**Data Challenge**

Based on the data available in .csv file provided, please try and answer the questions below:

1. Explain what you understand from the data provided, provide exploratory insights.

The goal of the data analysis is to investigate the relationships between the collected data features and the in-flight wifi prices. With better understanding of the relationship, we can have a more sophisticated strategy to set the price and generate more revenue.

The overall approach of the exploratory data analysis includes the following steps:

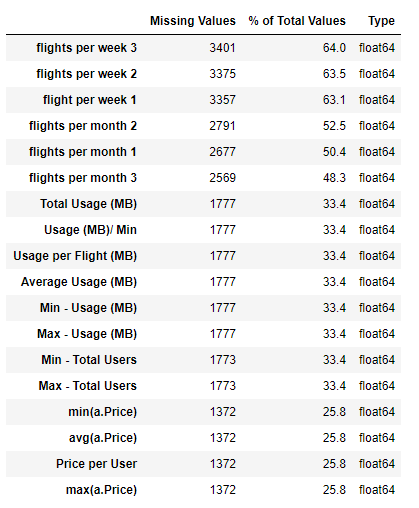
1. Basic data information
2. Missing value check
3. Plot the statistics of the key features (10 key features)
4. Summary of the insights from EDA
5. **Basic data information**

Before we dive in the the details of the data, we need to get some basic information of the data set itself to have an intuitive of what kind tools could be used to perform the data analysis. In total, there are 5316 samples and 33 features in the provided data.

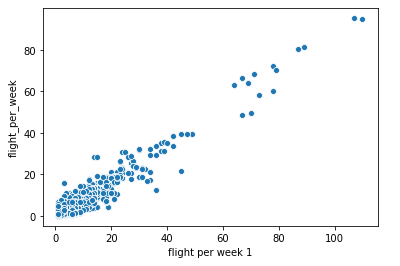
1. **Missing value check**

Figure 1 presents the number of the missing value and the fraction of the missing value as to the total number of samples. About 25.8% price related values and about 33.4 wifi usage related values were missing values. The cause of the missing value was uncertain, and may due to the fact that no every flight provided the wifi service. Since we are working on the the problem of setting up the wifi price, I am going to remove those data from the training data set for further analysis.

Another group of missing data came from the flight per week 1-3 and flight per month 1-3. The document did not provide much description about those features. I assume that those features are used to describe the average flight count. Since the total flight count and First/Last flight time were provided, I am going to create an new feature “flight per week” by dividing total flight count to number of weeks to replace those features. The scatter plot (Figure 2) between the newly created “flight per week” and existing “flights per week1” indicates that the newly created feature was a good replacement.



**Figure 1 Missing value statistics**



**Figure 2 flight per week vs flight per week1**

1. **Plot the statistics of the key features**

In total, there are 33 features in the data set, but not all of the features are self-explained. Based on my understanding, I divided them to four categories and the features marked as bold are the features I think would impact the In-Flight price. We will examine them one by one.

### Route Features:

*****'Route'******,******'Flight Count'******, 'flight per week 1', 'flights per week 2', 'flights per week 3', 'flights per month 1', 'flights per month 2', 'flights per month 3'*

### Flight Related Features:

*'First Flow', 'Last Flown', 'Airline Count', 'First Airline', 'Last Airline', 'Aircraft Type Count', 'First Aircraft Type',******'Average of Avg - Flight Duration (MB)'******, 'Min of Min - Flight Duration (Hrs)', 'Max of Max - Flight Duration (Hrs)',******'Avg - Seat Count'******, 'Min - Seat Count', 'Max - Seat Count'*

### In-flight WiFi Price:

*'Price per User',******'avg(a.Price)'******, 'min(a.Price)', 'max(a.Price)'*

### In-flight WiFi Usage:

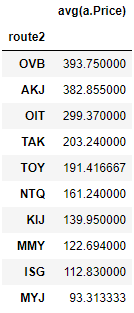
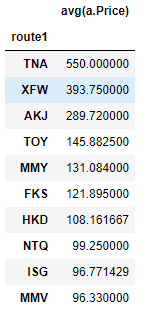
*****'Average Usage (MB)'******, 'Min - Usage (MB)', 'Max - Usage (MB)',******'Total Usage (MB)'******,******'Usage per Flight (MB)'******,******'Usage (MB)/ Min'******, 'Min - Total Users',******'Max - Total Users'*****

1. **Route: Flight routes including departure and arrival airports.**

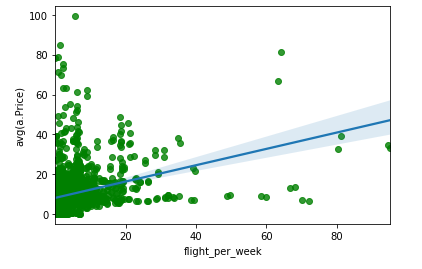
The departure and arrival airports may impact the demand of the in-flight usage. The more popular destinations draw more flight customer and drive more wifi usage demands. Flight customers from certain city may tend to use wifi more, for instance, business people or younger people. Figure 3 listed the top 10 departure and arrival airports with highest wifi prices. Some of the prices seemed unreasonably high and investigation is needed.

1. **Flight Count: flight per week**

Figure 3 shows the scatter plot of flight per week and average price. Despite the large disperse of average price when flight per week number is small. We can arguably observe a increasing trend of price versus flight per week. The possible explanation is that larger number of flight per week indicates popular destination and more flight customers, and that can translate to higher demand of in-flight wifi usage, and thus higher wifi price.



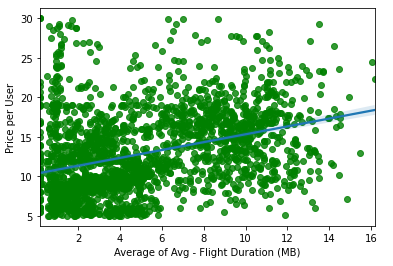
**Figure 3. The top 10 highest prices of departure (route1) and arrival (route 2)**



**Figure 4. the scatter plot of flight per week and average price**

1. ****'Average of Avg - Flight Duration (MB)': Average flight duration in Hour?****

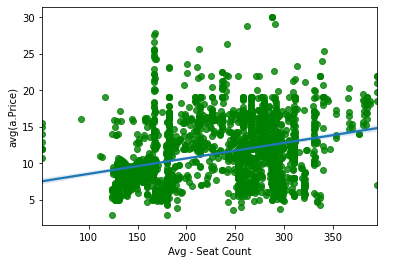
It is common to think that the longer the flight is, the more time people would spend on wifi and thus pay higher price on wifi service. Figure 5 plots the average flight duration and price per user. Even though the data is quite noise, an increasing slope of price per user on flight duration is observed.



**Figure 5, scatter plot of average flight duration and price per user**

1. *****'A*vg - Seat Count': Average seats sold per flight****

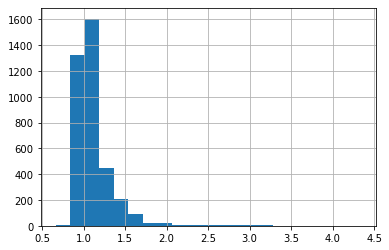
**Average seat count is directly related to the number of customer on board and the wifi usage demand is related to the number of customers. Figure 6 shows that the prices per customer also increase with average seat count.**



**Figure 5, scatter plot of average seat count and price per user**

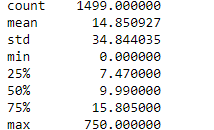
1. ****Price per user and average price.****

**There’s no clear definition of how those values are calculated. My guess is price per user is the total sale of wifi package divided by the total number of users. The average price is the total sale of wifi divided by the total number wifi packages sold. Figure 6 shows that the ratio of prices per user versus average price. The ratio larger than 1 may indicate that some users purchased more than one packages to meet their usage demand. Very small number of ratio with value smaller than 1 may due to the discount use.**

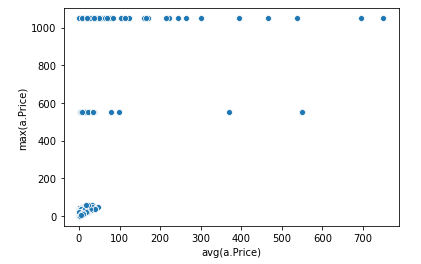


**Figure 7 Ratio calculated by price per user / average price**

**The statistics of the average price also shows some extreme values (Figure 8). The max value is 750, which means people payed 750 dollar for on-board wifi. Except we can find some reason to explain this high value, most customer may feel unconformable to pay such large wifi bill. Figure 9 indicates that the large average price may related to the max price.**

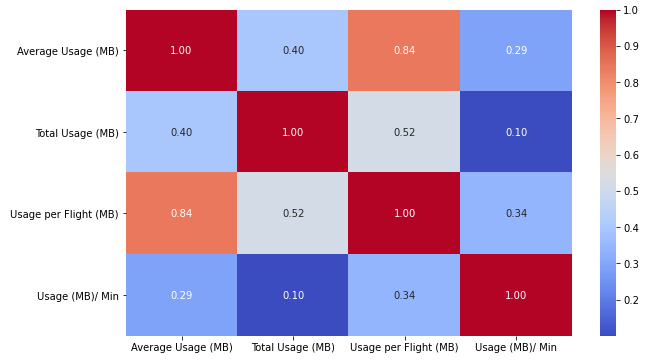


****Figure 8 the statistics of average price****



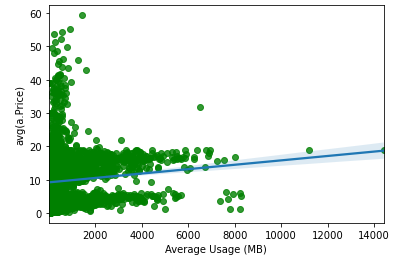
****Figure 9 scatter plot of average price and max price****

1. ****'Average Usage (MB)'****, ****'Total Usage (MB)'****, ****'Usage per Flight (MB)'****, ****'Usage (MB)/ Min'****,  ****'Max - Total Users'****

**There are a few features related to the wifi usage. Unlike the prices related features, a correlation matrix (figure 10) shows they are not strongly correlated with each other and each of the features may contain some unique information. The red and blue color represents the positive and negative con-variance between two features. It’s straight forward to see the total usage is positively correlated with usage per flight.** 

****Figure 10 Correlation matrix of wifi usage features****

**In Figure 11, by plotting the relation between average usage and average price, we observed that the average prices increase with the wifi usage. But we also notice that the dispersion of data points at low average use and at low average price. The dispersion at the lower ends of both price and usage shows that some customer paid really high price for some casual use. This may cause customer dissatisfaction on the value of their paid bills. On the other hand, some customer used large amount volume of data, while paying very little to the service. In the long-term, the service may generate less revenue than it supposed to. Thus, a good price strategy is to match both demand and price.**



****Figure 11 scatter plot between average usage and average price****

1. ****Summary:****
2. **Some extremely high wifi average prices were found and they might relate to the max price**
3. **Flight count, average flight duration, and average seat counts all contribute to drive the demand of wifi usage, thus drive higher wifi package price**
4. **Generally, the average wifi price increases with the average usage volume of wifi. However, in certain routes, customers paid higher bills for very small data usage, while some customer paid cheap price for excessive wifi use. In the long term, this mismatch between usage and price may cause customer dissatisfaction and reduced revenue for the service provider.**
5. What is your approach if you have to dynamically change the pricing for In-Flight WiFi? – Technical approach will suffice.

The goal of dynamic pricing is to set the price of In-Flight WiFi to match the ever-changing WiFi usage demand. The demand varies from flight to flight. Thus if I have the chance to design the approach. I’ll use all the data available and build a model based on the machine learning algorithms to predict the reasonable price for each flight.

For the data part, I would like to collect the data for the the following features:

1. Airports/citys on the flight routes
2. Averge flight per week/month for that route
3. Flight date, month of the year, week of the month, day of the week, holiday…
4. Airline
5. Aircraft type
6. Flight duration
7. Average seat count
8. Seat count for sold tickets
9. Average fair price
10. Historical data of wifi usage. Average usage, total usage, usage per user, usage/min, average number of user

I’ll collect the data from flights the wifi price is well matched with demand. Those data would be used for the predictive model training.

The collected dataset would be divided into training and test sets by a ration 8:2. The training set would be used for model building.

For the model building part, I am going to test a few models based on linear regression, random forest, gradient boosting trees and neural network. I don’t expected linear regression to work since the relationship between the listed features and wifi demand or price are non-linear. The final model would picked from the tree-based model or neural network. The predicted prices can first evaluated on test set by comparing the predicted price to the true price. The model with the best performance should bring least root mean square error. Then the model results should be tested on the future flights to check the if the performance are consistent and if the revenue is increased by using the fixed rate or prices predicted from other method.

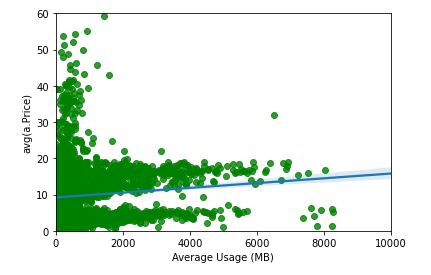
1. What are the statistical/ predictive methodologies that could be applied with the data you have – Please implement one model & explain the output.

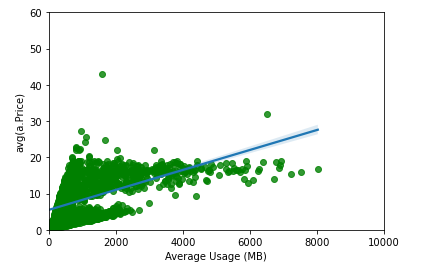
To build a dynamic pricing model, we might need more data. For the current data, the most important thing is to update the prices where customer may feel robbed for paying too much for little use and the prices where the service won’t generate any revenue since customer paid too little.

I’ll implement the follow steps to achieve the goal:

1. Identify the data set with the reasonable price range
2. Build a model based on the selected data set
3. Predict the price for the routes with unreasonable price using the model created.
4. Summary
5. **Identify the data set with the reasonable price range**

The unreasonable price may cause the price paid per 1 MB data is too high or too low. I picked two threshold to define the reasonable range of price per MB use. And Figure 12 showed the data before and after the threshold applied.





**Figure 12 The average price vs. Average usage. Upper: no data selection, bottom: after data selection**

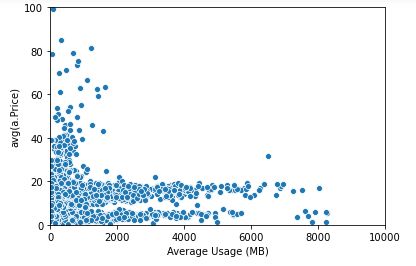
1. **Build a model based on the selected data set**

The training data set were selected based on the price per MB threshold and the following features were included in the training set: 'Flight per week', 'Average of Avg - Flight Duration (MB)', 'Avg - Seat Count', 'Average Usage (MB)', 'Total Usage (MB)', 'Usage per Flight (MB)'.

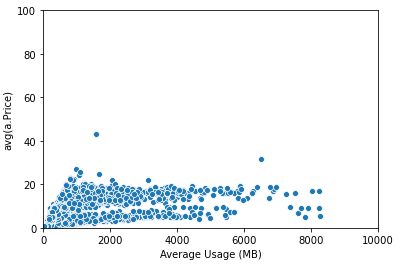
Random forest model was adopted to build the predictive model. The default parameters were used in the training.

1. **Predict the price for the routes with unreasonable price using the model created.**

Once the model was trained, the average price would be predicted using the same features for the routes where price per MB is not in a reasonable range. The following Figure compares the price before and after prediction.



Original avg price



After predicted price

Figure 13 The average price over average usage. Top: original price. Bottom: Predicted Price

1. **Summary**

The average price predictive model was able to provide a reasonable In-Flight WiFi price reference for routes the price is either too high or too low. The result is pretty preliminary, not much models and parameters were tested. In the future, we can include more data if available or improve the result by using testing different models and fine-tuning the parameters.

Please answer (visuals/ insights, theories if any) in this document.