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 #: 6

Team Members:

Name: Joshua Farina; EdX Username: joshuapfarina

Joshua Farina, a Supervisory Data Analyst at the New York State Department of Tax and Finance, is driven by a unique fusion of data analytics and social activism. In his professional role, he's made notable contributions such as analyzing the migration patterns of taxpayers during the COVID-19 pandemic and providing invaluable insights for policy decisions.

On a personal level, he’s pursued a project to algorithmically identify and spotlight food deserts, demonstrating his dedication to addressing societal challenges. Known for mentoring rising analysts and staying at the forefront of data science advancements, Joshua actively promotes using data as a powerful tool for social change within the community.

Name: Henri SALOMON; EdX Username: henri_salomon

Henri Salomon is a data analyst at the United Nations in New York where he specializes in data-driven insights about internal operations. With a background in management and consulting, Henri has worked with the World Food Programme (WFP) and PwC's Strategy&, demonstrating interest and expertise in innovation, efficiency, and strategic analysis.

Name: Raajitha Middi; EdX username: Raajitha_Middi

Holds engineering and MBA degrees. Worked as Data Analyst in various domains including IT, e-commerce, and healthcare. In my professional capacity, I worked on multiple projects like creating interactive dashboards on PowerBI, procurement data analysis etc.,

Name: Ryan Chandler; EdX Username: zujin87

I obtained my electrical engineering degree from Tennessee Tech University in 2008. I currently serve as an airborne instrumentation subject matter expert, performing research and development in airborne instrumentation, telemetry, and data processing for experimental flight testing.

Name: Alejandro Martinez; EdX Username: amzeta

Alejandro has an Aerospace Engineering degree and has been working in the Automotive industry for over 10 years. He first started out as a design engineer and is currently working as a Project Manager. He first learned about ML and AI through work and took some Coursera courses and is now enrolled in OMSA starting Fall of 2023.

OBJECTIVE/PROBLEM (5 points)

Project Title: Predict Customer Satisfaction in the Airline Industry

Background Information on chosen project topic:

Customer satisfaction is a key factor in attracting and retaining business. Identifying factors that have the strongest effect on customer satisfaction will provide key insights for airlines to enhance their services, improve the customer experience and optimize business operations.

By understanding what passengers truly care about, an airline can increase investment to the areas that have the biggest impact and reduce expenses in areas that aren't as highly regarded.

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

We intend to determine which factors drive customer satisfaction (or dissatisfaction) so that we can provide actionable recommendations on how 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. Does grouping variables by category (demographics, in-flight service quality, timeliness and delays) provide additional insight?
2. Is there a correlation between satisfaction and specific airports (if we can identify airports)?
3. Can we predict airline passenger satisfaction based on a limited set of factors?
4. Do characteristics which would imply a higher level of service (business class) result in higher levels of satisfaction?
5. Are there any noteworthy interactions which would affect the probability of a satisfied customer?

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.)

The airline industry has evolved around different trends, marked by the hybridization of business models – the traditional distinction between full-service and low-cost carriers has been blurring – and by the frequency of financial distress due to internal (mismanagement) or exogenous (e.g., oil prices). In this competitive context, improving air passenger satisfaction can have a positive impact on both companies' revenues and costs. First, air companies can focus on operational efficiency while addressing customers' real and personalized needs to increase satisfaction levels and reduce efforts on unnecessary needs ("value for money"). Second, improving customer satisfaction will reduce customer churn and improve customer loyalty, which has become a barrier to entry with higher customer acquisition costs, consequently increasing their revenues. Some estimates indicate that increasing customer retention by as little as 5% increases profits by 25%-95%. Providing an analysis of the key factors involved in predicting customer satisfaction will allow the airline to identify specific areas of improvement which would lead to profitability.

Finally, on the marketing side, improving customer experience will contribute to building a strong brand image and reputation. For example, passengers would be more inclined to share their experience (e.g., word of mouth, social media), which could attract new passengers.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

Our primary source of data will be the following:

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

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.

We have three other potential sources of data that we intend to evaluate in relation to the business problem.

1. The first dataset is found here: https://data.mendeley.com/datasets/64xkbj2ry5/1/files/e4e4b8b0-0d7d-41a9-a2be-c1e586897d7e

   This dataset includes customer satisfaction scores from four industries (one of which is airlines). Although the dataset uses demographic variables, it might help validate the coefficients of the analogous variables in our primary dataset.

2. The second dataset is found here: https://www.kaggle.com/datasets/open-flights/flight-route-database

   This dataset contains flight route information and airport code. We will investigate whether there's a reasonable process for relating each flight to the airport based on the travel distance between airports, thus allowing us to augment our dataset by including origin and destination.

3. The third dataset is found here: https://www.kaggle.com/datasets/lapodini/british-airway-reviews?select=rating_data.csv.

   This dataset contains online reviews for British Airways. It likely suffers from self-selected response bias, but may provide some utility in corroborating other analysis.

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

Our primary data source contains 130k rows, one id column, 22 independent variables and one response variable. Here's a snip of the metadata:

```
str(k.aps)
```

```
'data.frame':   103904 obs. of  24 variables:
 $ id                              : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ Gender                          : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
 $ Customer.Type                   : Factor w/ 2 levels "disloyal Customer",..: 2 1 2 2 2 2 2 2 2 1 ...
 $ Age                             : int  13 25 26 25 61 26 47 52 41 20 ...
 $ Type.of.Travel                  : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 ...
 $ Class                           : Factor w/ 3 levels "Business","Eco",..: 3 1 1 1 1 2 2 1 1 2 ...
 $ Flight.Distance                 : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
 $ Inflight.wifi.service           : int  3 3 2 2 3 3 2 4 1 3 ...
 $ Departure.Arrival.time.convenient: int  4 2 2 5 3 4 4 3 2 3 ...
 $ Ease.of.Online.booking          : int  3 3 2 5 3 2 2 4 2 3 ...
 $ Gate.location                   : int  1 3 2 5 3 1 3 4 2 4 ...
 $ Food.and.drink                  : int  5 1 5 2 4 1 2 5 4 2 ...
 $ Online.boarding                 : int  3 3 5 2 5 2 2 5 3 3 ...
 $ Seat.comfort                    : int  5 1 5 2 5 1 2 5 3 3 ...
 $ Inflight.entertainment          : int  5 1 5 2 3 1 2 5 1 2 ...
 $ On.board.service                : int  4 1 4 2 3 3 3 5 1 2 ...
 $ Leg.room.service                : int  3 5 3 5 4 4 3 5 2 3 ...
 $ Baggage.handling                : int  4 3 4 3 4 4 4 5 1 4 ...
 $ Checkin.service                 : int  4 1 4 1 3 4 3 4 4 4 ...
 $ Inflight.service                : int  5 4 4 4 3 4 5 5 1 3 ...
 $ Cleanliness                     : int  5 1 5 2 3 1 2 4 2 2 ...
 $ Departure.Delay.in.Minutes      : int  25 1 0 11 0 0 9 4 0 0 ...
 $ Arrival.Delay.in.Minutes        : num  18 6 0 9 0 0 23 0 0 0 ...
 $ satisfaction                    : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1 2 1 2 1 1 2 1 1 ...
```

Most of the numeric columns are on a 1-5 scale. Some, such as Age and Flight distance are continuous integer fields. A few describe qualitative labels, such as gender and class. The response variable is binary, where values can be “neutral or dissatisfied” or satisfied.

Here is a sampling of the rows/columns in the dataset:

	id <int>	Gender <fctr>	Customer.Type <fctr>	Age <int>	Type.of.Travel <fctr>	Class <fctr>	Flight.Distance <int>	Inflight.wifi.service <int>
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3
1	5047	Male	disloyal Customer	25	Business travel	Business	235	3
2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2
3	24026	Female	Loyal Customer	25	Business travel	Business	562	2
4	119299	Male	Loyal Customer	61	Business travel	Business	214	3
5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1180	3
6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276	2
7	96462	Female	Loyal Customer	52	Business travel	Business	2035	4
8	79485	Female	Loyal Customer	41	Business travel	Business	853	1
9	65725	Male	disloyal Customer	20	Business travel	Eco	1061	3

1-10 of 10 rows | 1-9 of 24 columns

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?)

Independent	Gender	Less Important
Independent	Customer Type	Somewhat Important
Independent	Age	Less Important
Independent	Type of Travel	Somewhat Important
Independent	Class	Somewhat Important
Independent	Flight distance	More Important
Independent	Inflight wifi service	More Important
Independent	Departure/Arrival time convenient	More Important
Independent	Ease of Online booking	Less Important
Independent	Gate location	Less Important

Independent	Food and drink	More Important
Independent	Online boarding	Less Important
Independent	Seat comfort	More Important
Independent	Inflight entertainment	More Important
Independent	On-board service	More Important
Independent	Leg room service	More Important
Independent	Baggage handling	Somewhat Important
Independent	Check-in service	Somewhat Important
Independent	Inflight service	More Important
Independent	Cleanliness	Somewhat Important
Independent	Departure Delay in Minutes	More Important
Independent	Arrival Delay in Minutes	More Important
Dependent	Satisfaction	Response Variable

Part of our analysis will look at binning variables like age into categories to see if they offer improved predictive power. Similarly, flight distance may be logarithmically distributed.

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?))

- **Data collection, cleaning and transformation:** After loading the data sets, missing values, outliers and data quality issues will be identified. Categorical variables would be encoded, and any necessary data transformations or scaling would be applied.
- **Exploratory Data Analysis:** Preliminary and descriptive analyses will be conducted on the data sets, providing already useful insights. For example, we will look at correlation matrix and distributions but also at average satisfaction across different groups (e.g., class type, flight type, loyalty program, gender, age,…).
- **Modelling:** Different models will be tested and assessed against performance. To compare models, the data set will be split into training, validation and testing with a 60:20:20 ratio, and the best model chosen best on performance on test data. The analysis of coefficients will help us reply to the research question 1. These models would include:
  - Regression: Linear, Log-linear, Log-log
  - Random forest
  - Feature selection including Lasso
  - PCA
  - Logistic Regression - Analysis of Distributions for possible transformations, Consider Regularization
  - Support Vector Machines - Analysis of Distributions for possible transformations, Center and Scale Data, Optimal Choice of Kernal, Optimal Value of C/lambda.
  - K-Nearest Neighbors - Center and Scale Data, Optimal choice of K
  - Classification Trees - Identify key splits. Prune tree. Possibly use for variable creation.
  - Random Forest – Discover relative feature importance.

- The models will be compared using ROC-AUC on a reserved validation set.

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:

- Factors with a high impact on passenger satisfaction will have to be prioritized as an improvement of the factor could yield a significant increase in customer satisfaction.
- Segmenting customers into different categories may lead to insights resulting in more productive marketing campaigns.

PROJECT TIMELINE/PLANNING (2 points)

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

June 18th - June 22nd: Data Collection and Preparation

- Collect and clean the airline satisfaction data, and other relevant datasets.
- Start on literature review and background research.
- Begin work on the Project Proposal Video.

June 23rd - June 27th: Exploratory Data Analysis

- Conduct an exploratory data analysis on the collected data.

- Continue literature review, identify potential models and methodologies.
- Finalize and submit the Project Proposal Video by June 27th.

June 28th - July 3rd: Model Building

- Build initial models, run preliminary analyses.
- Evaluate model results, adjust and refine as needed.
- Prepare and submit the Progress Report by July 3rd.

July 4th - July 10th: Model Refinement

- Continue refining the models, run additional analyses if necessary.
- Start preparing for the final report and presentation.
- July 11th - July 20th: Final Report Writing
- Collaborate on writing the Final Report, submit by July 20th.

July 21st - July 23rd: Final Video Preparation

- Collaborate on the Final Video Presentation, submit by July 23rd.

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