0. Executive Summary Presentation

3-Minute Video Link

https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=305f7bcf-0e11-4900-9ea2-b387011fef43

This short presentation provides an overview of the project's goal, approach, key findings, and recommendations for POGOH's next 50-station expansion. It highlights how data analysis supports smarter station placement, equitable community access, and effective e-bike deployment.

1.1 Translating Technical Results to Organizational Impact

The linear regression model for this project predicts the logarithm of average monthly rides at each POGOH station using four explanatory variables: neighborhood population (in thousands), distance to the nearest university, number of nearby stations within one mile, and whether the station offers electric bikes. The model achieved an R^2 of 0.678, explaining about 68 percent of the variation in station-level ridership. This indicates that the model effectively captures meaningful behavioral patterns that drive demand across Pittsburgh's bike-share network.

From an organizational perspective, this level of accuracy provides POGOH with reliable insight into where future stations are most likely to perform well.

- The most influential factor is distance from a university, which has a coefficient of –0.59 on the log scale. In practical terms, each additional mile away from a university corresponds to roughly 45 to 50 percent fewer monthly rides. This effect is expected since students and university employees ride for free, and many stations are located within or near campus areas. As POGOH plans its next expansion phase, placing new stations near student housing and areas that match students' after-school travel interests and traffic patterns—particularly in the Squirrel Hill neighborhood less than two miles from the university—should yield substantial increases in ridership.

- The number of nearby stations also has a positive relationship with usage. Each additional station within one mile is associated with about 13 percent higher ridership. Denser station clusters make it easier for riders to locate and return bikes conveniently, increasing overall trip frequency. The electric bike variable shows the strongest positive effect.
- Stations with e-bikes record approximately six times more rides than non-electric stations. This finding aligns with observed behavior that community riders, especially in a hilly city like Pittsburgh, prefer e-assist bikes for longer or uphill trips. Expanding e-bike availability could therefore attract more users and encourage longer rides.
- Intercept and seasonality.
  The model's intercept is 5.7586 on the log scale. As a baseline, exp(5.7586) ≈ 317 rides per station-month for a non-electric station when other predictors are at their reference values. Because we did not include calendar controls, the intercept absorbs average seasonal effects across stations. In practice, demand is higher during university semesters (roughly September–November and January–April) at campus-adjacent stations, and lower in winter (December–February) across the network due to cold weather. For planning and forecasting, a simple seasonal adjustment (e.g., month or semester indicator) can be applied so winter projections reflect this cold-weather dip while semester months near campus reflect higher activity.

- Although the population variable was not statistically significant in the model, its positive coefficient is consistent with the expectation that higher residential density supports higher utilization. The lack of significance may reflect data bias, since free university rides mask underlying population effects. When analyzing community stations separately, the relationship between density and ridership becomes clearer.

Operationally, these findings have clear planning implications. The model allows planners to estimate potential monthly ridership for proposed sites and compare alternatives before installation. For example, if a site near a university is predicted to generate 1,200 rides per month while a suburban site predicts only 600, decision-makers can prioritize higher-impact areas such as South and North Squirrel Hill. Evidence-based placement could reduce low-performing stations by 25 to 30 percent, saving approximately $30,000 to $40,000 annually in underutilized infrastructure and rebalancing costs.

Confidence is highest for established stations with consistent historical data and lowest for new or atypical sites. The model should be retrained annually as new data become available. Despite these limitations, the regression framework provides a transparent and interpretable foundation for decision support, helping POGOH expand strategically and deliver equitable, sustainable mobility solutions across Pittsburgh.

Confidence is highest for established stations with consistent historical data and lowest for new or atypical sites. The model should be retrained annually as new data become available to keep predictions aligned with changing travel patterns, seasonal trends, and infrastructure updates. Retraining will also allow the inclusion of richer spatial and demographic variables once they become accessible. Because this project worked with limited geographic data, such as block-level population density, employment centers, and detailed points of interest like transit hubs or malls, the current model can only provide a general estimate of ridership potential.

To achieve higher precision in future iterations, the management team should consider integrating finer-grained geographic and mobility data sources. Incorporating land-use data, employment density, or transit connectivity could help identify the underlying reasons for ridership differences between neighborhoods. These improvements would strengthen predictive accuracy and support more targeted station placement decisions.

While the present analysis offers valuable insights and a transparent foundation for decision support, it should be viewed as an overall planning tool rather than a final detailed deployment model. It could be a much better model if I can have more information from local authority and talk with the operation team. In the future, continued data collection, annual retraining, and collaboration with city planning departments will allow POGOH to refine its forecasts, validate assumptions, and align expansion strategies with both community needs and long-term sustainability goals.

1.2 Insights and Recommendations

The linear model results identify clear geographic and operational priorities for POGOH's upcoming expansion. Validation performance ($R^2 = 0.66$, MAE ≈ 215 rides) indicates the model predicts community-level demand with good reliability, the predictions highlight neighborhoods where additional stations are likely to achieve high utilization and strengthen the network.

Key Insights

1. Predicted demand is concentrated around the Squirrel Hill North, Squirrel Hill South, West Oakland, and Point Breeze corridors. These areas already show strong potential ridership (650–960 rides per month) and sit within one mile of existing stations. They are ideal for network densification. Reddit discussions and local observations confirm residents in these high density neighborhoods feel underserved and want bike access expand to their neighborhoods.

2. Outside the current core, several high-density communities—such as Greenfield, Duquesne Heights, and Upper Lawrenceville—show moderate predicted demand (450–750 rides per month). Adding stations in these areas would extend equity and improve coverage for local residents who rely on short, intra-neighborhood trips. These neighborhoods also feature local attractions, such as the Pittsburgh Zoo and nearby parks, which could attract more weekend and leisure riders. Expanding to these destinations would not only serve daily commuters but also broaden options for tourists and occasional riders seeking accessible day-trip routes.

3. Low-performing stations in isolated neighborhoods highlight the importance of clustering. More than ten existing sites show persistently low ridership, primarily because they lack nearby stations within convenient riding distance. To address this, future expansion will develop mini-clusters of at least three stations within each community hub. This network approach will make it easier for riders to start and end trips locally, encourage round-trip usage, and strengthen connectivity across the system.

3. Seasonal effects remain substantial. Ridership declines sharply from December through March due to cold weather. Seasonal rebalancing—temporarily storing surplus bikes in winter—will reduce maintenance costs and prevent unnecessary wear. The intercept analysis suggests roughly 300 baseline rides per station even in low-activity months, so winter storage should target the lowest-use zones.

Strategic Recommendations

1. Pittsburgh's urban core has a population of roughly 300,000 residents, Pittsburgh metropolitan area (Greater Pittsburgh): about 2.35 million residents and attracts more than 10 million visitors annually. However, the current POGOH system primarily serves areas surrounding the University of Pittsburgh and Carnegie Mellon University, reaching only a fraction of the city's total population. Implementing **Phase Three Expansion** should therefore be a top priority to extend access and equity across more residential neighborhoods. The plan calls for constructing approximately fifty new stations, starting with the top ten candidate sites identified by the model. Deployment should be completed within six months—before next spring's peak riding season—to ensure readiness for increased demand. Priority should be given to Squirrel Hill North, Squirrel Hill South, West Oakland, and Point Breeze and surrounding communities which combine strong predicted ridership with proximity to existing infrastructure and community activity centers.

   Except that, Outside the current core, several high-density communities-such as Greenfield, Duquesne Heights, and Upper Lawrenceville-show moderate but promising predicted demand ranging from 611 to 745 rides per month. Duquesne Heights currently averages about 612 predicted monthly rides with a 0.79-mile distance to the nearest existing station, highlighting a clear accessibility gap. Upper Lawrenceville shows even higher predicted demand of approximately 745 monthly rides, located only 0.35 miles from the nearest station, suggesting strong potential for improved network connection and community use. Both neighborhoods have moderate population density and are situated in active local zones that would benefit from more convenient bike access. Greenfield, another dense residential area, demonstrates similar potential due to its proximity to recreational corridors and key local attractions such as the Pittsburgh Zoo and nearby parks. Expanding stations in these communities would not only promote ridership equity and reduce accessibility gaps but also broaden mobility choices for local residents, tourists, and leisure riders seeking accessible day-trip routes.

2. **Promote Community Adoption.** To increase local participation, POGOH should offer introductory discounts or free-ride credits for residents in newly served neighborhoods. From the study, it is clear that cost remains a major barrier—over 90% of current riders are subsidized through city programs, major employers, or universities. Providing targeted price incentives can encourage more community riders to try the service and develop regular usage habits. These efforts would make the system more inclusive, helping POGOH attract a broader base of local users beyond its existing institutional partnerships.
3. **Sustain Financial and Operational Support.** Each e-bike costs over $2,000 to purchase and hundreds of dollars annually to maintain. Expanding the program will require partnerships with large employers and local business to co-fund stations and support long-term operations. Employer sponsorship not only offsets expenses but also integrates the bike-share system into the city's wider mobility ecosystem.
4. **Address Technical Reliability.** About 4 percent of total rides in the past two years recorded technical issues, including broken bikes, docking malfunctions, and unlock failures following the system's integration with the Lyft app in spring 2023. These issues likely reflect normal operational adjustments during the early phase of the new platform. Continued hardware maintenance and software updates should further reduce these errors, improving trip completion rates and overall user satisfaction. Ongoing monitoring of technical reliability is recommended to quickly identify and address recurring equipment or system issues across the network.
5. **Enhance Data and Model Granularity.** The current model relies on community-level averages. To improve accuracy, future iterations should work together with management

team in order to integrate block-level population density, employment centers, public-transit nodes, and retail or recreational points of interest. Combining these finer-scale data sources with updated ridership logs will allow more precise demand forecasts.

6. **Annual Retraining and Evaluation.** The model should be retrained each year as new trip and demographic data accumulate. This will align forecasts with shifting travel patterns, seasonal variations, and network changes.

These recommendations translate analytical findings into a concrete action plan. By densifying existing high-demand corridors, expanding into underserved neighborhoods, engaging local communities through targeted discounts, and maintaining partnerships for funding and operations, POGOH can strengthen both ridership and equity outcomes. Phase Three implementation within the next six months will position the network to meet next spring's ridership surge while ensuring long-term sustainability for Pittsburgh's bike-share system.

1.3 Risk Assessment and Mitigation Strategies

Every large-scale expansion comes with some level of risk, and POGOH's Phase Three plan is no exception. While the project will bring major benefits to accessibility and ridership, there are still uncertainties that need careful planning. This section outlines the main risks identified for the expansion—such as operational challenges, financial concerns, and safety issues—and suggests practical ways to reduce or manage them. By planning ahead, POGOH can handle these challenges smoothly and keep the project on track for success.

1. Incomplete Station Siting Plan

Risk: The current model provides community-level predictions rather than finalized station locations. Many details—such as parcel availability, land-use permissions, and infrastructure compatibility—still require detailed analysis by the management team. This introduces uncertainty in the final siting process.

Mitigation: Update demographic, employment, and transit data annually. Use actual ridership metrics from new stations to recalibrate the model and improve accuracy for future planning cycles. Conduct site assessments in coordination with the City of

Pittsburgh’s Department of Mobility and Infrastructure (DOMI). Engage with community stakeholders to validate location choices and confirm that selected sites support both operational efficiency and local travel needs.

2. Expanded Operational Complexity

Risk: Expanding the system from 60 to 110 stations will substantially increase the service area, creating new challenges for maintenance, bike rebalancing, and logistics. The larger network will demand higher staffing levels and more complex coordination.

Mitigation: Establish regional service zones to decentralize operations and reduce travel time for maintenance crews. Utilize predictive rebalancing models to optimize bike distribution. Consider partnerships with local universities, nonprofits, or community hubs for satellite maintenance and storage support.

3. Rising E-Bike Operating Costs

Risk: Overestimating e-bike demand or underestimating maintenance costs could create financial pressure. Each e-bike costs over $2,000 to purchase and several hundred dollars annually to maintain, and battery servicing adds additional expense.

Mitigation: Deploy e-bikes in phases and monitor usage data monthly to compare predicted versus actual ridership. Adjust fleet composition based on real demand. Implement a differential pricing strategy to encourage cost-sensitive riders to choose pedal bikes more often, helping balance operating costs while maintaining service flexibility.

4. Traffic, Safety, and Equipment Risks

Risk: Expansion into denser, lower-income neighborhoods may increase exposure to traffic conflicts, particularly in areas with limited cycling infrastructure. Additionally, higher density may lead to increased equipment wear, loss, or vandalism if preventive measures are not in place.

Mitigation: Collaborate with the City’s transportation department to identify high-risk areas and improve safety infrastructure through dedicated bike lanes, signage, and lighting. Launch rider safety campaigns in new communities and establish a consistent equipment tracking and rapid-repair protocol to minimize downtime and losses.

5. Financial Sustainability and Funding Gaps

Risk: Rapid network expansion without corresponding new funding sources could create budget shortfalls. Community-based discounts and expanded coverage may raise costs before additional ridership and sponsorship revenue stabilize.

Mitigation: Strengthen partnerships with large employers, universities, and healthcare organizations to co-fund stations and operations. Pursue federal and state sustainability grants, and explore tiered membership pricing to maintain affordability while supporting financial balance.

6. Community Engagement and Equity Challenges

Risk: If community voices are not incorporated into planning, new stations may fail to reflect actual travel needs or gain strong adoption in underserved areas.

Mitigation: Conduct early engagement with neighborhood associations and community leaders to identify preferred locations and build local ownership. Co-design outreach and promotional programs to build trust and long-term participation.

7. Regulatory or Policy Change Risk

Risk: Shifts in city or state transportation policy—such as new parking regulations, street redesigns, or changes in subsidy eligibility—could affect where stations can be placed or how pricing is structured.

Mitigation: Maintain active communication with city departments and review mobility policy updates quarterly to adjust siting or pricing accordingly.

8. Reputation or Public Perception Risk

Risk: If early expansion stations experience operational issues or community complaints (e.g., blocked sidewalks, vandalism), it could harm POGOH’s reputation and slow adoption.

Mitigation: Launch public education and outreach campaigns in each new community before rollout, and monitor social feedback channels (like Reddit and Nextdoor) for early warning signals.

Proactively identifying and managing these risks will ensure that POGOH's Phase Three expansion proceeds in a structured, sustainable, and equitable way. By balancing growth with operational readiness, cost control, and community engagement, the program can deliver measurable long-term value for residents, businesses, and the city's sustainable transportation network. Continuous monitoring and adaptive management will help the organization refine strategies as new data and conditions emerge.

2.1 Project Retrospective

1. What Went Well

The database in this project was relatively simple and clean, which made it easier to identify patterns and insights from both the trip and station datasets. I successfully applied Random Forest and Linear Regression models, and after plenty iterations, I achieved an R^2 of about 0.68, which was a satisfying result to me. Through these models, I identified five key factors that influenced ridership.

Although I experimented with other models such as Poisson regression, I wasn't yet skilled enough to produce acceptable results, so I decided to remove it. This experience gave me valuable lessons in model selection and evaluation. In future projects, I will have a clearer idea of how to choose models, interpret metrics, and recognize what a meaningful output should look like before spending time testing multiple options.

2. Challenges Faced

One of the biggest challenges was obtaining geographic and census block data. At the start, I hoped to include data about major employers, business centers, transportation hubs, and other institutions information to understand how geography affects ridership and potential new station placement. However, these datasets were difficult to locate, and many were either incomplete or unavailable.

When I finally downloaded the block data, I struggled to interpret the column names and spent an entire day trying to fit it in the models—only to realize later that I had misunderstood the data structure. This mistake halted my modeling progress, forcing me to adjust my approach.

Although my final model results aligned closely with POGOH's real-world expansion outcomes, they were much less detailed than I had originally planned. The lack of comprehensive geographic and census data meant I could only identify top candidate communities rather than specific new station coordinates. It was a lesson in how missing data can shape and limit project outcomes.

3. Alternative Approaches

If I could restart this project, I would spend more time on feature engineering and systematically reviewing model types before beginning the business modeling section. I now realize I rushed into analysis without a clear picture, which sometimes felt like walking through a maze without direction.

I also experienced confusion between assignments—my understanding evolved during the second phase, which made Assignment 1 and Assignment 2 feel little disconnected. Next time, I would plan the entire project flow more carefully, ensuring that early decisions align with later modeling and evaluation steps.

4. CRISP-DM Reflection

I followed the CRISP-DM framework step by step, which helped ensure that my work was structured and comprehensive. The phases of Business Understanding, Data Cleaning, and Insight Generation worked particularly well for me.

However, I found that Modeling and feature select were more challenging because CRISP-DM doesn't provide detailed technical guidance for coding or model selection close to my case—it focuses more on project management and documentation. Still, it served as a strong organizational framework, especially for keeping my process consistent and logically connected.

5. Time Allocation

I spent more time than expected on the Business Understanding phase, as there were no stakeholder interviews available. I started with limited background information from the official website and dataset, so I spent about three to four days researching how the overall system operates before I felt confident enough to start writing Assignment 1.

Data cleaning and exploration went faster than expected since the dataset was straightforward and clean. However, the Modeling phase was the most time-consuming and frustrating. I spent about three days searching for additional geographic and population-density data online, but I found very little usable information.

If I could do this project again, I would connect with other students to form a study group and exchange insights about model selection, metrics, and external data sources. Despite the challenges, I’m proud that my linear regression output closely matched the real-world decisions published on POGOH’s website, which confirmed that my analytical approach was realistic and meaningful.

Overall, this project was a valuable learning experience that strengthened both my technical and analytical skills. It helped me understand how different stages of a real-world data science project connect and how planning, data quality, and model selection influence final outcomes. Completing this capstone also gave me more confidence to handle future projects with better preparation, teamwork, and problem-solving strategies.

2.2 Lessons Learned and Future Applications

1. **Technical Lessons**

In this project, I learned how to use and compare linear regression and random forest models. I also tried Poisson regression but didn’t get satisfying results. Linear regression helped me understand how each variable affects ridership, while random forest captured more complex patterns. Both models had similar R^2 scores, but the validation results for random forest were unstable, so I decided to use multiple linear regression. I realized that sometimes a simpler model works better, especially when explaining results to stakeholders.

I also learned how important it is to compare metrics like R^2 and MAE to evaluate model performance instead of just looking at predictions. The dataset was clean, which made it easier for me to focus on testing and improving the models. I also learned how to combine multiple datasets to get better results.   My final regression model used three levels of data — trips, stations, and communities — which made the analysis achieved high R^2 score.

These experiences helped me understand when to use different models, how to evaluate them properly, and how to balance accuracy with interpretability. I now feel more confident about choosing and explaining models in future projects.

2. **Process Insights**

This project helped me understand how important it is to plan the workflow before jumping into modeling. At first, I was too focused on get the first assignment done and I realized later that I should have spent more time reviewing model types and metrics early on. Now I understand that building a clear structure step by step can save a lot of revises later. Slow is smooth, and smooth is fast.

Through this project, I learned how valuable documentation and notes are. After many iterations, I often got lost about what worked and what didn't. Writing down what I tried, what failed, and why helped me stay organized and avoid repeating mistakes. It also made it easier to explain my process when preparing the report.

Another lesson was the importance of iteration. I had to go back and adjust my model several times as I gained a better understanding of the data. It taught me that data projects are rarely linear — you often need to revisit earlier steps to improve results. In the future, I will plan more flexible timelines and expect a lot of back-and-forth process.

3. **Domain Knowledge**

I gained a better understanding of how geography, population density, and nearby stations affect bike ridership in Pittsburgh. I also learned how bias in data can influence future development if not carefully considered. Clustering stations close to each other helps increase ridership, while isolated stations often perform poorly. I now understand how local community patterns, weather, and e-bike access all shape usage.

These insights helped me see how data connects to real-world transportation planning and how modeling can support city decisions. I also learned a lot about Pittsburgh's downtown neighborhoods while collecting core community information. This helped me understand how each area's location and environment influence ridership and transportation needs, as well as how local residents think and what they need in their communities. I also realized that there are different types of stakeholders, each with different priorities and interests. Different stations serve different types of riders, and each group of riders has its own preferences and pattern.

4. Tool Proficiency

In this project, I got more practice using Python, especially pandas, NumPy, and scikit-learn. I learned how to clean and combine datasets, build regression models, and adjust and improve models step by step. I also learned how to debug and compare model performance using different metrics like R^2 and MAE. It took plenty tries to get the models to work well, but every mistake helped me understand more about how each part of the code and data connects. I used Jupyter Notebook to organize my work and keep all my notes and code in one place, which made it much easier to go back and review what I did, especially when I needed to explain my process later in the report.

I also learned how to merge different data sources and deal with errors in a more logical way. Now I have learned to slow down. Overall, this project helped me become more comfortable using Python Modeling for real data science work. I feel more confident about data cleaning, merging, and modeling, and I can see how much smoother my workflow has become compared to when I first started. It also gave me more patience and problem-solving habits that I know will be useful for future projects.

5. **Future Applications**

This project gave me a lot of practical experience in how to put everything together and how to start a study from square one. I now have a clearer idea of how to approach a new data science project—from exploring the data to choosing the right model and explaining the results. I also learned how important it is to keep testing and improving models instead of rushing to the further step. Having a comprehensive plan before making progress is essential because it saves time and reduces confusion later. These lessons will help me handle real-world projects with more confidence, better organization, and a stronger problem-solving mindset.

In the future, I plan to apply what I learned to next projects. The same methods—data cleaning, feature selection, and linear modeling—can also be used in business cases like sales forecasting or customer analysis.

I also realized how much value communication adds to technical work. Being able to explain the meaning behind the numbers and connect results to real decisions is just as important as building accurate models. As I continue learning, I want to improve this part even more, along with gaining stronger SQL and machine learning skills to prepare for my next goal—finding a data-related job and applying what I've learned in a professional environment.

References

POGOH. (2025). *POGOH Annual Report 2023.* Pittsburgh Bike Share. https://pogoh.com

POGOH. (2024). *Pricing and Membership Overview.* Pittsburgh Bike Share. https://pogoh.com/pricing

U.S. Census Bureau. (2020). *American Community Survey: Population and Housing Data.* https://data.census.gov