MGT 6203 Group Project Proposal Template

**Please edit the following template to record your responses and provide details on your project plan.**

**TEAM INFORMATION (1 point)**

**Team #:**

**Team Members:**

1. Team Member 1 Name; GT Id (OMSA) or EdX username (MM)

[Insert background information: Name, professional background, education background, previous analytics related projects you have worked on]

1. Team Member 2 Name; GT Id or EdX username
2. Team Member 3 Name; GT Id or EdX username
3. Team Member 4 Name; GT Id or EdX username
4. Team Member 5 Name; GT Id or EdX username

**OBJECTIVE/PROBLEM (5 points)**

**Project Title:**

**Background Information on chosen project topic:**

We plan to use the Airline Passenger Satisfaction dataset from Kaggle (<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>) to understand what are the main drivers for creating a positive experience for customers. The dataset consists passenger feedback taken from a survey and contains 24 features and a final feature for whether that customer was satisfied or not. By gaining a better understanding of what airline passengers truly care about we could help determine where to invest additional money and also where to cut unnecessary expenses. A big gap is that there is no information regarding the airports involved. We may try to infer this data based on flight distance if possible.

**Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):**

Understand what are the major factors that lead to airline passenger satisfaction to understand where money should be invested or re-allocated to keep passengers satisfied.

**State your Primary Research Question (RQ):**

What are the most important factors in ensuring that an airline passenger is satisfied?

**Add some possible Supporting Research Questions (2-4 RQs that support problem statement):**

1. Do different customer segments value different parts of the flight experience? Potential segments: gender, class, type of travel, flight distance.
2. Is there a correlation between satisfaction and specific airports?

**Business Justification:** **(Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)**

By identifying which parts of the flight experience are more important to create customer satisfaction we can help drive decision on which areas are worth spending additional money on and which areas investments could be reduced. From an operations standpoint this is related to efficient allocation of capital. From a marketing standpoint this would help increase customer satisfaction which could lead to greater customer loyalty and help secure future business as well as the potential to be able to charge higher rates.

**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

Main dataset: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>

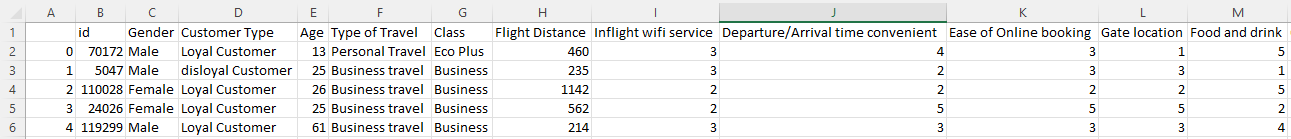
Secondary dataset (Flight routes): <https://www.kaggle.com/datasets/open-flights/flight-route-database>

One thought is that we could try to figure out which airports the passenger interacted with by cross referencing the travel distance in the main dataset to a matching distance in the secondary dataset. The idea is that the distances may be different enough that we can reasonably narrow it down to the correct airport. We may be able to find a correlation between certain airports and customer satisfaction.

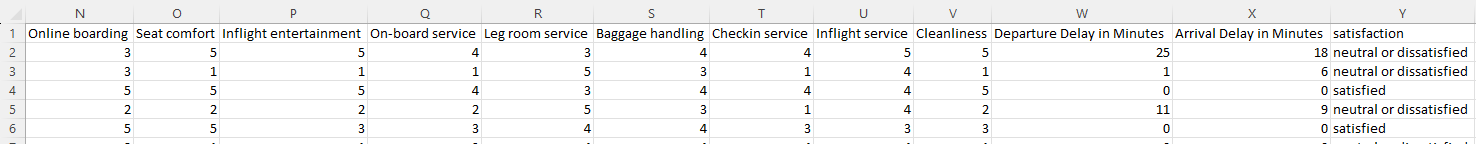
**Data Description (describe each of your data sources, include screenshots of a few rows of data):**

Main dataset is from Kaggle and is taken from surveys from a US airline. It includes user ratings for multiple aspects of air travel such as food and drink and seat comfort, as well as customer information such as age and gender. The last column specifies whether the customer was satisfied or dissatisfied.

Main dataset columns 1-12:



Main dataset columns 12-24:



Secondary dataset is also from Kaggle and was taken from openflights.org which is a tool that has nearly every airport route mapped.

Secondary dataset all columns:

A screenshot of a computer

Description automatically generated with medium confidence

**Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables?** **What variables do you hypothesize beforehand to be most important?)**

All the columns in the primary dataset with the exception of Satisfaction are independent variables. We will attempt to create a new variable which is airports interacted with which will be a cross reference between flight distance and the flight route database. We won’t know the departure and arrival airport but we will at least have a pair of airports and see if there is a difference between different airports in customer satisfaction.

We might also create a new variable that is simply a summation of ratings for multiple columns and see if it serves as a good predictor.

We may decide to breakdown age into buckets such as by decade or life or classes such as: minor, young adult, middle age, senior citizen, etc.

I hypothesize that the most important variables will be seat comfort, flight distance, class, leg room service, cleanliness, departure delay in minutes, and arrival delay in minutes.

**APPROACH/METHODOLOGY (8 points)**

**Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))**

In theory there are three ways that a customer can feel about a flight: satisfied, neutral, or not satisfied. Since our goal is to understand what makes a customer satisfied, we will treat any other customers (i.e. neutral or satisfied) as the same group. The data is already organized this way. Given that there are only two groups and we are trying to understand a state of being instead of predicting an actual value we will implement a logistic regression model with multiple predictors. This should give us a good understanding of which predictors are statistically significant so that we can narrow our focus for further analysis.

Additionally we will run separate logistic regression models on subsets to see if there are certain factors that each subset values differently. Some subsets are already part of the data such as: gender, type of travel, and class. Other subsets can be created by us such as: age (divide into buckets for each decade) and flight distance (can be divided into short, medium, and long).

Once we determine which classifiers are most important we will simplify the model by removing the statistically insignificant classifiers and tuning the cutoff value of p using the ROC curve.

Another algorithm that we will try is a random forest algorithm. We will tune the following hyperparameters: number of trees in the forest, max number of features considered for splitting a node, and max number of levels in each decision tree. The goal will be to minimize the mse. Once the model has been trained we can get the importance of each variable using the varImpPlot function.

We will compare the results of the random forest to the logistic regression model by evaluating the variables that the random forest shows as being important against the variables that the logistic regression model considers statistically significant. We will also compare the magnitude of each coefficient of the logistic regression model with the importance of the random forest model to see if both models consider certain variables to have approximately the same importance.

**Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement**

At the end of the project we expect to have a list of attributes that are statistically significant for predicting customer satisfaction. We will then be able to use the magnitude of the coefficients in the logistic regression model and the importance of the random forest model to rank them in order of importance.

Once we have a list of attributes ranked by relative importance we can reasonably determine through some research how much control an airline would have over said attribute. If it’s an attribute that can be affected by the airline then we can try to evaluate the magnitude of investment and complexity to make changes to that attribute. Based on overall complexity and investment required we can make recommendations on where airlines should focus enacting changes. The opposite approach also works, by looking at the least important factors we can determine if airlines could “cut corners” and save money without affecting customer satisfaction.

We expect that some of the factors that lead to customer satisfaction are going to be factors that are impossible or prohibitively expensive to fix such as departure and/or arrival delays. Delays are often times out of the control of an airline since they can be weather related or related to how a specific airport operates.

We expect that some factors are easily controlled by airlines such as leg room but it may not be beneficial for airlines to change since they may get more revenue from the additional seating even if it leads to less satisfied customers. Such a conclusion will likely be evaluated qualitatively since we won’t be able to determine real costs for changing leg room on an aircraft vs. the change in realized revenue from adding or removing seating. We can reasonably expect that most airlines have done this analysis and are running close to maximum efficiency.

**What business decisions will be impacted by the results of your analysis? What could be some benefits?**

We can help determine where additional money should be invested and where airlines can spend less money.

**PROJECT TIMELINE/PLANNING (2 points)**

**Project Timeline/Mention key dates you hope to achieve certain milestones by:**

Project Proposal – Submitted by 6/20

Multiple Logistic Regression Models – completed by 6/30

Proposal Video – Submitted by 6/30

Progress Report – Submitted by 6/30

Random Forest Models – completed by 7/12

Final Report – Submitted by 7/17

Final Video – Submitted by 7/20

**Appendix (any preliminary figures or charts that you would like to include):**