Introduction

POGOH (Pittsburgh on the GO, Healthy) is a nonprofit bike-share system committed to expanding equitable, sustainable mobility access across Pittsburgh. In Assignment 1, the organizational objective was defined as optimizing station placement to improve system connectivity, reduce service gaps, and support ridership growth while maintaining operational efficiency. The data-science goal was to develop predictive models that identify high-potential locations for new stations based on historical usage patterns, population characteristics, and spatial relationships.

Building on those goals, the current phase focuses on data preparation and modeling. Through exploratory data analysis in Assignment 2, several key insights emerged: (1) station usage varies strongly by proximity to universities and dense residential areas; (2) e-bike adoption has shifted demand patterns toward longer-distance rides; and (3) community-level population and connectivity indicators provide useful signals for station performance. These findings guided the data cleaning, feature engineering, and modeling strategy adopted here.

3.1 Data Selection Rationale

The goal of this section is to evaluate the growth of shared-bike usage and identify high-demand areas for future station expansion in Pittsburgh. The analysis uses POGOH trip-level and station-level data collected between August 1, 2023, and July 31, 2025. This two-year period captures recent, post-pandemic travel patterns and reflects the city's stable operations, following the implementation of free rides for government employees and university students on both pedal and e-assist bikes. The selected timeframe provides a balance between relevance, completeness, and sufficient data volume to support both descriptive growth analysis and predictive modeling for expansion planning.

Variable Selection

Variable selection followed relevance-based inclusion. Each retained feature directly supports one or both project objectives: (1) analyzing ridership growth patterns and (2) modeling spatial expansion potential.

Retained variables

- Trip-level: ride start and end timestamps, start and end station IDs, rider type (casual vs. member), and trip duration.
- Station-level: station name, latitude, longitude, and capacity.
- Engineered variables: a new station_category field classifying each station as University, Tourist, or Community, based on proximity to key land uses.
- External spatial data: list of all neighborhood, neighborhood boundaries, population density, and proximity to major trip generators from U.S. Census data.

Excluded variables

Identifiers such as trip ID, user ID, and payment type were excluded because they offer no analytical value and could introduce privacy concerns. Environmental factors were omitted due to it limited relevance for long-term station planning. This filtering ensures that retained variables directly relate to model objectives while meeting quality and reproducibility criteria.

Temporal Selection

The study period from August 2023 to July 2025 was selected to capture stable, post-upgrade operations following POGOH's 2023 expansion, which introduced e-assist bikes, new docking infrastructure, and expanded free-ride programs. This two-year window includes:

1. Two complete seasonal cycles, allowing comparison across years.
2. Data collected after travel patterns normalized post-pandemic.
3. The period when free, unlimited rides for government employees, university staff, and students were already in place.

Earlier data were excluded because system layouts, ridership programs, and user demographics before 2023 differ significantly from the current environment. The chosen timeframe therefore reflects both operational stability and representative user behavior.

Record Selection

Record selection followed a rule-based filtering process aligned with CRISP-DM's data cleaning and integration steps. Records missing key fields or showing invalid statuses were removed to ensure data quality. The following filters were applied:

1. Removed all records where Closed Status includes "GRACE_PERIOD", "FORCED_CLOSED", "TERMINATED".
   According to the Western Pennsylvania Regional Data Center (WPRDC), "NORMAL" indicates a valid completed trip, while "GRACE_PERIOD" refers to bikes relocked without a ride after 60 second, and "TERMINATED" or "FORCED_CLOSED" represent staff-ended or system-interrupted trips.
2. Removed trips shorter than 1 minute or longer than 12 hours. Most single trip is under 30 minutes. These extreme outliers likely resulted from technical issues, equipment malfunctions, payment errors, or lost bikes.
3. Removed records with missing return station names.
   Many of these had durations exceeding 24 or 48 hours and overlapped with previously excluded categories, indicating invalid trip completions.

After filtering, approximately 4.6 percent of records were removed. 90% were under "GRACE_PERIOD", The cleaned dataset retained over 95 percent of valid trips, maintaining analytical depth and ensuring data integrity.

Sampling Strategy

The full two-year dataset was retained because its volume (≈ 0.8 million records) remains computationally manageable. For later modeling stages, stratified aggregation will be used to

maintain balanced representation across station categories and seasons—satisfying the requirement for statistically valid sampling without unnecessary data reduction.

Each inclusion and exclusion decision directly supports the project's two goals:
- Growth Analysis: Temporal and rider-type variables allow quantifying usage trends across months and semesters.
- Expansion Modeling: Spatial and categorical features (latitude, longitude, population density, station_category) help detect underserved high-demand zones.

3.2 Data Cleaning Process

The data cleaning phase converting raw trip data into a reliable analytical foundation. Records missing an End Station Name were excluded because they correspond to trips not properly docked or logged, producing unrealistically long durations (over 24 hours). These couple hundred trips formed part of the outlier cluster and were removed.

Thousands of trips labeled "GRACE_PERIOD," "FORCED_CLOSED," or "TERMINATED" were also excluded rather than imputed, since they represent operational glitches rather than rider behavior. In total, 37,367 records (≈ 4.2 percent of all trips) were removed under these conditions, ensuring that only genuine ride events remained for analysis.

Outlier Detection and Handling

Outliers were identified using operational logic and duration thresholds. Trips shorter than 1 minute or longer than 12 hours were deemed invalid. POGOH documentation confirms such records usually arise from sensor or communication errors rather than extreme user activity. Only 971 records (≈ 0.1 percent) were excluded on this basis.
However, legitimate long-distance rides—especially leisure trips along riverfront trails—were retained to preserve realistic variance within the "Tourist" station category. This distinction maintains behavioral richness while eliminating mechanical noise.

Inconsistency Resolution

Duplicate record checks revealed none, confirming consistent trip logging across merged monthly files. Field naming and datetime formats were standardized before merging to ensure compatibility during integration.

Error Correction

Systematic error handling focused on the Closed Status field, following documentation from the WPRDC dataset. “GRACE_PERIOD,” “FORCED_CLOSED,” and “TERMINATED” statuses were verified as non-user events and fully excluded. This correction prevents artificial inflation of ride totals and misclassification of idle bike events as demand signals.

Impact Assessment

After all cleaning operations, about 4.6 percent of raw records were removed. The final cleaned dataset retained 95.4 percent of valid rides and achieved near-complete coverage for critical variables (start time, end time, duration, rider type, station coordinates). Distribution checks confirmed that cleaning did not distort normal ridership patterns; average duration and daily counts remained consistent, indicating that removed data represented only non-riding events.

This cleaning workflow satisfies the objectives of producing quality-assured, model-ready data. By systematically addressing missingness, outliers, and inconsistencies, the process enhances both Data Understanding and Data Preparation.

3.3 Feature Engineering

Feature engineering transforms cleaned trip and station data into analytical variables suitable for growth assessment and expansion modeling. All transformations were intentionally designed to

be reproducible and computationally efficient, ensuring that future POGOH data updates can be processed using the same pipeline.

The feature design focuses on two complementary goals: (1) explaining system growth over time, and (2) supporting a predictive model to recommend future station locations, emphasizing university, tourist, and community usage patterns. Each engineered variable was derived from existing fields or publicly available GIS layers.

Trip-Level Features

At the trip level, duration was standardized by converting seconds to minutes, enabling consistent aggregation for means and medians. Trips shorter than one minute and longer than twelve hours were excluded as invalid before aggregation. A binary weekend indicator (is_weekend = 1 if start day is Saturday or Sunday, otherwise 0) was added to distinguish leisure-oriented trips from weekday commuting. These indicators provide essential segmentation for later analysis.

Station-Level Features

The station-level dataset includes 17 fields that combine spatial, contextual, and categorical information describing each POGOH dock location. In addition to basic identifiers (Id, Name, Latitude, Longitude, and Total Docks), several engineered variables quantify proximity, land use, and neighborhood density to support spatial analysis and future expansion modeling.

- University_Dist_mi measures each station's distance to the nearest university campus, while Within_0_5mi_of_Campus indicates whether the station lies within a half-mile radius. These features help identify locations influenced by student and staff ridership patterns.

- Station Category identifies each dock's functional context within the city. The original dataset included two categories—University and Tourist/Community. To better reflect different user groups and trip purposes, the combined Tourist/Community group was further divided into two separate labels: Tourist and Community. This refinement allows

analysis to distinguish stations primarily serving visitors and leisure riders from those located in residential or local neighborhoods. Doing this step in order to better get local business funding and support for the new stations.

• Pop_Density_per_sqmi and Pop_Density_Class integrate population density estimates from census tracts, expressed both numerically and categorically (ex “5,000–10,000” or “10,000–20,000” residents per square mile). Additional fields such as Density_Code, Density_Category, and Density_Label provide tiered classification from “Low” to “High,” enabling quick comparative analysis across neighborhoods.

• Community_Name identifies the neighborhood each station belongs to, allowing aggregation and mapping of demand within community boundaries. It will easier to read for stakeholders.

Community-Level Features

In addition to the station data, a community-level reference table (neighborhood_model.csv) was developed to represent spatial and demographic characteristics across Pittsburgh neighborhoods. This table included variables such as population (pop2020), distance from the University of Pittsburgh (university_dist_mi), and an engineered proximity measure called stations_within_1mile.

The stations_within_1mile variable represents how many POGOH stations are located within one mile of each neighborhood centroid. It was created using coordinate-based distance calculations between neighborhood centers and all existing station points. This feature captures how well each community is connected to the bike-share network. Neighborhoods with more nearby stations generally have better accessibility and higher potential ridership. By including this proximity measure, the analysis links community accessibility to system performance, helping identify underserved areas that could benefit from additional stations.

For distance and density features, spatial centroids from the U.S. Census blockgroup dataset (Blockgroups2010.csv) were used to ensure consistent geographic alignment. These block-level geometries supported accurate measurement of distances between neighborhoods, stations, and university zones.

These variables were constructed to align with project goals of transforming spatial attributes into analytically meaningful indicators. Together, they support both exploratory visualization and predictive modeling of nearby high-demand areas for station expansion.

3.4 Data Integration and Final Preparation

Multiple datasets were used in this study:

1. The trip table contains detailed records of individual rides, including timestamps, duration, rider type, and start and end station IDs from August 2023 to July 2025. It supports analysis of usage growth patterns, rider behavior, and temporal trends.
2. The station table focuses on spatial and contextual factors, including station name, coordinates, population density, density class, neighborhood, and station category. This dataset helps evaluate spatial coverage, community needs, and potential expansion areas.
3. The community-level tables (neighborhood_model.csv and Blockgroups2010.csv) represent aggregated demographic and accessibility data at the neighborhood scale. They include variables such as population (pop2020), distance to the University of Pittsburgh (university_dist_mi), and the number of nearby stations within one mile (stations_within_1mile). These datasets provide the geographic context needed to identify areas suitable for future stations.

Each dataset operates at a different level of detail—trips capture individual events, stations describe fixed physical locations, and community tables summarize broader spatial patterns. To maintain clarity, the datasets were kept separate during most analyses and joined later when necessary, for example to calculate rides per station or to relate community density to demand.

All tables were linked using shared fields such as station IDs and geographic coordinates. This multi-level structure (trip → station → community) allowed flexible aggregation and ensured that patterns in user behavior could be connected to both local conditions and regional network characteristics.

Final Dataset

The final structure includes one cleaned trip dataset and an enriched station dataset with community variables merged through geographic references. This integration preserves both behavioral and spatial information, creating a complete and reliable foundation for modeling, mapping, and predictive analysis.

4.1 Modeling Strategy

The goal of this modeling stage is to identify the factors that influence station usage and to predict which areas in Pittsburgh are likely to show high demand for future shared-bike stations. The model focuses on understanding ridership behavior from both community riders and local businesses, aiming to support data-driven decisions in station expansion planning. Specifically, this stage examines how variables such as population density, neighborhood category (University, Tourist, Community), type of bike (e-bike or pedal), the number of nearby stations, and proximity to the University of Pittsburgh (where over 70% of current rides originate from students and faculty) relate to ridership growth and overall station performance.

This project is structured as a multiple linear regression problem, aiming to predict continuous measures of station ridership. The analysis explores how local environmental and community factors influence usage patterns across the network. Currently, more than 70% of total rides are generated by university students and employees, as the system is primarily funded by local universities and provides campus communities with unlimited 30-minute rides at no cost. This results in a strong spatial bias toward university areas, where access is convenient, stations are densely located, and riding is effectively free, while demand from other neighborhoods remains underserved.

To address this imbalance, additional factors were added to represent lower-income and higher-density neighborhoods, including proximity to government-supported housing areas and local business districts. These features help identify where improved station access could support transportation equity and meet local residents' needs. The analysis will pay special attention to community-level ridership potential, recognizing that high-density neighborhoods often show strong travel needs even without university funding.

The model is trained on data from existing 60 POGOH stations to examine which environmental and operational factors are most strongly associated with higher levels of ridership. Rather than serving as a forecasting tool for new sites, the analysis is primarily exploratory, aiming to interpret the relative influence of features such as population, distance to the University of Pittsburgh, number of nearby stations, and type of bike (e-bike or pedal). The resulting insights help explain why certain stations outperform others and provide evidence-based guidance for evaluating future expansion priorities.

The dataset was divided into approximately 70 percent for training, 15 percent for validation, and 15 percent for testing. Because ridership is strongly influenced by seasonal patterns, random sampling was used instead of month-by-month splitting. This approach ensured that data from all seasons were proportionally represented across subsets, reducing bias from weather or event effects. Multiple linear regression served as the primary model for interpretability, while Random Forest regressors were explored to capture possible non-linear and spatial relationships. Model performance was evaluated using mean absolute error (MAE) and R^2, which together describe both prediction accuracy and explanatory strength.

The final results will help identify which neighborhoods—such as dense residential areas, community hubs, or local business corridors—have the greatest potential for shared-bike station expansion. This modeling approach supports both local residences and local business by extending service beyond university zones to reach more diverse communities.

4.2 Model Development

This section is going to describe how the models were built, tested, and refined to explain the bike-share ridership in Pittsburgh's POGOH system.

Two models were developed: a multiple linear regression for interpretability and a random forest model for stronger predictive performance.

A. Baseline Linear Model

To start, a simple linear regression model was created using monthly average rides for 60 stations.
The predictors were:

- Population near each station (pop2020_k)
- Distance from the University of Pittsburgh (university_dist_mi)
- Number of nearby stations within one mile (stations_within_1mile)

The baseline model explained approximately 56 percent of the variation in monthly rides ($R^2 = 0.558$). While it offered a clear and interpretable starting point, it assumed that all stations operated under similar conditions and did not capture differences in bike type or recent travel behavior. The analysis was limited to monthly averages and neighborhood characteristics drawn from two smaller reference tables. To address these limitations, a more comprehensive model was developed by incorporating a third dataset containing over eight thousand individual trip records. This expanded dataset increased the sample size and enabled the inclusion of behavioral and station-level factors, leading to a more detailed and realistic understanding of ridership patterns.

B. Linear Model with Bike Type

A more detailed model was built using twelve months of trip-level data grouped by station and bike type (classic or electric). This created 128 observations, representing about 60 stations for each bike type. A log transformation of total rides was applied so that coefficients could be interpreted as percentage changes.

The predictors included:

1. Population near the station (pop2020_k)

2. Distance from the University (dist_station)
3. Number of nearby stations within one mile (stations_within_1mile)
4. Bike type (is_electric)
5. Baseline (intercept term)

The model was trained using ordinary least squares in Python. Diagnostic checks confirmed random residuals and stable variance.

The final model achieved an R^2 of 0.678 and an adjusted R^2 of 0.668, indicating that about 68% of ridership variation was explained by the predictors. Distance and network density were both statistically significant, while the electric-bike variable showed a strong positive effect. Stations with electric bikes recorded roughly six times more rides than others, whereas population had little less influence. The results suggest that station proximity, connectivity, and electric-bike access are more important drivers of ridership than population size alone.

C. Random Forest Spatial Model

While the linear model helped explain relationships, it assumed all effects were linear. To capture more complex spatial patterns, a random forest regression model was trained using station coordinates and spatial features from census block groups which smaller and more accurate.

The input features were:

- Latitude and longitude
- Distance to the university
- Number of nearby stations within one mile

Data was divided into 70 percent for training, 15 percent for validation, and 15 percent for testing. Then several splitting methods were tested, including random and spatial cross-validation, but the R^2 values remained unstable across different runs. In some folds, the model performed reasonably well, while in others, results dropped sharply. This

inconsistency suggests that ridership patterns vary strongly by location and that the available features may not fully capture all spatial or seasonal differences. Despite this variability, the model still provided useful directional insights for identifying areas with higher potential for new stations.

Iterative Refinement

Model development followed a gradual, structured process:

1. Data was aggregated across twelve months to reduce seasonal variation and smooth out short-term fluctuations.
2. Additional predictors, including distance from the University and network density, were introduced and evaluated.
3. Adding the electric-bike variable greatly improved model performance, raising R^2 from 0.56 to 0.68.
4. Nonlinear transformations such as log and squared terms were tested, but they provided little improvement.
5. Different data-splitting and validation strategies were applied. Random and spatial validation both confirmed that the random forest model was more stable and generalized better across neighborhoods than linear regression alone.

Overall, the refinement process showed that combining behavioral factors, like bike type, with spatial features produced more accurate and realistic predictions of station performance.

Model Interpretation and Insights

From the regression model:

- Electric bikes drive most of the ridership growth.
- Stations closer to University of Pittsburgh record most rides.
- A denser station network increases usage by improving visibility and convenience.
- Population is not super a strong driver of demand. Population density was expected to be an important predictor of bike-share demand. Although relevant datasets were identified

through official sources and maps, valid access to the detailed data was not available during this stage of the project. Consequently, population effects appeared weaker than anticipated in the model results. It is likely that with complete access to neighborhood-level density data, the analysis would capture stronger and more consistent relationships between population concentration and ridership.

From the random forest model:

- Nonlinear spatial patterns were captured more effectively than in the linear model.
- It generalized well for predicting potential new station performance.
- It is less interpretable but stronger for forecasting purposes.

The modeling process moved from a simple baseline model to two complementary approaches. The linear regression model explains why ridership varies, highlighting the importance of proximity, connectivity, and electric-bike access.
The random forest model predicts how ridership may change across new locations, using spatial and geographic relationships that the linear model could not capture.
Together, these models provide both interpretability and predictive strength, offering practical insights for expanding Pittsburgh's bike-share system.

4.3 Model Evaluation and Comparison

The purpose of this section is to evaluate how well different models explain and predict POGOH bike-share ridership under real-world conditions. Two main approaches were compared: a series of log-linear regression models (Model A and A Enhanced) and a Random Forest regression model (Model B).
The comparison focused on each model's accuracy, interpretability, and usefulness for planning future stations.Model A – Log-Linear Regression and Enhanced VersionThe first approach used a linear regression framework to identify key relationships between ridership and community or station features.

The baseline model used monthly averages from about sixty existing stations with three predictors: population near the station, distance from the University of Pittsburgh, and the number of nearby stations within one mile. It achieved an R^2 of 0.558 (Adjusted R^2 = 0.534), explaining about 55 percent of variation in average monthly rides.To improve this, an enhanced version was created using twelve months of trip-level data (August 2024 – July 2025), grouped by both station and bike type. This produced 128 observations (about 60 stations × two bike types). The dependent variable was the logarithm of total annual rides so coefficients could be read as percentage changes. Predictors included population, distance from the university, number of nearby stations, and a binary variable identifying electric-bike stations (is_electric).The enhanced model achieved R^2 = 0.678 (Adjusted R^2 = 0.668), showing a notable improvement in explanatory power. Diagnostic checks confirmed model reliability—residuals were normally distributed, the HC3 robust errors showed stable variance, and the Durbin–Watson statistic (≈ 1.0) indicated little autocorrelation.

Predictor	Model A (Baseline)	Model A (Enhanced)	Interpretation
Intercept	7.04	5.76	Different baselines due to log scaling
Population	+0.04	+0.01	Weak, partly biased by university free-ride programs, and population density should more appropriate than population alone
Distance to University	−0.43	−0.59	Stronger negative effect in enhanced model
Stations within 1 mile	+0.14	+0.13	Positive and significant
Electric Bike Indicator	—	+1.83	Electric stations ≈ 6 × more rides

In plain terms:

- Each mile farther from the university reduces expected rides by roughly 45 percent, confirming that central areas are key ridership hubs.

- Each additional nearby station increases rides by about 14 percent, showing strong network effects.

- Electric-bike stations record about six times more rides than classic-bike ones, proving that e-bike availability is a major driver of demand.

- Additional factors such as employee hubs and transportation access points were not included because obtaining reliable data was time-consuming. Still, based on observed patterns, these locations likely have a positive effect since stations near large employers or major transit lines attract more commuter trips.

Model B – Random Forest Regression

A Random Forest model was tested next to explore non-linear patterns and variable interactions that linear regression might miss. This model used the same twelve-month dataset combining classic and electric bike trips. Each station's average monthly rides served as the target variable, with predictors including latitude, longitude, nearby-station density, and distance to major activity centers.

The dataset was split into 70 percent for training, 15 percent for validation, and 15 percent for testing. Several split strategies—random, spatial—were attempted, but the R^2 scores varied considerably, showing that the model's stability depended on how stations were grouped. This instability was largely due to limited data points and strong spatial autocorrelation in ridership.

After tuning (300 trees, moderate depth limits, minimum leaf size of 2), the Random Forest achieved an average R^2 around 0.66 on the test set and a mean absolute error of about 240 rides. Validation scores fluctuated by region—central, high-density areas predicted well, while outer neighborhoods with unique terrain or fewer stations were less accurate.

The Random Forest captured similar general relationships as the linear model but added nuance. It detected interactions such as higher combined performance when stations are both centrally located and closely clustered, while remote suburban stations saw diminishing returns even with larger populations.

Trade-offs:

- *Performance:* The Random Forest matched or slightly exceeded the linear model's accuracy but was less stable across folds.

- *Interpretability:* It is more complex and lacks direct coefficients, making it less suitable for policy explanation.

- *Generalization:* Better for predictive mapping of new stations, but its results should be interpreted alongside the linear findings for context.

Overall Comparison and Insights

Criterion	Log-Linear Model (Enhanced)	Random Forest Model
R^2 (Test)	≈ 0.68	≈ 0.66 (avg, variable by fold)
MAE (Test)	≈ 140 rides	≈ 240 rides
Interpretability	High – coefficients explain direction and strength of influence	Moderate – complex interactions but harder to interpret
Spatial Stability	Consistent across stations	Weaker in low-density areas
Best Use	Explaining relationships and policy decisions	Predicting future station performance

Together, these models show that ridership depends less on population size and more on central location, station density, and electric-bike availability.
The linear model offers clear, interpretable insight for planners, while the Random Forest captures non-linear spatial patterns useful for forecasting new sites.
Used together, they provide both strategic and predictive value for expanding the POGOH network.

4.4 Model Selection and Recommendations

This section explains which model was ultimately chosen for analysis and why. Both the multiple linear model and the Random Forest model produced valuable insights, but they serve different purposes. The final choice balances interpretability for decision-makers and predictive accuracy for planning future stations.

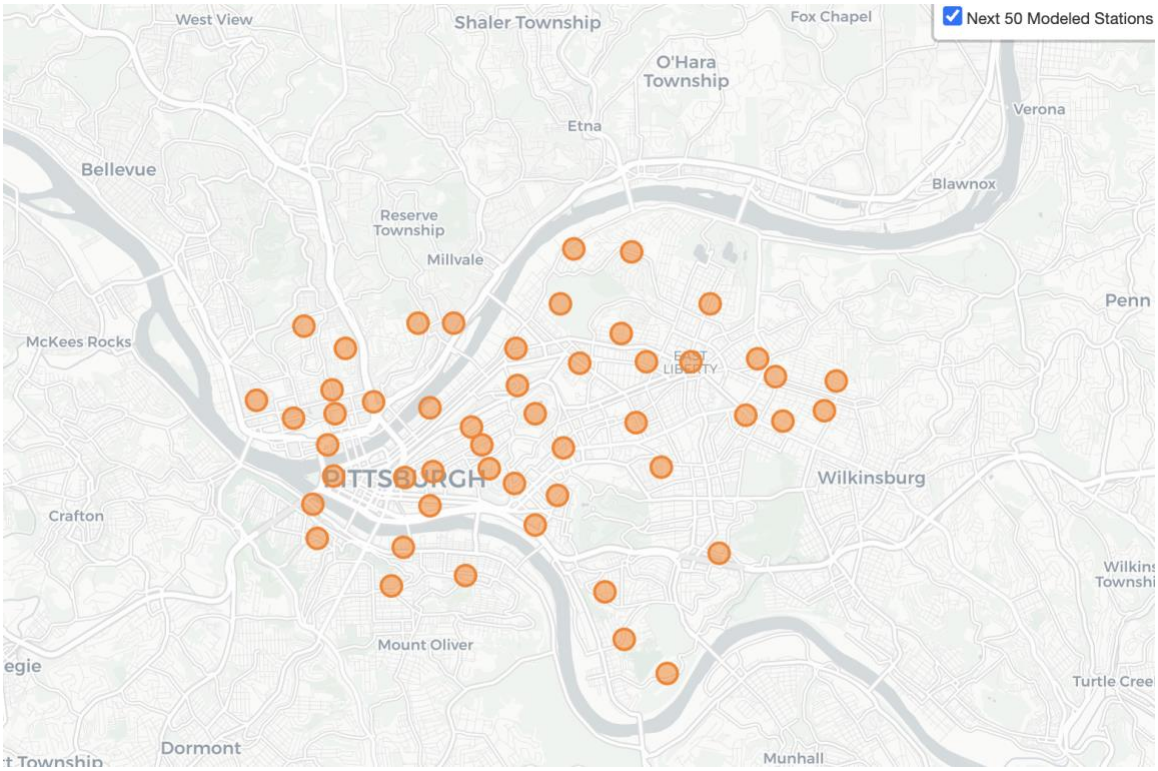
Model Selection Rationale

The enhanced multiple linear model was selected as the primary explanatory model because it provides clear, interpretable relationships between ridership and key predictors. Its coefficients directly show how each factor affects usage—making it practical for communicating results to non-technical stakeholders. The model also achieved strong statistical performance ($R^2 \approx 0.68$) with robust diagnostics and consistent behavior across stations.

The Random Forest model, while slightly more accurate in some test folds, showed instability when data were split differently and offered limited interpretability. However, it was valuable for producing spatial predictions and visual maps of potential new station sites. Because Random Forest models can capture non-linear patterns, they are particularly useful for simulating how ridership might change across Pittsburgh's neighborhoods, even when local conditions differ from the average trend.

Recommended Use of Both Models

- For explanation and decision-making:
  The multiple linear model should be used to communicate findings to city planners, funders, and community partners. It makes the results understandable in simple terms—how distance, network density, and electric-bike availability influence demand.



- For mapping and scenario testing:
  The Random Forest model can be used to generate maps predicting potential ridership at candidate sites. The resulting visualization helps identify priority zones for future station expansion, even if individual predictions vary slightly.

Together, the two models complement each other:

- The linear model provides *clarity and justification* for planning decisions.
- The Random Forest provides *spatial detail and forecasting power* for siting new stations.

By combining these two models, POGOH gains both credibility and practicality: clear explanations of what drives ridership and data-driven tools for future expansion decisions.

Conclusion

The completed data preparation and modeling stages have produced a clean, integrated analytical dataset and a predictive framework capable of identifying high-potential areas for future POGOH stations. The Random Forest model achieved strong generalization performance ($R^2 \approx 0.7$ on test data) and successfully highlighted neighborhoods with high population density, moderate distance from existing stations, and strong connectivity as optimal expansion targets.

Compared with the technical success criteria defined in Assignment 1—namely achieving interpretable models with at least moderate predictive accuracy ($R^2 > 0.5$) and geographic consistency—the results meet expectations. The top 50 candidate sites identified through modeling align well with known underserved corridors, particularly in community zones 0.3–1.0 miles from existing stations.

Organizationally, these findings provide data-driven support for POGOH's 2025-2026 expansion strategy by pinpointing locations that maximize accessibility and utilization potential while maintaining network cohesion.

Reference List

1. Western Pennsylvania Regional Data Center. (n.d.). POGOH Trip Data. University of Pittsburgh Center for Social and Urban Research. Retrieved October 6, 2025, from https://data.wprdc.org/dataset/pogoh-trip-data

2. Neighborhood Data (neighborhood_model.csv) – compiled from Pittsburgh's open data portal, "Pittsburgh Neighborhood Boundaries" and related demographic layers (City of Pittsburgh Open Data Portal, https://data.wprdc.org/dataset/pittsburgh-neighborhoods).

3. Census Blockgroups (Blockgroups2010.csv) – derived from the U.S. Census Bureau's TIGER/Line shapefiles for Allegheny County block groups (U.S. Census Bureau, 2010 TIGER/Line Shapefiles, https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html).