* **Problem Statement**

Taking a look at all of the Different Stores for this Project It is the responsibility of sales departments all around the world to forecast their daily revenue in advance. Store sales are affected by a wide variety of factors, including seasonality, location, promotions, competition, school and state holidays, and holidays observed by individuals. With thousands of sales records, we will be making predictions about future sales depending on the specifics of each customer's situation; hence, the accuracy of our findings can vary quite a little. The objective at hand is to make a projection for the "Sales" column.

* **Literature Review**

1. **Retail Sales Forecasting Using Deep Learning: Systematic Literature Review**

In this comprehensive overview of the relevant literature, we focus on deep learning (DL) models for retail sales forecasting. The precision of a retail sales estimate is a critical factor in ensuring that business operations will go without interruption. Accuracy in the retail industry involves reducing the expenses associated with the supply chain and storage, preventing any products from being out of stock, and ensuring that promotional activities run smoothly. This study conducts an analysis of the DL frameworks that were utilised in the previously evaluated literature. DL models that have been tested, in addition to other machine learning and linear models that were utilised for evaluation comparison, are included here. In addition to that, the review provides a presentation of the metrics that the authors utilised when evaluating the model. This essay comes to a close by discussing the advantages of using DL models for sales forecasting as well as their drawbacks.

1. **Sales Forecasting for Fashion Products Considering Lost Sales**

For businesses in the fashion retail industry, predicting sales of new products with a high degree of accuracy is essential for enhancing management effectiveness and customer happiness. Forecasting is challenging due to the substantial censored demand issues caused by the low inventory strategy of fashion products and the low stock levels at each physical and mortar store.

In order to forecast the overall sales of new products, this research proposes a two-layer (TLs) model. Demand is calculated at the top layer using linear regression (LR). Sales are represented in the second layer as a function of both demand and inventory. The feature selection process for the TLs model is carried out using the gradient-boosting decision tree approach (GBDT). A mixed k-mean approach is used for product clustering, and a genetic algorithm is used for parameter estimation in each cluster, taking into account the heterogeneity in products. Our model outperforms LR, GBDT, support vector regression (SVR), artificial neural network (ANN), and real-world data from a Singaporean corporation in most circumstances, according to experimental findings. Additionally, two metrics are created: the average conversion rate and the marginal conversion rate, which examine the ideal inventory level and quantify the competitiveness of the products, respectively. These metrics enable fashion industry managers make decisions.

* **Datasets**
* This data set is from a Kaggle. The goal is to forecast product sales for Corporation Favorita, an Ecuadorean Grocer with hundreds of stores and about 200,000 unique products. The dependent variable will be unit sales by store and item.

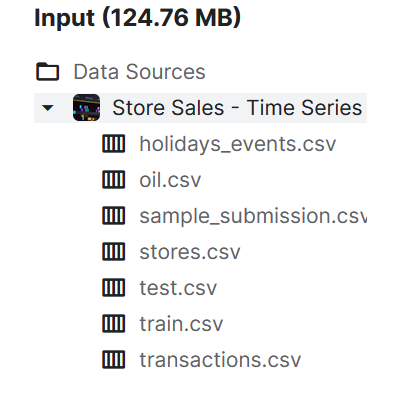


Fig: Dataset CSV files

* The 6 data sets used for exploratory analysis are shown in the table below with their variable names, a count of the variables, and the number of observations in the data set. Corporation Favorita has provided several data sets to help predict sales. The table below shows some basic information about each data set.

| Dataset | variables | Variable.count |
| --- | --- | --- |
| train | id, date, store.nbr, item.nbr, unit.sales, onpromotion | 6 |
| oil | date, dcoilwtico | 2 |
| holidays | date, type, locale, locale.name, description, transferred | 6 |
| items | item.nbr, family, class, perishable | 4 |
| stores | store.nbr, city, state, type, cluster | 5 |
| transactions | date, store.nbr, transactions | 3 |

**train** - The primary data set is *train*, with over 125 million observations (10% used - shown here). This is the most basic sales data, with a date/store/item, how many were sold, and whether the item was on promotion when it was sold. The included data begins on January 1st, 2013, and ends on August 15, 2017.

**oil** - Daily oil prices. Ecuador’s economy depends on oil prices - lower oil prices have been a challenge for Ecuador, and we’ll use oil prices to help predict grocery sales.

**holidays** - A list of all holidays, including local holidays, wherever Corporation Favorita has stores. The first holiday in the data is on March 2, 2012, and the last is on December 26, 2017. The type and transferred variables are of particular interest. A transferred holiday is a holiday that falls on a weekend and so is moved to another day. So - a transferred holiday is not a holiday at all. However, if the type is “Transfer,” the date had a holiday transferred to it from the weekend. Generally, this is a weekend holiday being moved to a Monday or Friday, as we do in the United States. There is one slightly more complicated type that we don’t see much in the United States. Bridge days are days added to holidays to extend a long weekend. These days are made up by working on a day that would typically be a day off. This could get fun!

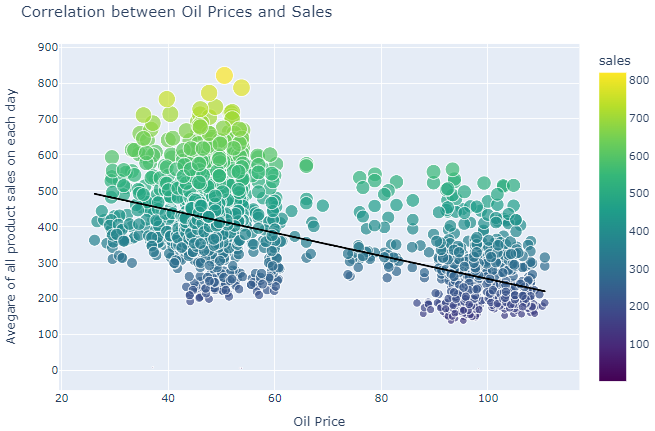
**items** - A listing of each item sold in the stores (with an item number, not a name), the “family” (grocery, cleaning, bakery, and others), class (unclear what this means), and whether or not the item is perishable. We’re using a subset of items for the analysis to reduce memory requirements.

**stores** - A list of stores by location, type, and cluster. It’s not clear what type means. The group is a pre-selected clustering of stores based on their characteristics, though no detail indicates what parts were used in the clustering. Whoever did likely the clusters had more detail than was provided with this data. However, we may try to cluster separately or at least identify which characteristics were used in clustering.

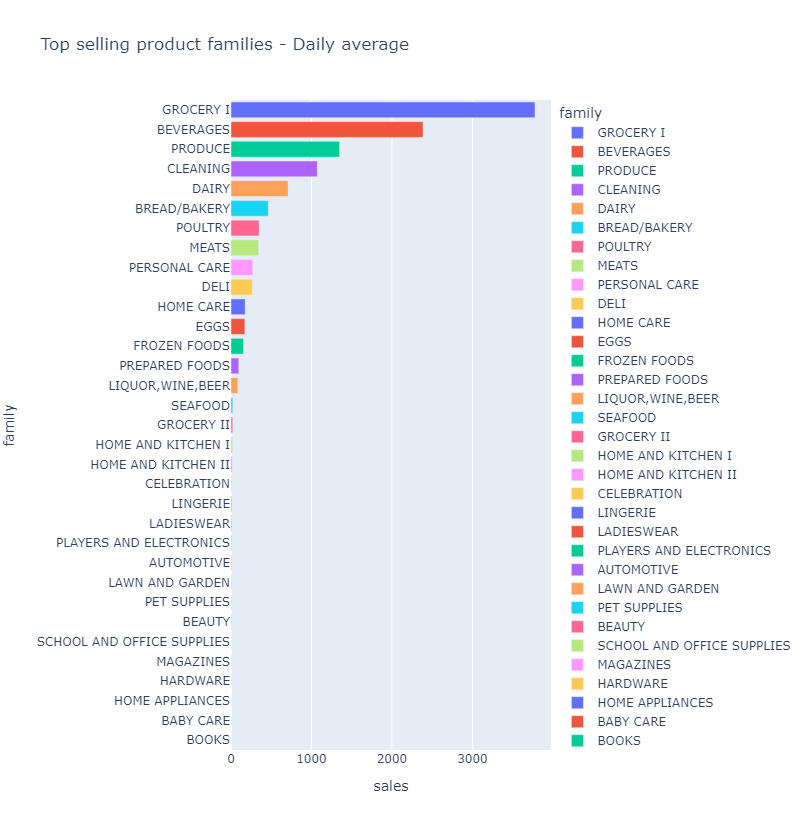
**transactions** - These are not individual transactions (though that would be nice!). Instead, it counts the number of transactions by store, by day. The data beings on January 1st, 2013, and ends on August 15, 2017.

* **Methodology**

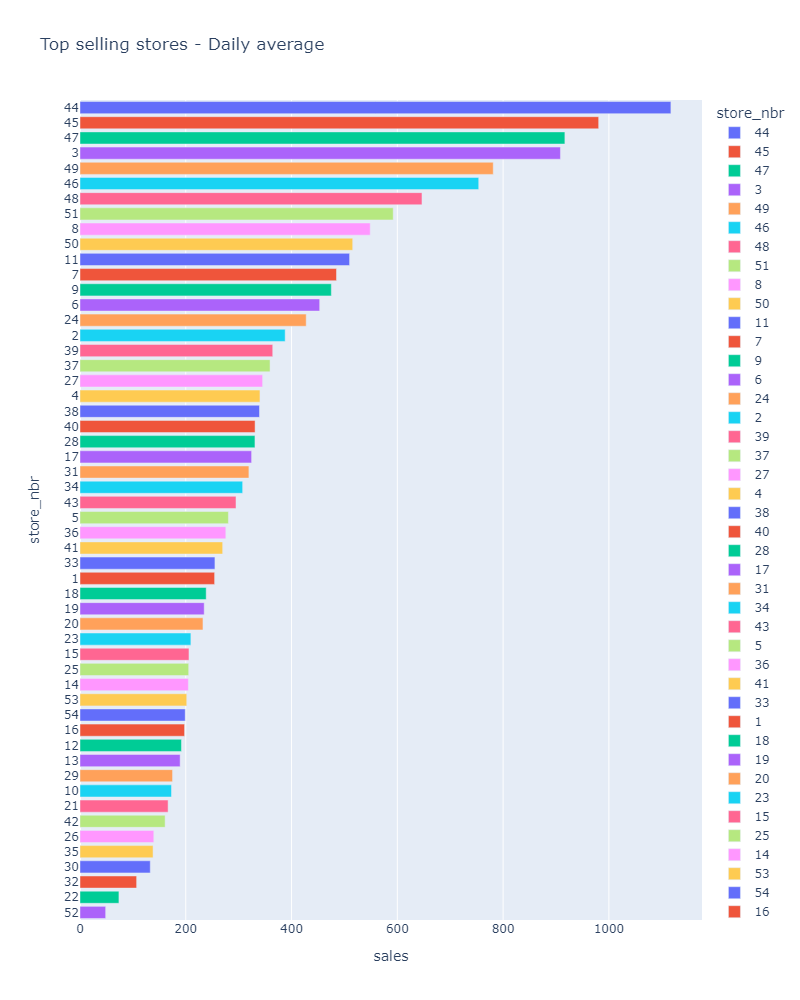
An initial step is to bring in the library and the data set. Given that the imported datasets originally contained multiple files, a merge operation is performed. In this case, the oil dataset is combined with the sales dataset because of the strong correlation between them. The correlation is as shown below:



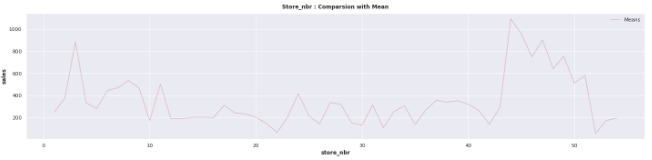
As can be seen in the diagram below, the next step is to conduct exploratory data analysis.

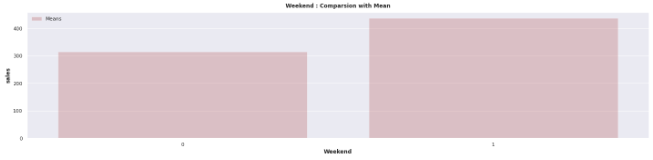


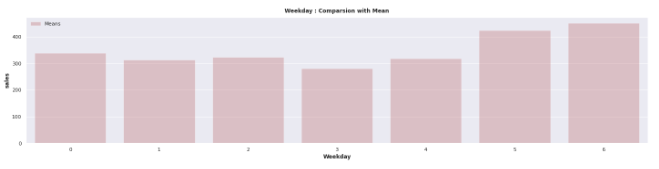
The highest volume of sales were observed for Grocery 1, Beverages, and Produce. Finally, the revenue of each shop is tallied, and the top five moneymakers turn out to be Shops #44, #45, #47, #3, and #49.



The average sales are then calculated and correlated with factors such as store ID, store family, Onpromotion, Dcoilwtico, holiday type, name, description, state, shop type, cluster, year, month, day, and weekday. It is depicted below.







So, we turned to the XGBoost Model to forecast store sales. This machine learning library, known as XGBoost (Extreme Gradient Boosting), is a scalable, distributed gradient-boosted decision tree (GBDT).

The premise on which gradient boosting is based is that using the best possible next model in conjunction with the best possible previous models will result in the lowest possible prediction error. The key is to define the desired results of the next model run so as to restrict the error as much as possible.

Faster execution and better model performance are the main draws to XGBoost. When it comes to classification and regression predictive modelling problems, XGBoost is unrivalled when using structured or tabular datasets.

The model's accuracy in both training and testing is detailed below.



Below is a graph representing the XGBoost Regression model. Predicted value is compared to a fixed value (reality) to demonstrate the relationship between the two.

