DS 5110: Traffic Congestion Prediction

Adit Vakil, NUID: 002539939 | Jainil Desai, NUID: 003152660 | Kalpan Shah, NUID: 002563256

1 Project Kickoff

What are the specific goals and expected outcomes of this project?

The primary goal of our project is to analyze traffic congestion patterns, particularly in Boston, and provide actionable insights. The following are the expected outcomes:

- Identifying peak traffic congestion times and their locations.
- Developing and using a model that predicts travel time between two points (A to B) under various conditions.
- Quantifying the correlation between weather conditions and traffic slowdowns.
- Comparing Boston's traffic patterns to global benchmarks.

How can we clearly define the project scope?

- **Acquiring Data:** Utilizing open, free data sources like Kaggle, the Boston Open Data Database, MassDOT traffic data, and the Google Traffic APIs (free tier).
- **Storage of Data:** Design and implement an SQL database to store location data, timestamps, traffic speed, and weather conditions.
- **Data Analysis:** Focusing on the Boston metropolitan area.

What are the key deliverables for each project phase?

- **Phase 1: Data Acquisition & Database Design:** A data source will be finalized, and a complete SQL database schema with loaded tables will be created.
- **Phase 2: Data Cleaning & Preprocessing:** A documented data cleaning pipeline (e.g., Python script) will be created and used that handles missing values, merges time-series sources, and encodes categorical data.
- **Phase 3: Analysis & Reporting:** A series of SQL analytical reports (as specified in iteration 1 feedback) and a final report summarizing key findings.
- **Phase 4: Predictive Modeling (Initial):** A baseline predictive model for congestion and a report comparing predicted vs. actual output.

What major milestones and deadlines should be established to track progress?

- **Milestone 1:** Database schema being finalized and data sources confirmed.
- **Milestone 2:** Data acquisition and database population complete.
- **Milestone 3:** Data cleaning pipeline developed and executed.
- **Milestone 4:** Advanced SQL reports (8 queries) completed.
- **Milestone 5:** Final report and presentation of insights.

Do the team's current capabilities align with these objectives, or are there gaps that need to be addressed early?

Yes, the team's current capabilities in data analysis, SQL, and Python align with the project objectives.

Is there an existing dataset available for this project, or is no dataset required for the current iteration?

Yes, existing datasets will be used. We will use data from open and free platforms, which include Kaggle, Boston Open Data Portal, MassDOT traffic data, and the free tier of Google Traffic APIs, as specified in the iteration 1 feedback.

2 Team Discussions

What core skills does each team member contribute?

Each team member possesses a strong foundation in core data science skills, allowing collective ownership and supervision of different project modules, including Exploratory Data Analysis (EDA), Python scripting for data cleaning, and SQL database management.

How will each member's expertise support specific tasks or components of the project?

The team operates on a collaborative model. Members with more extensive experience in specific areas (e.g., advanced SQL, time-series analysis) will lead those tasks while also mentoring other members. This ensures best practices are followed and facilitates knowledge sharing, allowing all members to contribute effectively.

Are there missing skills that could create challenges or delay completion?

No, the team's combined skillset covers all requirements of the project as defined in the current scope. We don't think there are any skill-based gaps that would create challenges.

What tools and technologies does the team already have experience with, and what must be learned?

The team is proficient with the primary tools for this project, including Python (with libraries like Pandas, NumPy, etc) and SQL. While we have foundational experience with geospatial data, this project will be an opportunity to strengthen our skills in optimizing geospatial queries and handling large-scale, location-based datasets.

Based on the project's needs and the team's background, which programming languages and platforms should be used?

Based on the project's requirements for data cleaning, analysis, and database management, the primary technologies will be:

- **Python:** For all data preprocessing, cleaning, and merging tasks.
- **SQL:** For database creation, data storage, and executing all analytical reports.

We will use a relational database platform such as PostgreSQL, which offers robust support for geospatial data.

3 Skills and Tools Assessment

Are there external resources that can assist in areas where the team lacks expertise?

Yes. While the team is confident in its core abilities, we will look to leverage external resources for guidance. This includes our Professor and TAs for guidance, as well as referencing already built models and published reports on similar traffic analysis projects for best practices.

Which tools, frameworks, and libraries best fit the project's scope and objectives?

- **Database:** PostgreSQL, as its PostGIS extension provides robust support for geospatial queries.
- **Data Handling:** Python, using pandas for data manipulation, numpy for numerical operations, and psycopg2 to interface with the database.
- **Data Acquisition:** Python libraries such as requests will be used to interact with the Google Traffic APIs and other web portals.
- **Future Modeling:** scikit-learn will be used for developing baseline predictive models.

How can we ensure all team members are comfortable and proficient with the selected tools?

The team is already comfortable and proficient with the core technologies we selected (Python and SQL). This allows us to focus directly on project implementation rather than starting from scratch.

Have we assigned tasks strategically based on individual strengths, and is everyone clear on their roles and responsibilities?

Our approach is balanced. While initial tasks are assigned based on known strengths to ensure project momentum, all tasks are distributed to provide every member with an equal opportunity to learn and grow. This ensures every member becomes well-versed in all project concepts, from data cleaning to advanced SQL reporting

4 Initial Setup

What development environment is required for this project?

- **Python:** We will be using Anaconda distribution for package management.
- **IDE:** Visual Studio Code or JupyterLab for script development and analysis.
- **Database:** A local or cloud-hosted instance of PostgreSQL.
- **Version Control:** Git will be used for version control.

Has version control been configured successfully, and does everyone have access to the repository?

Yes, a Git repository (on GitHub) has been created, and all team members have access to it.