TASK 2 — BUSINESS UNDERSTANDING

Project title: Analyzing how different factors affect Uber ride demand and patterns

Team members: Ralf Andreas Vendel, Georg Lumila

Github repository: https://github.com/GeorgLumila/IDS2025-E6-UberDataset.git

1. Identifying Business Goals

Background

The Uber dataset we selected has a variety of trip details, such as when a ride started and ended, the pickup and drop-off locations, whether the ride was business or personal, the distance traveled, and sometimes even the ride purpose. This information helps show how people move around at different times and in different areas. Because the dataset includes time-related fields and ride details, it helps us understand when people use Uber and why.

Business Goals

Our project is based on the three main goals:

Goal 1: When do people use Uber the most?

Analyze temporal patterns in ride demand identify peak hours, days of the week, and months with the highest number of trips.

Goal 2: Where are Uber rides most common?

identify which areas appear most often as the rides start or end points.

Goal 3: Can we predict Uber ride demand?

Build a predictive model to estimate future ride counts based on time and location.

Business Success Criteria

Our project will be considered successful if we can clearly:

- Show the main peak times for Uber usage
- Identify which locations are the most active
- Build a simple model that shows at least some predictability in ride demand

The goal is not to create a perfect prediction model but to demonstrate that the dataset contains patterns that can be understood and used.

2. Assessing the Situation

Inventory of Resources

We are using the Uber dataset available on Kaggle. For analysis, we will use Python libraries such as Pandas, NumPy, Matplotlib, and possibly scikit-learn for prediction. GitHub is used as our project repository. The team consists of two members (Ralf Andreas Vendel, Georg Lumila), and the work will be split evenly.

Requirements, Assumptions, and Constraints

Requirements:

- Understand temporal trends (hours, weekdays, months)
- Identify location based hotspots
- Attempt a basic prediction of ride demand

Assumptions:

- The timestamps are good to find meaningful patterns
- The location names are good enough to detect common areas
- Distance and purpose values are most of the time correct

Constraints:

- The dataset is relatively small.
- Some fields (like PURPOSE) might be missing frequently
- Location names may not always be standardized.

Risks and Contingencies

A possible risk is that some data might be missing or inconsistent, which could limit parts of our analysis. For example:

- If PURPOSE is missing too often, it may not be usable

- If location fields are messy, we may need to group or simplify them
- If predicting demand turns out unreliable, we might focus more on descriptive trends rather than predictive modeling

If issues arise, we will adjust the analysis to rely more on the most reliable data fields (mainly timestamps and location names).

Terminology

- **Pickup:** Where the ride starts
- **Drop-off:** Where the ride ends
- **Hotspot:** A location with a lot of activity
- **Demand:** How many rides happen at a certain time

Costs and Benefits

There are no money costs for this project. The only real cost is the time we spend cleaning and analyzing the data. The benefit is that we learn how Uber rides behave and get practice using the CRISP-DM process. The results could also be useful for anyone interested in how people move around, like drivers or planners.

3. Defining Data-Mining Goals

Data-Mining Goals

- Look at how many rides happen at different times.
- See which areas show up the most as starting or ending points.
- Try to put together a basic model that can roughly predict when people will need more rides.

Data-Mining Success Criteria

We succeed if we have:

- Clear plots and summaries showing the main patterns
- A reasonable (not necessarily perfect) prediction model

- Insights that match the business goals stated above

Task 3 — DATA UNDERSTANDING

Project title: Analyzing how different factors affect Uber ride demand and patterns
Team members: Ralf Andreas Vendel, Georg Lumila

1. Gathering Data

Outline data requirements

For this project, the main analytical goal is to understand patterns in Uber ride behavior and prepare the dataset for later modeling tasks. To achieve this, the required data must include detailed trip-level information. This includes precise timestamps for both the start and end of each trip in order to analyze travel frequency, duration, and temporal trends. Location data is also essential for identifying common routes. Trip distance in miles is needed for understanding travel intensity. Finally, categorical labels such as trip category (Business/Personal) and purpose (e.g., Meeting, Errand, Customer Visit) are needed to help us understand the riders behavior.

Verify data availability

The provided dataset, UberDataset.csv, includes 1156 rows and seven fields: START_DATE, END_DATE, CATEGORY, START, STOP, MILES, and PURPOSE. All required attributes identified in the project plan are present. Data absents is minimal for most fields (one missing value in END_DATE, CATEGORY, START, and STOP), except for PURPOSE, which contains 503 missing entries. This is expected because purpose information often relies on user input and may be optional in ride-tracking systems. All variables appear in readable formats, although timestamps must be converted to datetime objects.

Define selection criteria

Since every column contributes meaningful context for understanding user travel behavior, no variables are excluded at this stage. All 1156 rows are retained to preserve data completeness and prevent loss of temporal patterns. No geographic or temporal filtering is applied because the dataset already represents a certain period of 2016–2017 and focuses on trips in a consistent region.

2. Describing Data

The dataset consists of individual Uber rides taken over a multi-month period. START_DATE and END_DATE are expressed in string timestamp format, representing the exact moment each trip begins and ends. CATEGORY mostly contains Business, suggesting that the dataset may reflect travel logs for reimbursement or reporting. START and STOP contain city level location names, which allow geographic trend analysis. MILES is a numerical field representing the trip distance, ranging from short trips to long journeys. PURPOSE contains trip reasons such as Meeting, Meal/Entertainment, or Customer Visit. Many entries are missing, which will require attention in data preparation. Overall, the fields are relevant and consistent with the analytical goals.

3. Exploring Data

Initial exploration shows that timestamps span most of the year 2016 and parts of 2017. Formatting appears consistent, although change to datetime will be necessary. The distribution of MILES indicates that most trips fall between 1 and 10 miles, with some long-distance exceptions such as 60-mile travel. No zero or negative values are present, and no unrealistic outliers were noticed. CATEGORY is dominated by Business rides, with very few Personal entries, supporting the assumption that the data was collected for corporate purposes. PURPOSE has many different values, but lots of them are missing. Frequent origins and destinations include Fort Pierce and West Palm Beach, indicating localized travel patterns.

4. Verifying Data Quality

Data quality is generally acceptable for analysis, with no critical issues preventing progress. The primary quality concern is the large proportion of missing PURPOSE values. Because so many PURPOSE values are missing, it may not be useful for prediction and might need to be filled in, labeled as “Unknown” or left out of some analyses. The timestamps need to be converted into a consistent format, but the actual dates and times look correct. Minor missing entries in END_DATE, CATEGORY, START, and STOP affect less than 0.1% of the dataset and can be safely cleaned or removed. No duplicated rows or corrupted entries were detected during initial checks. Overall, the dataset is complete enough to analyze and prepare for later modeling steps.

Conclusion

This data understanding phase confirms that the Uber trip dataset is appropriate for the project's goals. The dataset has all the needed columns, is organized well, and only has a few quality issues that can be fixed later during data preparation. Now that we understand what each field means and how the data is structured, we can move on confidently to cleaning and preparing the data.

TASK 4

1. **Data Cleaning & Preparation 4/4h**
   - ○ Convert timestamps, fix missing values, standardize location names, and create new features (hour, weekday, month).
2. **Exploratory Data Analysis (EDA) 9/9h**
   - ○ Analyze temporal patterns, distances and identify high-frequency pickup and drop-off locations using visualizations.
3. **Predictive Modeling 8/8h**
   - ○ Build a simple model to estimate ride demand based on time-related features.
4. **Interpretation of Results 5/5h**
   - ○ Relate analytical findings to the business goals (peak times, hotspots, predictability) and evaluate model usefulness.
5. **Presentation and Poster 4/4h**
   - ○ Compile results into a clear report following the CRISP-DM structure and create a poster summarizing the insights we gathered.

For analysis, we will use Jupyter Notebook, Python libraries such as Pandas, NumPy, Matplotlib, and scikit-learn for prediction.