Blue Bikes Final Report

Vita Khan, Quinn Reilly, Aarushi Attray

DS3000

Github repo: https://github.com/vitakh/DS3000-Project

Dr. Mohit Singhal

Abstract

This project aimed to predict Blue Bike-sharing trip durations using machine learning (ML) models based on the time of day, user demographics, and station locations. Linear Regression proved to as the best-performing, though it explained less than 6% of the trip duration variability, highlighting significant limitations. Polynomial Regression and enhanced models with interaction terms performed even worse, emphasizing the need for more complex approaches and the challenges inherent in real-world predictive modeling. Despite efforts to address multicollinearity using Principal Component Analysis (PCA), the results remained unsatisfactory. The analysis depicted consistent violations of critical assumptions, including linearity, independence, and constant variance, suggesting that trip durations are influenced by factors not captured in our cleaned dataset. Overall, our findings accentuate the challenges of predicting real-life trip durations and the need for expanded datasets and more advanced modeling techniques for future research.

Introduction

Urban traffic congestion poses a growing challenge, driving the need for sustainable alternatives like bike-sharing programs, which offer an eco-friendly and efficient travel method. Analyzing Blue Bikes (Metro Boston) system data can uncover valuable insights into how bike usage patterns vary across different times of day, locations, and user demographics. This analysis will focus on understanding peak usage times, identifying high-traffic stations, and exploring how trip duration varies based on user type, time of day, and station locations. Additionally, we aim to predict trip duration through machine learning models, which can optimize bike availability and improve system efficiency. The key questions are as follows:

1. Are there peaks in bike usage during certain times of the day?
2. Which stations have the highest traffic, and how does the distribution of bike usage vary geographically?
3. How does bike trip duration vary by user type (eg. age, gender, membership status)?
4. Can we predict the trip duration based on factors such as the time of day, user demographics, and starting/ending stations?

Data Description

Summary of the Data Processing Pipeline:

1. Web scrape to get the raw data
2. Clean the data to prepare the data frame for visualization and analysis

3. Visualize using plotting libraries, such as Seaborn, Plotly, and Matplotlib
4. Implement Machine Learning models and perform regression analysis

To begin analyzing the data, we began by acquiring the Blubikes dataset through webscrapingTo process the data, we first acquired the Bluebikes datasets. This process involved collecting the raw data, performing data cleaning, and saving the cleaned datasets as .csv files, which were subsequently imported into a Jupyter Notebook. The cleaning process entailed removing invalid values, such as NaN, n/a, and 0 (where applicable), addressing missing or inconsistent values across key columns like start station ID, end station ID, and bikeId, converting starttime and stoptime columns to DateTime format, and excluding unanalyzable records, such as trips with negative tripduration.

Next, we focused on addressing our key questions. This involved extracting the hour and day of the week from the starttime column and calculating users' ages based on their birth year to facilitate an analysis of usage patterns. Since the data spanned multiple files, we combined them using common identifiers, such as start station ID and end station ID, to create a unified dataset. Further analysis included generating basic statistics and creating visualizations, such as time series plots to identify peaks in bike usage and geographical heatmaps (using Seaborn, Matplotlib, and Plotly) to visualize station traffic based on start and end station names. These steps provided insights into peak usage times, station demand, and trip duration patterns across various user demographics. Finally, the cleaned data was prepared for machine learning by selecting relevant features for predictive blue bike trip duration modeling. Due to the large amount of data in all our cleaned CSV files, we selected two datasets representing two distinct seasons—summer and winter—from random months (January and July). This approach allowed for possible analysis of seasonal usage patterns while ensuring smoother graphing processes.

Head of cleaned January data (csv_files/cleaned_jan_202401-bluebikes-tripdata.csv):

	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual	tripduration
0	D2F4A4783B230A84	electric_bike	2024-01-31 12:16:49	2024-01-31 12:21:02	Ames St at Main St	M32037	Central Square at Mass Ave / Essex St	M32011	42.362357	-71.088163	42.365070	-71.103100	member	4.216667
1	D305CEFFD4558633	classic_bike	2024-01-12 08:14:16	2024-01-12 08:19:48	Ames St at Main St	M32037	Central Square at Mass Ave / Essex St	M32011	42.362500	-71.088220	42.365070	-71.103100	member	5.533333
2	02009BB4EBA0D1F6	electric_bike	2024-01-29 15:00:05	2024-01-29 15:05:47	One Memorial Drive	M32053	Kennedy-Longfellow School 158 Spring St	M32065	42.361697	-71.080273	42.369553	-71.085790	member	5.700000
3	04C230C1C39071F7	classic_bike	2024-01-09 16:33:40	2024-01-09 17:00:41	Ames St at Main St	M32037	Brookline Town Hall	K32005	42.362500	-71.088220	42.333765	-71.120464	member	27.016667
4	CEAFE67E28B43852	classic_bike	2024-01-23 10:19:21	2024-01-23 10:31:39	Mass Ave T Station	C32063	Chinatown T Stop	D32019	42.341356	-71.083370	42.352409	-71.062679	member	12.300000

The example of cleaned data for January 2024

Methods

To prove the soundness of our chosen ML methods, we carefully considered the mathematical foundations, underlying assumptions, and potential pitfalls that could impact our project. Our primary research question asked, "Can we predict trip duration based on factors such as the time of day, user demographics, and starting/ending stations?" Based on this, we initially developed models using Linear Regression for three features and Polynomial Regression for one feature to explore the relationships in our dataset. For the first ML model, we employed Linear Regression, which estimates the relationship between independent variables—such as time of day, user demographics, and starting/ending stations—and the dependent variable, trip duration. This method is widely used due to its simplicity and interpretability. However, Linear Regression relies on several critical assumptions: linearity of the relationship between predictors and the response variable, independence of residuals, constant variance of residuals (homoscedasticity), and normality of residuals. However, upon evaluating our dataset, we observed significant violations of these assumptions, including nonlinear relationships and inconsistent residual patterns, limiting the accuracy and reliability of the model's predictions, as detailed in the Results section.

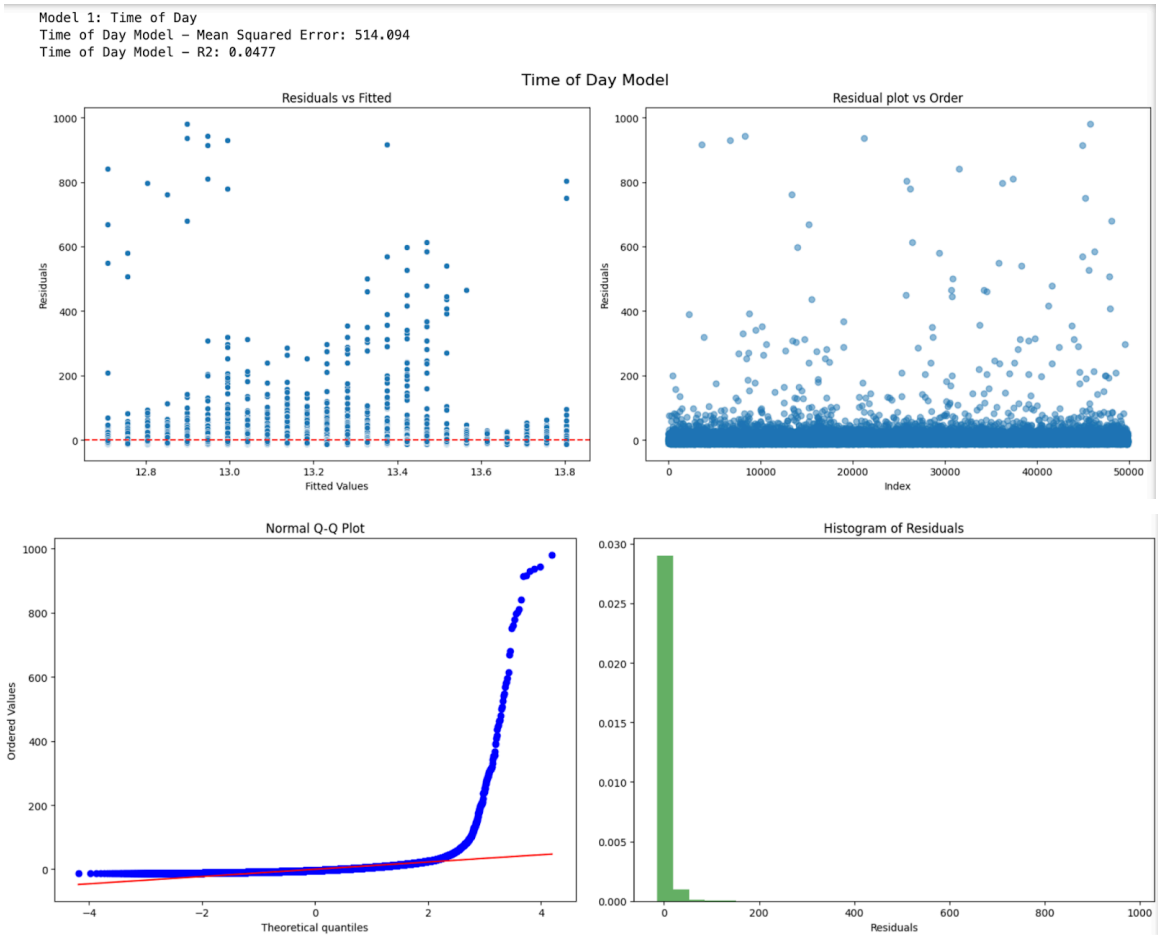
To address potential nonlinear relationships, we implemented Polynomial Regression for our second ML model, focusing on the time of day feature. We proceeded with two polynomial regression models: a basic model and a more complex model incorporating interaction terms, such as membership type and bike type. While the model with interaction terms demonstrated slightly better performance, neither polynomial regression model outperformed the Linear Regression model significantly, underscoring the limitations of polynomial regression in efficiently modeling the complexities of our dataset (Please refer to the Results section to see the graphs and values).

Recognizing these challenges, we incorporated PCA as our final ML model to refine our approach. Performing PCA is helpful when dealing with multicollinearity, as it allows us to keep as much information about the X features before using them to predict y or tripduration. By applying PCA before rerunning our regression analysis, we retained the maximum variance in the data while simplifying the relationships among predictors. For this revised analysis, we focused on key features, including time of day (hour), member_casual, and rideable_type.

Results

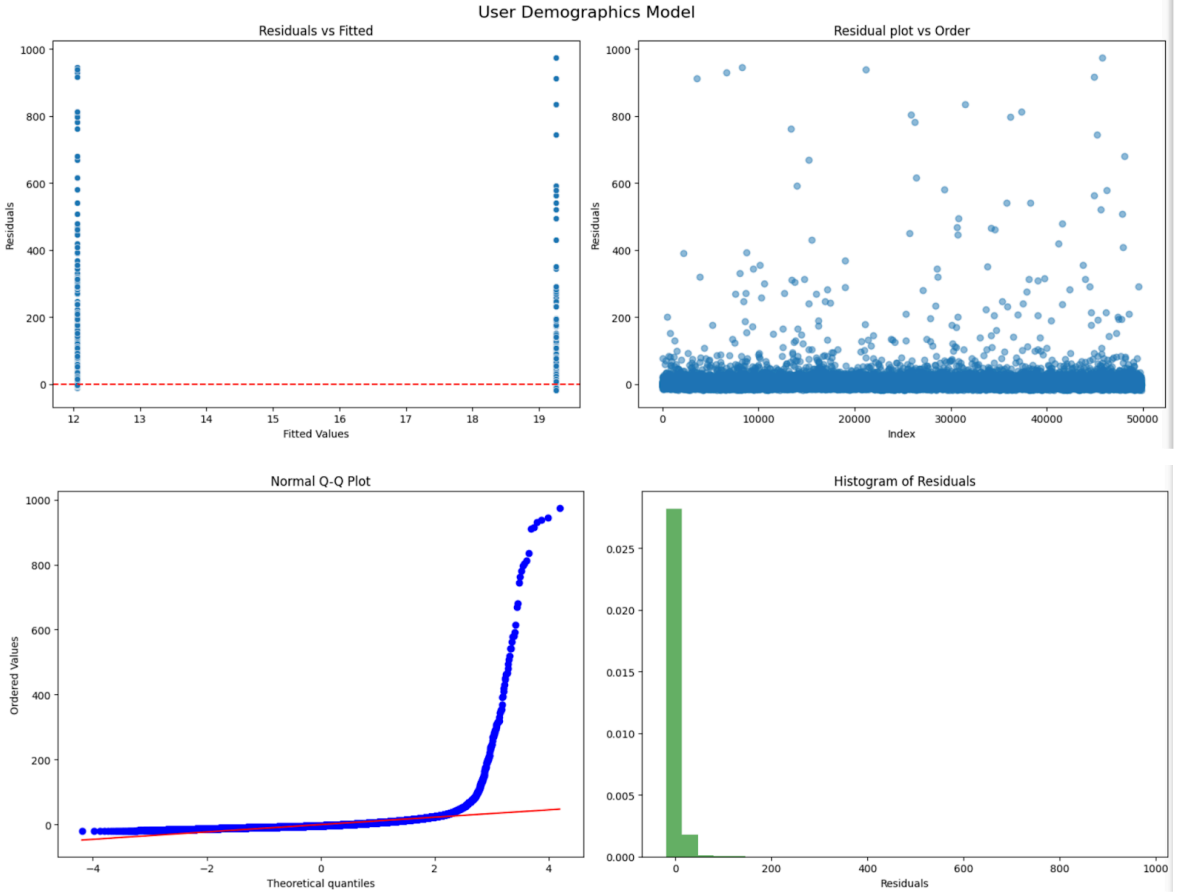
Firstly, we analyzed **Model 1: Time of Day** using a Linear Regression model. The calculated Mean Squared Error (MSE) was 514.094, and the R^2 value was 0.0477. These results indicate that while the MSE was lower than those of the Polynomial Regression models, the R^2 value was slightly lower than the second Polynomial Regression model. A closer examination of the

Residuals vs. Fitted plot revealed violations of the linearity and constant variance assumptions, as the residuals failed to follow a linear trend around zero and displayed multiple outliers. The Residuals vs. Order plot further suggested a potential violation of the independence assumption, with a concentration of data points near y=0 and several outliers. The Histogram of Residuals and the Q-Q plot showed significant right skewness, confirming a lack of normality and thus a violation of the normality assumption.

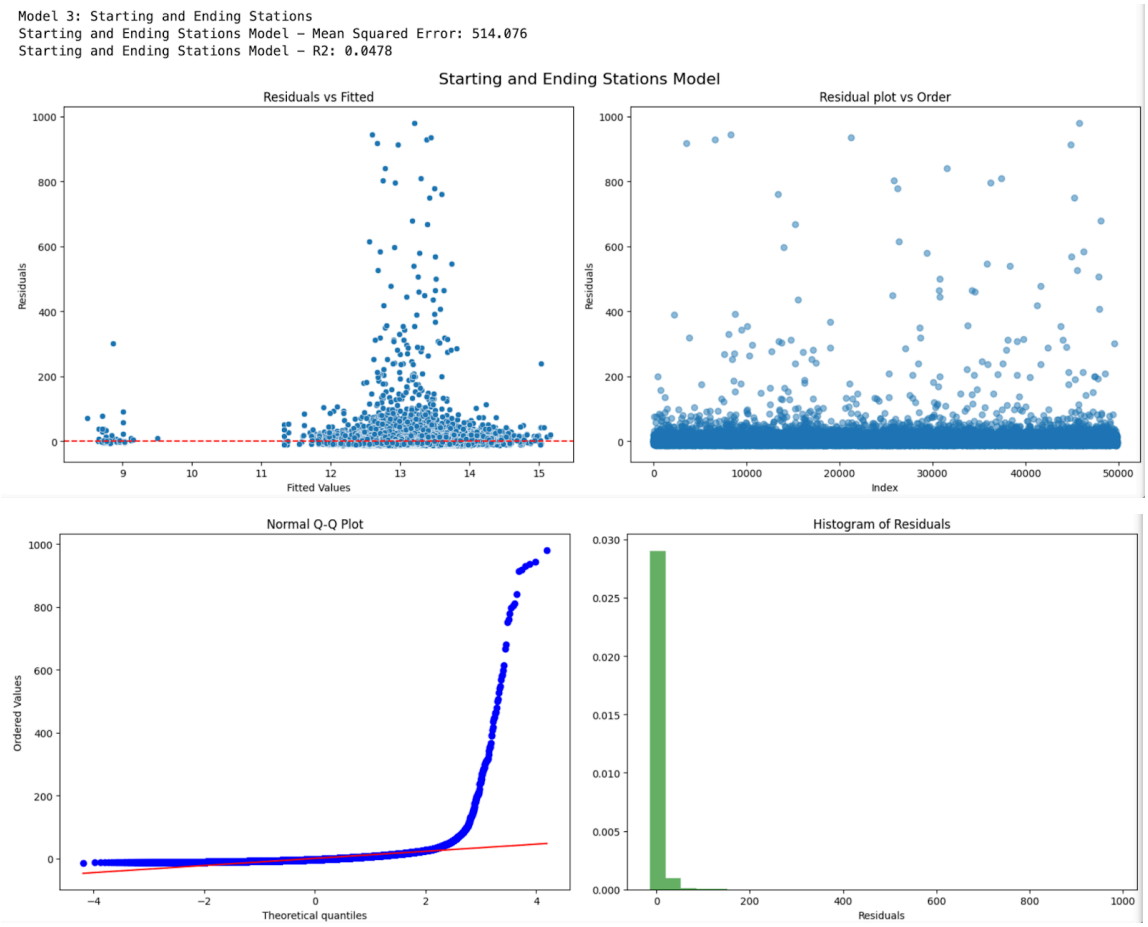


Secondly, we analyzed **Model 2: User Demographics** using another Linear Regression Model. This model achieved an MSE of 508.001 and an R^2 value of 0.059, the highest R^2 among all tested models. Despite the improved metrics, the model still violated key assumptions. The Residuals vs. Fitted graph demonstrated an unequal distribution of data points around the y=0 line, with noticeable gaps forming vertical patterns, indicating violations of linearity and constant variance. The Residuals vs. Order plot also suggested a potential independence violation due to the clustering of data points above the y=0 line. Lastly, the Histogram of Residuals and the Q-Q plot showed right skewness, highlighting the absence of a normal distribution.

```
Model 2: User Demographics
User Demographics Model - Mean Squared Error: 508.001
User Demographics Model - R2: 0.059
```



For **Model 3: Starting and Ending Stations**, the results were similar to Model 1, with an MSE of 514.076 and an R^2 value of 0.0478. The Residuals vs. Fitted graph again highlighted violations of linearity and constant variance assumptions, as data points were not evenly distributed around the y=0 line and showed occasional gaps and outliers. Moreover, the Independence assumption might be violated due to the density above the y = 0 line. Finally, we see that the residuals are right skewed on the Q-Q plot and Histogram of residuals, indicating violations of the normality assumption.



We then applied Polynomial Regression models to explore potential nonlinear relationships. Initially, this approach yielded an MSE of 627.869 and an R^2 value of 0.0035, performing worse than the Linear Regression models. To improve the model, we incorporated interaction terms and dummy variables (e.g., member_casual and rideable_type) with binary values of 0 and 1. This adjustment slightly improved the results, reducing the MSE to 599.699 and increasing the R^2 value to 0.0482. However, these metrics remained worse than those of the Linear Regression models, indicating limited improvement.

Lastly, to enhance the regression analysis, we incorporated the PCA via sklearn library modules. We selected three predictor variables (hour, member_casual, and rideable_type) and used tripduration as the dependent variable to predict the trip duration based on factors such as the time of day, user demographics, and the type of the ride. While PCA effectively simplified the regression model, the resulting analysis still performed poorly, as demonstrated by the unsatisfactory MSE and R^2 values (Please see image below). In conclusion, despite iterative refinements, including Polynomial Regression, interaction terms, and PCA, Linear Regression remained the most effective model in this context. However, all models exhibited significant limitations due to violations of fundamental assumptions.

```
# Preprocess the data
df = preprocess_data(df)
df['member_casual'] = df['member_casual'].astype('category').cat.codes
df['rideable_type'] = df['rideable_type'].astype('category').cat.codes

# Separate features and target variable
X = df[['hour', 'member_casual','rideable_type']].values
y = df['tripduration']

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Apply PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# Perform regression on principal components
model = LinearRegression()
model.fit(X_pca, y)

# Predictions and evaluation
y_pred = model.predict(X_pca)
mse = mean_squared_error(y, y_pred)
r2 = r2_score(y, y_pred)

print(f"Mean Squared Error: {mse:.2f}")
print(f"R^2 Score: {r2:.2f}")
```

```
Mean Squared Error: 604.51
R^2 Score: 0.03
```

Discussion

These findings underscore the complexities of addressing the bike-sharing prediction challenge. While Linear Regression provided our best results among the attempted approaches, the consistent violations of key assumptions across all models reveal fundamental challenges in our modeling strategy.

The pervasive assumption violations highlight systemic challenges in our approach to prediction. For example, the clear patterns observed in the residual plots suggest that the relationship between predictors and trip duration is more intricate than our linear models assume. For instance, the impact of time of day on trip duration might follow a more complex pattern influenced by rush hours, weekends, or seasonal effects that our linear approach can't capture. Similarly, the relationship between station locations and trip duration may be influenced by unmodeled factors such as geographic features, traffic patterns, or neighborhood characteristics, all of which require more sophisticated analytical techniques.

The heteroscedasticity observed in our models (non-constant variance in residuals) suggests that our predictions' accuracy varies significantly across different conditions. This issue implies that the accuracy of our predictions varies considerably under different conditions. For instance, trip

durations may be more predictable during regular commuting hours but considerably less so during leisure times. This varying predictability challenges our model's ability to provide reliable estimates across all scenarios.

The notably low R^2 values (ranging from 0.0477 to 0.059) are particularly striking. These values indicate that our models explain less than 6% of the variance in trip durations, leaving over 94% of the variation unexplained. This substantial unexplained variance likely results from missing critical features, complex interactions, or latent variables that our current models do not address.

Future research directions might include but are not limited to:

- Exploring more sophisticated modeling techniques such as mixed-effects models to account for hierarchical structure within the data
- Expanding feature engineering to capture additional predictors that better explain variability in trip durations, such as weather conditions, socioeconomic factors, etc.
- Considering non-parametric approaches that don't rely on the strict assumptions of linear regression, offering greater flexibility in capturing non-linear relationships
- Implementing geospatial modeling techniques to better capture the impact of station locations

Understanding these limitations and potential improvements is crucial for developing more effective prediction models in future iterations. The complexity revealed by our analysis suggests that successful trip duration prediction might require a more nuanced, multi-model approach that can adapt to different conditions and user patterns. Our results demonstrate the challenges of real-world predictive modeling and the importance of thorough diagnostic testing, even when working with seemingly straightforward prediction tasks. Simultaneously, it also offers valuable insights for future efforts aimed at more effectively predicting bike-sharing dynamics.