**Outline of Analysis**

Name: **Kevin Patel**

**Siva Ram Kottapalli**

I. Description of Data

1. **What data is** – The data is an ecommerce public dataset of orders made at [Olist Store](http://www.olist.com). Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. It also contains a a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.
2. **Where it comes from** – The dataset is picked up from Kaggle and shared below is the link.

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?datasetId=55151

1. **Extent (size, number of years covered, etc.)** – The dataset has information of about **100k** orders from **2016** to **2018** made at multiple marketplaces in Brazil.

The dataset has 9 csv files with and below is a brief description of the same.

1)**olist\_orders\_dataset**: Contains all the key details about the orders. Includes fields such as order\_id, customer\_id, order\_status etc.  
2) **olist\_order\_items\_dataset**: Contains the details of items purchased such as item\_id, product\_id, and price.   
3) **olist\_order\_reviews\_dataset**: Contains details related to reviews posted by the customer.  
4) **olist\_products\_dataset**: Contains details related to a product such as the ID, category name and measurements.  
5) **olist\_order\_payments\_dataset**: Contains payment details for an order and includes details such as payment\_type, payment\_installments & payment\_value.  
6) **olist\_customers\_dataset**: Contains details about customers and includes fields such as customer\_id, zip\_code etc.  
7) **olist\_geolocation\_dataset**: Contains geographical information related to both the sellers and customers and contains fields such as State, City, Latitude, longitude etc.  
8) **olist\_sellers\_dataset**: Contains the seller information and includes fields such as seller id, zip-code and city.

1. **Question 1**— **Payment value prediction for orders based on various parameters.**
   * + 1. **The data that will be used for this question (specific identification of fields, especially target)**

**Target variable**: payment\_value (olist\_order\_payments\_dataset)

**Predictors**: product\_type, delivery\_hours, freight\_value, review\_score, payment\_sequential, payment\_installments, product\_weight\_g, product\_length\_cm, product\_height\_cm, product\_width\_cm (this we are using for base-line model)

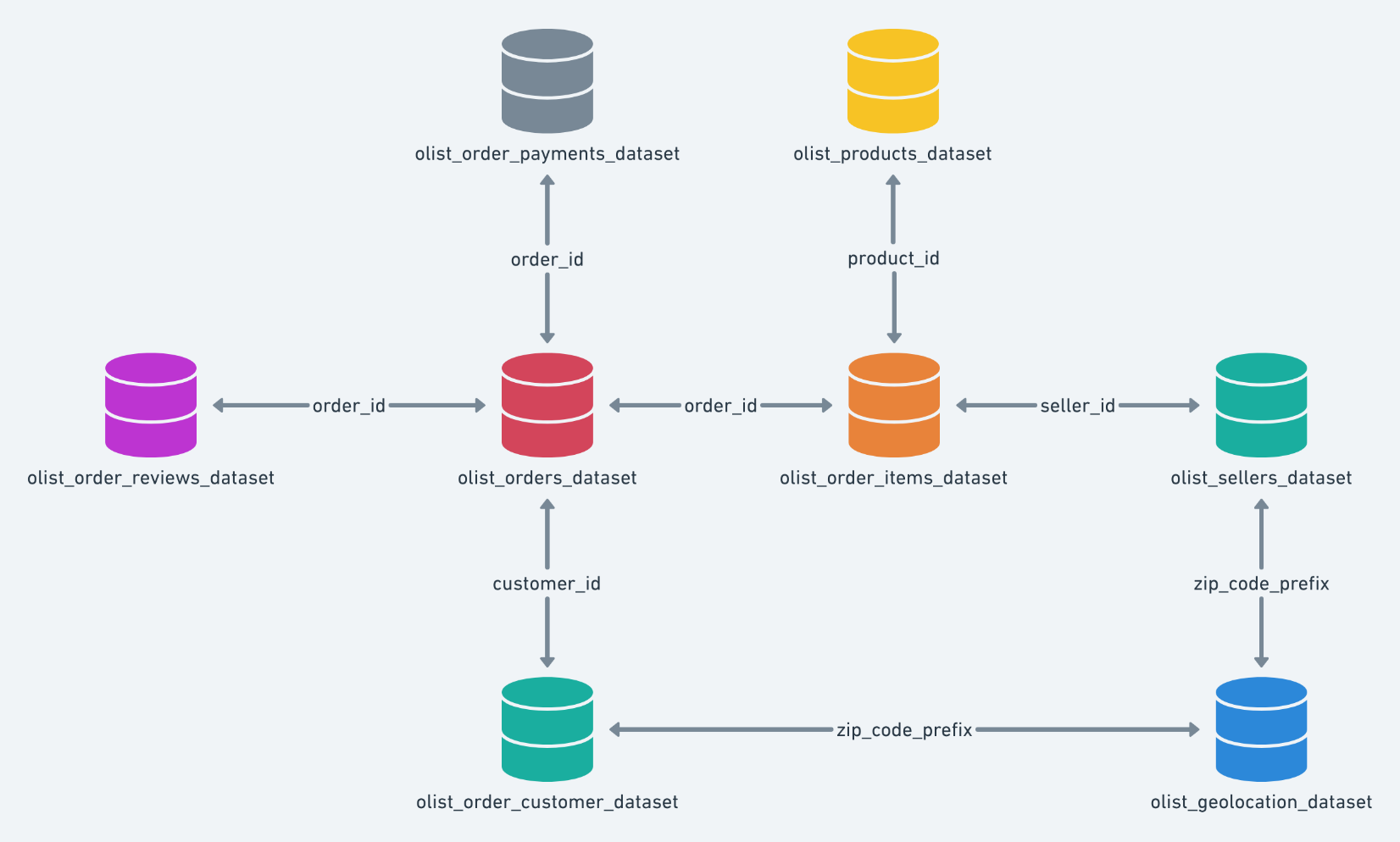
The predictor parameters are from various datasets and can change subject to the evaluation during the course of project implementation.

* + - 1. **Description of what you do with the data before modeling**

1. **Merging of datasets**

The requisite data is spread across multiple csvs. The first step is to primarily merge all the tables into a single dataset. We will mostly use inner joins to merge the datasets.

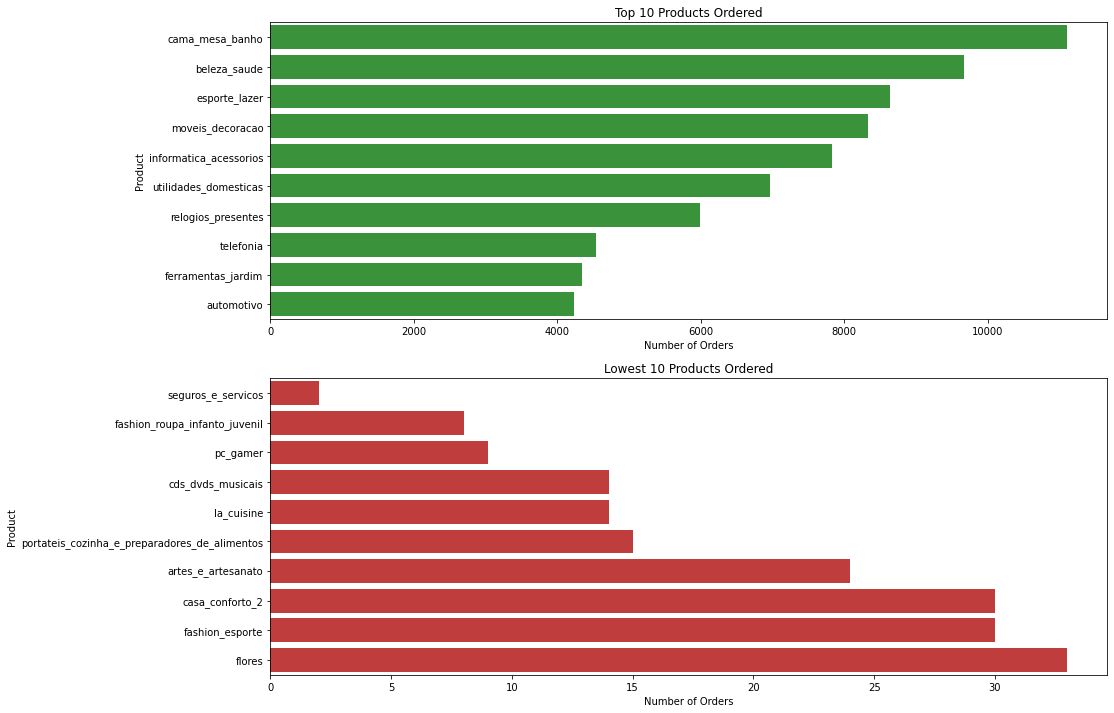
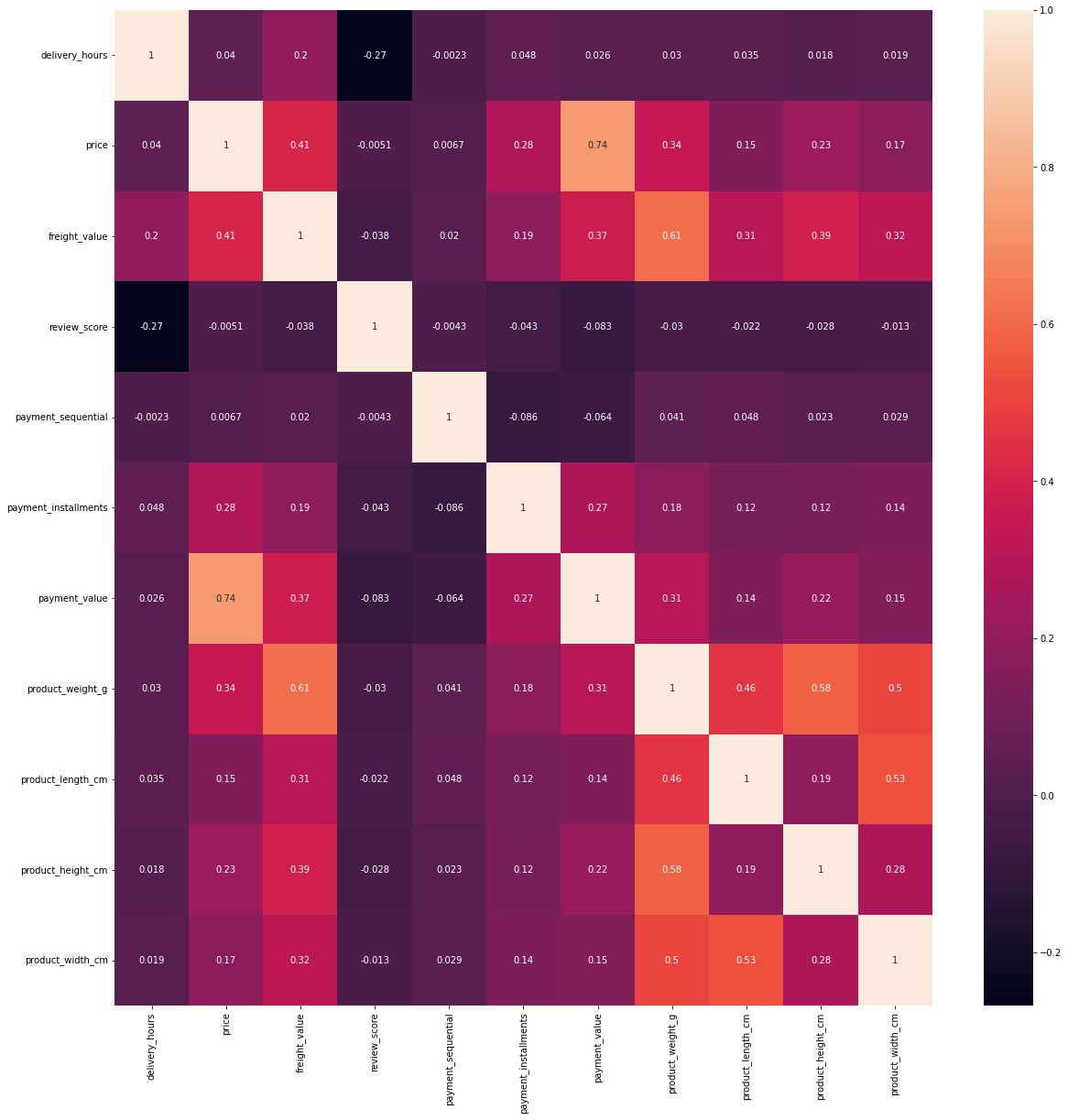
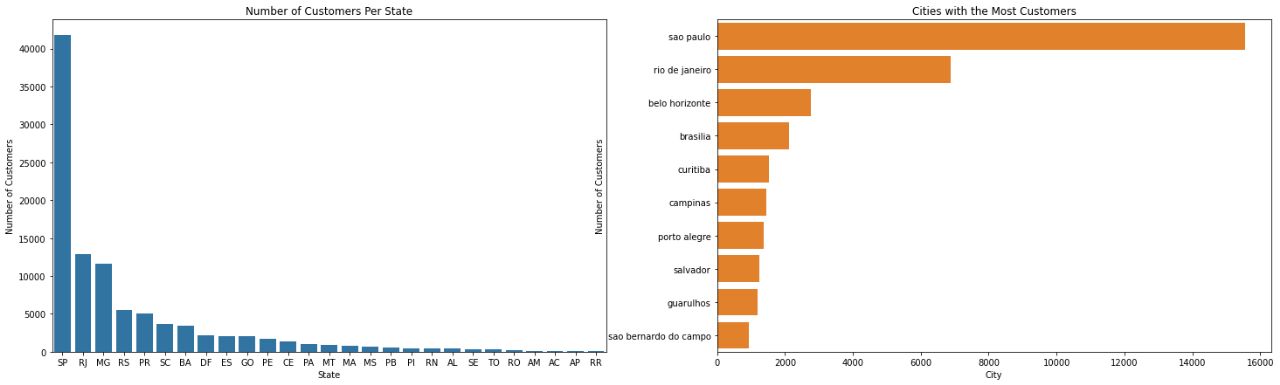
Below is the data schema representation and the corresponding foreign key-parameters (order-id, customer-id, product-id, seller-id).



1. **Exploratory Data Analysis**

As part of the exploratory data analysis, we will perform the below activities (and a few more).

1. Check for irrelevant columns. The dataset is huge and most of the columns linked to geospatial data and dates are redundant and/or not be relevant for the problem we are trying to solve. Some of the examples include order\_delivered\_carrier\_date", "order\_delivered\_customer\_date", "order\_estimated\_delivery\_date", "delivery\_duration", "shipping\_limit\_date", "review\_creation\_date", "review\_answer\_timestamp", "order\_approved\_at", “geolocation\_lat”, “geolocation\_lng” etc. We will clean the data of these columns so as to have more relevant features for analysis.
2. Check for missing values and outliers in the predictor variables. Some of the parameters used for evaluation like product\_weight\_gm, payment\_installments etc have multiple outliers and we will try to remove outliers so as to avoid any bias while building the model.
3. Investigate the relationship between the predictor variables based on correlation matrix, scatter plots etc.
4. Some of the prediction parameters like product\_weight\_gm have big values while others like freight\_value, payment\_installments etc are small. So as to avoid issues with the imbalance in values, we plan to use Scaling/Normalization.
5. Also, we have some categorical variables like City, State names and these may require to be changed to Numerical values in case we use the same for our model implementation. We will probably use encoding methods like one-hot encoding etc for this purpose.
6. We may also introduce new features as part of the feature engineering. For now, we could think of 1 feature - delivery duration, and feel it as relevant to payment\_value.
7. Below are few plots specific to exploratory data analysis done for the project.

1. **Clustering and PCA**

For the problem statement (payment\_value prediction), clustering may not be relevant and out of scope for our project. Some features like review\_score can be assessed using clustering.

* + - 1. **Description of Modeling to be done**
    1. **Type of Model:**

Our target variable ‘payment\_value’ is a continuous variable and so its reasonable to use regression models for prediction purpose. We will start with a simple linear regression model and also explore other regression models such as polynomial regression, decision tree regression, random forest regression etc. We also plan to use ensemble models for the project purpose.

* + 1. **What will be used as a baseline model:**

Our baseline model will be linear regression model and we will try to use different features to predict and improve overall accuracy.

* + - 1. **Analysis of Models:**
    1. **Error analysis to be done:**

For error analysis, we will use root mean squared error (RMSE) as the evaluation metric. We will calculate the RMSE of the predicted values against the actual values for the test set.

We will also explore:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. R-squared value
4. Residual plots
   * 1. **Comparison to baseline**

We will compare the various models’ performance to a simple baseline model (e.g. mean, median, or linear regression) using metrics like Confusion matrix, Accuracy, Precision, Recall, F1 Score.

* + 1. **Visualization of results**

For visualization, we will use various scatter plots, line plots etc to depict the metrics details and comparisons of models predictions.