Project Summary

1. The problem statement is to identify the qualities of the facilities provided. Since the given data does not have an explicit variable defining the quality of facilities, I have considered it as an unsupervised problem and fit K-Means to it.
2. The implicit features that give an idea about the quality are ‘p4w_total_gross_revenue', 'rental_count', 'p4w_new_drivers', 'p4w_return_pct', 'searches', 'utilization'. Other features are considered properties of the facilities.
3. An Error in the definition in ‘rental_count’ column. It has been mentioned in *EDA_&_Data_Preparation.ipynb*. So, ‘rental_count’ is replaced with the sum of ‘count_repeat’ and ‘count_first_rentals’.
4. There were some redundant features that meant the same thing but on different timelines, one for the past four weeks and the other had an unknown past. In such cases, features with an unknown past are dropped.
5. Correlated features have been visualized and dropped.
6. Missing values have been imputed with selected values from data analysis. Also, imputations are made with **KNNImputer** where missing values could not be imputed after analysis. All necessary details are provided in Jupyter Notebooks.
7. For feature selection, as features are of mixed types, **PFA** is used after referring http://venom.cs.utsa.edu/dmz/techrep/2007/CS-TR-2007-011.pdf research paper.
8. For fitting K-Means, an optimal number of clusters is chosen and visualized with inter-cluster distances. The clusters seem to be well-separated and variable-sized.
9. Cluster Profiling is done to understand the clusters. The data frame is constructed with clusters in columns and displays their characteristics in rows. For categorical variables, the mode is considered and for numerical values the median.
10. The composition of every cluster is demonstrated using Pie Charts. Each row of pies in *Model_Development.ipynb* (Cell 32) represents core properties for a single cluster.
11. Clusters are ranked based on cluster profiles created. Considering facilities falling in each cluster, quality is determined along with categorical features responsible for the average and poor quality.

Next Steps:

In the developed code, categorical features responsible for poor and average qualities of

facilities were identified. The next step will be to identify the roleplay of numerical features like technical issues, user product issues etc., in determining the quality of facilities.

Also, In the next steps, we can determine R,F,M (Recency, Frequncy and Monetary) scores for the facilities using features like ‘count_repeat’, ’rental_count’, ‘p4w_total_gross_revenue', ‘p4w_new_drivers’, ‘p4w_repeat_drivers’.
This will be done to segment facilities and see which facilities have a higher revenue-generating probability. Such facilities will be targeted and features to be improvised will be determined to increase revenue