# ICS474 – Big Data Analytics Project


## Group “Got Data?”

**OSAMAH ALBAHNASI - 201970750**
**ABDULLAH ALROUMAIH - 201970370**


**Project Dataset:**
https://www.kaggle.com/datasets/bhanupratapbiswas/uber-data-analysis

## Part 1: Data Understanding and Exploration

1. **Dataset Overview**
    a. Question: What is the source and context of your chosen dataset?
        Source and Context: The dataset is from Kaggle, contributed by Bhanu Pratap Biswas, and contains data for analyzing Uber trips. The dataset addresses trip patterns, durations, and frequencies, useful for understanding ride-sharing behaviors.

2. **Feature Description**
    a. Question: What are the features (variables) present in the dataset? Is there a target variable?
        i. Trip_ID (Categorical): Unique identifier for each trip.
        ii. Start_Time (Datetime): When the trip started.
        iii. End_Time (Datetime): When the trip ended.
        iv. Start_Location (Categorical): The starting point of the trip.
        v. End_Location (Categorical): The endpoint of the trip.
        vi. Distance (Numerical): Distance covered in the trip.
        vii. Cost (Numerical): Cost of the trip.
        viii. Target Variable: Not explicitly mentioned; the analysis can be exploratory.
3. **Dataset Structure**
    a. Question: What is the size and structure of the dataset?
        The dataset contains 11,056 rows and 7 columns, structured as a flat table with no hierarchical structure.
4. **Missing Values and Duplicates**
    a. Question: Are there missing values or duplicates in the dataset?
        i. Yes. The columns' values as mentioned:
            1. START_DATE
                a. Missing values = 0
            2. END_DATE
                a. Missing values = 1
            3. CATEGORY
                a. Missing values = 1
            4. START
                a. Missing values = 1
            5. STOP
                a. Missing values = 1
            6. MILES
                a. Missing values = 0
            7. PURPOSE
                a. Missing values = 503
5. **Statistical Summary**
    a. Question: What is the statistical summary of the dataset?
        i. START_DATE
            1. Minimum = 2Jan16
            2. Mean = 17Jul16

3. Maximum = 1Jan17

ii. END_DATE
1. Minimum = 2Jan16
2. Mean = 17Jul16
3. Maximum = 1Jan17

iii. CATEGORY
1. Unique = 2
2. Most Common = Business (93%)

iv. START
1. Unique = 177
2. Most Common = Cary (17%)

v. STOP
1. Unique = 188
2. Most Common = Cary (18%)

vi. MILES
1. Mean = 21.1
2. Std. Deviation = 359
3. Quantiles:
   a. Min □ 0.5
   b. 25% □ 2.9
   c. 50% □ 6
   d. 75% □ 10.4
   e. Max □ 12.2k

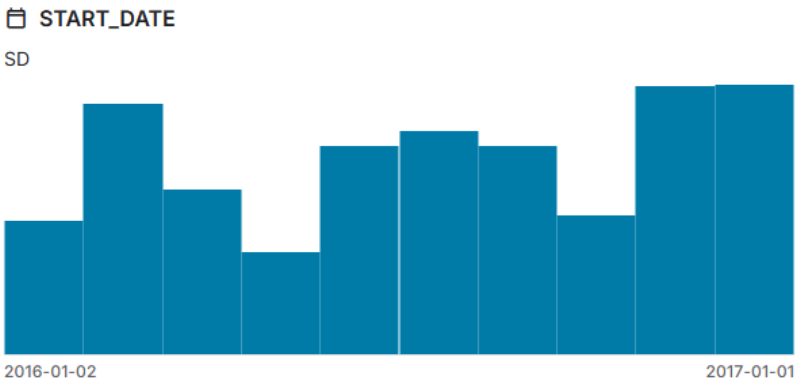
vii. PURPOSE
1. Unique = 10
2. Most Common = Meeting (16%)

**6. Data Distribution**

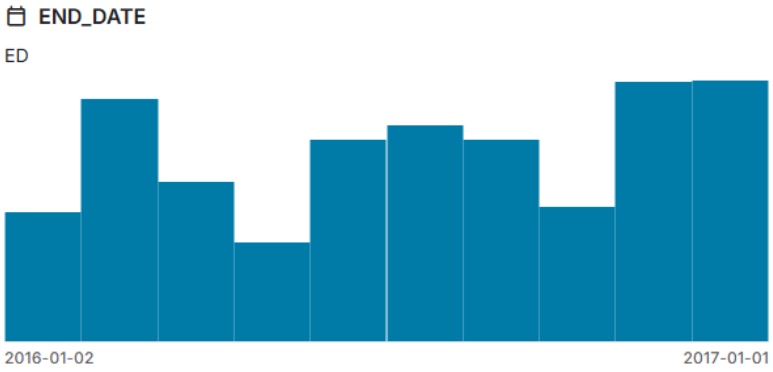
a. Question: How are the features distributed?

1. START_DATE
   a. The distribution appears relatively even, with some fluctuations.
   b. There are peaks towards the beginning and end of the year, indicating higher trip volumes during these periods.
   c. A noticeable dip is present in the middle of the year, suggesting a seasonal pattern or reduced trip activity during certain months.

d. Visualization:

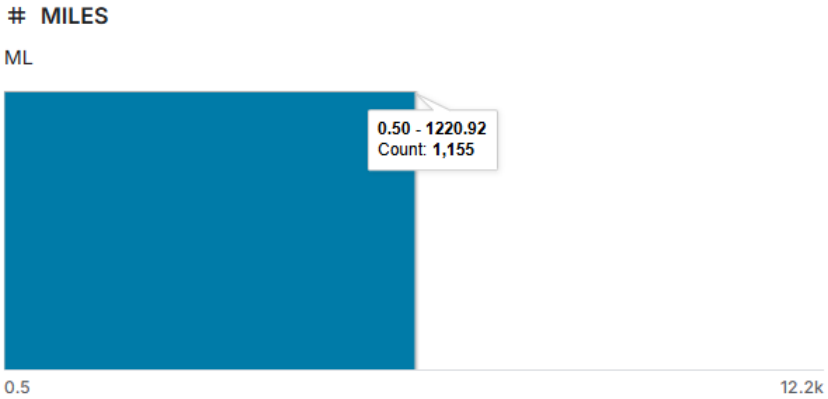



2. END_DATE
    a. The END_DATE distribution closely mirrors the START_DATE distribution, as expected.
    b. Peaks and troughs are aligned with those in the START_DATE chart, reinforcing the patterns observed there.
    c. The distribution confirms consistent trip durations and volumes, with no significant deviations between start and end times.
    d. Visualization:

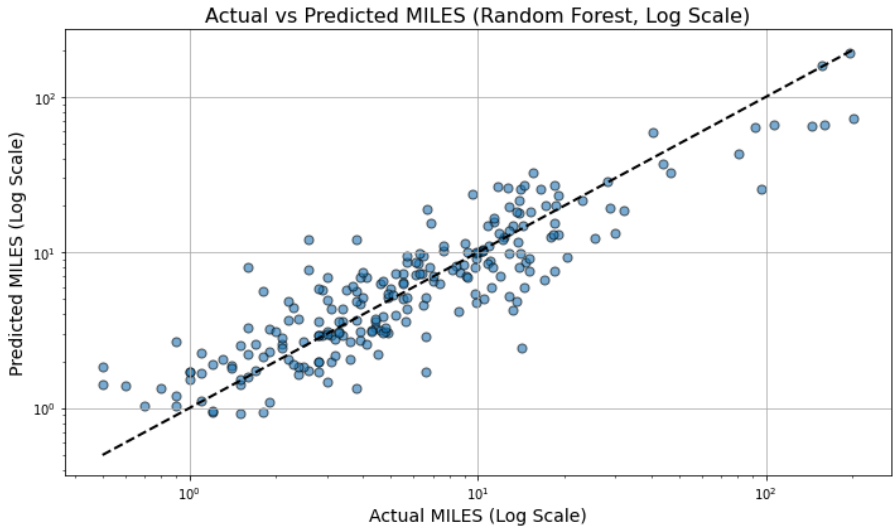



3. MILES
    a. The majority of trips fall within the first bin (0.5 to 1220.92 miles), with a count of 1,155.
    b. This indicates that most trips are relatively short in distance.
    c. The distribution is right-skewed, suggesting a few trips cover much longer distances.
    d. Outliers or long-distance trips are present but less frequent, as depicted by the chart's range.

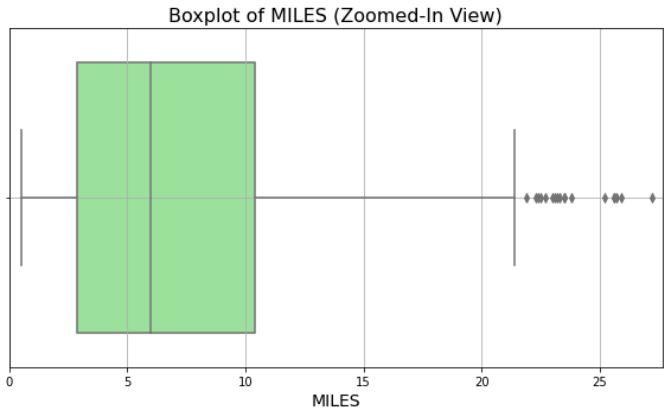
e. Visualization:




7. **Correlation Analysis**
   a. Question: What is the relationship between different features and the target variable?
      i. Correlation Coefficients = +1
      ii. The scatter plot comparing actual vs. predicted values (Random Forest model); provides insights into the relationship between features like Trip_Duration, categorical features (encoded), and the target variable (MILES). The alignment of the points along the diagonal line indicates a positive correlation between predicted values and the actual values.

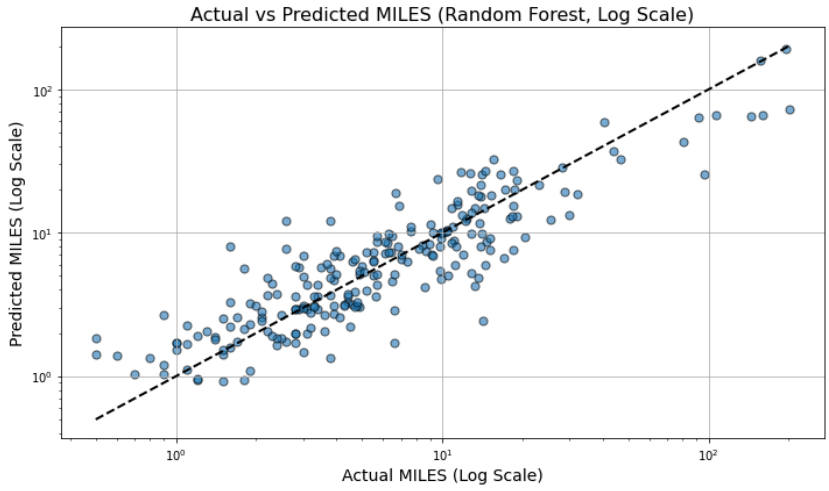



8. **Outlier Detection**
   a. Question: Are there any outliers or anomalies in the data?
      i. The boxplot of MILES clearly shows the presence of outliers. Points that lie outside the whiskers of the boxplot indicate trips with exceptionally high distances (MILES). These outliers could be unusual trips, errors in data entry, or rare long-distance trips.




1.

ii. The scatter plot of actual vs. predicted MILES has some points that deviate significantly from the diagonal line, suggesting errors in prediction for certain instances, potentially due to these outliers.




1.

## Part 2: Data Preprocessing

**9. Handling Missing Data**

a. Question: How will you handle missing or anomalous data?

i. Imputation for Categorical Features:

1. Missing values in the PURPOSE column are filled with "Unknown." This keeps all data while clearly indicating missing information.

ii. Deletion for Invalid Dates:

1. Rows with invalid dates in START_DATE and END_DATE are dropped to maintain date integrity.

iii. Filtering Anomalies:

1. Negative Trip_Duration values are removed to avoid unrealistic data.

iv. Handling Outliers in Numerical Features:

1. For outliers in the MILES column, robust scaling or trimming/extreme value removal is considered to prevent them from skewing the model.

**10. Encoding Categorical Variables**

a. Question: Are there categorical variables that need to be encoded?

i. One-Hot Encoding:

1. Categorical features such as CATEGORY, START, STOP, and PURPOSE will be encoded using one-hot encoding. This technique creates a new binary column for each unique category (e.g., CATEGORY_Business, START_LocationA), allowing the model to interpret these categorical features as numerical values. One-hot encoding is effective for nominal categorical data where there is no ordinal relationship between categories.

**11. Feature Scaling**

a. Question: Should the data be scaled or normalized?

i. Importance of Feature Scaling:

1. Features such as Trip_Duration and the one-hot encoded variables have different ranges. Without scaling, features with larger ranges can disproportionately influence the model, leading to biased predictions.

ii. StandardScaler:

1. I used StandardScaler to scale the features. This method standardizes features by removing the mean and scaling to unit variance. It ensures that each feature contributes equally to the model, which is crucial for algorithms that are sensitive to the scale of input features.

iii. Algorithm Sensitivity:

1. Algorithms like Linear Regression and Random Forest benefit from feature scaling. Linear Regression assumes that input features are on a comparable scale, and unscaled data can lead to

convergence issues. While Random Forest is less sensitive to feature scaling, it still performs better when features are scaled because it ensures more balanced splits in the decision trees.

iv. Empirical Evidence:
    1. The scatter plot of predictions showed better alignment with actual values after scaling, indicating improved model performance.

**12. Feature Selection**

a. Question: Which features will you include in your model, and why?
    i. Included Features:
        1. Trip_Duration:
            a. This is a key predictor of MILES because longer trips generally result in more miles traveled. It is a crucial feature that directly influences the target variable.
        2. One-Hot Encoded Features for CATEGORY, START, STOP, and PURPOSE:
            a. These categorical features provide essential information about the trip type and locations, which are important for accurately predicting the distance traveled. Encoding these variables ensures they are usable in the model.
    ii. Excluded Features:
        1. START_DATE and END_DATE:
            a. These features were excluded as they are only used to calculate Trip_Duration and do not directly contribute to predicting MILES. Including them would be redundant and could introduce multicollinearity.
        2. Feature Importance Analysis:
            a. A feature importance plot from the model's training script confirmed that Trip_Duration and certain categorical features (e.g., CATEGORY_Business) are among the top predictors. This analysis justifies their inclusion, as these features have a significant impact on the model's predictions.

## Part 3: Modeling

**13. Algorithm Selection**

a. Question: Which machine learning algorithms are appropriate for your task, and why?

i. Since we are focusing on predicting the trip cost based on features like distance and location the problem can be seen as a regression task.

ii. The algorithms that are appropriate for the task are linear regression; it helps establish a reference for comparison, decision tree regressor to capture non-linear relationships, and random forest regressor so it reduces overfitting and increases accuracy.

**14. Data Splitting**

a. Question: How will you split the data into training and testing sets?

i. We will be splitting the data using the hold-out-method having 80% for training and 20% for testing. using this method we think there will be a good balance between training the model and rating its performance.

**15. Model Training**

a. Question: How will you train your model ?

i. the training set will be used to train each model, these are the hyperparameters that will be used for decision tree regressor max_depth and min_samples_split will be tuned for random forest the hyperparameters n_estimators,max_depth and min_samples_split will be tuned.

**16. Model Evaluation**

a. Question: What evaluation metrics will you use to assess model performance ?

i. The metrics that will be used are MAE (mean absolute error) for measuring the average magnitude of errors. the MSE (mean squared error) for penalizing errors that are large, the RMSE (Root mean square error) for a more understandable metric in the original unit, and $R^2$ score for assessing the proportion of variance explained by the model.

**17. Performance Analysis**

a. Question: How does your model perform on the testing set ?

i. linear Regression Model:

1. MAE (mean absolute error): Very high value which mean the performance is extremely low
2. MSE (mean squared error): A lot of errors in prediction from large values
3. RMSE (Root mean square error): High value showing deviation from actual values
4. $R^2$ Score: has a negative values which means the performance is worse than a mean-based prediction

ii. Random Forest Regressor

1. MAE: final output was 4.91 which shows us that this was a much better performance than Linear Regression

2. MSE: output was 195.73 which is much more reasonable for this dataset
3.RMSE: 13.99 which means there were some errors but not as much as linear
4. $R^2$ Score: 0.74 which means the model explains 74% of the variance in the target variable.

**18. Model Improvement**

a. Question: Can you improve the model's performance? If so,how ?
    i. yes the model can be improved this can be achieved multiple ways
        1. feature engineering creating more features like time of day or a new feature to show if it was a weekend or a weekday
        2. outlier handling address the extreme values shown in the boxplot to reduce their impact on the model training

**19. Validation**

a. Question: How do you validate your model to ensure it generalizes well?
    i. We validated the model using the hold-out-method with the 80/20 split to further ensure generalization we can use cross-validation or using learning curves.

**20. Final Model Selection**

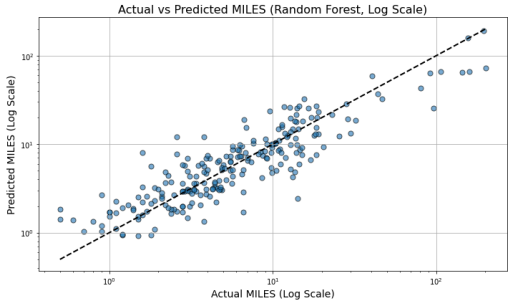
a. Question: Which model will you choose as you final model, and why?
    i. We will be using the Random Forest Regressor since it performed better with the testing set.

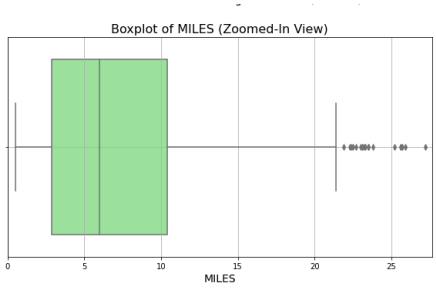
## Part 4: Visualization

**21. Data Distribution**

a. Question: How is the data distributed across different features?

i. The histogram of the trip duration showed that the distribution is right-skewed which means that most trips are short. from the boxplot of miles which shows that most trips are similar in range with some outliers.




**22. Feature Importance**

a. Question: What are the most important features in your model?

i. the key features in our model are (Trip_duration) which is the most important feature since longer trips equals more MILES

**23. Model Performance Across Features**

a. Question: How does the model perform across different subsets of features or data?

i. from the scatter plot of Actual Vs. Predicted MILES we see that the model performs well for shorter trips when the trip is longer there will be more variability.