Introduction

As part of our final mini project, we have used the data set from Kaggle events, "Predict Future Sales", provided by one of the largest Russian software firm - 1C Company [1]. The data set consists of historical daily sales data for different stores and items from the year 2013 to 2015. For 2015, sales data is not available for the months of November and December. Our goal for this project is to use statistical forecasting methods and predict both **short term forecast - the sales volume of last two months in 2015** and **long term forecast - the sales for next two years** based on the data set.

Data

The data set is distributed in different files but we have mainly focused on **sales_train.csv** for our analysis. The data set contains 6 variables:

- date: Date is a factor variable. It is in format dd/mm/yyyy and takes values from Jan-2013 to Oct-2015.
- date_block_num: A consecutive month number counting from January 2013 with 0. It is formatted as integer variable.
- shop_id: Unique identifier of a shop. It is formatted as integer variable.
- item_id: Unique identifier of a product. It is formatted as integer variable.
- item_price: Current price of an item. It is formatted as integer variable.
- item_cnt_day: Number of products sold. It is formatted as integer variable.

As date is a factor variable we have converted it to date type and also divided into year, month and day. These new variables are stored in new columns for future analysis.

The summary statistics of the data can be observed in the Appendix. There are a total of **2,935,849 observations** in this data set.

We have checked for missing values and the data set has no missing values. We have also observed that the data is tidy. Hence, we can proceed with this data set.

Initial Analysis

We will try to answer below questions in the preliminary analysis:

- Are there consistent patterns?
- Is there a significant trend?
- Is seasonality important?

We first take a look at the total sales from Jan-2013 to Oct-2015 and total sales by month. From Figure 1a, we can observe that the total sales are decreasing over the year. The total sales in 2015 decreases by 50% compared with sales in 2013. From figure 1b, it is hard to conclude about the monthly pattern of total sales. Thus, we next try to see if there is indeed any monthly pattern by grouping the data both by year and month wise.

Figure 1a: Yearly Sales Data

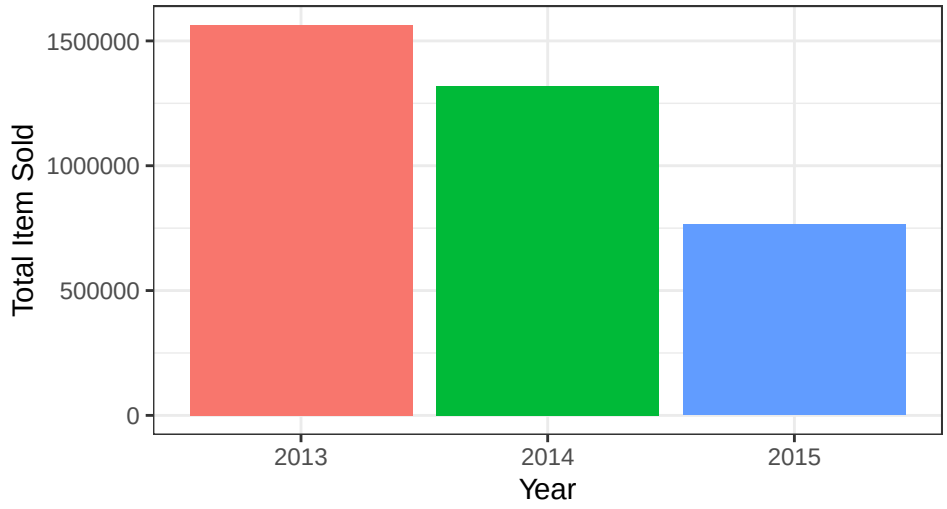
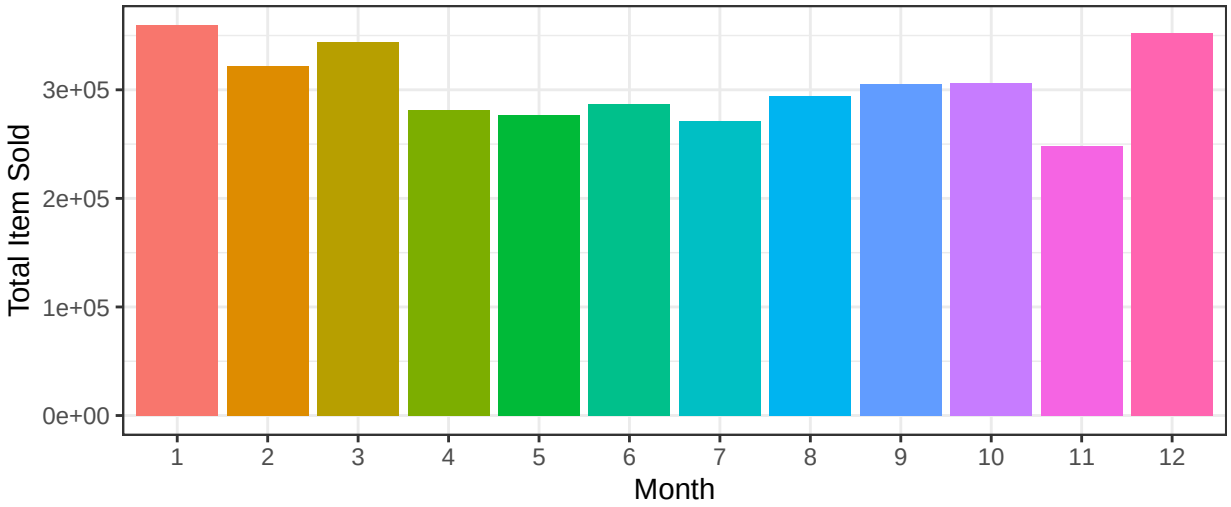


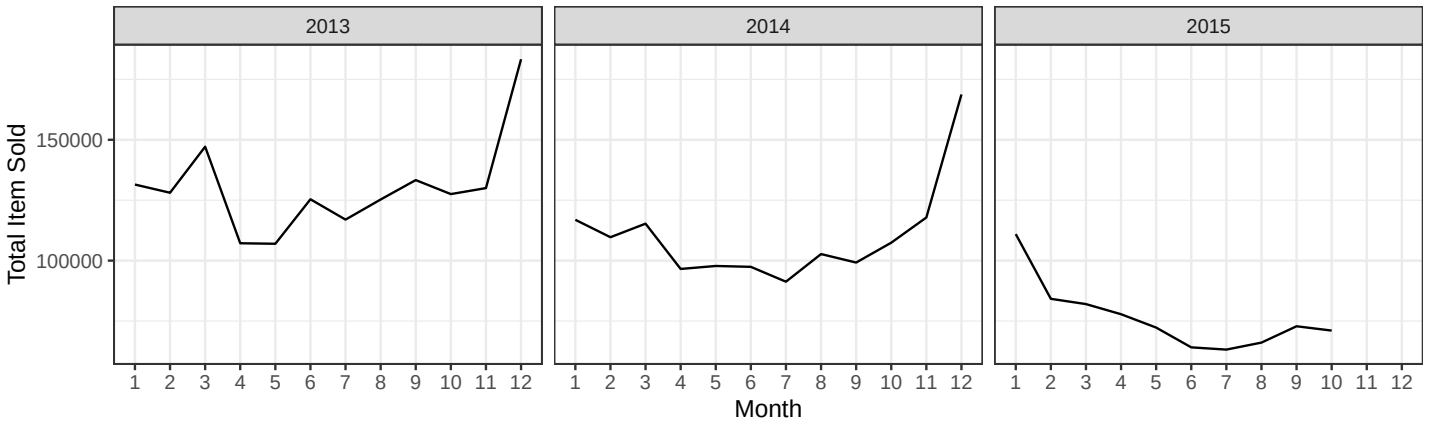
Figure 1b: Monthly Sales Data



From Figure 2, we have evidence to conclude that **there is a seasonal pattern** in sales volume. In both 2013 and 2014 there is a spike of sales in the month of December and there is a dramatic drop in the sales in the month of April (see Appendix). We also observe that in the month of January there is always a drop in sales of items and in December we have spike in no of items sold. This may be because post Christmas the sales generally tend to drop.

We also conclude that **there is a decreasing trend** in sales volume as the total sales volume decreases over the years. These plots are helpful for us to identify the patterns in the data. Next, we want to check if these patterns are significant.

Figure 2: Total sales by month and year



Planning

We are trying to forecast the future - getting an idea of total sales amount for the last 2 months of 2015 as well as next two years by identifying the trends and seasonal variances. For this purpose, it will be ideal to use time series analysis as it has the advantage of understanding the past as well as predicting the future.

The first step is to convert data to a time series object. We have grouped the total sales value by ordering by each year and grouping by each month. The summary of time series object data is shown in the Appendix. Next, we want to check if the trends and seasonality in this data set are significant.

The scatter plot in Appendix shows lagged values of the time series. The colours indicate the month of the sales variable. The relationship is strongly positive at lags 1, 2, 3, 4, 5, 6, 7, 8, and 9. This reflects **a strong seasonality in the data**.

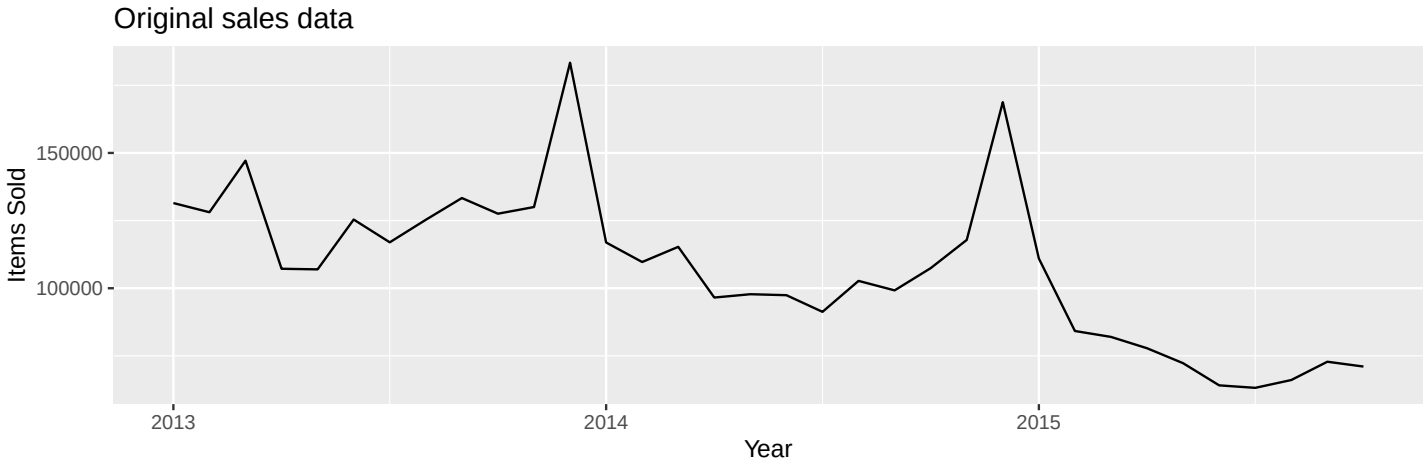
As our goal for this report is to use statistical forecasting methods to predict sales in future, we want our data set to be stationary based on the assumption of forecasting models. This is because a non stationary time series will affect the value of the time series at different times [4].

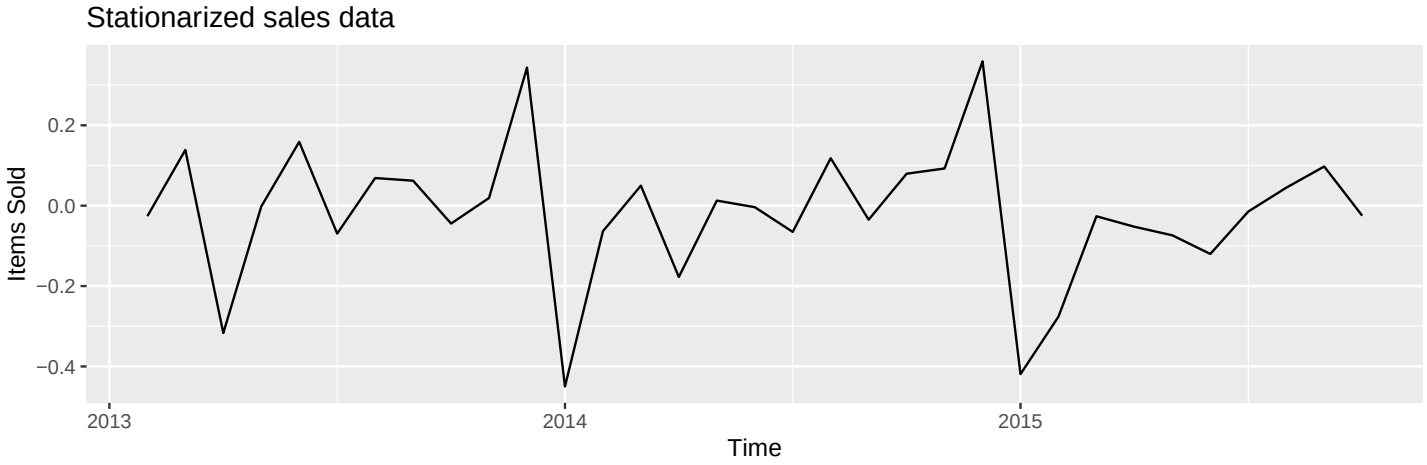
We will use **Dicky-Fuller test** to determine if the data set is stationary. Below is our hypothesis for this test:

- *Null Hypothesis*: It has some time dependent structure.
- *Alternate Hypothesis*: It does not have time-dependent structure [2].

The test output shows that p-values(0.9835)>0 (see Appendix). We reject the null hypothesis and conclude **there is a significant evidence that our data set is not stationary**.

As our data set is not stationary, we will convert it to stationary by applying log function to remove unequal variances. Then, we use differentiation function to get constant mean [3]. Below is the comparison between plot of the original data and stationarized data. We observe that our seasonal properties have been simplified.





Modelling and fitting analysis

After simplifying the seasonal pattern in our data set by removing the variations, our next step is to use different forecasting methods to predict the future sales using our stationarized sales data from the planning part.

We first create a benchmark for our upcoming forecast models and for that we choose **seasonal naive method**. The output of this model is shown in Appendix.

As we have our base model now, we can check the accuracy of this model which will give us an idea of how fit our model is. The function *accuracy* gives the following measures of accuracy of the model fit: mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE) and the first-order auto-correlation coefficient (ACF1). Out of these values, we are using **MAPE** here to validate if our model is a good fit or not.

The residual standard deviation of naive model is **0.1022** with MAPE is **346.9721%**. To check if auto-correlation exists in this model, we use *Ljung-Box test* with below hypothesis.

- *Null Hypothesis*: The data are independently distributed or the model does not show lack of fit.
- *Alternate Hypothesis*: The data are not independently distributed or the model does show a lack of fit.

With $\mathbf{p = 0.2635}$, there is no evidence against the null hypothesis that the residuals are independently distributed.

As MAPE value is too high for this base model, our next question is: **Can we improve this base model?**

As **ARIMA model** is one of the most popular approaches in time series forecasting, we have tried to use this as our next model for forecasting.

We first find P & Q values required for ARIMA(P, d, Q) model. We did ACF and PACF plots to find the values. From the plots in the Appendix, we can observe that PACF cuts off at lag 0 ($\mathbf{P = 0}$) and ACF tails off at lag 0 ($\mathbf{Q = 0}$). As we have already differenced our data, we take $\mathbf{d = 0}$ as our model parameter. Our second model is **ARIMA(0,0,0)**. This is a special case of ARIMA models called white noise.

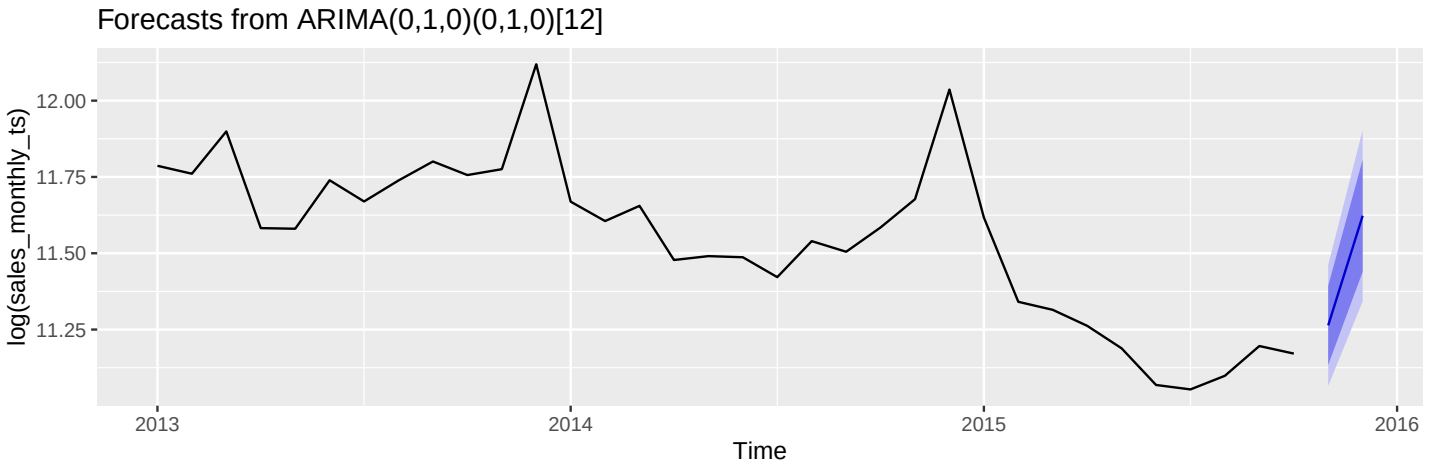
The output of ARIMA(0,0,0) model shows that the residual standard deviation has reduced to **0.1705286** which is slightly higher from our benchmark model. However, MAPE value is greatly improved - **128.1393%** compared with **346.9721%** from naive model (see Appendix). The AIC value for the ARIMA(0,0,0) model is **-19.1**. The Ljung-Box test output shows $\mathbf{p = 0.3015}$ hence we do not reject the null hypothesis and conclude that the residuals are independent.

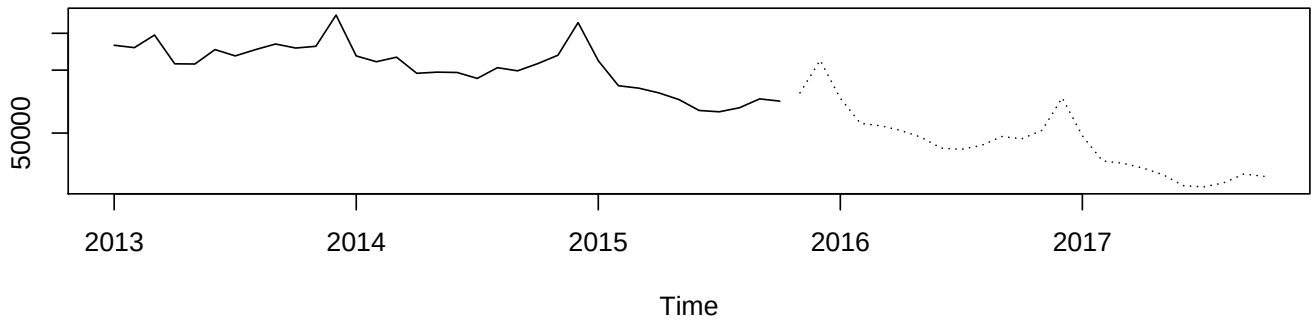
To check if we can further improve the residual standard deviation and AIC, we use the auto.ARIMA(). The algorithm uses a step-wise search to traverse the model p and q values. Hence, we expect it would provide a better fit for our time series data.

The automatic ARIMA model suggests us **ARIMA(0,1,0)** for this data set. This is a special case of ARIMA models called random walk as it has no constant. The results shows that residual standard deviation has reduced to **0.1011435** from the previous model (see Appendix). The AIC value has also reduced to **-34.81** which is better fit than AIC of ARIMA(0,0,0). Ljung-Box test output shows $\mathbf{p = 0.1058}$ hence we do not reject the null hypothesis. However MAPE shows **0.4859299%** which means we can say that the accuracy of this model is **99.5140701%**.

Now , we can test our auto ARIMA model on the given historical data to check if the model is a good fit. From the output seen in the Appendix, we can say that the predicted values are very close to the actual data. This proves that **ARIMA(0,1,0)** is a good fit for our data and can be used for further forecasting. Our model can be written as: $\hat{x}_t = x_{t-1}$

Forecasting





The prediction intervals for ARIMA models are based on assumptions that the residuals are uncorrelated and normally distributed [4]. For this reason, we plot the ACF and histogram of the residuals as below. We observe that the assumptions are satisfied. We can continue with our forecasting.

Now, we use auto ARIMA(0,1,0) to forecast the trend of the total sales for the last 2 months of 2015 and the years of 2016 and 2017. The forecast for last 2 months of 2015 can be seen in the Appendix. The forecasting trend shows that, similar to 2013 and 2014, there will be **increases** in the total sales in the months of November and December for the year 2015. We can also predict that the total sales trend will **continue to decrease** over next two years.

Conclusion

As part of this report, we begin with exploring the historical sales data for 1C company for the year 2013 to 2015. We observed that there is indeed a seasonal trend as well as overall decreasing trend in the total sales. Next, we have tried to predict the future sales trend for the months of November and December of 2015 as well as next two year's total sales using different forecasting methods. Our base model by using naive method has shown high MAPE value which leads us to try the popular ARIMA model. The final model ARIMA(0,1,0) depicts a better MAPE value as well as the residuals have been reduced and we selected that model to fit our sales data. Finally, our forecasting shows that similar to the historical data, we will observe the decreasing trend of total sales in upcoming two years and also notice a seasonal trend of a spike of sales during the month of December (before and around X-mas) followed by significant drop on January (post X-mas).

For future work, we may work on to predict sales of different shops and different items which can provide us an estimate of marketing strategy - items less popular can use some advertising or discount options and will boost up the sales of the company.

References

Shown in Appendix.