Assumptions

We make a few assumptions about Citi Bike trip data. First, we assume that all the record trips represent legitimate bike rentals and system-generated test or rides or maintenance are negligible. In addition, we assume that the timestamps are accurate and consistently recorded by the system across months, which allow for reliable calculation of trip duration and other patterns. We also assume that names are correct, despite the fact that stations could be added, removed, or relocated over time. Finally, we assume that missing values (i.e. extremely short or long trips) are rare and don't bias results, as well as the fact that all the labels are accurate and stable.

Data Ethics

This dataset represents human mobility behavior, and we have to analyze this with care to avoid harmful or misleading conclusions. Interpretations about rider behavior shouldn't be stereotyped or stigmatized towards specific groups of people. Comparisons across user types, time periods, or locations should be framed descriptively and subjectively instead of judgementally. Finally, the findings should not be used to support discriminatory policies or any surveillance usages. The goal of this analysis is to improve system efficiency instead of profiling or penalizing riders.

Privacy and Security Notes

The Citi Bike trip data doesn't contain direct personal identifiers such as names, addresses, or payment information. However, the trips do include precise timestamps and station locations, and, combined with external data, could be used to infer individual behavior. Analysis should avoid attempting re-identification or tracking of individual riders across trips, and any visualizations or examples should be aggregated rather than focusing on single trips. This dataset needs to be stored securely and away from personally identifiable datasets. When publishing results, we should only publish summary statistics and anonymized insights.

Failure Playbook

When the model or analysis fails to perform as expected, we have several failure modes to consider. First, poor predictive performance may indicate unhandled seasonality, such as weather effects or major city events not in the dataset. Second, if we inadvertently use future information to predict outcomes that should only rely on start-time data, we may have data leakage. Third, structural changes such as new pricing models, station expansions, or policy changes may cause our set patterns and models to break down. In these cases, we would need to revisit feature selection, retrain models, validate assumptions, and clearly document our limitations. If the results become unstable or misleading, we need to default the analysis to descriptive insights rather than predictive claims.