**Goal**

The aim of this project is to Allocate drivers efficiently where demand is likely to be high**.**

**Libaries used in this project include;**

pandas, numpy, matplotlib, seaborn

sklearn: StandardScaler, KMeans, train\_test\_split, LinearRegression, RandomForestRegressor

**Step 1: Feature Selection & Preprocessing**

* Extracted **hour** and **day of week** from the datetime column.

(We extract temporal features (hour, dayofweek) to capture patterns in ride requests that vary based on time of day and day of the week.)

* Selected key features: temp, humidity, windspeed, hour, dayofweek, and count.
* Applied **mean imputation** for missing environmental values.
* Scaled the data using StandardScaler to ensure all features contribute equally.

**Step 2: K-Means Clustering**

* Removed count (target variable) and retained only input features for clustering.
* Used **K-Means** algorithm to group data points with similar environmental and time conditions.
* Ran the **Elbow Method** to determine optimal number of clusters by plotting inertia across values of k (1 to 10).
* This helps identify the best number of groups in the data — likely between **k = 3–5**.

**Process:**

* Selected numerical features (temp, humidity, windspeed, hour, dayofweek, count).
* Scaled features using **StandardScaler**.
* Used the **Elbow Method** to determine optimal clusters (k=4).
* Applied **KMeans clustering** and assigned each data point to a cluster.

**What the Plot Shows:**

* **X-axis:** Number of clusters (k)
* **Y-axis:** Inertia (sum of squared distances from points to their assigned cluster center)

**Elbow Point Analysis:**

* From the plot, there is a **steep drop** in inertia from **k=1 to k=3**.
* After **k=4**, the rate of decrease becomes much more gradual and flattens.

**The elbow (bend) appears clearly at k = 4**, suggesting that using **4 clusters** is the best trade-off between reducing inertia and keeping the model simple.

**Key Insights for Driver Allocation**

**1. Demand Clusters Identified (via KMeans Clustering)**

Using KMeans, i grouped regions/zones with **similar demand behavior**.

**Insights:**

* **Cluster 0** (e.g.): High ride requests in early mornings and evenings → likely **residential zones**.
* **Cluster 1** (e.g.): Peak ride requests around midday → possibly **shopping or business districts**.
* **Cluster 2** (e.g.): Consistently low demand → **rural or less busy zones**.

**Step 3 . Machine Learning Model Building**

**Features Used:**

* temp, humidity, windspeed, hour, dayofweek, cluster

**Models Trained:**

* **Linear Regression**
* **Random Forest Regressor**