**RAW:**

1. **cohort\_sessions\_raw** 🡪 Session level cohort table, containing session level data filtered using for the cohort definition

**PROCESSED:**

1. **cohort\_sessions\_cleaned** 🡪 Session level table after data cleaning (negative nights, duplicate sessions, discount anomalies)
2. **cohort\_sessions\_capped** 🡪 Session level table with outliers handled using Winsorization
3. **cohort\_sessions\_encoded\_booleans\_fixed** 🡪 session level table where categorical columns are one-hot encoded, this file is the input for feature engineering. This file contains cleaned, capped and encoded features ready for feature engineering.
4. **user\_features\_final\_corrected** 🡪 User level table with engineered features (metrics). This csv is the input for Data Transformation (Log Transform + Robust Scaling)
5. **user\_features\_scaled** 🡪 User level table with scaled input features. This csv is the input for Modelling and Segmentation (Baseline Model)

**OUTPUT:**

**Baseline Model**

1. **user\_features\_clustered\_Baseline\_5\_Clusters** 🡪 contains clustering results of baseline model with 5 clusters. Contains all the input features along with a cluster label for each user

**Iteration 1**

1. **df\_true\_cluster\_k5** 🡪 contains clustering results of Iteration 1 model with 5 clusters. Contains all the input features along with a cluster label for each user
2. **cluster\_profiles\_k5\_true** 🡪 contains cluster summary (profile) of iteration 1 with 5 clusters

**Iteration 2**

1. **user\_features\_iteration2\_input.csv** 🡪 contains input features for iteration 2 (with “dollars\_saved\_per\_km “feature)
2. **user\_features\_iteration2\_scaled.csv** 🡪 contains scaled input features for iteration 2
3. **df\_true\_k6** 🡪 🡪 contains clustering results of Iteration 2 model with 6 clusters. Contains all the input features along with a cluster label for each user

**Iteration 3**

1. **user\_features\_iteration3\_input** 🡪 contains input features for iteration 3
2. **user\_features\_iteration3\_scaled** 🡪 contains scaled input features for iteration 3
3. **df\_true\_cluster\_k7\_iter3** 🡪 contains clustering results of Iteration 3 model with 7 clusters. Contains all the input features along with a cluster label for each user
4. **cluster\_summary\_k7\_iter3** 🡪 contains cluster summary (profile) of iteration 3 with 7 clusters
5. **cluster\_k7\_demographics\_percentages\_iter3** 🡪 contains demographic distribution of users by cluster for iteration 3
6. **perk\_mapping\_k7\_iter3** 🡪 contains Perk Mapping strategy for iteration 3 with 7 clusters

**Iteration 4**

1. **user\_total\_spend** 🡪 contains money spent by each user on hotels and flights
2. **user\_features\_iteration4\_scaled\_final** 🡪 contains scaled input features for iteration 4, this file is the input for iteration 4 clustering
3. **df\_true\_cluster\_k7\_iter4** 🡪 contains clustering results of Iteration 4 model with 7 clusters. Contains all the input features along with a cluster label for each user
4. **cluster\_summary\_k7\_iter4** 🡪 contains cluster summary (profile) of iteration 4 with 7 clusters
5. **perk\_mapping\_k7\_iter4\_final** 🡪 contains Perk Mapping strategy for iteration 4 with 7 clusters
6. **cluster\_k7\_demographics\_percentages\_iter4** 🡪 contains demographic distribution of users by cluster for iteration 4