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:**

1. Joshua Farina;

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.

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: 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.

**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 make provide actionable recommendations to the airline.

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

What factors have the strongest impact on customer satisfaction?

**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. Can additional datasources be used to enhance insights generated by our analysis?

**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.)**

It’s generally understood that the cost of retaining existing customers is substantially lower than that of acquiring new customers. [Some estimates indicate](https://www.forbes.com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/?sh=45bfd27d1c7d) 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.

**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>

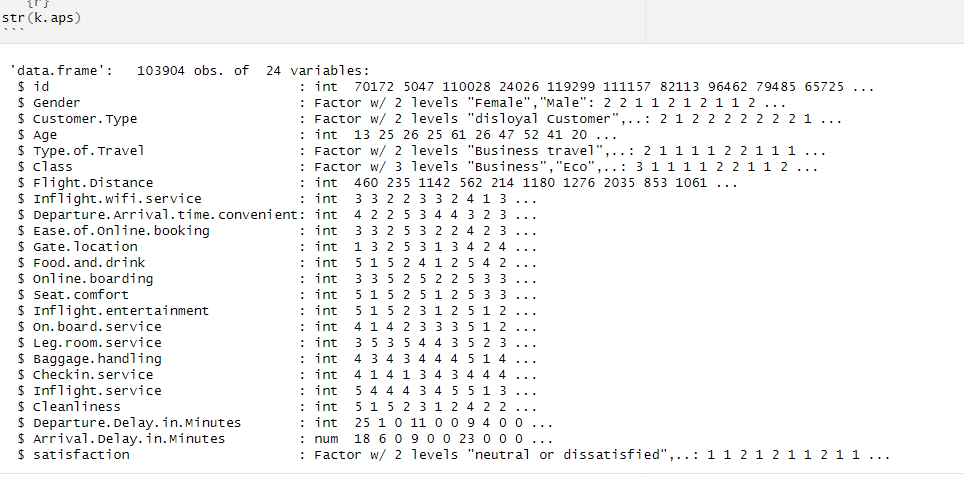
We have two other potential sources of data that we intend to evaluate in relation to the business problem.

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.

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

Our primary data source contains the following columns:

****

Most of the numeric columns are built on a 1-5 scale,

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

**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 | Less Important |
| Independent | Age | Less Important |
| Independent | Type of Travel | Less Important |
| Independent | Class | Less Important |
| Independent | Flight distance | Somewhat 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 |

**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?))**

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.

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**

We expect the factors which directly impact in-flight comfort to have the greatest significance. Since these factors are likely to be the most expensive to change, our analysis will attempt to identify improvements that can be made efficiently.

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

The analysis would guide the airline on how to allocate resources. Ideally, it would identify the most cost effective areas for improving the customer experience, but ascertaining cost data could prove problematic. Still, the research could be used as part of a future analysis where proprietary information is available.

**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):**