Technical Challenges

EDA:

- Excel does not read file with 2 million records.→ Open file for analysis directly in Python.
- GitHub does not allow to load files bigger than 25 Mb.→ Upload files on GoogleDrive.

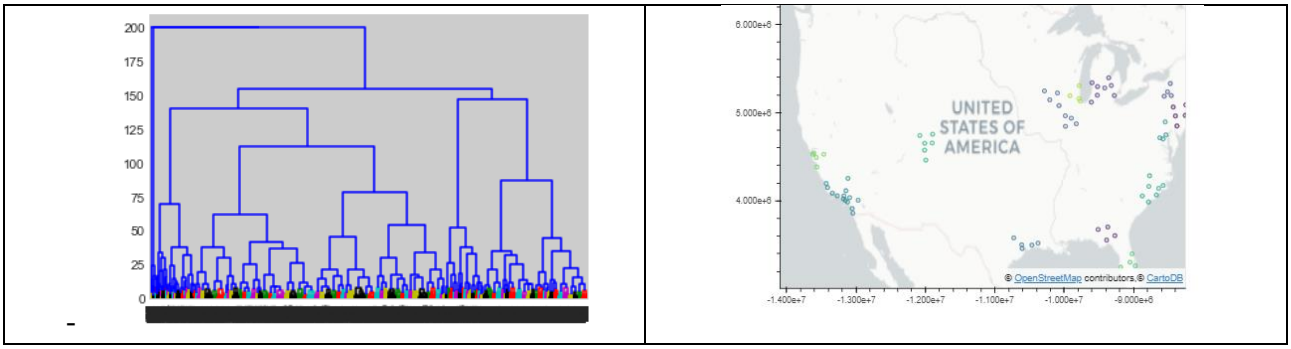
Model Regression:

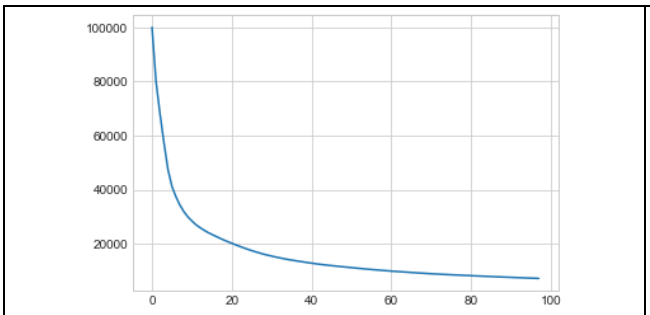
In theory regression models on categorical data are quite simple. However, in practice,

- first, one have to parse categorical variable to numerical data type;
- second, computational mechanism is memory consuming. In my case I had do drop one variable due to high variety of its unique values (which leads to lack of possible important information while making decision and prediction).

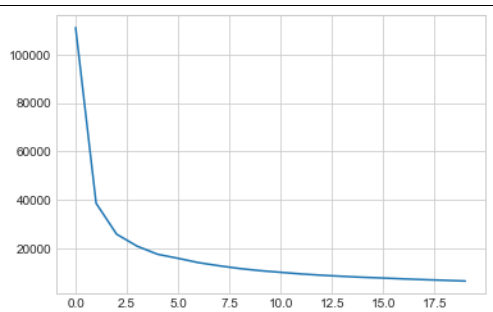
Clustering:

- While feature engineering loading of high volume datasets and high variety (around 300 unique values) of some variables leads to crashing program and auto laptop restarting/or "memory error". → To use smaller datasets (2000 records instead of 2000000) for clustering and avoid complex variables (more than 50 unique values).
- Incovenience with clustering of categorical data – when you look at clusters instead of categorical values you see its numerical representation, which is not informative at first.→ To find way to replace numerical values back by its initial categorical ones.
- Hierarchical and K-means algorithms are inappropriate for my data (see Fig.1). in the former case inter and intra cluster distance are close, in the latter one – intra cluster is even higher, that indicates that both mechanisms are not sufficient for our data. Although, one can increase number of cluster, but it will lead to higher processing time. .→ To choose another clustering method.
- DBSCAN was used only for practice (see Fig.1) – we clustered US airport based on its location. The challenge would be to combine this geospatial data with the average flight delay time, carriers and tickets prices.





Inter Cluster distance 0.6410762635050768
Intra Cluster distance 0.5303611531145348
Inertia 57277.65505278631



Inter Cluster distance 0.11826508214201123
Intra Cluster distance 0.8252159233181614
Inertia 52835.9162160502

Fig.1. *Top left – dendrograms of hierarchical clustering; top right – map of US airport created via DBSCAN; bottom left – inertia of hierarchical clustering; bottom right – inertia of K-means clustering mechanism.*

APP: Some issues with coding and libraries that were not obvious at the beginning and were solved with external help.