.**
Concordance ratios										
Concordance	*45%*	*91%*	*53%*	*89%*	*39%*	*71%*	*54%*	*99.1%*	*99%*	*99%*
Discordance	*55%*	*9%*	*47%*	*11%*	*61%*	*29%*	*46%*	*0.09%*	*0.1%*	*0.1%*
Tied	*0%*	-1.3878e-17	5.5511e-17	2.77558e-17	0	*0*	*0*	3.9894e-17	3.6429e-17	-5.204e-17

The highlighted cells are for the best preforming criteria and model.

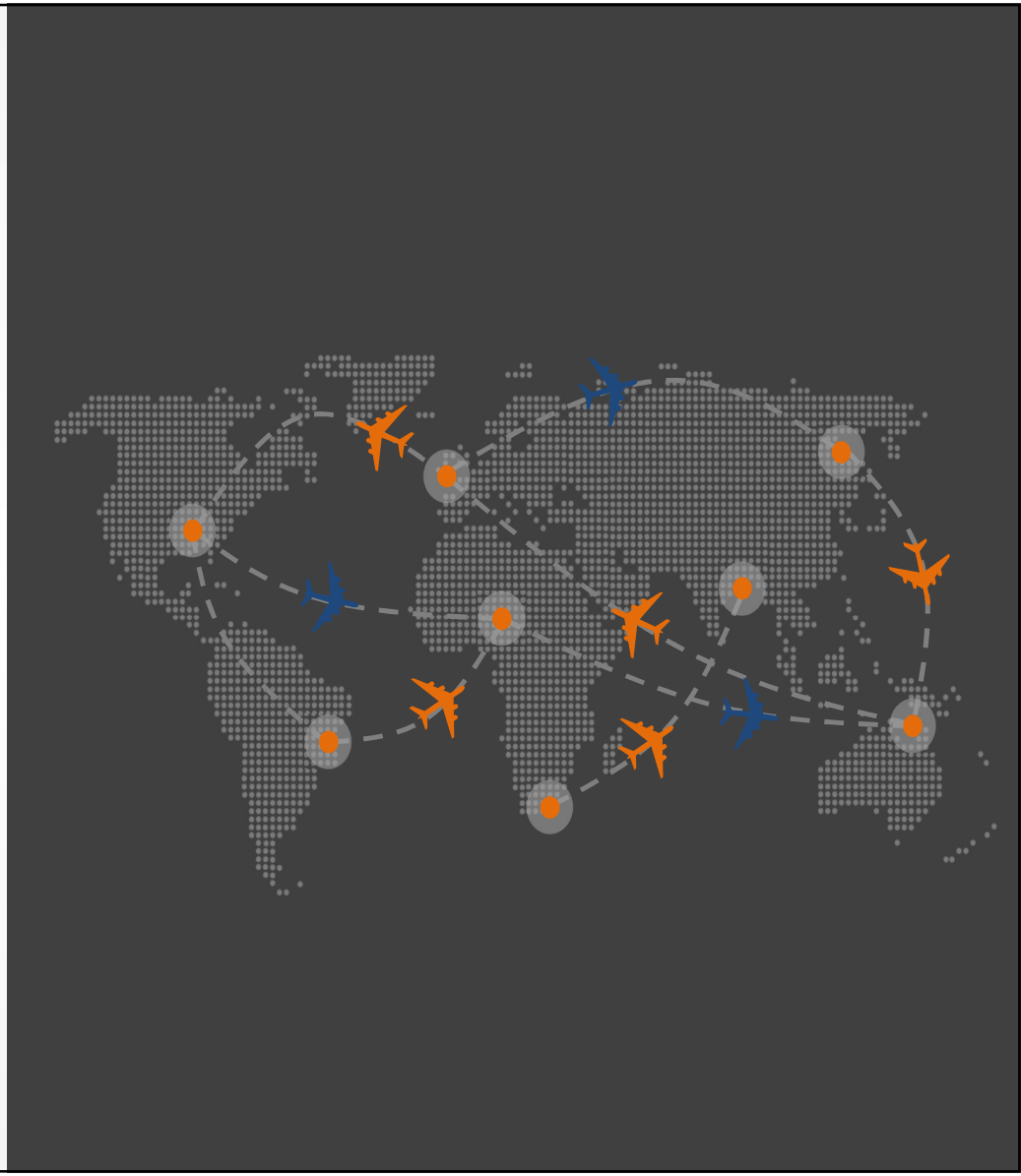
FALCON AIRLINES
NONSTOP YOU

Section 7:

Conclusion and key findings

FALCON AIRLINES
NONSTOP YOU

Table of Contents



The Project key objectives…

… was to support ***Falcon Airlines respond to their recent issues faced around Passengers' Satfication level drop***. Such, issue is even posing a greater challenge to Falcon Airlines, as it is directly correlated to customers churn to competition. Which might affect the airlines on a number of potential fronts, such as;

1. Reputational damage: **Driven by the lack of Satfication and poor reviews for the airlines.**

Reputational damage is no easy matter, it directly affect the airlines financial positions, back in 2017 United Airlines shares have dropped by ~4% due to a viral video of mistreating a passenger.

2. Market Share loss: Reputational damage leads to market share loss, as clients use other airlines.

~45% of our sample are not satisfied with our services, which put nearly 50% of the business at risk and is expected to impact our market share with a similar percentage.

3. Financial loss: Market share loss can be translated to loss of revenue and an increase in operating costs, leading to an overall drop in net come.

We tried to countify the impact of revenue loss from our sample, using ticket average market prices from similar airlines. The total loss from Business class was around 64m, while reached 34m for Eco and Eco plus.

As we understand the severity of the impact of the aforementioned challenges, we tailored the scope of this project to achieve two key objectives;

1. To understand which parameters play an important role in swaying a passenger feedback towards 'Satisfied'.

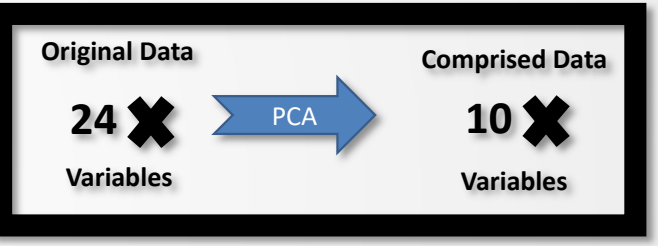
2. To predict whether a passenger will be satisfied or not given the rest of the details are provided.

In our attempt to achieve project objective one…

… which focus of understanding the most important variables that affects overall Passengers satisfaction, we've used number of statistical methods such **Variable importance & Weight of Evidence and PCA**.

*In **PCA** we reduce high dimensional data like ours, into a smaller one while retaining data properties and preventing information loss. In layman terms PCA reduces the number of variables in a data set, by grouping the highly correlated ones. By doing do we are able to understand what are key areas that affected overall satisfaction.*

Our PCA analysis has reduced 24 original variables and comprised them into only 10 variables that explains 86% of Satisfaction variance.



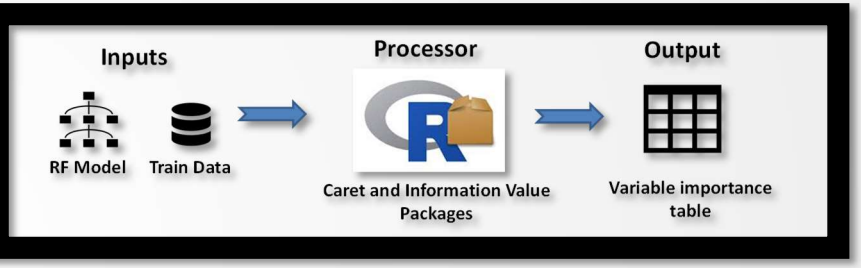
PCA 1	PCA 2	PCA 3	PCA 4	PCA 5
Online Support	*Boarding Ease*	*Services & Facilities*	*Timeliness*	*Entrainment*

PCA 6	PCA7	PCA 8	PCA 9	PCA 10
Check-ins	*Age*	*Distance*	*Seating*	*onboard service*

We also used Vimp & Woe to achieve project objective one…

*In Variable importance "**Vimp**" and Weight of Evidence "**WOE**" we used number of R packages to score the important variables based on their impact to the overall predictions of our top preforming models, RF and XGB. Below a flow chart the explain the steps followed in this process..*

Our Process outputs a sorted list of variables as per their importance to the model's prediction, the most important variable is Seat comfort, followed by entertainment

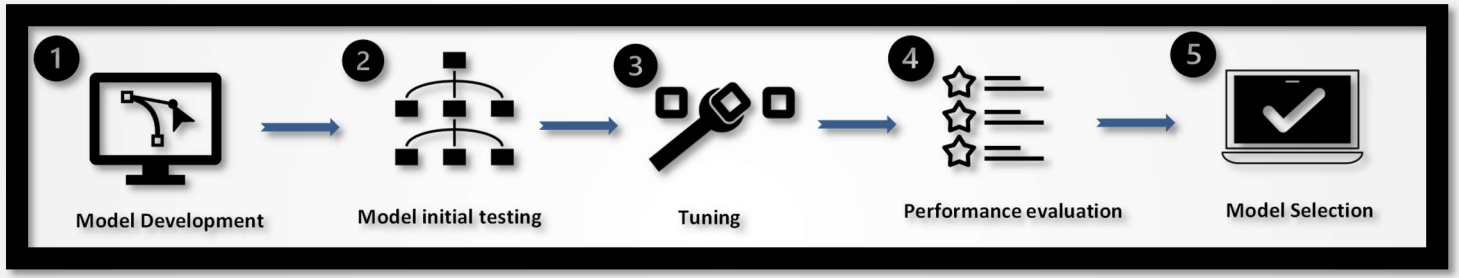


Ranking	Variable	Imp. Score
1	Seat_comfort	216
2	Inflight_entertainment	119
3	CustomerType	116
4	Checkin_service	112
5	TypeTravel	84
6	Online_support	82
7	Gender	69
8	Baggage_handling	67
9	Departure.Arrival.time_convenient	65
10	Class	62
11	Cleanliness	62

Ranking	Variable	Imp. Score
12	Age	58
13	Leg_room_service	57
14	Onboard_service	54
15	Ease_of_Onlinebooking	48
16	Flight_Distance	44
17	Food_drink	42
18	Online_boarding	41
19	ArrivalDelayin_Mins	40
20	Gate_location	36
21	DepartureDelayin_Mins	32
22	Inflightwifi_service	30

In our attempt to achieve project objective two…

… which focuses on providing Falcon Airlines the ability to predict passengers Satisfaction. In order to bring the most value to the Airlines we aligned the solution developed to business problem, which revolve around classification of passengers. Consequently, we've selected to build around ten machine learning classifiers and compared their performance to select the best for Falcon Airlines. Below is flow of the process implemented



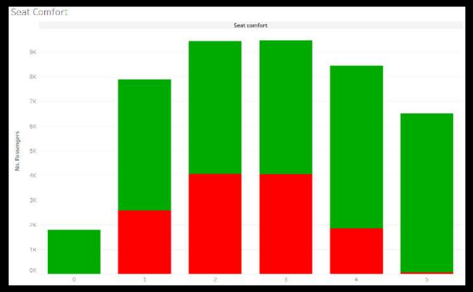
Based on the performance of the different model, the final results are very close nearly a tie between Random Forest "RF" and XGBoost "XGB". Both are an ensemble leaner, which built more than one model to find the best results. Their performance can be attributed to fact that these algorithms develop a huge number of weaker models to learn and improve performance so they can finally build a stronger one. Results are very close in every aspect, especially models ran on regular data both would be good model to predict overall Satfication for Falcon Air.

In order to create even further value we wanted to….

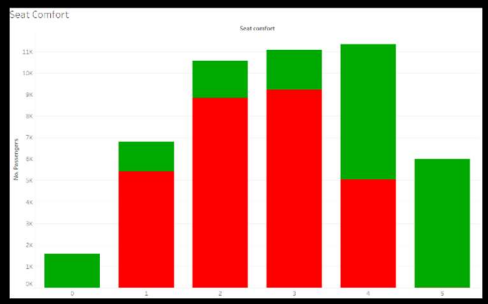
… analysis further the important variables as shown on slide 85, to shed more light on they change based on passengers' demographics. Let's start off by our first variable from Importance standpoint **"Seat Comfort".**

By Class

In Business class seat comfort is not really affecting overall Satfication, this is due to the fact business class seats are more comfortable by nature.

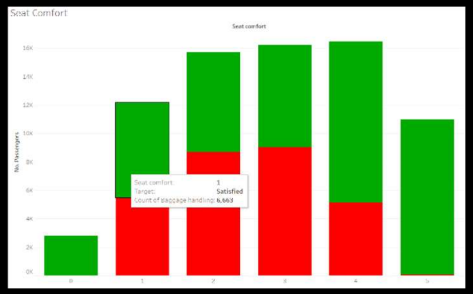


While in Eco and Eco Plus classes seat comfort a huge issue that is really affecting overall passengers Satfication, you see mostly who score 1 to 4 for seat comfort are not satisfied with Falcon Airlines.

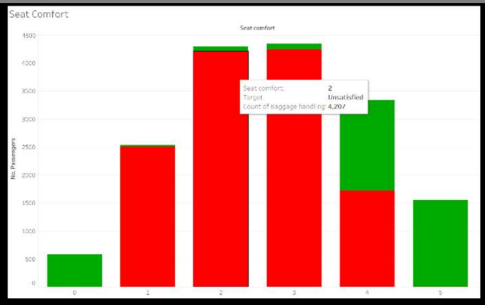


By Customer

For frequent flyer "Loyal customers" seems that seat comfort is not as severe. Overall rating for seat comfort is better compared to nonfrequent flyers.



Nearly all of nonfrequent flyers are dissatisfied with seat comfort and scored between 1 to 3. Most of nonfrequent flyers are also not satisfied with Falcon Airlines.

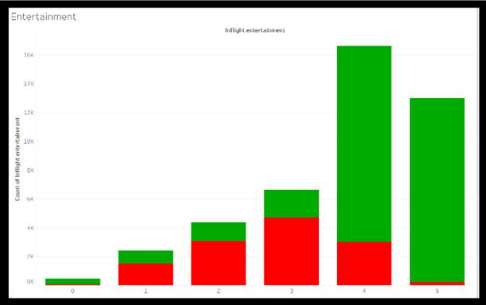


It also worth mentioning that business traveler are less affected by seat comfort.

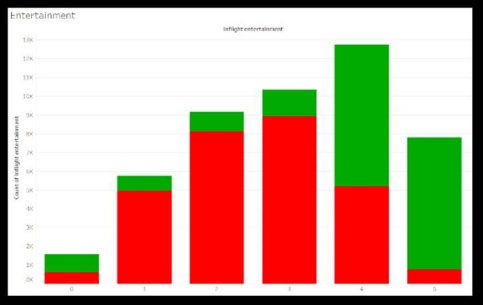
Let's look at Entrainment…

By Class

In Business class Entrainment seems to a good indicator for overall Satisfaction, passengers who scored entrainment either 4 or 5 are highly likely to be satisfied with the Airlines services.

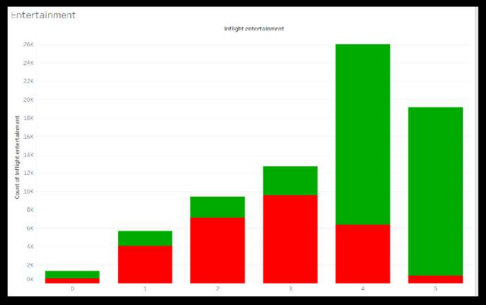


Lesser impact is witnessed in ECO and Eco plus classes. Where you passengers score are more spread, but again the ones scored 4 or 5 are likely to be satisfied with Falcon services.

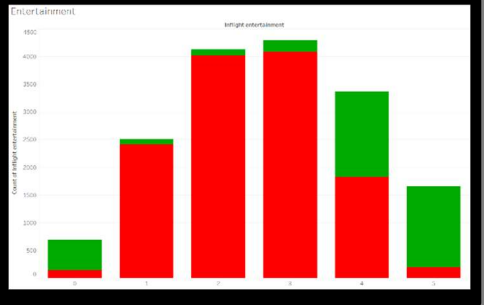


By Customer

For frequent flyer "Loyal customers" seems to be happy with our entertainment and overall are likely to be satisfied with Falcon services.



Our flight entertainment are not very popular in nonfrequent flyers. But again the ones score 5 are likely to be satisfied with Falcon services.

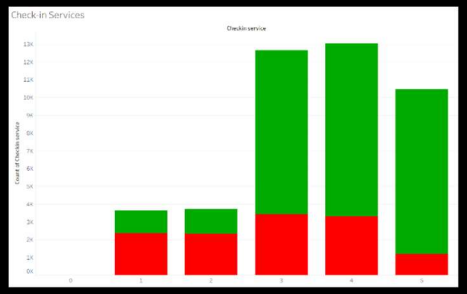


It also worth mentioning that type of travel and age have less affected by entrainment.

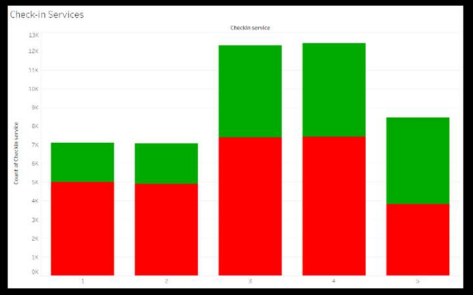
Let's look at Check-in Services…

By Class

In Business class overall positive trend and scoring for check-in services. The likely hood of a person who scored 3 or above are likely to be satisfied with Falcon airlines.

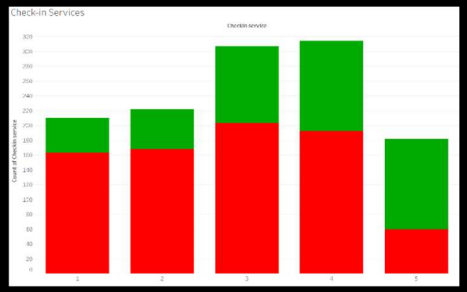


Score are more spread in ECO and Eco plus classes, with concentration for 3 or 4. Again likely hood of being satisfied with Falcon Airlines seems low.

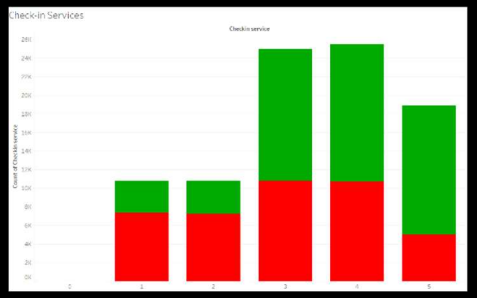


By Age

Age group above 60 years are not happy with our Check-in services and more likely to be unhappy with overall Falcon Airlines



Other Age groups showed a consistent trend and scored mostly between 3 to 4.

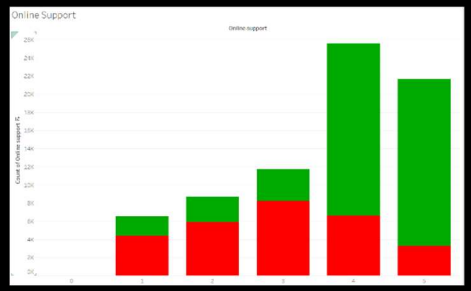


It also worth mentioning that type of customer type and travel type have less affected by entrainment.

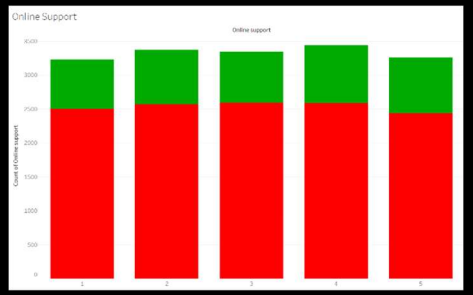
Let's look at Online Support…

By Customer

Frequent Flyers has scored our online support highly between 4 to 5 and most likely to satisfied with Falcon Airlines.



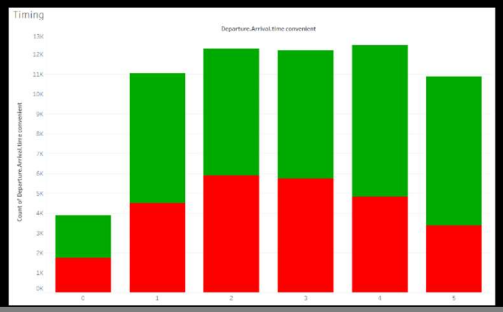
While nonfrequent flyers have mixed reviews of Falcon Airlines online support and are highly likely to be an unsatisfied.



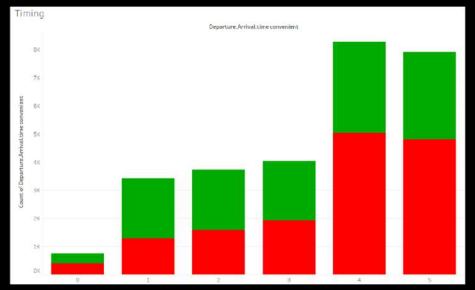
Let’s look at Timing…

Traveler type

Our Business Travelers are not very happy with timings and have mixed reviews, but this has not materially affected their overall satisfaction.



While personal travelers seems to be happy with our timings but still likely to be unhappy with our overall services.



Overall Satisfaction demographic…

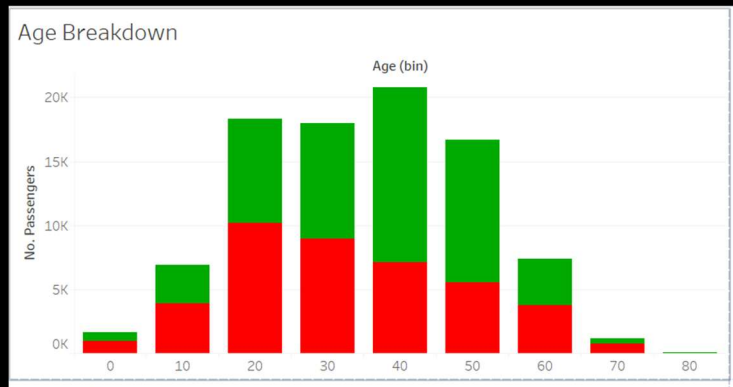
Satisfaction demographic

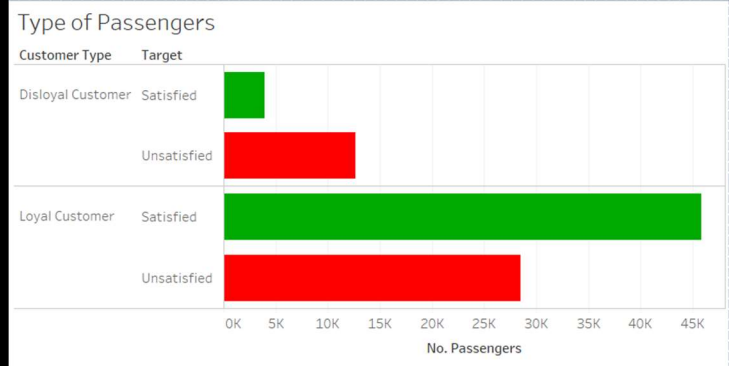
Overall the riskiest **Age** groups are between 0-30 years and 60+ years old, that are not Satisfied with Falcon services.

While from a **Class** perspective, the most problematic is Eco.

Most of **nonfrequent flyers** are unhappy with our services.

And **Personal Traveler** have a 50% chance of being unhappy with Falcon Airlines, similar trend is witnessed for **Male travelers.**





Summary of business insights and Recommendations

Project Objective no. one outcomes

The top 6 variables in predicting satisfaction are:
1- Seat comfort
2- *Entrainment*
3- *Checking-services*
4- *Customer type*
5- *Travel type*
6- *Online support*
If Falcon airlines works to improve these top services, we believe they might see a noticeable improvement in Passengers' overall Satisfaction.

PCA resulted has in reducing the data set from 24 to 10 dimensions. These ten dimensions captures c.86% of overall variance of passengers' satisfaction. These PCAs are;

- *PCA1: Online Support*
- *PCA2: boarding ease*
- *PCA3: Services & Facilities*
- *PCA4: Timeliness*
- *PCA5: Entrainment*
- *PCA6: Check-ins*
- *PCA7: Age*
- *PCA8: Distance*
- *PCA9: Seating*
- *PCA10: onboard service*

These are the most important variable to predict Satisfaction.

Project Objective no. two outcomes

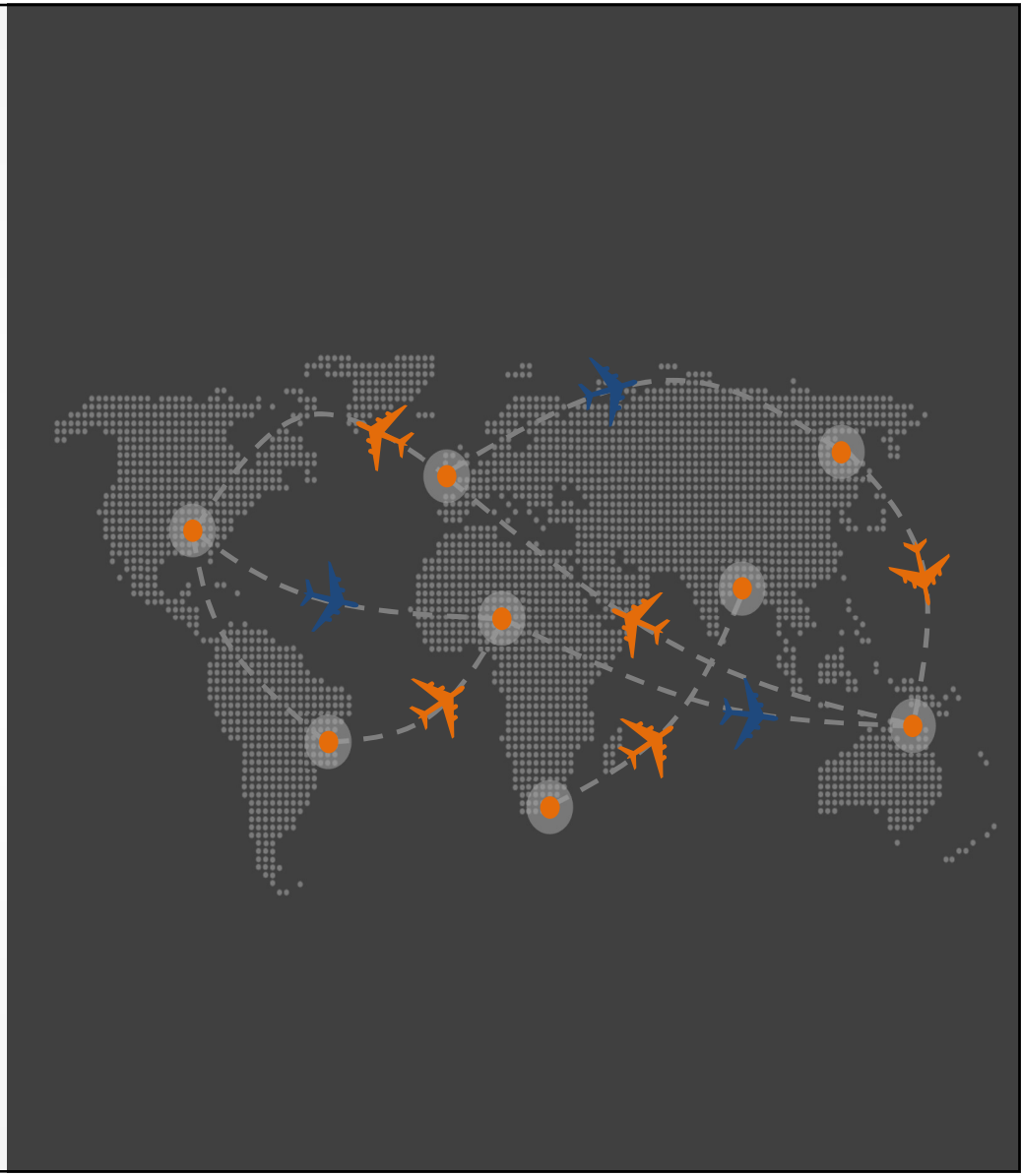
Overall there **two** great **models** that can be used to predict Passengers Satisfaction were ensemble models. Meaning that regular models wont work as effect for predication. These model might be little more complex to tune and maintain, but they provide much accurate results as we seen in the performance evaluation. I don't think PCA is not best used with ensemble model.
So, We recommend using either of these two model as they provide impressive performance with both high accuracy *95%+* and precision *95%+.* We would recommend using either **RF** or **XGBoost** in predicting passengers' satisfaction.

FALCON AIRLINES
NONSTOP YOU

Appendix:

FALCON AIRLINES
NONSTOP YOU

Table of Contents



Capstone Project – Guidelines

In this project we aim to show case our knowledge of the following key areas;

1. Experience end-to-end problem solving using a combination of tools and techniques in analytics.
2. Learn practical implementation of various analytical techniques and choose the one which gives results most appropriate for business.
3. To understand the trade-offs that need to be made when solving a problem in real life.
4. To develop better presentation and report writing skills.

The project consists of three submissions, with the following goals';

1. Problem framing and data explorations;
2. Data preparation;
3. Modeling and insights.

Key deliverables consists of the following;

1. Notes I, II & III submissions – 17 May to 22 June
2. Final Report – 6 July
3. Final Presentation - 9 to 11 July

The R Code used for Submission Notes III

Below Attached copies the R code used

The R Code

Falcon Airlines - Capstone Project - Notes III - Mohanned Gomaa.Rmd

Training Dataset

Microsoft Excel Worksheet

Testing Dataset

Microsoft Excel Worksheet

The R Code used for Submission Notes II

Below Attached copies the R code used

The R Code

D:\UT Autin\ one\Project Notes

Cleaned Master Dataset

Microsoft Excel Worksheet

The R Code used for Submission Notes I

Below Attached copies the R code used

The R Code

Falcon Airlines - Capstone Project - Notes 1 - Mohanned Gomaa.Rmd

Tableau Dashboard Link

Below Attached link to my Tableau Dashboard

Link

https://public.tableau.com/profile/mohanned.gomaa#!/

FALCON AIRLINES
NONSTOP YOU

Thank You