CMPT 741: Data Mining (Fall 2019) – Feature Engineering

Date: 8th October 2019

Predict Future Sales:
Final project for "How to win a data science competition"

Team DataScavengers: Anurag Bejju – abejju@sfu.ca, Savitaa Venkateswaran - savitaav@sfu.ca

---

Problem Statement:

The motive behind this challenge is to *predict future sales* (sales for the upcoming month) for every shop and item pair, given 3 years' historical sales data ranging between years, 2013 to 2015 (including). This historical data has been generously provided by a Russian software firm named, *1-C company*. However, the dataset has sales figures depicting per shop per item per day in contrast to the ask of the challenge (to predict monthly sales). The highlight about the dataset presented is that the list of shops and items vary every month (i.e.) the shops and items present in the dataset aren't bound to have sales every single month. Adding on, the item name, item categories, etc., are in Russian- making it harder to interpret in the first glance. All these cumulatively add up to the complexity of the challenge.

Exploratory Data Analysis:

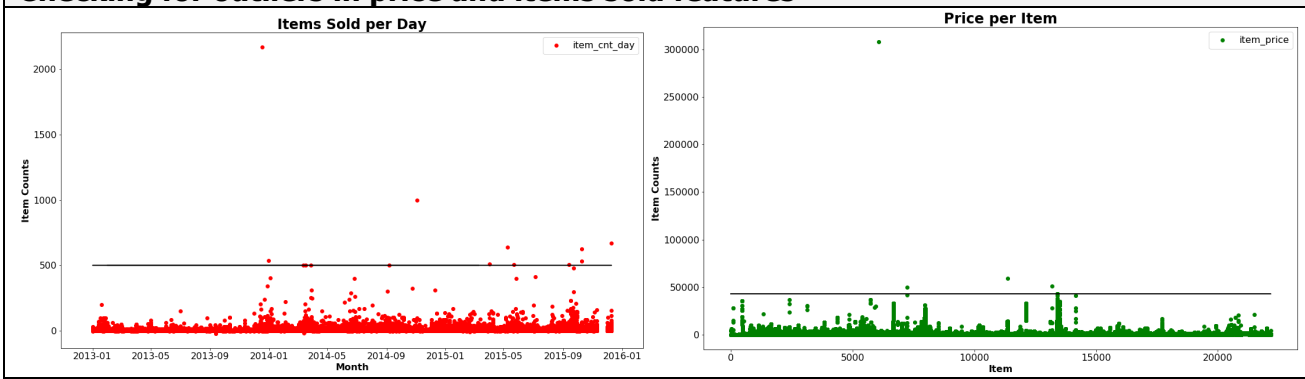
Using the presented data at its earliest stages, we have performed basic data analysis, including plotting sum and mean of item_cnt_day for each month to find some patterns, exploring missing values, inspecting test set etc. Here are some ETL tasks as well as exploratory data analysis that was performed to engineer meaning features for our prediction model.

Understanding the Dataset

```
Total Shops in the given training dataset:  60
Total Items in the given training dataset:  21807
Total Categories in the given training dataset:  84
Timeline
Start date for the training dataset:  2013-01-01 00:00:00
End date for the training dataset:  2015-12-10 00:00:00
Item Price
Min item Price in the training dataset:  -1.0
Max item Price in the training dataset:  307980.0
Item Sold
Min item sold in the training dataset:  -22.0
Max item sold in the training dataset:  2169.0
```

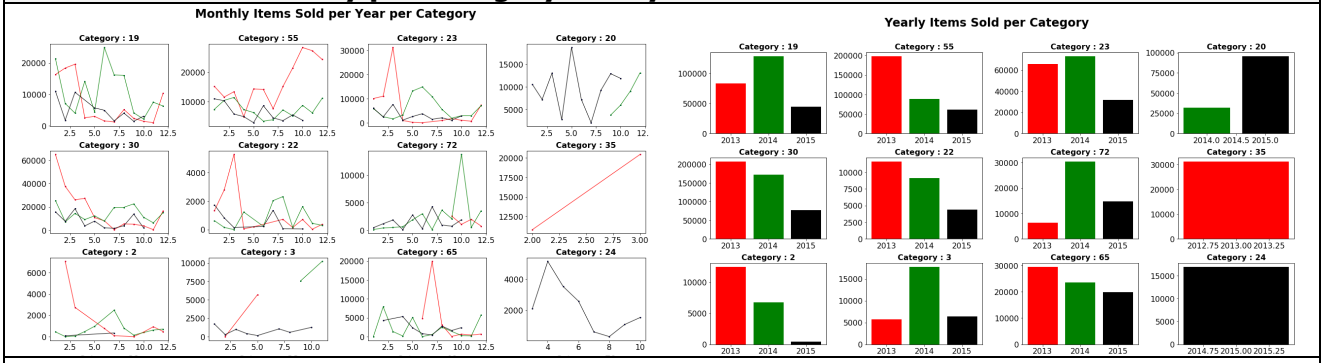
There are a total of 60 shops, 21807 items classified into 84 categories. This data is for three years starting form 1st January 2013 to 10th September 2015. Talking about min-max, the cheapest item is -1 and the expensive item was for 307980. We have also observed that there were a min of -22 items sold (could be returns) and a maximum of 2169 on a day.

Checking for outliers in price and items sold features



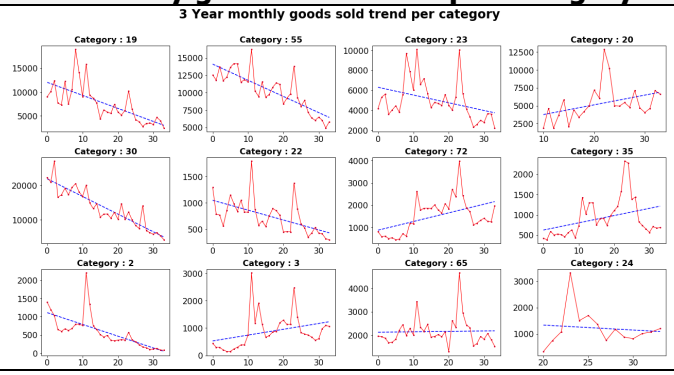
From the above plots, we have decided to remove outliers from df. Our selected range for item price is between 0 and 43000 and for item count per day is between 0 and 500.

Items are sold monthly per Category each year



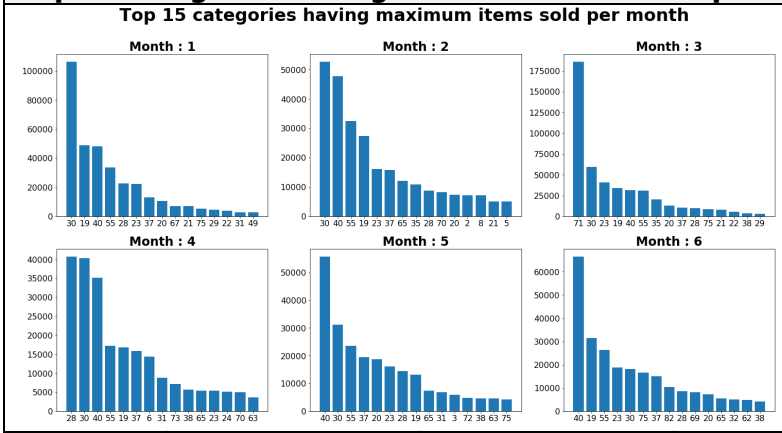
From the first plot, we couldn't really figure out how volume of goods sold is behaving with respect to each category. There is no clear trend between the three year. So from the second plot, we got to see the bigger picture year wise. There is trend of decline or increase per category yearly

3 Year monthly goods sold trend per category



It was really interesting to observe the trend line for the last three years in each category. We have assessed and found some steep decline with every passing year for some categories and for few we can see an increase in goods sold.

Top 15 categories having maximum items sold per month



It was intriguing to see such large volume of items are being sold in categories like 40, 30 and 55 in every month. These could be day to day products which are needed throughout the year. It would also be good to see if we could group such products together.

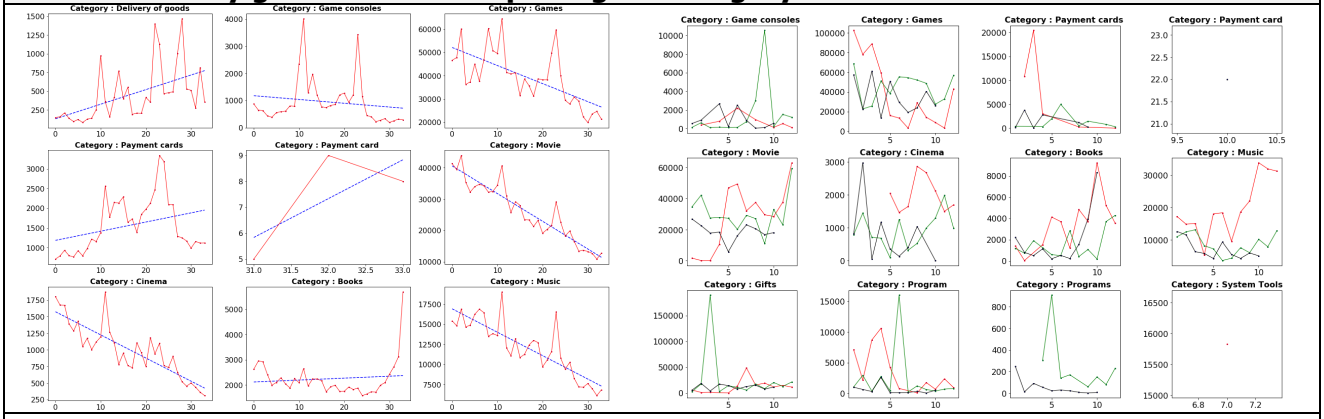
Category Analysis

The 19 categories are

Category 1	PC	Category 10	Cinema
Category 2	Accessories	Category 11	Books
Category 3	Tickets (digits)	Category 12	Music
Category 4	Delivery of goods	Category 13	Gifts
Category 5	Game consoles	Category 14	Program
Category 6	Games	Category 15	Programs
Category 7	Payment cards	Category 16	System Tools
Category 8	Payment card	Category 17	Utilities
Category 9	Movie	Category 18	Net carriers
		Category 19	batteries

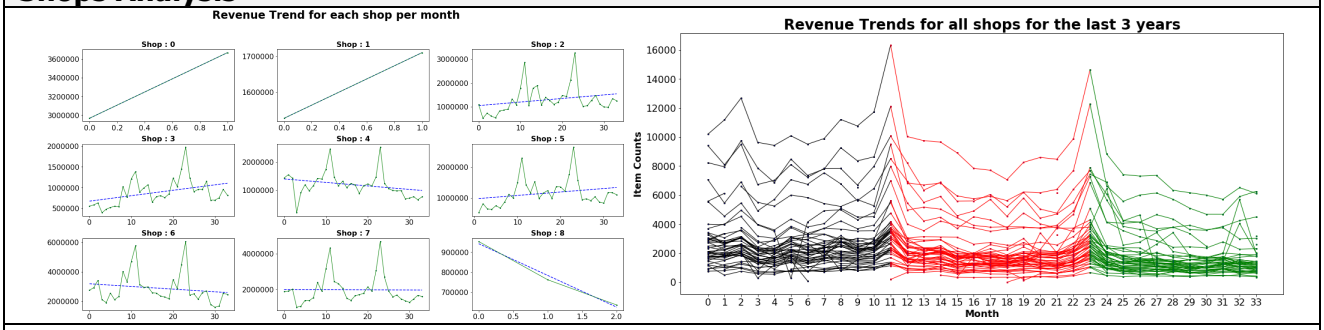
Since the categories are in Russian, we have used google's googletrans package to convert it to English. After that we were able to derive 19 higher categories from the present 84 sub categories.

3 Year monthly goods sold trend per higher category



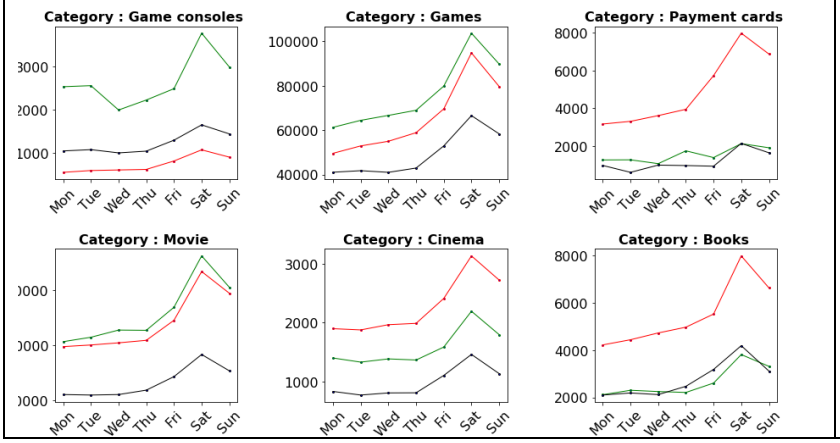
These trend lines are more clear with the higher categories. We have observed some recurring peaks in the time line. It would be interesting to figure out the reason for these peaks

Shops Analysis



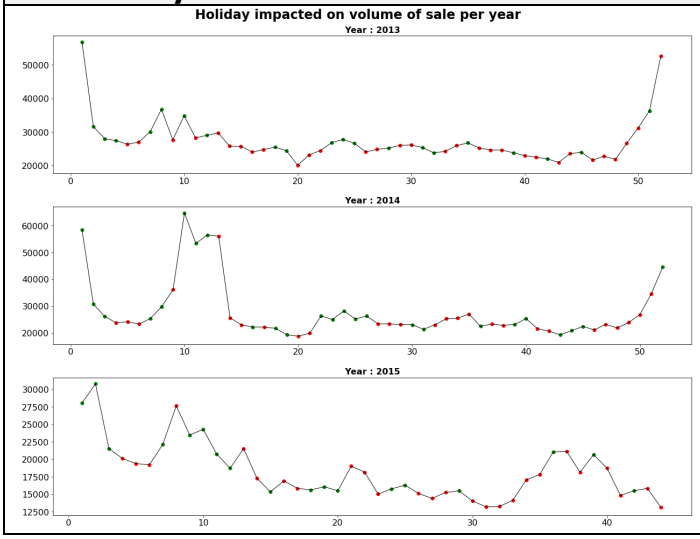
These trend lines are more clear with the higher categories. We have observed the revenue trend for each shop to assess their revenues as well as check which shops are actively selling goods. We have also stacked revenue trends of all shops for the last three years and found two recurring peaks.

Weekend vs Weekday



One really interesting feature we found is weekend vs weekday. We were able to establish some really meaningful information with this EDA. Almost all categories showed a significant spike during weekends as compared to weekdays.

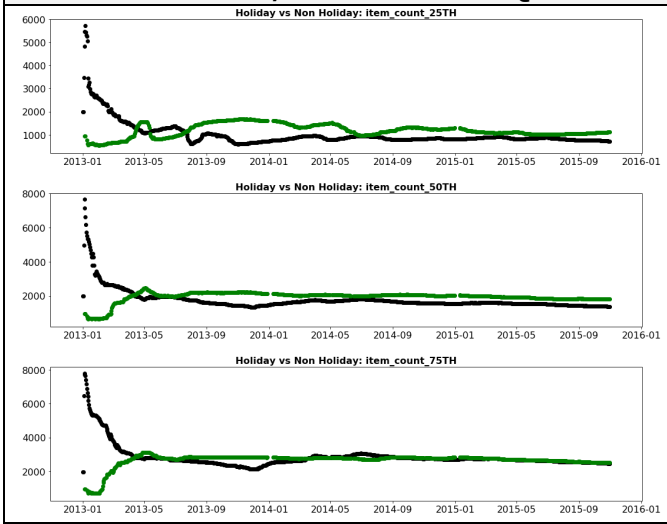
Seasonality



First we have web scrapped all the Russian holidays for these three years. Then we have used an assumption that people would be buying products throughout the week till actual holiday. Once we had all the holidays extracted we have created new features that says if week has a holiday or not.

The interesting thing we have found was that the volume of sales during the holidays spiked and during no holidays it declined. It was also accurate because some of these categories like music, gifts, games do have seasonality associated with it.

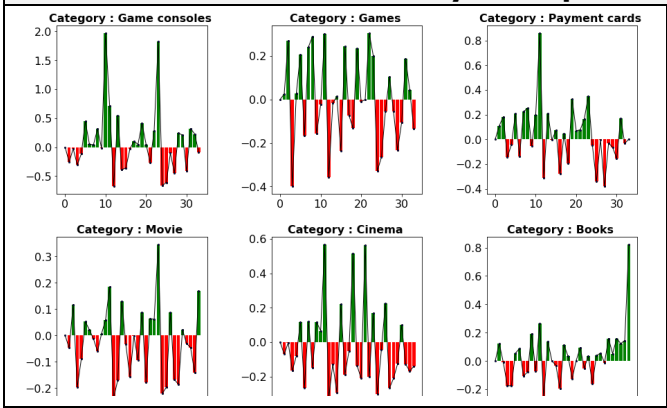
Observe the 25th, 50th and 75th Quantile for volume of items w.r.t holidays



This method lets us assess item count with respect to the quantity of items sold w.r.t to holidays. We wanted to check if specialty items are sold more during holidays and general category items are sold in a similar irrespective of holiday or not.
We were able to establish a clear distinction. *item_count_25th*: As you can see the items that have less volume (i.e specialized items) are sold more during holidays.

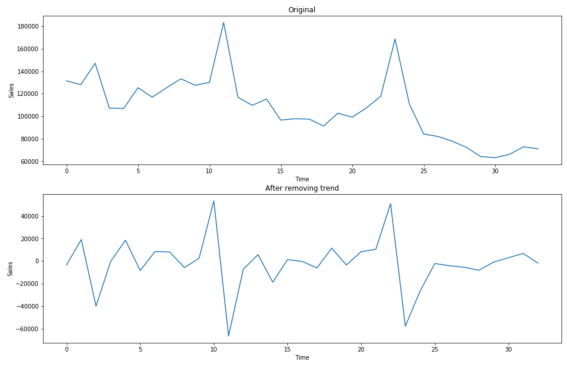
item_count_75th: Similarly, that have high volume (i.e more general items) are sold pretty much same on all days.

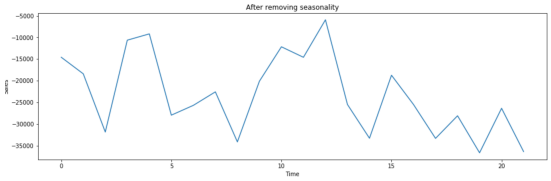
Differential revenue - week by week per category



With this feature, we could see how the past week's performance impacts next week in terms of revenue per category. This was really effective as you can proper periods of up fall and downfall in each category.

Stationarity Analysis

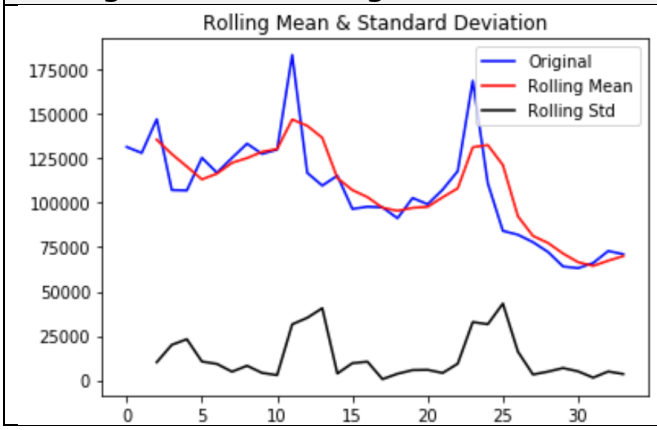





```
Results of Dickey-Fuller Test:
Test Statistic                  -3.270101
p-value                          0.016269
#Lags Used                       0.000000
Number of Observations Used     21.000000
Critical Value (1%)             -3.788386
Critical Value (5%)             -3.013098
Critical Value (10%)            -2.646397
```

In this feature analysis, we have explored the stationarity of the time-series dataset for its entirety. This helps us to understand the dips and ups better and also helped us to model simple moving average models like ARIMA, SRIMA, etc., on the training dataset to understand the impact of moving average(s) as a feature for predicting the response variable (monthly sales). The statistics shown are the test performed to check the stationarity of the waves and the p-value < 0.05 clearly concludes it.

Rolling Mean and Rolling SD



The rolling mean and rolling standard deviation curves shown in the figure to the left when compared with the original curve, clearly shows there is a strong positive correlation between them. Especially, with the rolling mean- we can see that the overall trend is showing a downward movement which is in close call to what the original (3 years historical data) time-series data shows. Hence, these are good statistical features to be used.

Feature Engineering:

Using the above exploratory data analysis, we have come up with the following features.

Feature Name	Feature Type	Description
mean_monthly_trend_per_category	numeric	This gives us the trend of sale per category
Higher_category	categorical	Use the new 19 categories
Weekend_weekday	boolean	Check if it's a weekend or weekday
Holiday_week	boolean	Check if it's a holiday week or not
city	categorical	Use the city from shop name
25th_quantile_volume_of_items (50th and 75th)	numeric	This gives us the trend based on the volume of items sold.
differential_revenue	numeric	Calculate differential_revenue based on the previous week.
item_id, shop_id, date_block_num	numeric	Basic given features
Total_price	numeric	Item_price * item_cnt_day
Stationarity_analysis	numeric	stationarity of the time-series dataset
Rolling_mean , Rolling_SD	numeric	Doing rolling mean and rolling standard deviation wrt time.