6101 - Introduction to Data Science Midterm Project Proposal

Team Members

First Name	Last Name
Aaron	Yang
Jianjun	Gao
Luhuan	Wang

Possible Research Topics

The two main factors affect the final price of Uber and Lyft in Boston.

Smart Questions

- Is there any relationship between price and hour?
- Is there any other factors to affect the final price of Uber and Lyft?
- Do the weather factors, like temperature, wind, and sunset affect the final price?

Processes

1. Create a GitHub Repository;
2. Confirm an appropriate topic;
    - Understand the project purpose
    - Seek for a potential dataset/database/data sheet
    - Confirm the topic based on the dataset
3. Confirm the whole project processes;
4. Make a clear plan and timeline;

5. Prepare the first draft and slides;
6. Make a summary and final modification.

Possible Data

The data is from Kaggle.com. The data is about Uber and Lyft Dataset Boston, MA. This is a very beginner-friendly dataset. It does contain a lot of NA values. It is a good dataset if you want to use a Linear Regression Model to see the pattern between different predectors such as `hour` and `price`.

A really amazing part of this dataset is that I have included the corresponding `weather data` for that `hour` with a short summary of the `weather`. Other important factors are `temperature`, `wind`, and `sunset`.

The original data can be available on the Kaggle.com.

Link to GitHub Repo

DATS 6101- Midterm Project- Group 1

6101 - Presentation Preparation

Slides Outline

Brief

Introduction to the topic and team contributions

Why this topic?

1. **Ride-sharing Popularity**: With the increasing popularity of ride-sharing platforms like Lyft and Uber, understanding the determinants of pricing can be valuable for consumers, policy-makers, and urban planners.
2. **Easy Data Access**: It is easy to find the dataset related to Uber and Lyft
3. **Dynamic Pricing Models**: Both Uber and Lyft use dynamic pricing algorithms that consider various real-time factors. Unpacking these can provide a deep understanding of tech-driven pricing models, their effectiveness, and their impact on consumers.

4. **Comparative Analysis**: Since Lyft and Uber are the major ride-sharing platforms but have slightly different operational and pricing strategies, a comparative analysis can provide insights into how different business strategies play out in the same market.

Team Contributions

1. **Aaron**:

   Outline & Proposal

   Data Analysis (with Jianjun)

   Presentation

   Paper Co-author
2. **Jianjun**:
   - Data Collection
   - Data Analysis (with Aaron)
   - Presentation
   - Paper Co-author
3. **Luhuan**:
   - Data Summary
   - Slide Creation
   - Presentation
   - Paper Co-author

Introduction to the Background Knowledge

Data Science and Statistical Methods:

Data Collection & Cleaning

Exploratory Data Analysis (EDA)

Regression Model

Programming:

R

Python

Visualization:

Graphical Representation

The below is not in the slide:

- **Data Collection & Cleaning**: You might use web scraping tools or APIs provided by services like Lyft and Uber. Familiarize yourself with R packages like `rvest` for scraping and `tidyverse` for data manipulation.
- **Exploratory Data Analysis (EDA)**: Before diving into advanced analysis, understanding the basic trends, outliers, and patterns in your data is crucial.
- **Regression Analysis**: To study how different factors influence pricing, regression models (like linear regression) can be beneficial.
- **Machine Learning**: More advanced models like decision trees, random forests, or gradient boosting machines could be useful if you have ample data and want to predict pricing.

*

- **R**: As your primary language, you'll need to be familiar with various packages for data analysis, visualization (like `ggplot2`), and modeling (like `lm` for linear models or `randomForest` for random forests).
- **Python**: Even if R is your main tool, Python has powerful libraries like `pandas` for data manipulation and `scikit-learn` for machine learning.

- **Graphical Representation**: Use R's `ggplot2` or Python's `matplotlib` and `seaborn` to visualize trends and insights.
- **Interactivity**: If you want to create more interactive visualizations, consider using tools like `shiny` in R.

Introduction to the overall steps

Step No.	Phase	Tasks & Details
1.	Topic Decision	- Team Meeting - Choose a topic based on available data
2.	Data Collection	- Identify Data Sources - Gather data - Initial Data Cleaning
3.	Research & Background	- Literature Review - Industry Knowledge
4.	Data Processing & EDA	- Data Cleaning & Transformation - Descriptive Statistics - Visualization
5.	Data Analysis & Modeling	- Hypothesis Testing - Regression Analysis - Advanced Modeling - Model Evaluation
6.	Interpretation of Results	- Discuss Findings - Statistical Significance
7.	Visualization & Reporting	- Data Visualization - Write the Report/Paper
8.	Presentation Preparation	- Slide Development - Rehearsal
9.	Presentation & Feedback	- Deliver Presentation - Collect Feedback
10.	Project Conclusion	- Team Debrief - Documentation

Introduction to the data collection, mining, and analysis

Summary

Hour and Mean Price

Based on the graph you provided, which illustrates the relationship between the hour of the day and the mean price, we can make the following observations:

1. **Price Fluctuations Throughout the Day:** The price does not remain constant; it shows significant fluctuations as the day progresses. This is expected for services like Uber and Lyft since pricing can vary based on demand, traffic, or other dynamic factors.
2. **Possible Peak Hours:** There are noticeable peaks in the price at certain hours, which could correspond to increased demand. For instance, the first peak close to the 5th hour might indicate higher prices during morning rush hours when people are heading to work or school. The subsequent decrease might signify a lull during mid-morning.
3. **Midday Drop & Evening Rise:** The graph indicates a noticeable drop in prices around the middle of the day (approximately between the 10th to 15th hours) followed by a rise. This could indicate a lower demand during midday and a surge in the evening, perhaps when people are returning home or heading out for the evening.
4. **Late Night/Early Morning Pricing:** The graph starts and ends at a lower mean price, suggesting that prices might be lower during the late night and very early morning hours, potentially due to decreased demand.
5. **No Data Beyond Hour 20:** It seems the graph does not provide data for hours beyond the 20th hour (or 8 PM), so we can't make observations for late evening or nighttime.

Remember, while these are initial observations based on the graph, deeper insights might require further statistical analysis and contextual information, such as events happening in the city, promotions, weather conditions, etc. that could influence the demand and price of Uber or Lyft rides.

Boxplot of Prices by Cab Type

The boxplot you provided showcases the distribution of prices for different cab types. Let's analyze the visual data:

1. **Cab Type Price Range**:
   - **Black, Black SUV, Lux, Lux Black, Lux Black XL**: These types seem to be premium services with a higher median price, particularly Black SUV which has the highest median price. They also have a wider price range, indicating variability in the fare based on certain factors (e.g., distance, demand, time).
   - **Lyft, Lyft XL, UberX, UberXL**: These are standard or slightly upgraded services. They have a median price lower than the premium services but higher than the shared ones. Lyft XL and UberXL, being larger vehicles, tend to be pricier than their standard counterparts.
   - **Shared, UberPool**: These services are shared rides, which naturally have the lowest median prices due to the cost split between multiple passengers.
   - **WAV**: This stands for Wheelchair Accessible Vehicles. Its price range seems close to the standard Lyft and UberX services, but it's essential to consider that these types of services might be priced based on accessibility rather than luxury or vehicle size.

2. **Outliers**: The dots above the boxes represent outliers, or prices that fall outside the typical range for each cab type. It's worth noting that almost all cab types have outliers, indicating occasional rides that are much more expensive than usual.
3. **Interquartile Range (Box Size)**: The size of the box represents the interquartile range (IQR) which is a measure of statistical dispersion. A larger box means greater price variability. Premium services like Black SUV and Lux Black XL have wider boxes, indicating greater variability in their prices.
4. **Median (Line in the Box)**: The horizontal line inside each box represents the median price for each cab type. This gives a good indication of a 'typical' price, with half of the prices being above this value and half below.

From a practical perspective, if you're looking to choose a cab type based on price:

- For premium services with higher prices, consider **Black SUV, Lux, or Lux Black XL**.
- For standard services with moderate prices, consider **Lyft, Lyft XL, UberX, or UberXL**.
- If saving money is a priority and you're okay with sharing the ride with others, consider **Shared or UberPool**.
- For those needing accessibility features, **WAV** is the choice.

Remember, prices can also vary based on other factors like demand, location, promotions, and more. This graph gives a general overview, but actual prices can differ.

Thank You page

Presentation Order

1-4 Aaron Yang

5 Jianjun Gao

6 Luhuan Wang