**Technical Summary of Nairobi Bus Ticket Sales Prediction Project**

**Framing the Problem:** The primary objective was to develop a model that predicts the number of tickets sold for different Bus rides in Nairobi.

**Data Preparation and Feature Engineering**

* Feature selection: For variables with little variation and meaning, we dropped them. In addition, we augmented this dataset with Longitude and Latitude data.
* Creation of new features: A pivotal aspect of the project entailed creating new features such as the distance to Nairobi, bus time gap intervals, time to reach origin, and speed.
* Categorical Encoding Mastery: One-hot encoding was applied to the categorical variables such as day\_of\_week and month.

**Model Selection, Hyperparameter Tuning, and Evaluation**

* XGBoost Efficacy: After investigation we discovered the dependent variable was nonlinear, therefore other supervised learning methods would give us unbiased estimates, unlike XBG boost that can model nonlinear relationships.
* Hyperparameter Optimization: Rigorous hyperparameter tuning was executed via GridSearchCV, a strategy for identifying optimal hyperparameters (e.g., max\_depth, learning\_rate, and n\_estimators) for the XGBoost model. Grid search was configured to minimize negative mean absolute error, a pivotal performance metric.
* Rigorous Evaluation Metrics: The Mean Absolute Error (MAE) emerged as the yardstick for assessing model efficacy. Beyond this, the R-squared (R2) metric was instrumental in gauging the model's capacity to explicate the variance in the target variable. Initially, after splitting the “training data”, both train and test datasets from the original training data were used to ensure comprehensive assessment. After training the model, the MAE (put the figure) and R2(Put the figure).

**Feature Importance**

* Feature Importance Insights: Based on the analysis of feature importance, we ranked list of the key factors that have the most significant influence on predicting ticket sales:

**Lessons Learnt**

XGB overfits the model – To resolve this we reduced on the ratios of the Train and Test Set

Creation of More Features such as hourly travelers and categorizing the months as high, low and medium to segment the period of traffic. This could have given us a lower MAE.