Predict Future Sales

Predict Future Sales

Collaborators in no order:

Temidayo Adejobi, Beniamkem Koffi, Haiming Luo, Karen Parra, Elanchezhian Vaithianathan

**Contents**

[1 Objectives 3](#_Toc26213368)

[2 About Dataset 3](#_Toc26213369)

[2.1 Dataset “sales\_train\_v2.csv” 3](#_Toc26213370)

[2.2 Other supplementary dataset. 4](#_Toc26213371)

[2.3 Data Analysis 4](#_Toc26213372)

[3 Feature Analysis 6](#_Toc26213373)

[3.1 Seasonality Summary 6](#_Toc26213370)

[3.2 Sales Distribution by Shop. 8](#_Toc26213371)

[3.3 Sales Distribution by Item Category 9](#_Toc26213372)

[3.4 Sales Distribution by Item Variety 10](#_Toc26213372)

[4 Prediction model implementation 12](#_Toc26213374)

[5 Execution of Model on Test dataset 12](#_Toc26213375)

[6 Conclusion 12](#_Toc26213376)

[7 Associated files 12](#_Toc26213377)

[8 References 12](#_Toc26213378)

**Report Version History**

|  |  |
| --- | --- |
| **Version No.** | **Description** |
| 0.1 | Draft version – introduction |
| 0.2 | Update the data loading and profiling section |
|  |  |

# Objectives

Objective of this report is to build a prediction model and predict the total sales for every product and store in the next month for the 1C Company. To tackle this problem, this requires data wrangling and cleaning, data transformation to make it stationary and supervised and model building. Predicting the future sales of one’s business can be used as a benchmark, budget planning and planning for demand and supply for specific product items and store.

The dataset chosen on which the analysis is "Predict Future Sales" dataset

The dataset was downloaded from the public dataset on Kaggle at the url, <https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data>

The dataset is being used under the terms of the license below.

License: This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/.

# About Dataset

Kaggle’s Predict Future Sales dataset is a time-series dataset consisting of daily sales data provided by one of the largest Russian software firms – 1C Company.

1C Company is a leading Russian software development firm specializing in development, distribution, publishing and support of mass-market software. They are known for video game development and have several internal studios. Most popular titles produced by the company are *Il-2 Sturmovik*, *King’s Bounty*, *Men of War* and *Space Rangers* series. 1C Company is the official distributor of top vendors such as Microsoft, Novell, Symantex, Borland and over 100 other software vendors

Any public user can download the Google Play store data from Kaggle at no cost. Users n needs to register with Kaggle and sign-in to access this dataset.

Brief descriptions of the column names for the datasets from Kaggle are outlined in Tables 1 and 2.

## Dataset “sales\_train\_v2.csv”

Sales\_Train\_V2 dataset is the core component of the dataset. It provides month-wise sales information of the shop with price as well. Refer to below table for the dataset info. This dataset contains close 3 million records.

|  |  |
| --- | --- |
| **Feature name** | **Description** |
| date | Date of the sales |
| date\_block\_num | Consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33 |
| shop\_id | Overall user rating of the app (as when scraped) |
| Item\_id | Number of user reviews for the app (as when scraped) |
| Item\_price | Size of the app (as when scraped) |
| Item\_cnt\_day | Number of user downloads/installs for the app (as when scraped) |

**Table 1 – Description of the “sales\_train\_v2” datasets**

## Other supplementary dataset.

In addition to sales\_train\_v2.csv, Predict future sales dataset includes following dataset

1. shops.csv - Shops id to shop name
2. item\_categories.csv – item name, item id and item category mapping.
3. items.csv – mapping of item name to item id
4. test.csv - dataset to the prediction model.

Following table captures the data elements available in the supplementary dataset.

|  |  |
| --- | --- |
| **Feature name** | **Description** |
| item\_name | Name of the item |
| Shop\_name | Name of the shop |
| Item\_category\_name | Name of the item category |

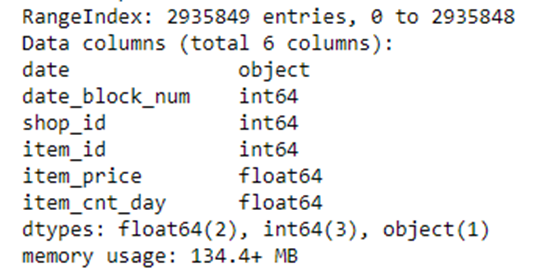
**Table 2: Description of "Supplementary dataset review" dataset**

## Data Analysis

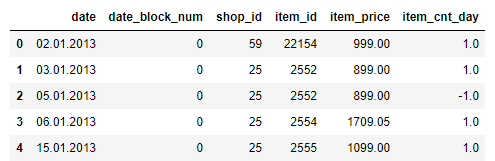
Leveraged Python to perform the data analysis. Python and its rich modules provide rich capabilities to analyze, transform and visualize observations. Some of the key packages of python includes Numpy, Pandas, Matplotlib, Seaborn, Sklearn. Most of these packages were used in the analysis presented in this report.

Basic analysis of the sales\_train\_v2 dataset is given below

**Info Summary**

****

**Head info**

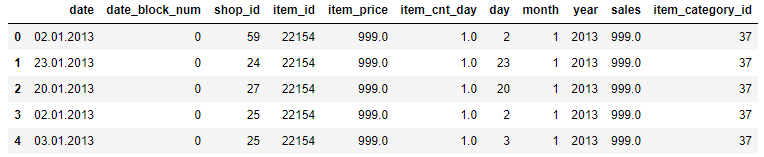
****

Initial analysis revealed the little need for the data cleansing. There is no null filed in the main sales dataset. Shop name, item name and all the text are in the Russian language. However, it doesn’t limit implementing and executing the prediction model. Date fields requires refinement to convert into a proper date format.

**New data elements**

* Splitting the date field into **Year, Month and Day** fields to help time series analysis.
* Add a new element - “Sales” using item\_price and item\_cnt\_day. It provides the net value of the sales for a given month and shop id.
* Merge “items\_category” field into the main data frame on item-id. This would provide category-based analysis of the sales.

View of the Train dataset post new data and merge is given below



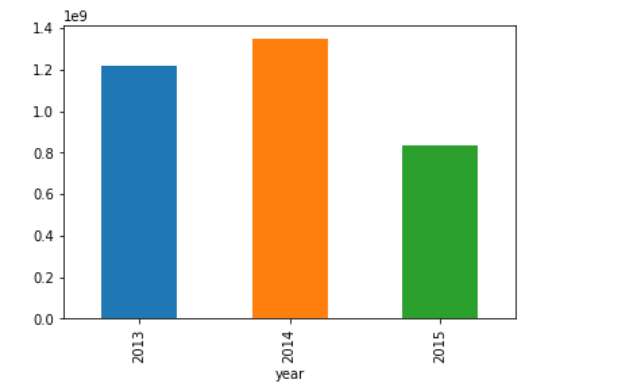
# Feature Analysis

## Seasonality Summary

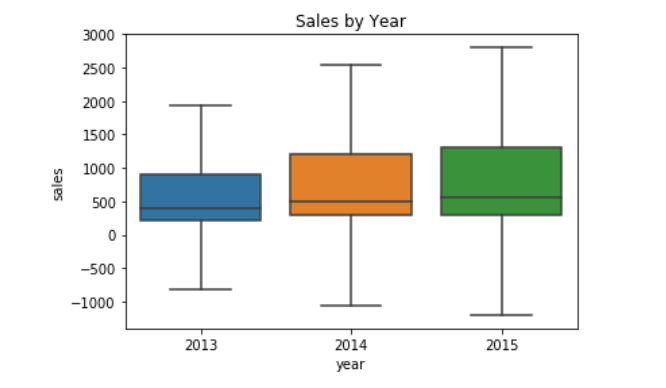
**3.1.1 Sales by Year**

* The year trend chart shows that 2014 total sales is higher than 2013. 2015 only contains 10-month data, so it is not comparable with the other 2 years.
* Judging from the plots, the Median, 1st quantile and 3rd quantile, as well as minimum and maximum for the daily sales for unique item for 2014 and 2015 are higher than 2013.

**Total Sales by Year ($)**

****

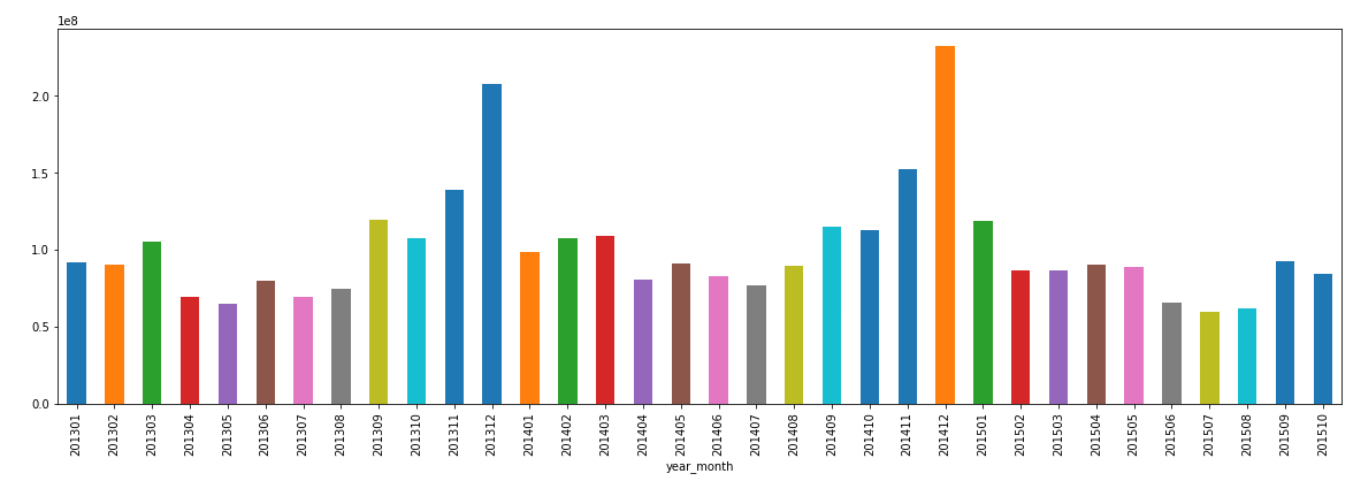
**Daily Sales for unique item by Year ($)**

****

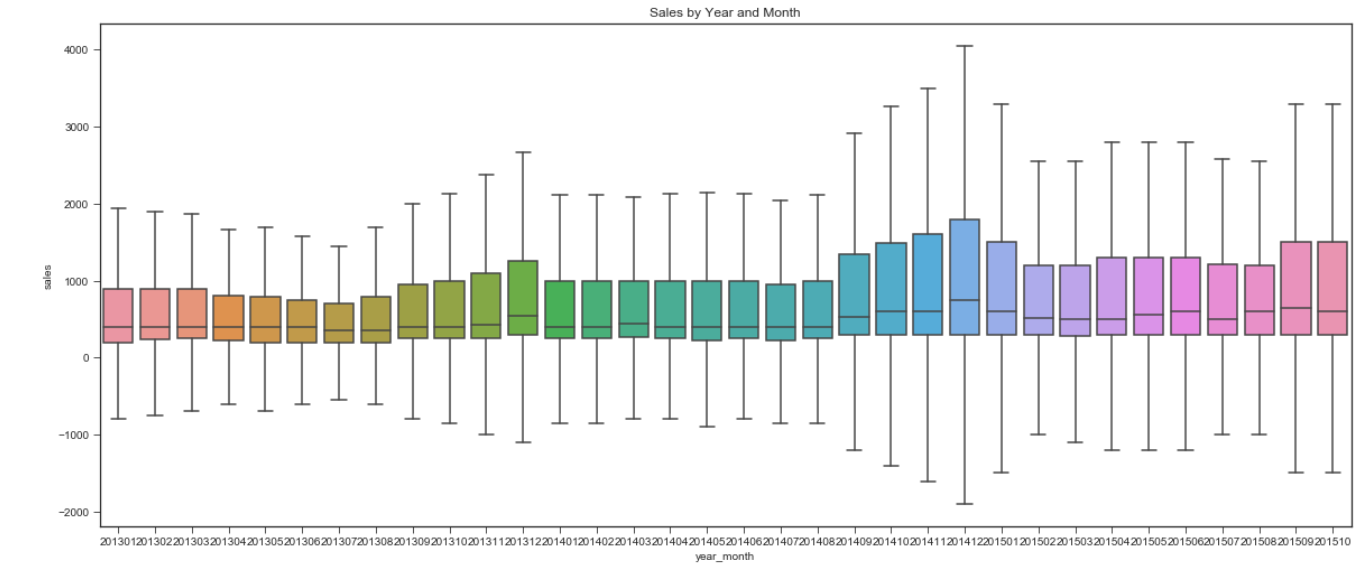
**3.1.2 Sales by Year and Month**

* Both the Bar Chart and the Boxplot show that the total sales or daily sales for unique item in the month of November and December are higher compared to other months.

**Total Sales by Year and Month**

****

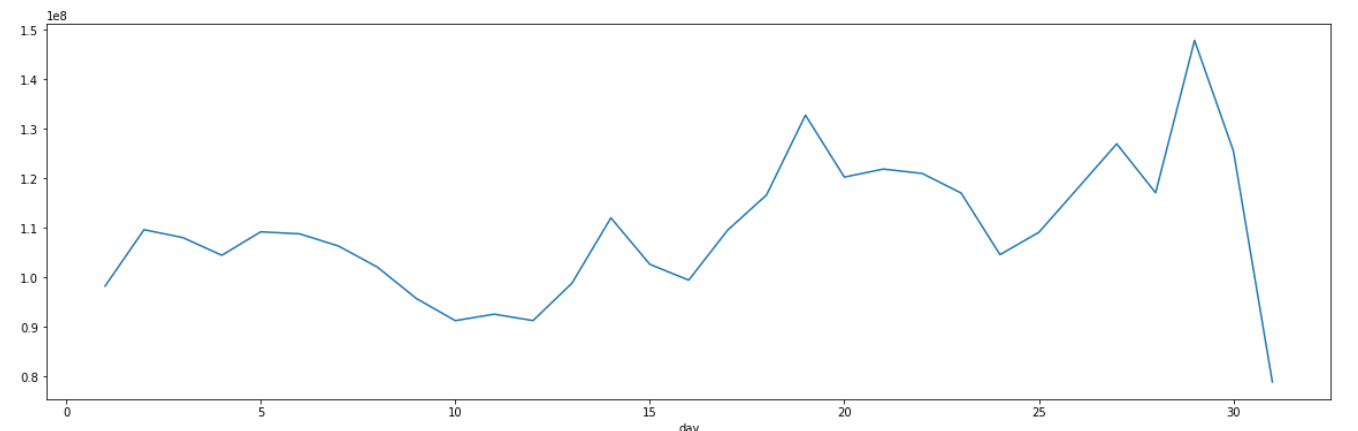
**Daily Sales for unique item by Year and Month ($)**

****

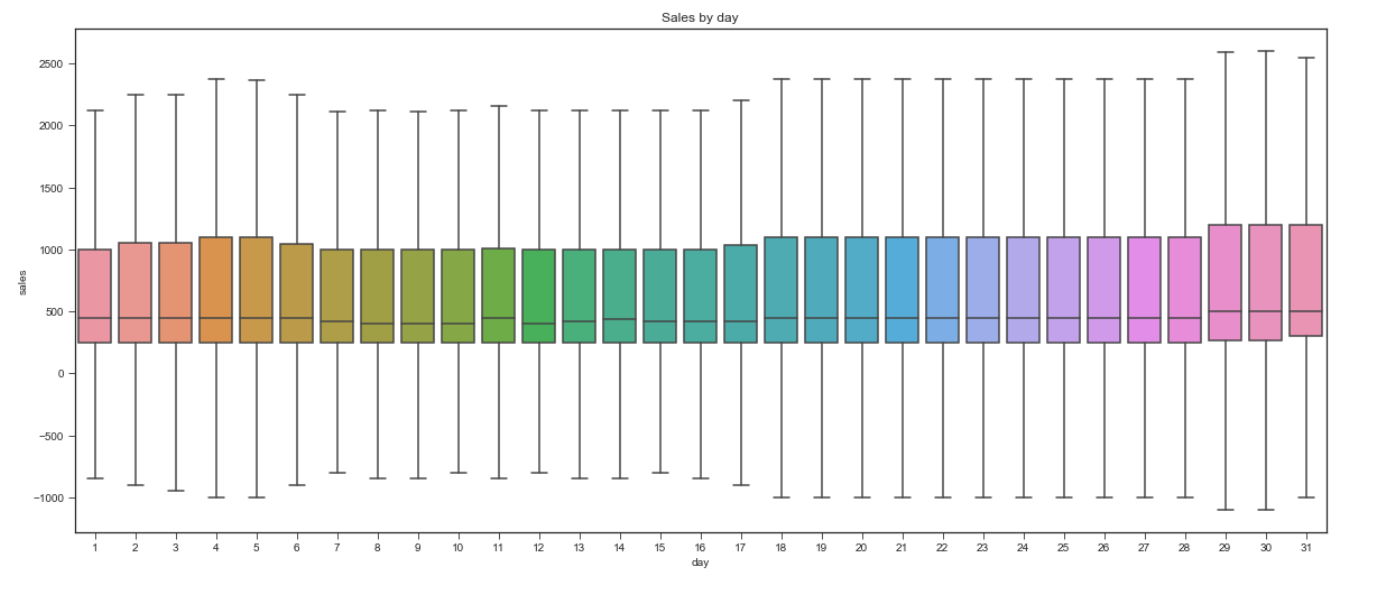
**3.1.3 Sales by Day**

* Bar Chart shows that the day 19 and 30 have outstanding peaks on total sales.
* Boxplot shows end of the month (day 29, 30, 31) has higher daily sales for unique item compared to other days within the month.

**Total Sales by Day ($)**

****

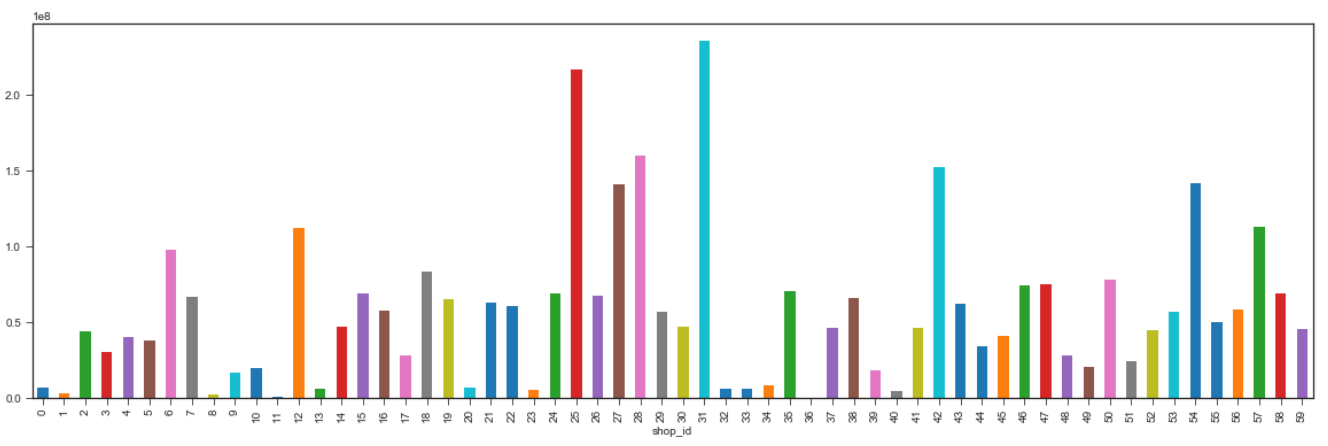
**Daily Sales for unique item by Day ($)**

****

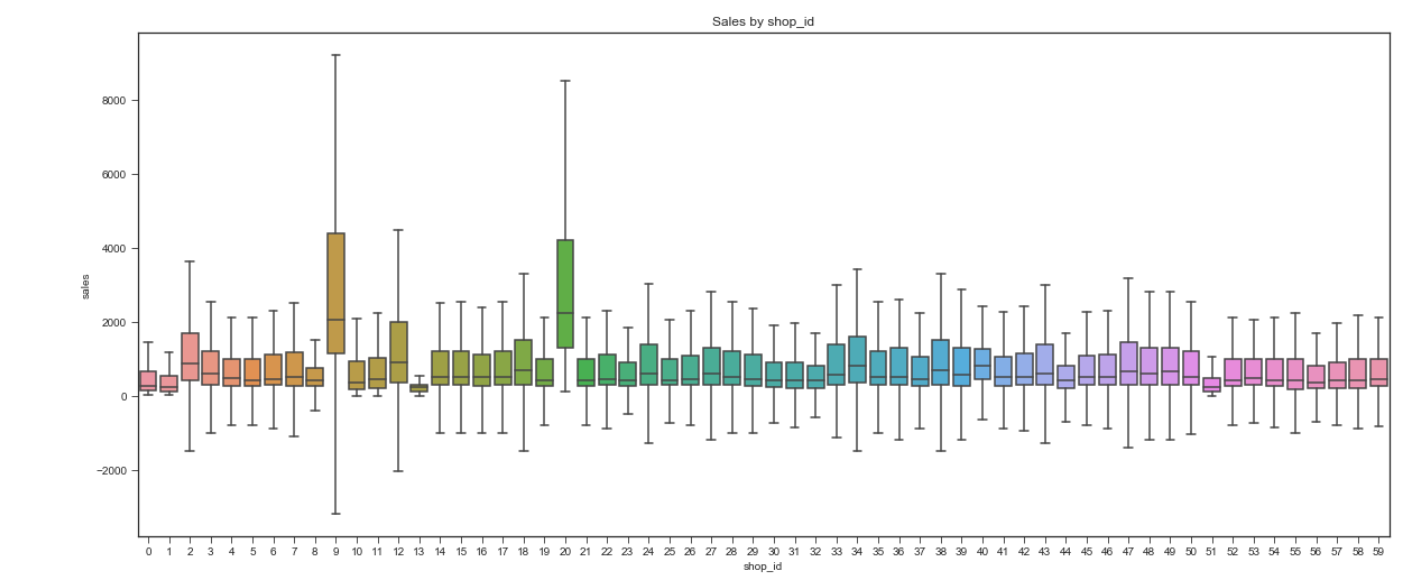
## Sales Distribution by Shop

* The Bar Chart shows that Shop ID 31, 25, 42 has the top total sales.
* The Boxplot suggests that Shop ID 9 and 20 has the highest daily sales for unique single item.

**Total Sales by Shop ID Day ($)**

****

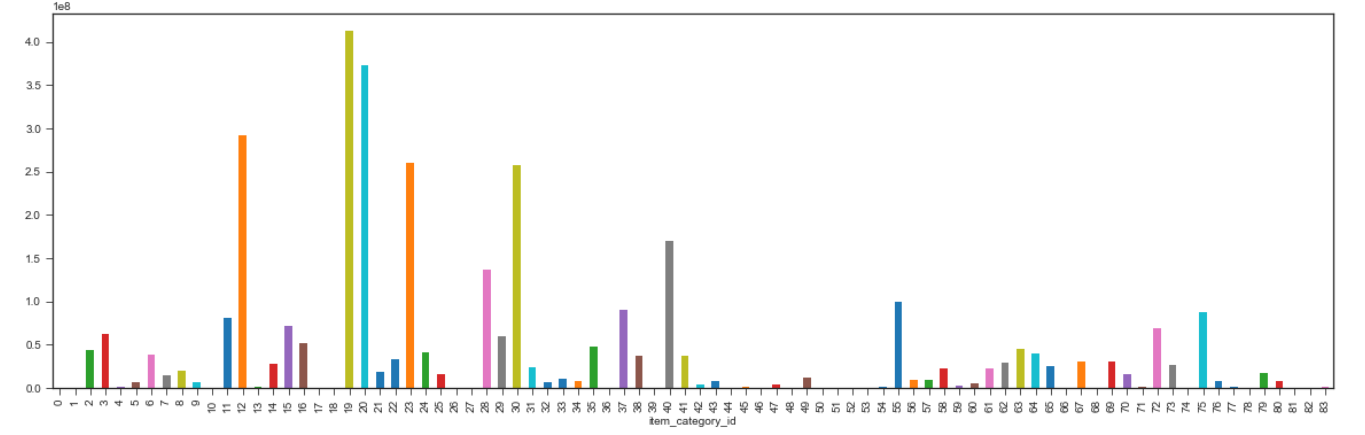
**Daily Sales for unique item by Shop ID ($)**

****

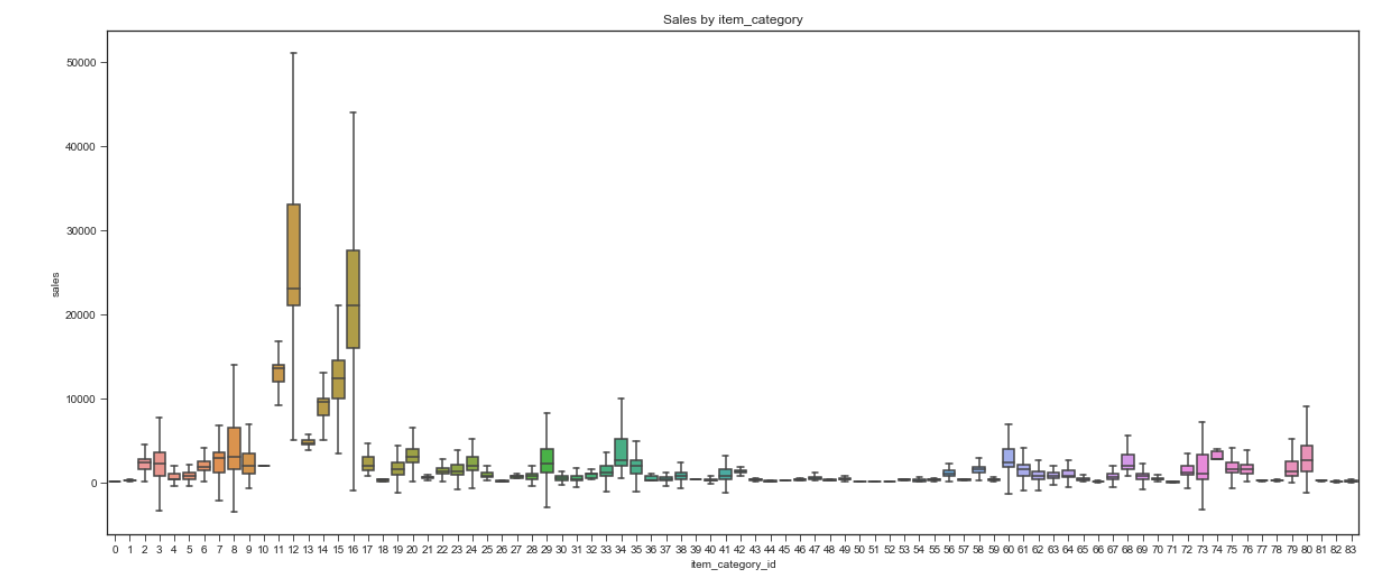
## Sales Distribution by Item Category

* The Bar Chart shows that Item\_catogory\_id 19, 20 has outstanding high total sales.
* The Boxplot shows that Item\_catogory\_id 12 and 16 has the highest daily sales for single item.

**Total Sales by Item Category ID ($)**

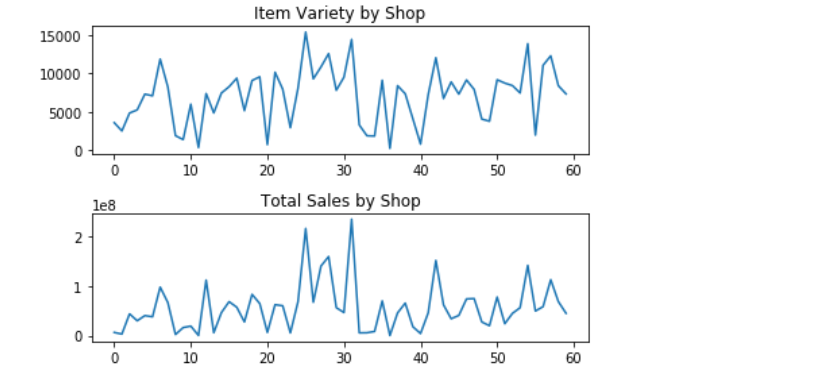


**Daily Sales for unique item by Item Category ID ($)**

****

## Shops and Item Variety

* Item Variety is calculated as the count if the unique item ID for Each Shop ID
* The Boxplot shows that he more variety of items the Shop has, the more the total sales.

****

* However, the following trend do not show obvious trend for the Shop that has more unique Item Category ID.
* In some shops, the more unique category ID, the less the sales.



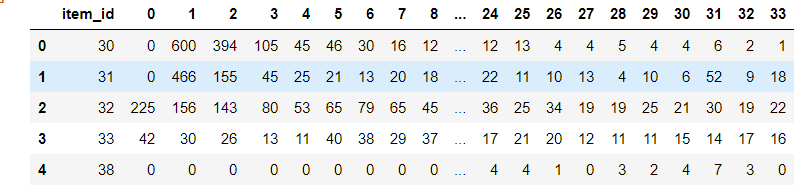
# Prediction model implementation

## Considerations over train and test set

The data set present conditions that will determine a better adjustment of the predictions, it was considered exclude those items that do not appear in the shop during the time of study on the train set but appear in test set because it do not have history these items will be excluded. In total are 102,796 items that do not have history in each respective store during Jan 2013 - Oct 2015.

Another important review is the variable 'item\_cnt\_day' is defined like the number of products sold by each shop during a day, there are 2,941 registers after evaluating the first consideration that present 'item\_cnt\_day' <0 these cases were excluded from the study.

Analogously, there are items that haven't been sold in a specific shop in the last 7 months, this can be reflected in the pivot table considering the variable 'date\_block\_num ' that indicates a unique number for month and year vs. 'item\_id' (See reference Table 4.1)



*Table 4.1 Pivot table to review last items sold*

After the analysis, were discarded 168 items outdated with no sells in the last 7 months.

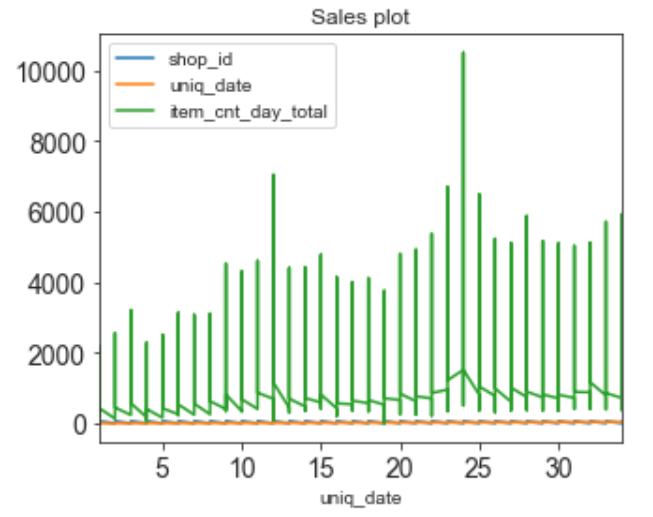
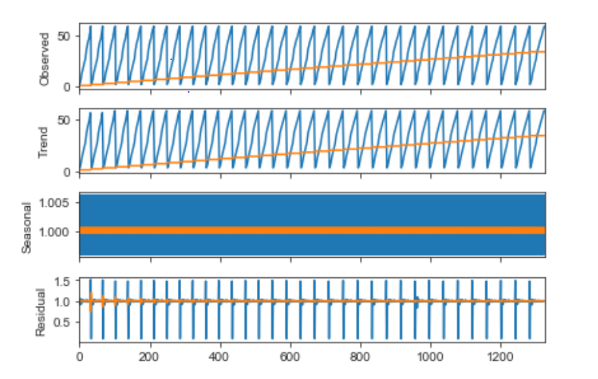
## Modelling

Since the point of view of the objective to predict the total items sold by shop for Nov 2015, the focus of the model prediction is considering at the level of shops instead of item\_id that could be an enhance of our approach. Both train and test set were grouped by month, shop. After seeing the summaries by month, it was detected some shops that does not have enough monthly history, it were considered only those shops with at least 4 months of history. The approach was considered in two ways considered a regression tree and an ARIMA adjustment by shop level defined by

The RMSE for the regression tree was 38.68 and for the ARIMA models was 37.25

# Execution of Model on Test dataset

The data were first observed through ‘seasonal\_decompose()’ , plot and tested through duckey-fuller where a small stationarity was discovered. It was a trend related to the variable “item\_count\_total” In order to address the issue, a differentiating technic (log\_diff(), log()) were used. The technics were given a better p-value but very underfitted model. The only technic that seems to work was the detrending through a linear regressor with gave an RMSE of 40.23. Finally, the data were fitted in an auto-ARIMA and Regression Tree. The ARIMA model showed the best accuracy with RMSE of 37.25



# Conclusion

The forecasting of the quantity of products sold by shop for Nov 2015 were predicted better with an ARIMA implementation with a RMSE of 37.25 instead of a regression tree with a RMSE of 38.68 since as the tree not detect patterns in trend if the serie is non stationary. Understanding in deep the approach of this model, it is important to remark that at the level of item\_id can be obtained better performances since a detailed analysis can be done including hierarchal models using bayes analysis could be another technique to forecast at the level of shop and item id, for time considerations it was considered to focus at the shop level, however, more extended techniques can be explored.

The data set even in training and test set were treated before applying the modelling phase related to some error factors in the original data set key to achieve better results like observe the shops with the items sold through the time and detect which items appear on the test set but do not have history, one approach that could be implemented is a cluster analysis at the level of shops to detect the more closer ones in groups and determine predictions according to each similar groups, including also the analysis of the categories and shop names initially are in russian language and can be used the library googletrans and clean the cases, in the project it was initially used the translations for shop names ought to the volume of products to categorize and cleaning of texts we decided not to use it for the project and work with the ID of products and shops.

# Associated files

The files associated to this report are,

|  |  |
| --- | --- |
| **File name** | **Description** |
| Group5-Assign\_PredictFutureSales\_Final.ipynb | All the code is in this Jupyter notebook |
| ‘sales\_train\_v2.csv | the training set. Daily historical data from January 2013 to October 2015 |
| Shops.csv | Supplemental information about the shops |
| item\_categories.csv | Supplemental information about the items categories |
| Items.csv | Supplemental information about the items/products |
| test.csv.csv’ | Need to forecast the sales for these shops and products for November 2015 |

# References

*Available at:* <https://www.kaggle.com/c/competitive-data-science-predict-future-sales/>