NYC Taxi Data Analysis

Team 30

Rishitha Komatineni | Suchitra Yechuri | Kartheek Bellamkonda | Chetan Reddy Bojja

Motivation/Introduction

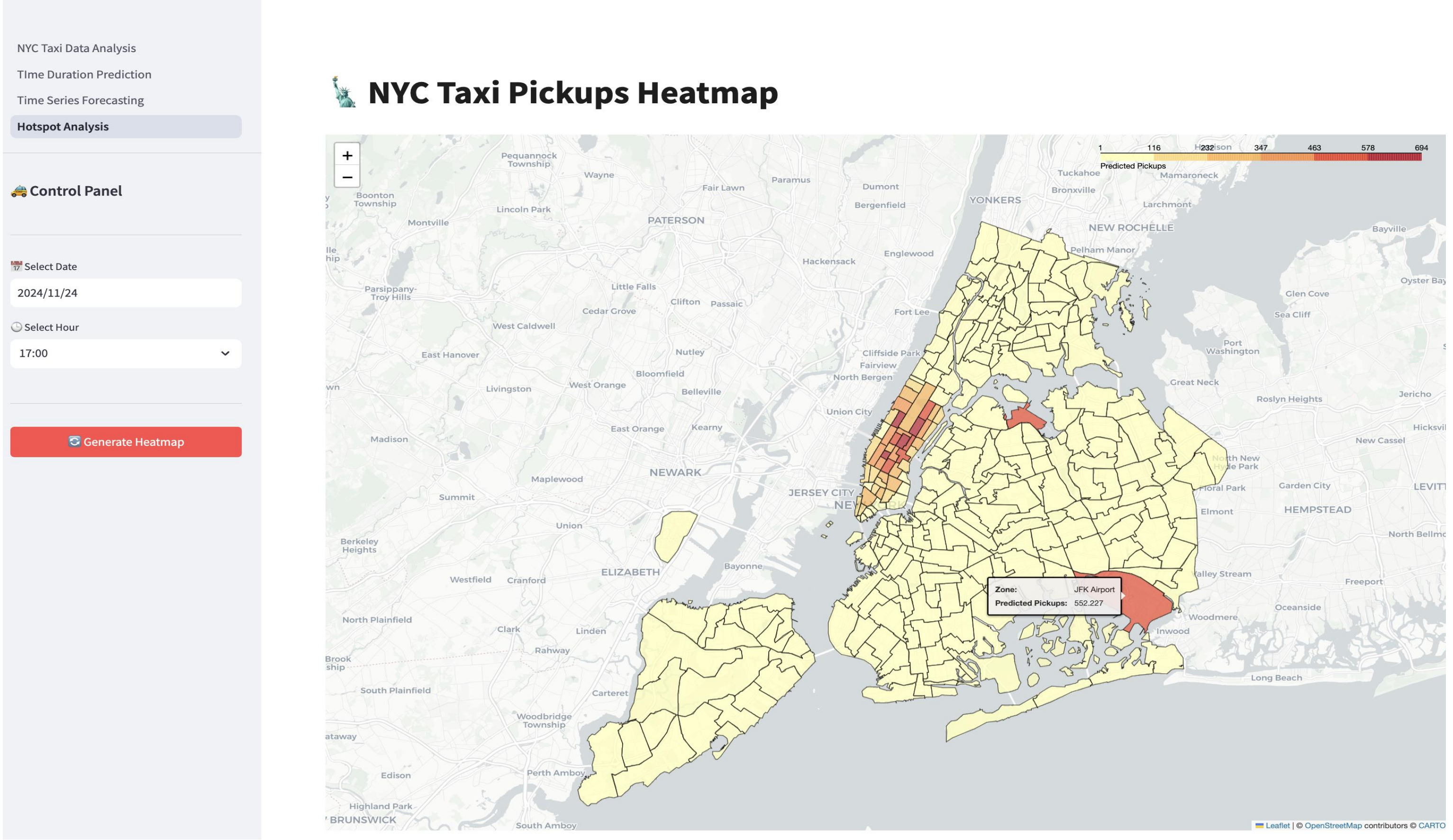
Managing NYC's vast taxi network is challenging due to fluctuating demand and inefficiencies in resource allocation. Accurate predictions of trip durations, demand trends, and high-demand hotspots can optimize taxi operations, reduce passenger wait times, and improve urban mobility. To address these challenges, we will develop an interactive visual dashboard to enable data-driven decision-making for taxi management.

Data

The NYC Taxi dataset was sourced from Kaggle, a reliable platform for public datasets. The dataset contains approximately 1.5 million records, including details such as pickup/drop-off times, locations, trip distances, and passenger counts. It is temporal, with a size of around 6.7GB, and provides rich insights for spatial and temporal analysis.

Approaches

- Pickup HotSpot Analysis:
  - Utilized Decision Tree to identify high-demand areas, visualized dynamically through heatmaps, highlighting demand variations over NYC zones.
- Passenger Forecast:
  - Employed the SARIMAX model to analyze time-series pickup data and forecast daily pickups using dynamic time-series graphs.
- Trip Duration Prediction:
  - Leveraged XGBoost to predict trip durations based on pickup/drop-off locations and time, with an NYC map interface for user input.
- We provide a unified, interactive dashboard that integrates advanced models to optimize taxi operations and enhance decision-making.



NYC Taxi Pickups Heatmap

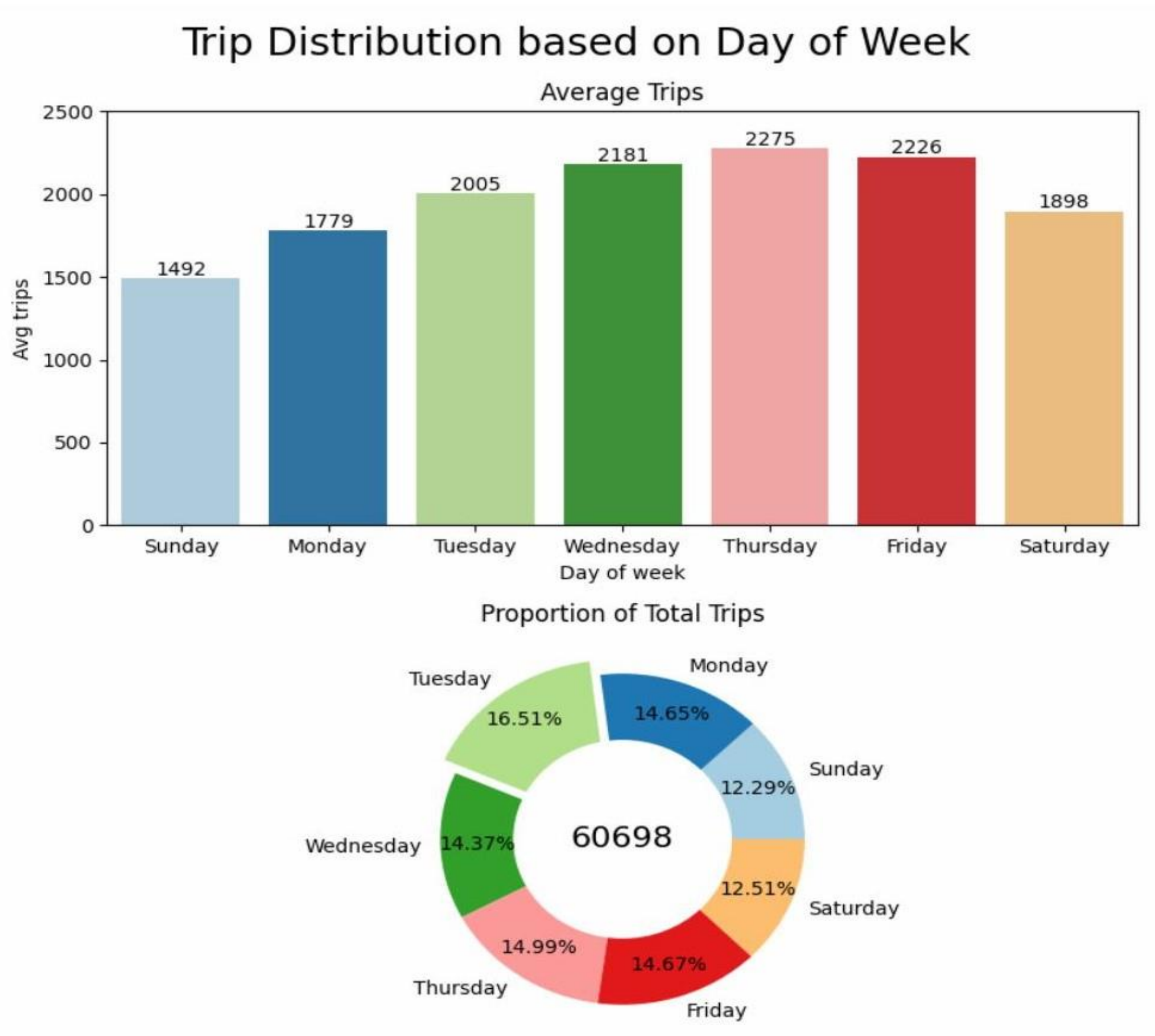
Control Panel

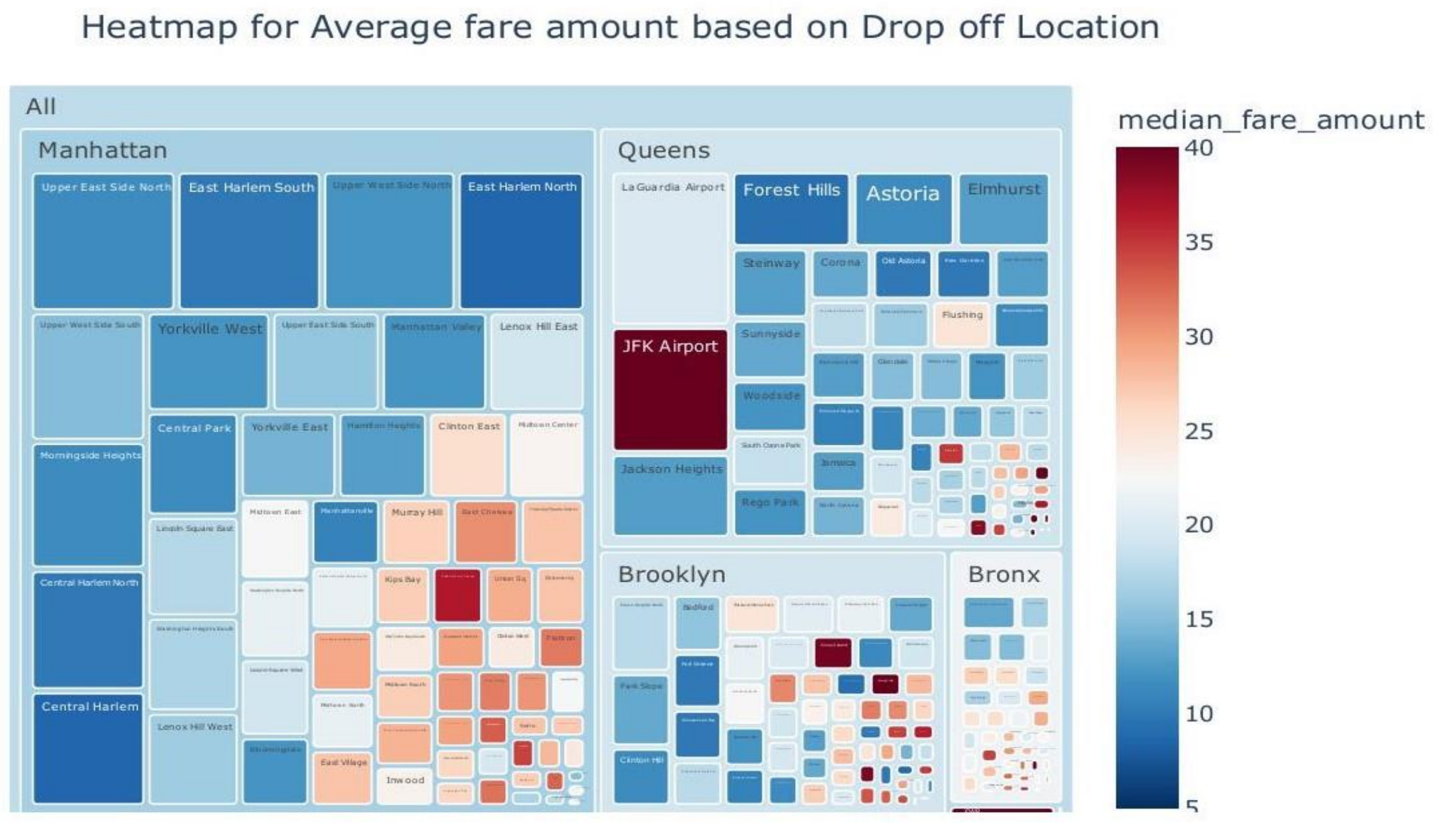
Select Date: 2024/11/24

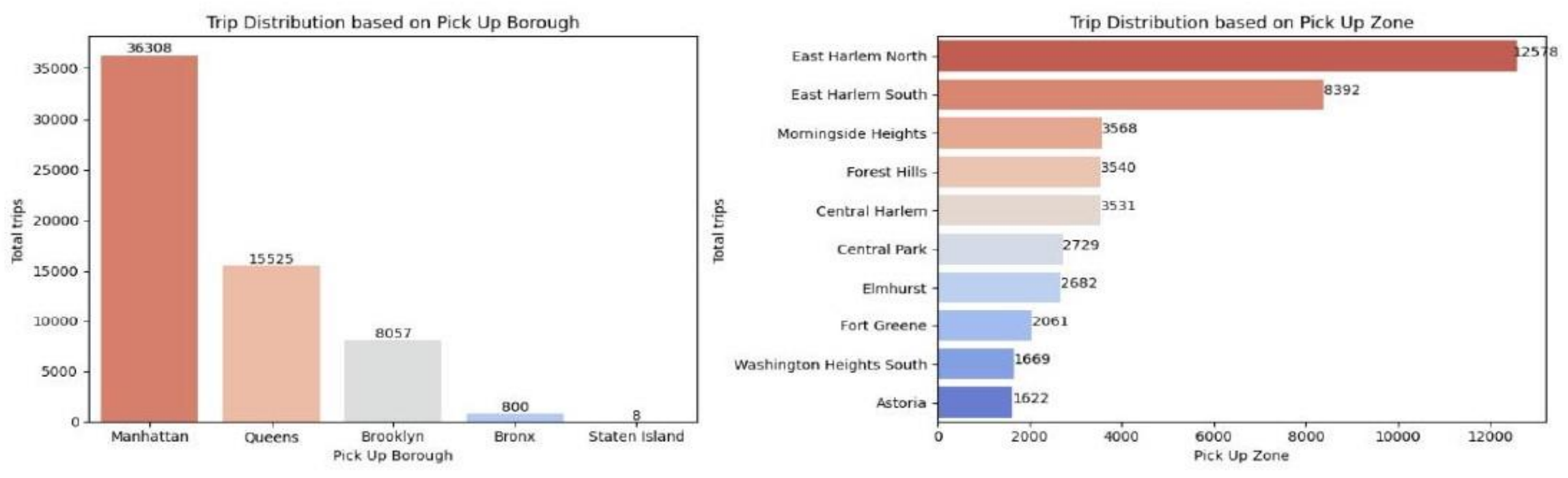
Select Hour: 17:00

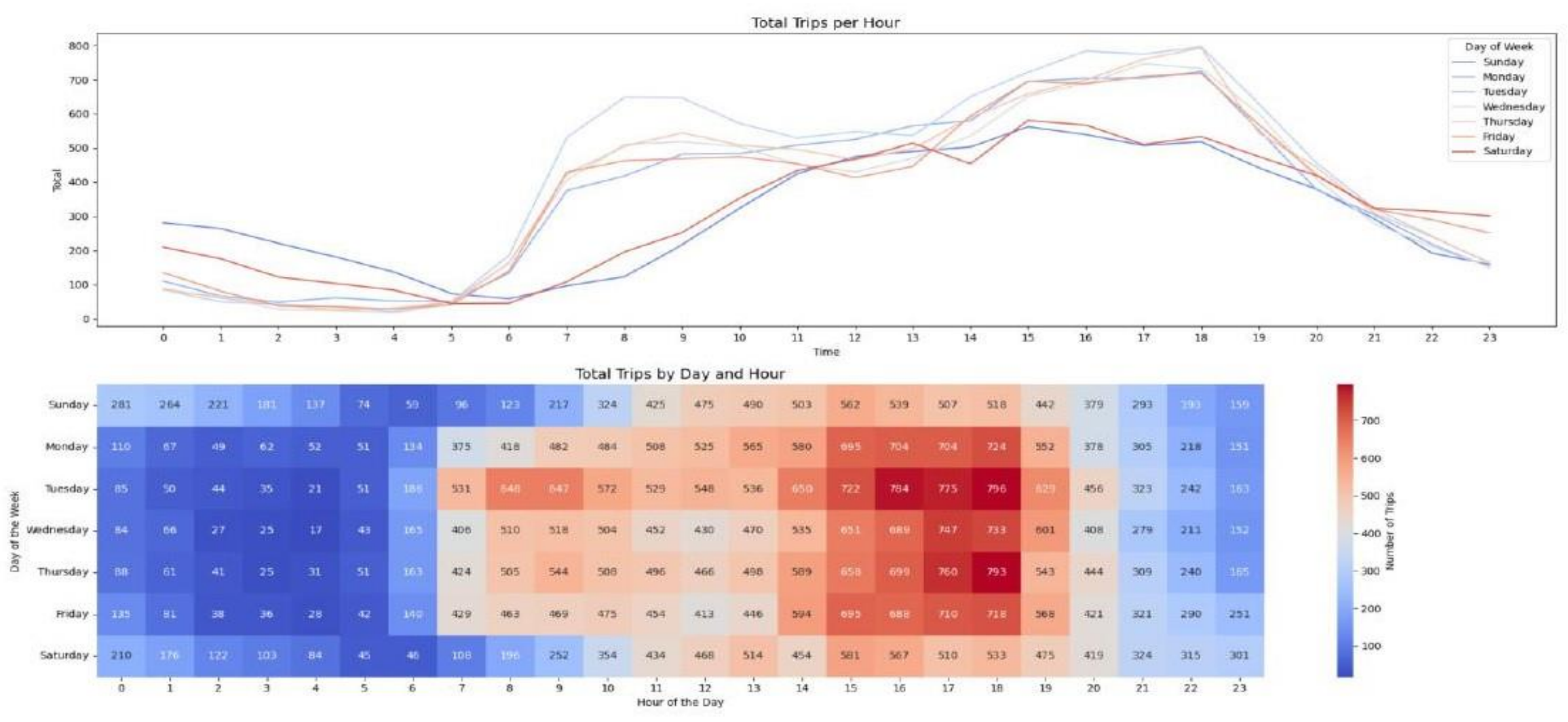
Zone: JFK Airport

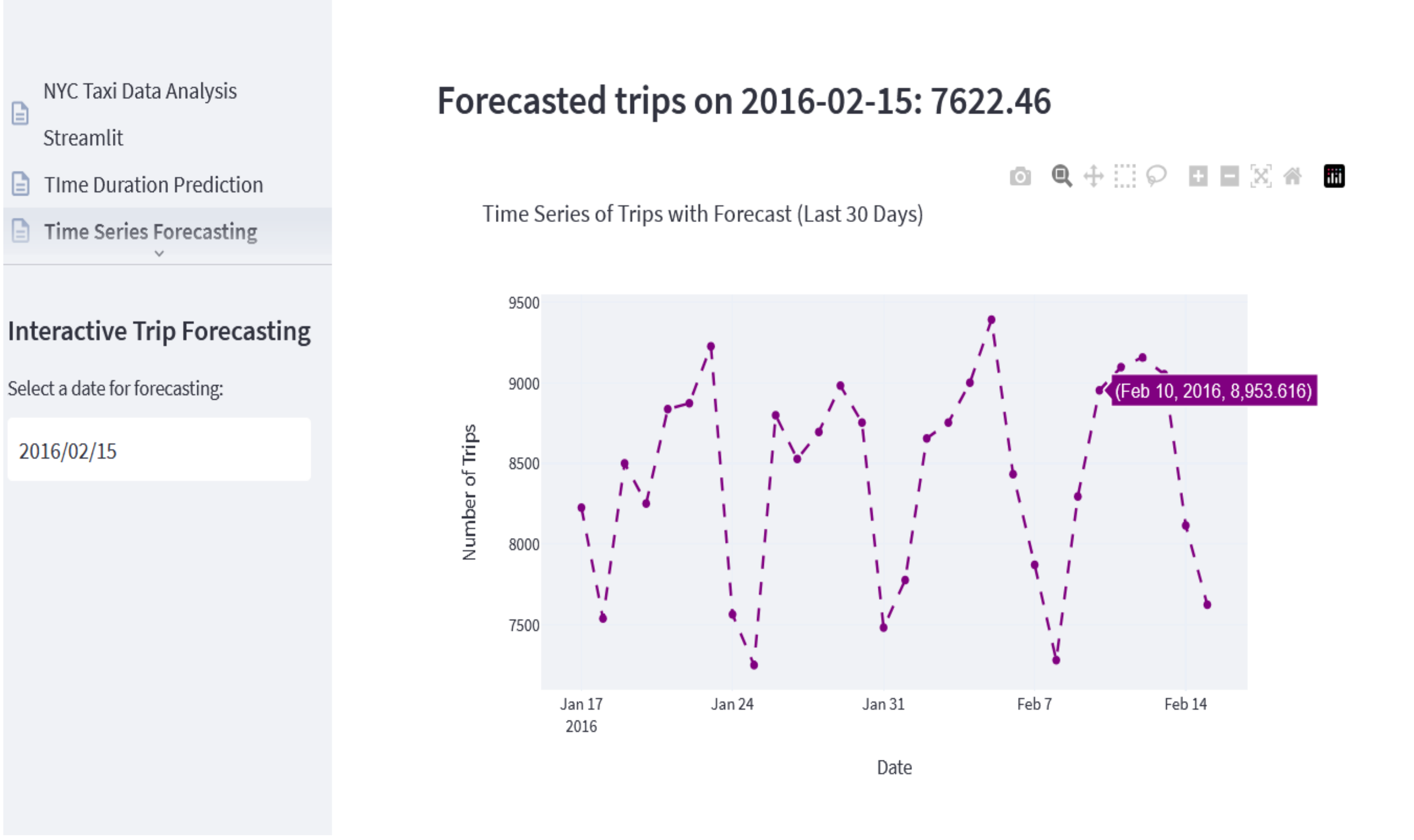
Predicted Pickups: 552.227







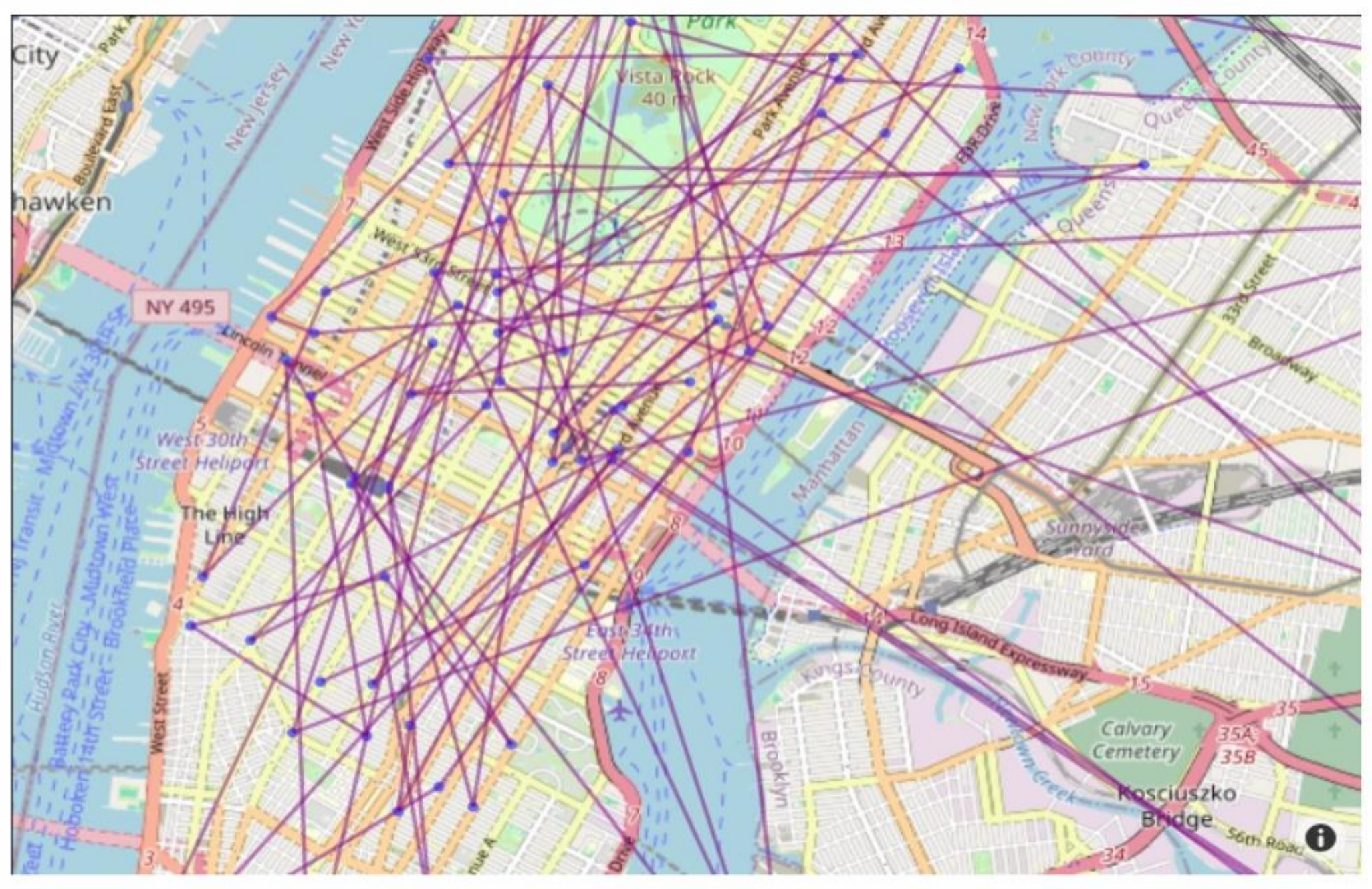




Forecasted trips on 2016-02-15: 7622.46

Interactive Trip Forecasting

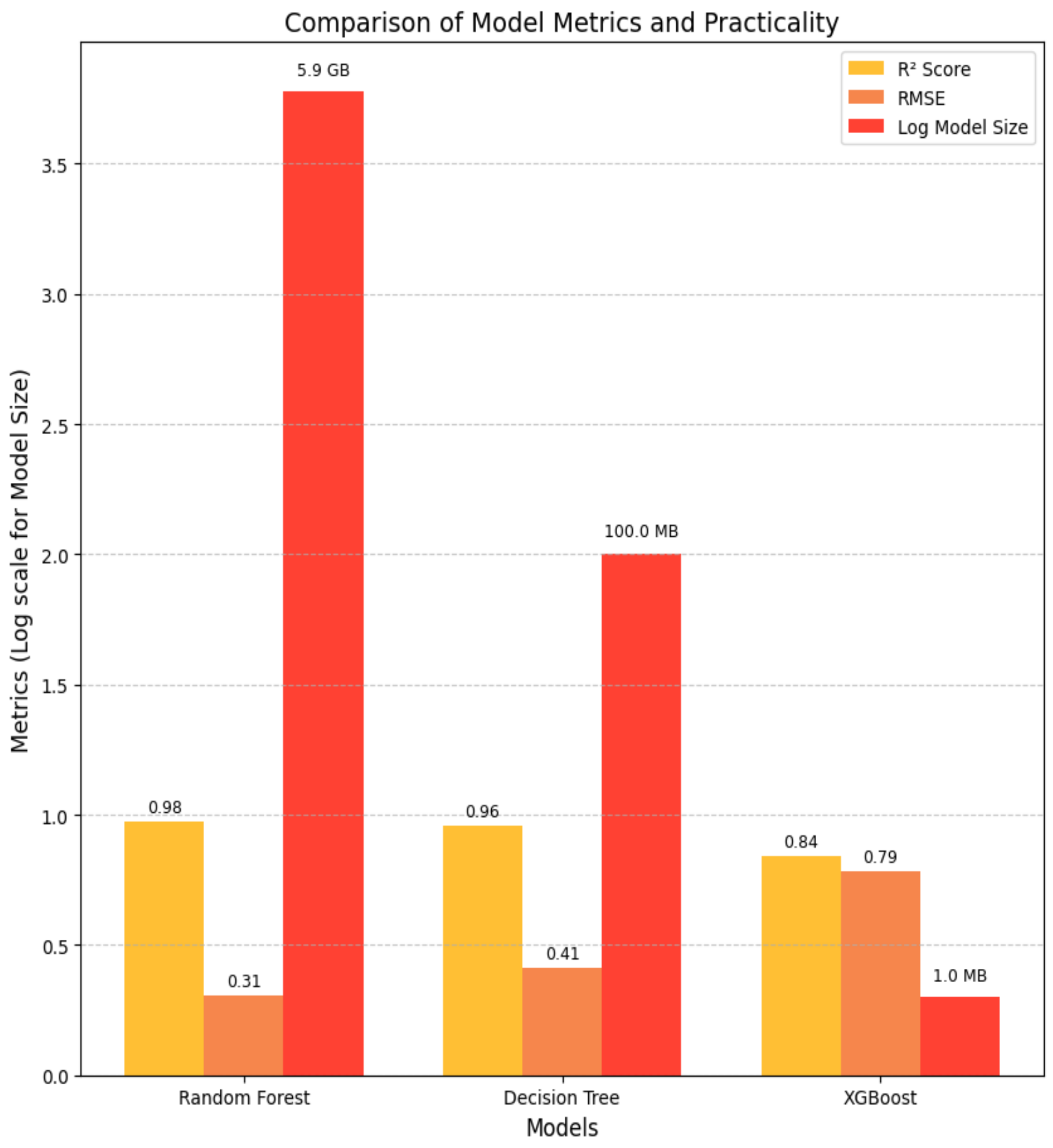
Select a date for forecasting: 2016/02/15

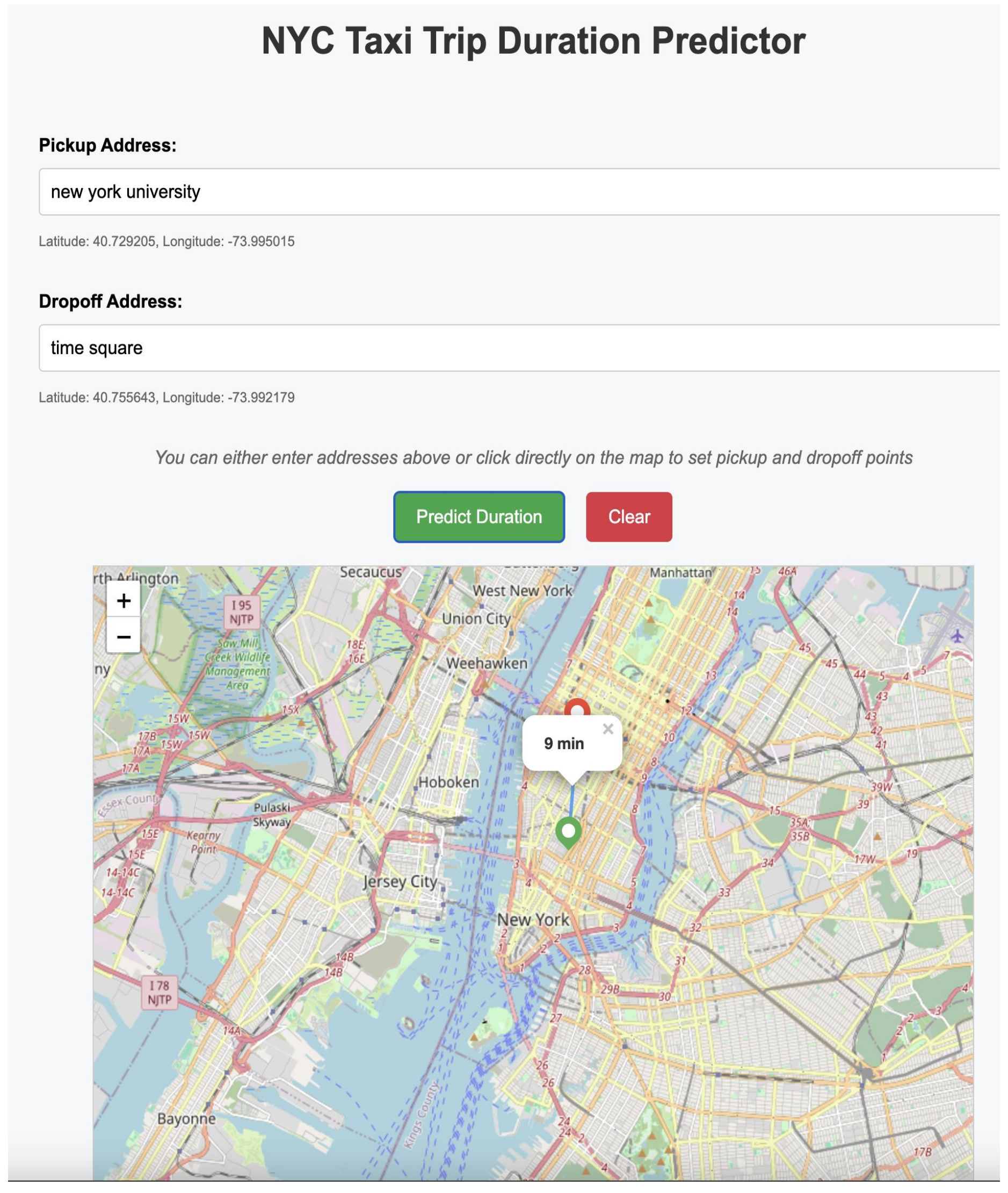


Interactive map showing 80 trips with their duration, distance, and timing patterns

Experiment and Evaluation

- **Evaluation**: Approaches were assessed using metrics like R², RMSE, and RMLSE for predicting taxi pickups and trip durations. User feedback evaluated dashboard usability for intuitive demand visualization.
- **Results**: Random Forest achieved the best pickup prediction (R²: 0.976, RMSE: 0.307); however, its 6GB model size was impractical for real-time systems. Decision Tree, with a much smaller size (100MB), was used for deployment (R²: 0.957, RMSE: 0.411). XGBoost excelled in trip duration prediction (R²: 0.72, RMLSE: 0.39).
- **Comparison**: Random Forest outperformed Decision Tree and XGBoost in pickup predictions, but Decision Tree was selected due to its practicality. XGBoost surpassed Random Forest in handling trip duration outliers.
- **Insights:** The dashboard provided actionable, user-friendly insights through heatmaps and statistical summaries.





NYC Taxi Trip Duration Predictor

Pickup Address: new york university

Latitude: 40.729205, Longitude: -73.995015

Dropoff Address: time square

Latitude: 40.755643, Longitude: -73.992179

You can either enter addresses above or click directly on the map to set pickup and dropoff points

Conclusion

Our integrated approach combines trip duration prediction, hotspot analysis, and demand forecasting with interactive visualizations. This enables efficient taxi allocation, reduces idle times, and improves passenger satisfaction. The dashboard empowers decision-makers with actionable insights, offering a comprehensive tool for optimizing urban mobility and addressing fluctuating demand.