**Predict Future Sales - Data Preparation and Exploratory Data Analysis**

This is an EDA  project using python  based on Kaggle competition: [Predict Future Sales](https://www.kaggle.com/c/competitive-data-science-predict-future-sales)

In this competition, time-series dataset consisting of daily sales data, is provided by one of the largest Russian software firms - 1C Company. We are provided with daily sales data for each store-item combination, but our task is to predict total sales for every product and store in the next upcoming month.

**1. Summary**

After merging data, we had 60 shops with 22170 unique items to work with. There is a clear weekly cycle, with more sales on Friday, Saturday and Sunday with Saturday being the favorite shopping day. When I grouped sales by month, we can clearly see that: December and January had the highest item\_cnt\_day. December also had the highest item\_sale (most likely as a result of christmas) November had the lowest item\_cnt\_day but July had the lowest item\_sale.

Also, there are a few peaks over the years, they are; End of November 2013, lots of revenues and for item\_cnt\_day; End of December 2013 had relatively lower item\_sale compared to Nov 2013 and End of Dec 2014 where there was a high number of item\_cnt\_day. There are also peaks around the end of May 2014 and 2015. We also see declines in item\_cnt\_day and item\_sale from 2013 to 2015

**FILES DESCRIPTION**

|  |  |
| --- | --- |
| sales\_train.csv | the training set. Daily historical data from January 2013 to October 2015. |
| test.csv | the test set. You need to forecast the sales for these shops and products for November 2015. |
| Sample\_submission.csv | a sample submission file in the correct format |
| items.csv | supplemental information about the items/products. |
| items\_categories.csv | supplemental information about the items categories. |
| shops.csv | supplemental information about the shops. |

**DATA DESCRIPTION**

|  |  |
| --- | --- |
| ID | an Id that represents a (Shop, Item) tuple within the test set |
| shop\_id | unique identifier of a shop |
| item\_id | unique identifier of a product |
| item\_category\_id | unique identifier of  item category |
| item\_cnt\_day | number of products sold. You are predicting a monthly amount of this measure |
| item\_price | current price of an item |
| date\_block\_num | a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33 |
| item\_name | name of item |
| shop\_name | name of shop |
| item\_category\_name | name of item category |
| date | date in format dd/mm/yyyy |

**3. Methodology**

We grouped sales by day, weekday, month so we can observe the revenue flow for each store, this led to an important discovery that some store weren’t bringing in any revenue which maybe because they were not in operation. Only functioning stores were analysed since zero revenue generating stores are undergoing investigation. Also, we excluded records where one or more fields were missing.

**Data Preparation and Exploratory Analysis**

Following steps were performed for data preparation and analysis to get the best predictions.

**STATISTICALLY EXPLORING DATA**

Checked:

* Data dimensions, rows and columns
* Columns names
* Data Type and data Info
* Unique values per column (to decide whether to make column Categorical?)

**CLEANING DATA**

* Looked for any missing Data
* Identified and converted categorical columns/values to Numerical representation using Dummy Variables if suitable for modelling
* Identified and converted numerical columns/values to categorical representation
* Checked for distinct values in Categorical columns

**STATISTICAL OVERVIEW OF DATA**

* Checked head, tail of data to see complete required data loaded
* Described data, columns
* Identified numerical columns and look for insights like mean, median, mode, etc.
* Understood the relationship of columns and how they are affecting each other.
* Applied feature engineering by creating new columns which may help in better understanding of data and predictions (grouped the date into weekday, month and year)

**RESULTS**

By item, the most popular were:

* By item\_cnt\_day was: **item\_id 20949,** with 187642 units sold, generated $929k
* By item\_sale was: **item\_id 6675**, worth $219M in revenue, with 10289 units sold
* The worst are: item 1590, 11871, 18062, 13474, 13477. Shops lost money on them.

**By shop:** **shop\_id 31** had the highestitem\_count day(310777 items sold), and also highest item sale($235M). **Shop\_id 36** had the lowest item\_cnt\_day(330), with a revenue of ($377k)

**Item\_category:** The most popular by item\_cnt\_day is **item\_category 40**, with 634171 units sold, generating $170M. While by item\_sale is **item\_category 19**, with 254887 units sold, generating $412M . The worst are **item\_category 51**, with only one unit sold ($129), and **item\_category 50**, with 3 units sold ($24).

**CONCLUSION**

We have a number of questions regarding the dataset

* Most of the shops have a similar selling rate, but 3 of them have a much higher rate, is this indicative of the shop size or location?
* Why are the products that aren’t bringing in revenue still stocked?
* Some shops are not in operation as they sell very few to zero items so why are they still in operation?
* Also noticed that the shops that have low revenue are not selling the items from categories that have high selling rate, is there a reason for this?