# **Group 5 - Assignment Summary**

## **Abstract**

Objective of assignment is to build a prediction model to predict the total sales for every product and store in the next month for the 1C Company based on the training data. The dataset chosen for the analysis is **"Predict Future Sales"** dataset. Dataset was downloaded from the public dataset on Kaggle at the URL, <https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data>.

Key dataset includes **“Sales\_Train\_V2.csv”** which provides item wise, month-wise sales information of the shop. It also includes price and the item category. Dataset contain close to 3 million records. Other support dataset includes, shops.csv, item\_categories.csv, items.csv, test.csv provides further elaboration of the main data.

**Data Analysis -** Initial analysis revealed the little need for the data cleansing. Shop name, item name and all the text within the dataset are in the Russian language. However, it doesn’t limit implementing and executing the prediction model. Created new data elements such **Year, Month and Day** fields to help time series analysis and **sales** elements from item\_price and item\_cnt\_day. Merger of the item\_category from from item\_categories.csv ensured the model can manipulate category wise sales.

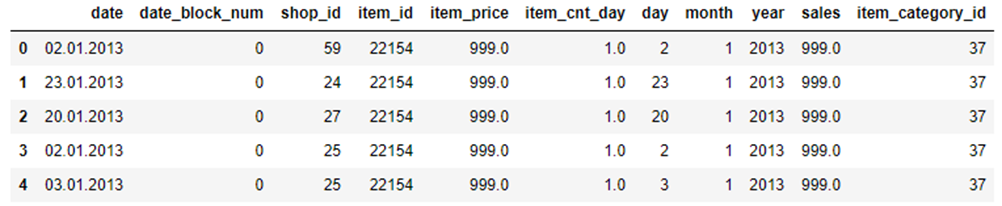


Figure : Sales\_Train\_V2 dataset after cleanup

**Feature Analysis**

Feature analysis of the time series data is provided below

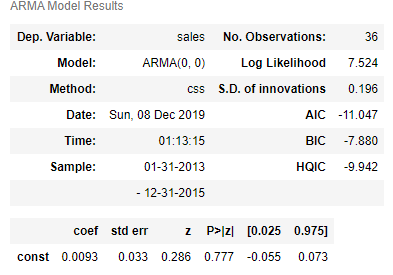
**Seasonality Summary**

* **Sales by year** - Year trend chart shows that 2014 total sales are higher. 2015 only contains 10-month data, so it is not comparable with the other 2 years.
* **Month wise Sales - Analysis** of e Bar Chart & the Boxplot show that the Total Sales in the month of November and December are higher compared to other months.
* **Sales by Day -** Both the Bar Chart and the Boxplot show that the day 19 and day 30 have highest total sales.

**Sales and volume by shops**

* The Bar Chart shows that Shop ID 31, 25, 42 has the top 3 total sales.
* The box plot suggests that Shop ID 9 and Shop ID 20 has the highest daily sales for single item
* The Bar Chart for volume shows that Shop ID 31, 25, 54 has the top 3 total sales volume.
* **Sales by Day -** Both the Bar Chart and the Boxplot show that the day 19 and day 30 have highest total sales.

**Prediction Model & Execution**

****We used **ARIMA model** to model the future sales prediction. **ARIMA** stands **‘AutoRegressive Integrated Moving Average’**, is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. We will focus on the **Sales forecast** for the performance

Execution of dickey-fuller test rules out the stationarity(seasonality). However, a trend was noticed with the decomposed plot. The P static value is less than the critical value, it demonstrates that the data is stationary.

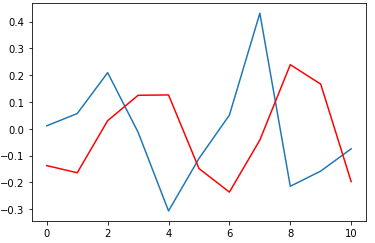
We implemented the ARIMA model using Python module **pmdarima.** By running ADF, KPSS and PP test, we landed on 0,1,0 parameters to feed into the Model.

Figure : ARMA Model Results

From the ARMA model result, one could infer, there is no AR values which means the series is stationary.

We also explored variations by tweaking the inputs parameter and model parameters to see the impact to the model result. However, all of them produced a P value significantly more than the earlier set value of 0.05.

**Test data vs Prediction Model Observation**

****Following diagram provides the view of the model comparison. We ran the model for the entire test dataset estimated around 215,000 records. Model predicted sales date (**Red)** closely followed the test data depicted in **blue**. Key model performance is given below

* Mape : 3.919
* Standard error : 3.919
* MPE : -3.561
* RMSE: 0.293

Figure : Test(Blue) vs predictions (Red)

**Conclusion**

We have successfully built a sales prediction model using ARIMA that predicts the future sales data for a given data set. From the **Test vs prediction graph**, one could infer the overall the predicted values are close to the test dataset. MSE implies on average the model was wrong 8.6% of the times. **MAPE at 3.19% implies that our model is 96.09% accurate and RMSE of 0.293 shows the errors are trending towards zero.** Model parameters could be further refined to reduce the variation between actual and predicted value for business usage.