**ML for Big Data Project**

**Dataset**

1. https://www.kaggle.com/marcio486/sales-data-for-a-chain-of-brazilian-stores?select=Sales+Report.csv
2. https://www.kaggle.com/codemysteries/salesdb/
3. https://www.kaggle.com/frtgnn/dunnhumby-the-complete-journey

**Imp links**

* https://www.neuraldesigner.com/blog/customer\_segmentation\_using\_advanced\_analytics
* https://www.neuraldesigner.com/learning/tutorials/data-set
* Feature hashing - https://dzone.com/articles/feature-hashing-for-scalable-machine-learning
* Categorical encoding - https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/
* Model selection - https://www.kaggle.com/junkal/selecting-the-best-regression-model
* RandomForest Hyperparameters tuning - https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

Project Idea

**Purchase prediction with dynamic pricing for behavioural consumer segments**

**Steps**

1. Dataset
2. Descriptive Analysis
3. Feature Engineering for Clustering -> RFM
4. Clustering into different segments -> K-means / DBSCAN with Sihoutte hyper-parameter optimization
5. Predict if a “customer segment” will purchase or not given product features, price, date, and other features. Is dimensionality reduction required? If yes then maybe build a pipeline. Can the model improve incrementally?
6. Change price to see the behaviour of the segments? **Target customers with right prices.**
7. Incremental nature:
   * Show how new customers can be added to segments by Clustering prediction model?
   * Prediction if that customer will buy or not?
   * Can we add a new product incrementally?

**Web References**

https://docs.dask.org/en/latest/dataframe.html