# Background

Since I was a little kid, I’ve always had a fascination for airports and airplanes. I remember wanting to stay at the airports all day long and just watch all the airplanes whenever we were dropping someone off or I was going on a trip myself. It wasn’t just the planes taking off which fascinated me, it was also the small things like watching the luggage/cargo being loaded onto the plane, seeing the pilots and flight attendants walking around the airport, and watching the planes arrive at and depart from the terminal gates. I haven’t even mentioned my curiosity of how these enormous things are made and how they’re able to fly so safely several thousand feet above the ground with hundreds of passengers onboard. Now, it’s probably easy to see why I was excited to work on this project.

# Question

Maven Airlines, headquartered in Boston, Massachusetts, is displeased to find that over 50% of their passengers are not satisfied with the airline overall, according to the latest customer satisfaction survey. In order to prevent losing a part of their customer base, leadership has brought me in as a Data Analyst to identify where things might be going wrong and to help provide a data-driven strategy which will bring those satisfaction levels back up. I’ve been given the option to present my findings in the form of a single-page report or dashboard.

# Data Collection

The dataset was provided by Maven Analytics and can be found at the following link under **Airline Passenger Satisfaction**: <https://www.mavenanalytics.io/data-playground>

Each record in the dataset represents a single passenger and contains information on overall passenger satisfaction with the airline and their ratings for different factors like cleanliness, comfort, etc. It also includes some passenger details and flight data such as gender, passenger age, flight distance, and any arrival/departure delays.

## Data Profile

# of Fields: **24**

# of Records: **129,880**

# of Text Fields: **5**

# of Numeric Fields: **16**

Structure: **Single Table**

File Type: **CSV**

# Importing Data into MySQL Database

I first created a database in my server named *maven\_airlines* where all the data will be stored. I decided to use the LOAD DATA LOCAL method to load the input CSV file into my database since it’s proven to be the most efficient method for such data sources.

To implement the method, I followed the below procedure:

1. Create a table named *passenger\_satisfaction* and define data types for all fields
2. Check server-side load local data capability. If OFF, turn ON
3. Write LOAD DATA command

1st Attempt:

Records: 129880 Deleted: 0 Skipped: 0 Warnings: 393

* Upon reviewing the log file, I found that the warnings were due to empty cells in the *Arrival\_Delay* field. So, I included a NULLIF as a preprocessing transformation to replace all empty input cells for this field with NULLs.

2nd Attempt:

Records: 129880 Deleted: 0 Skipped: 0 Warnings: 0

To confirm that all data loaded properly into the database, I checked a couple of records against the input file.

# Querying the Database

In order to devise a strategy that will help improve the overall passenger satisfaction rate, several different questions must be answered to gain a deeper understanding of where both the satisfaction and dissatisfaction are coming from.

The first thing I wanted to extract from the data was the overall passenger satisfaction and dissatisfaction rates. I found that **56.55%** of all passengers are either Neutral or Dissatisfied with the airline, meaning that only **43.45%** of passengers are Satisfied (confirming that the satisfaction rate has indeed dipped below 50%).

Then, I performed a deeper analysis through the following:

* Does gender have a significant effect on overall satisfaction rate?

Distribution of Gender:

Male: 63981 Female: 65899

* + Males and females are almost equally satisfied/dissatisfied with the airline overall (44% vs 43%)
* Are Returning customers more or less satisfied than First-time customers?

Distribution of Customer Type:

First-Time: 23780 Returning: 106100

* + Returning customers are a lot more satisfied than new customers (48% vs 24%)
* How does satisfaction vary by Travel type?

Distribution of Type of Travel:

Business: 89693 Personal: 40187

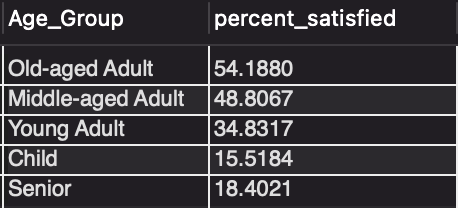
* + Passengers who are travelling for business purposes are a lot more satisfied than those who are travelling for personal reasons like vacation, visiting family, etc. (58% vs 10%)
* How does satisfaction vary by age?
  + For this analysis, I first wanted to group all the different ages because using the age numbers alone doesn’t seem to provide much value. So, I used CASE statements to categorize all the different ages into 1 of 4 categories: Child, Young Adults, Middle-Aged Adults, and Old-Aged Adults. This specific age grouping was extracted from the following source: <https://www.semanticscholar.org/paper/Human-Age-Group-Classification-Using-Facial-Bhat-V.K.Patil/19ddb412336ce633c1fe21544605c7bd65ff8d66/figure/0>

I then added a new column (*Age\_Group*) into the *passenger\_satisfaction* table which will hold the age classification for every passenger.

* + - Using the result of the CASE statement, I tried to fill the Age\_Group column but was experiencing a lot of trouble. I knew I wanted to use an UPDATE JOIN statement since I was essentially trying to update the values in one table using values from another table. To do so, I created a temporary table which held the result of the CASE statement and then used it as a second table in the UPDATE JOIN to fill the column. But I kept receiving “Lost connection to server during query” errors both when filling the temp table with the CASE statement result and when executing the UPDATE JOIN. I found that it had to do with the timeout settings for my server; after increasing the number under settings and repeating the procedure, it still didn’t work. Then, I tried creating a VIEW for the CASE statement result instead which, like the temp table, treats the result as a table. I then repeated the UPDATE JOIN clause under the modified timeout settings, but this time I noticed the query execution took much longer before spitting out the same error. After exiting and restarting MySQL, I noticed that the query had actually worked and the Age\_Group field was properly filled. I’m assuming the timeout errors must’ve been caused by the amount of work the query had to do for every row in the table because of the CASE statement.

Distribution of Age Groups:

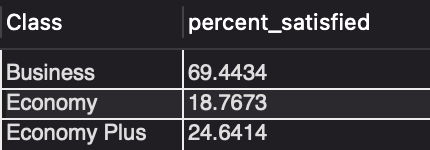
Old-aged Adults: 41535 Middle-aged Adults: 41482 Young Adults: 32258 Children: 8622 Senior: 5983



* + As it can be seen from the table above, it appears that Middle-aged and Old-aged adults between the ages of 31 and 64 are more satisfied with the airline overall than Seniors and the younger population
    - I realized at this step that I actually wanted to include a separate age group for seniors (ages 65+). So, I had to go back and edit my initial CASE statement, which meant I also had to edit my VIEW which was used to populate the *Age\_Group* field. Repeating the previous UPDATE JOIN procedure, I faced the same problem of losing connection to the server. I set the read timeout under the server settings to 0 so that the query can take as long as it needs to return a result and waited 1.5hrs without any results before realizing that this method won’t work. I came across a forum [<https://stackoverflow.com/questions/27900505/update-join-table-taking-a-very-long-time>] which mentioned setting indexes for the columns that are being joined on in both tables leading to much faster query execution. I successfully added an INDEX to the ID field of the *passeneger\_satisfaction* table but wasn’t able to do the same for the VIEW because an index can’t be added to them. So, I alternatively resorted back to the temporary table method I had previously employed since an INDEX can be added to such objects. Repeating the UPDATE JOIN with the newly created temp table, the *Age\_Group* field was successfully populated with the added Senior group in ~3s.
* How does satisfaction vary by Travel Class?

Distribution of Travel Class:

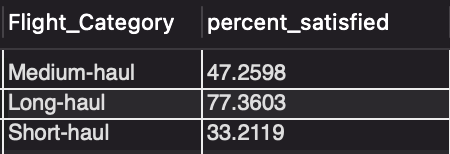
Business: 62160 Economy: 58309 Economy Plus: 9411

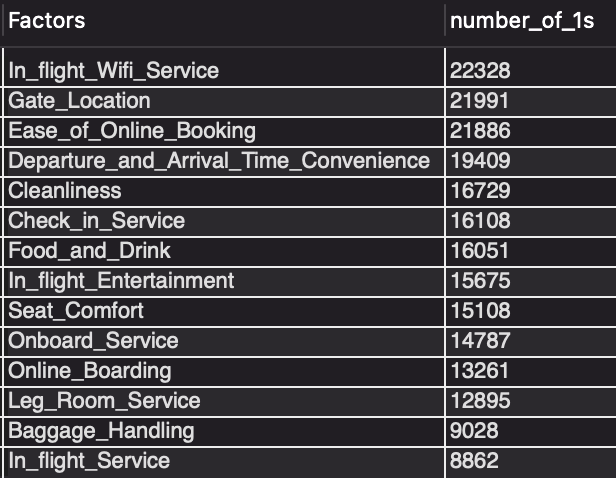


* + As it can be seen from the table above, passengers who travelled in Business Class were the most satisfied with the airline and also made up the majority of passengers.
* How does satisfaction vary by Flight Distance?
  + The shortest flight distance is **31** miles and the longest distance is **4983** miles. To get a better idea of how flights varied by distance, I categorized all flights into 1 of 3 categories according to United Airlines definition of flight length [<https://www.usatoday.com/story/travel/roadwarriorvoices/2016/04/22/united-airlines-mileage-plus-points/83365166/>] & [<https://united.mediaroom.com/2015-01-09-United-Airlines-Offers-Bonus-Miles-Promotion-for-Premium-Cabin-Travelers>].

Distribution of Flight Distance:

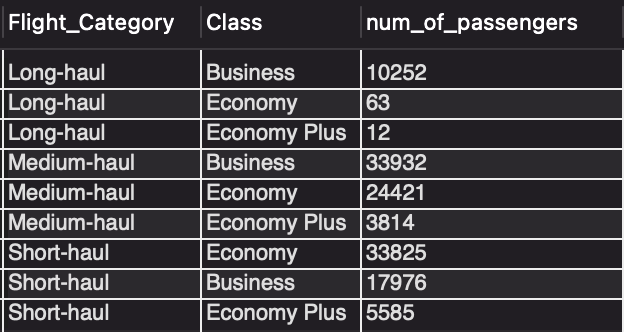
Short-haul: 57386 Medium-haul: 62167 Long-haul: 10327



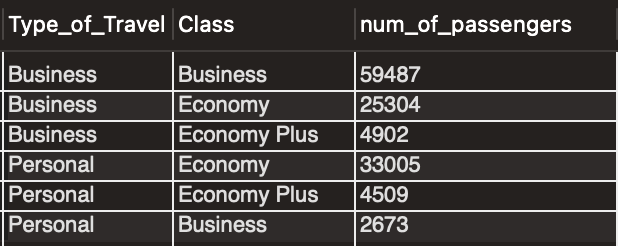
* + As it can be seen from the table above, passengers who flew long distances were the most satisfied while those who flew short distances were the least satisfied.
* Now, I’d like to analyze the various different factors which passengers rated on a scale of 1-5 (1-lowest to 5-highest).
* The first thing I wanted to find was the total number of lowest satisfaction score (1) each factor received. From there, I wanted to see which factors passengers were most displeased with.
  + According to the data table above, the Top 5 factors passengers are most displeased with based on their satisfaction score of 1 are **In Flight Wi-Fi Service**, **Gate Location**, **Ease of Online Booking**, **Convenience of Departure & Arrival Times**, and **Cleanliness**
    - to supplement this particular analysis, I also found the average satisfaction scores for each of the airline factors
* The next thing I wanted to analyze was how the satisfaction scores for all the different factors varied by the 3 main Travel Classes: Business, Economy, Economy Plus; specifically, which factors were the majority of passengers in each travel class least satisfied with
  + According to the results:
    - **Business Class** passengers are least satisfied with the **Gate Location**
    - **Economy Class** passengers are least satisfied with the **Ease of Online Booking**
    - **Economy Plus** passengers are least satisfied with the **Check-In Service**
* I then performed the same analysis as the previous one, but this time broken down by Customer Type
  + - The Top 3 factors **Returning** passengers are least satisfied with are **Gate Location**, **In-flight Wi-Fi Service**, and **Ease of Online Booking**
    - **First-time** passengers are least satisfied with the **Seat Comfort** on the airplane

NOTE: the majority of First-time passengers (57%) flew Economy Class while the majority of Returning passengers (50%) flew Business Class

* I’d like to see if flight distance affects customer preferences or flight patterns?



* + - 10252/10327 (99%) of **Long-haul** passengers choose to fly **Business Class**
    - 33932/62167 (55%) of **Medium-haul** passengers choose to fly **Business Class**
    - 33825/57386 (59%) of **Short-haul** passengers choose to fly **Economy Class**
* Does travelling for business reasons mean you’re more likely to fly in Business or Economy Plus Class? How about travelling for personal reasons like vacation?

****

* + - 66% of passengers travelling for **Business** purposes choose to fly in **Business Class** while only 28% fly in Economy Class
    - 82% of passengers travelling for **Personal** reasons choose to fly in **Economy Class**
* Created a separate script named VIEWS which creates VIEWS for all query result tables that will be exported into Excel and used for visualizations in Tableau (16 total tables). I then wanted to export all the VIEWS into Excel and found an efficient method to do so.
  + Using an Open Database Connectivity (ODBC) driver, I was able to connect Excel to my MySQL server. From there, I could easily transfer any data from the server into any Excel workbooks. However, since I didn’t have a license key, only the first 3 rows of any data tables or VIEWS from the server were being transferred. So, I had to resort to the manual method of selecting all the views and individually copying/pasting the results into separate worksheets.

# Visualization in Tableau

* Before connecting to the Excel file data source, I found opportunities where individual sheets could be combined into a single sheet based on common fields. This would save some cluster when working with the data in Tableau

1. Connected to the Excel file which contained each VIEW’s result in separate sheets
2. Upon dragging/dropping all the sheets into the logical layer on the Data Source page, I was having some trouble because all the sheets’ tables didn’t have related fields. To resolve this issue, I created an ID field for every table in Excel, reconnected to the data source, and repeated the initial procedure
3. I’d like to see the average satisfaction score for all factors and the total # of lowest satisfaction score (1) for each factor in the same visual; specifically, two bar graphs on each side separated in the middle by a list of all the different factors
4. I worked on creating visuals for how the satisfaction rate varies by different classifications of passengers. Each of the visuals are contained in different sheets.
   1. I was having some trouble on how I wanted to visualize all this data. I tried simple horizontal bar graphs for each classification which gave a basic idea of the story, but something was missing. The bars were displaying percentage of passengers satisfied that fell into each category, but I wanted to be able to show what percentage of all passengers each category represented on the same bar. That way, it would tell a better story and one can easily identify how a particular category within a classification is represented as a whole; therefore, making it easy to find which areas need special attention. I found Bullet Graphs to be the best visuals for such a vision as they display how one primary measure’s performance compares to another measure’s.
      1. To be able to show the percentage of all passengers each category represented in a classification, I first needed to go back and edit all the sheets in the original Excel data source to include this information. Using the distribution of passengers for each classification from above, I was able to create the desired fields.
      2. Refreshing the data source in Tableau, I created the visual for only one classification first, Gender. I noticed that the information wasn’t being displayed properly. I found it had to do with the number formats. In Excel, the new field created in step (i) was defined as a number between 0 and 1. When imported into the Bullet Graph, I changed the format to a percentage to match the percent satisfied field. However, this didn’t work. I soon realized it was because even though all the percent satisfied fields were exported from MySQL in the form 00.00, they weren’t actually percentages and still represented numbers. So, I was basically doing a number to percentage comparison. I then went back to all the sheets in Excel and converted all percent satisfied fields into a number between 0 and 1 to maintain a consistent format with the % of total passengers fields.
      3. Refreshed the data source in Tableau again and created a Bullet Graph for only the Gender classification as a test. The data was properly displaying now, so I repeated the same procedure for all the other classifications.
5. Displayed distribution of lowest satisfaction score broken down by Travel Class in a stacked horizontal bar chart.
6. Used pie charts to show how passengers were distributed by Flight Distance and Customer Type
7. Created BANS, which include key metrics, to be displayed across the top of the dashboard
8. Combined visuals from steps 3-7 into a single dashboard and added various design and format elements to create an aesthetic appeal

# Strategy

* The three major areas to focus on in order to improve overall satisfaction with the airline are In-flight Wi-Fi Service, Ease of Online Booking, and Gate Location.
* Some of the largest disparities in % of total passengers and % satisfied exist in Type of Travel, Flight Distance, Travel Class, and Customer Type
  + To address low satisfaction in travelers flying for personal reasons (~1/3 of all passengers), focus should be placed on Economy Class travelers because a massive 82% of Personal-purpose passengers fly in this class. A couple of relatively inexpensive factors which would significantly improve satisfaction for these passengers are Ease of Online Booking and In-flight Wi-Fi service as they seem to be the most displeased with them. Also, keeping the Economy Class section of the airplane clean for these passengers can be achieved with little to no additional expenses and will help ensure their vacation, exploration of a new city/state, or family visit is off to a great start.
  + Almost half of all passengers fly distances less than or equal to 700 miles and the majority of them (67%) are not satisfied. These travelers typically fly Economy Class which means improving factors like In-flight Wi-Fi service, the Entertainment systems within airplanes, and overall cleanliness in the section will keep them happy (similar tactic to Personal-purpose passengers).
  + Business Class travelers will expectedly be more satisfied with their experience due to added benefits/services included in their ticket, but such passengers on Maven Airlines didn’t shy away from expressing any of their dissatisfactions. Although the Economy Class accounted for the majority of total lowest satisfaction scores (52%), Business Class passengers held the Top 4 highest response rates for any single factor amongst all 3 travel classes, signifying they really want their dissatisfactions to be known. Though their top factor of dissatisfaction (Gate Location) may not be feasible to improve upon, improving the Wi-Fi service onboard, Easing the process of Online Booking, and ensuring flights departure/arrive on time will improve their experience even more. Time is of essence for these passengers, which is why long walks between terminal gates and departure/arrival delays can cause them to become very frustrated and leave bad reviews.
    - On the other hand, Economy Class passengers show a more widespread distribution of factors they are displeased with. They are most concerned about the Online Booking experience; so, ensuring the Maven Airlines travel site is easy to navigate, responsive, and catered with the end-user in mind will significantly improve their satisfaction levels. Additionally, improving things like the Wi-Fi service and In-flight Entertainment - better selection of movies, songs, games - of the Economy section (already mentioned in previous plans) will have a great impact. Making small improvements in the Food & Drinks offered onboard (more variety or better quality), Cleanliness of section, Check-In Service (greeting customers with a smile and maintaining positive engagement), and Onboarding Service (better organization) can go a long way in bringing that satisfaction level up.
  + A whopping 82% of passengers are Returning customers, but nearly half of them are not satisfied. This area certainly deserves our attention. Over 90% of them are either Business or Economy Class passengers which means their dissatisfaction stems from the same factors that affect these travel classes. So, improving upon the In-flight Wi-Fi Service, Easing the process of Online Booking, reducing flight departure/arrival delays, and making small improvements in previously mentioned areas will help retain this huge customer base and avoid possibly losing many of them to competition in the next round of surveys.