2015 Sales Trends In The Gaming Industry

1C Company growth trend in 2015

source: [Predict Future Sales Data Set]

Group 10

Achidi Kisob

Adila Islam

Natalia Restrepo

Qasim Malik

Table of content

Introduction

This report observes a time-series dataset consisting of daily sales data, from one of the largest Russian software firms - 1C Company. The company specializes in the development, distribution, publishing, and support of mass-market software. Using the dataset within a 3-year timeframe, we will explore and predict total sales of the company for the next six months.

Objective

To develop a predictive model that can forecast the number of items sold by 1C Company during November 2015 to April 2016.

The Predict Future Sales Data Set from Kaggle was utilized, which included extensive details from nearly every item sold by the company between January 2013 and October 2015. Time series forecasting is one of the critical building blocks of Machine Learning. We will use the Seasonal Autoregressive Integrated Moving Average (SARIMA) model to create a forecast and analyze the results of this model.

Data Preparation

- *Datasets*

Table 1. Relevant Variables for the Research

Dataset	Description
sales_train.cvs	The training set. Daily historical data from January 2013 to October 2015.
test.csv	Test set
items.csv	Supplemental information about the items.
shops.csv	Supplemental information about the item's categories.

- *Data Cleaning*

First, we did an exploratory analysis of the five .csv files described in Table 1. From this analysis, we observed no missing or invalid values on any of the datasets. Moreover, we observed that the files: 'items', 'item_category', and 'shops' had unique items listed in Russian that could also be identified in the dataset 'train_sales', thus, instead of merging all the data frames, we focussed our analysis and modeling only in the data frame 'train_sales', which contains 6 unique columns and 2' 935 849 observations.

From the results in *Figure 1,* we observed that the variable target ('item_cnt_day') is of type 'float' and contains negative values. We would have expected this variable to be of type 'int' and to only have positive values since it reflects the number of items sold every day. Moreover, we also observed that the variable 'items_price' also contained negative values, suggesting that there were errors in this column. We decided to remove these entries from the dataset as they could affect significantly the real amount of sales obtained by day and the accuracy of the model.

	date_block_num	shop_id	item_id	item_price	item_cnt_day
count	2.935849e+06	2.935849e+06	2.935849e+06	2.935849e+06	2.935849e+06
mean	1.456991e+01	3.300173e+01	1.019723e+04	8.908532e+02	1.242641e+00
std	9.422988e+00	1.622697e+01	6.324297e+03	1.729800e+03	2.618834e+00
min	0.000000e+00	0.000000e+00	0.000000e+00	-1.000000e+00	-2.200000e+01
25%	7.000000e+00	2.200000e+01	4.476000e+03	2.490000e+02	1.000000e+00
50%	1.400000e+01	3.100000e+01	9.343000e+03	3.990000e+02	1.000000e+00
75%	2.300000e+01	4.700000e+01	1.568400e+04	9.990000e+02	1.000000e+00
max	3.300000e+01	5.900000e+01	2.216900e+04	3.079800e+05	2.169000e+03

Figure 1. Describe function in dataset train_sales

In addition, we observed that the 75th quartile in 'item_cnt_day' is 1.00, which means that 75% of the items sold each day were <= to 1, whereas the max value is 2,169. We notice a similar pattern in 'item_price' where 75% of the price is <= 999, but the max value is 307 980. This indicates that we might have significant outliers.

From the histogram obtained for each column in *Figure 2*, we observed that the highest peak of items sold usually occurred around month one (January), this suggests that we might have a seasonal component in our dataset. Also, the variable 'item_price' presented a distribution significantly skewed to the left, implying that we present outliers in this column representing items considered luxury or collectibles. To avoid a significant impact on the outliers in our model, we removed all items (n=3) having a price higher than 50 000.

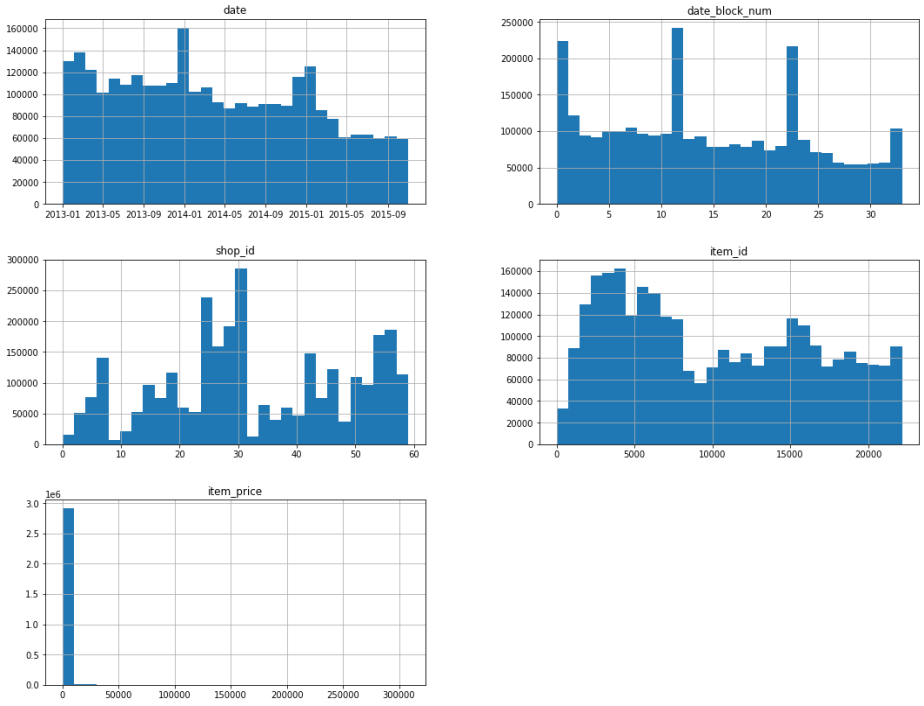

Figure 2. Histogram of data frame train_sales

- ***Data Exploration***

From the data exploration done in the data frame, we observed that items that were sold the most during the 3 years have a price under the average, except for one time (ID 5822). We also observed that the items that were sold the most during the 3 years are related to electronics and video games.

From *Figure 3*, we observed that the number of items sold by month for each year started increasing significantly around November, reaching a peak in December and decreasing in January of the next year. This behavior could be due to events like thanksgiving, black Friday, and Christmas when people tend to buy more things than what they usually buy like gifts and decorations. We anticipate having a similar behavior in the months that our model will try to predict.

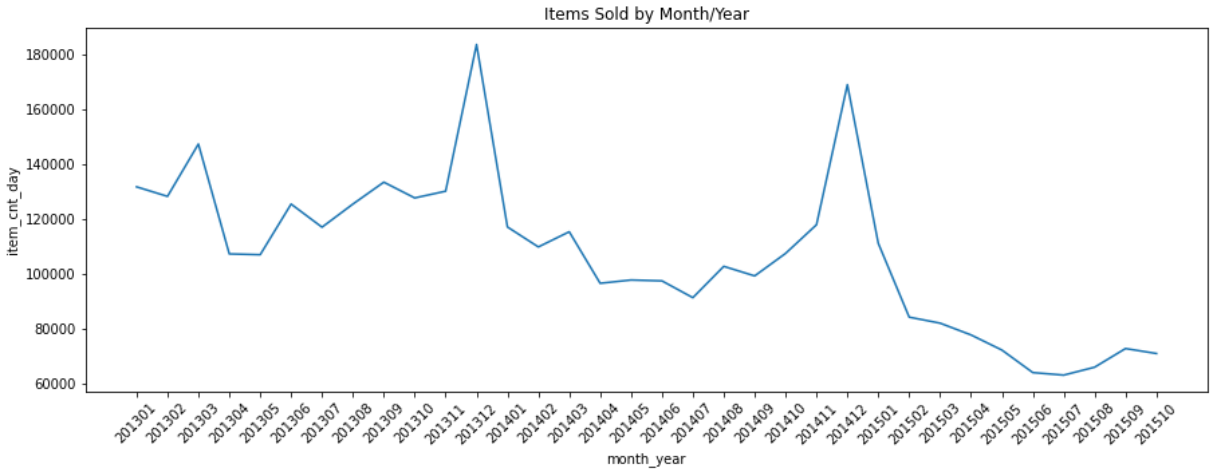

Figure 3. Items sold by Month/Year

From *Figure 4* we observed that the amount of money made in each month for every year seems to follow a similar behavior through the months of November and December. We can confirm from these plots that our dataset has a seasonal component, we will discuss the impact of this characteristic further.

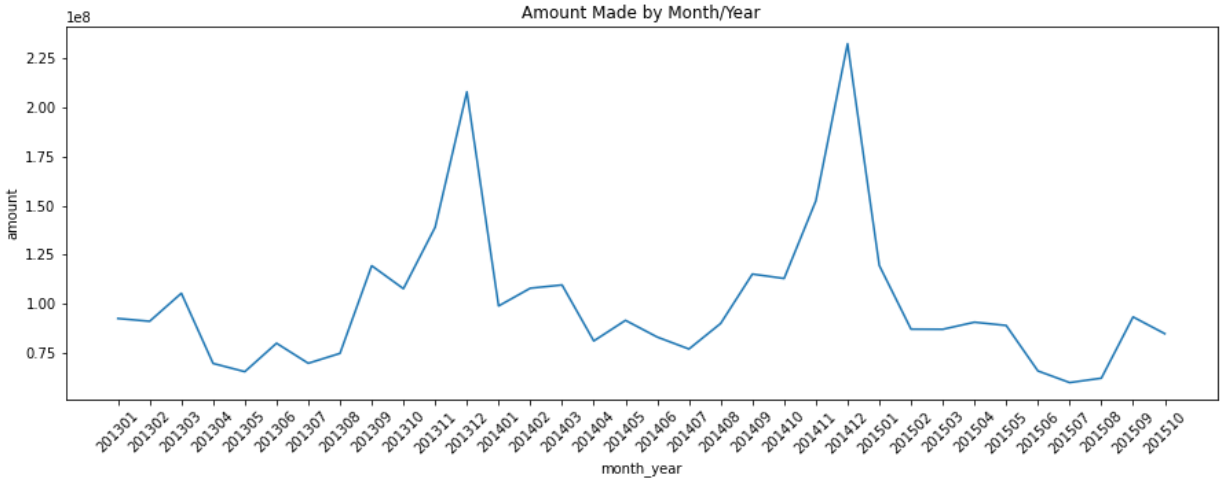

Figure 4. Amount of money made by Month/Year

In addition, we analyzed the correlation between the features and we arrived to the conclusion that there is no significant correlation between the variables. The highest correlation value obtained was 0.44 between column 'item_price' and 'item_count_day', see *Figure 5*.

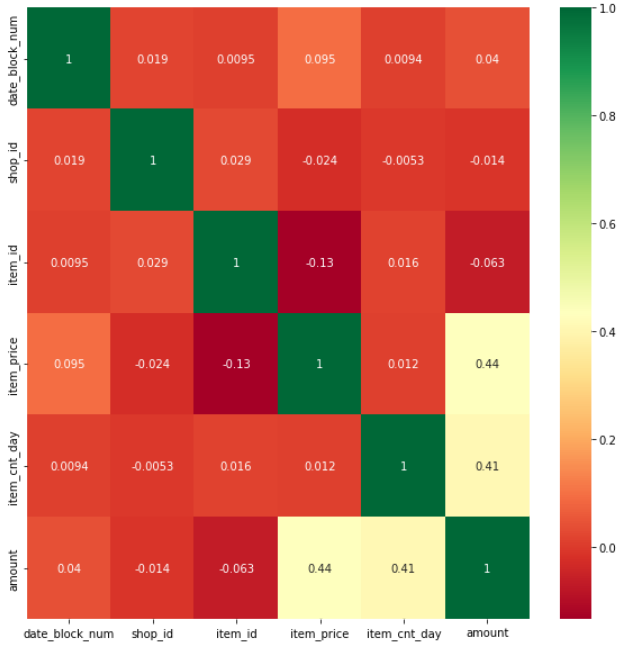

Figure 5. Correlation matrix between features.

Finally, for our exploratory analysis, we decided to cluster the items sold using the 'KMeans method' to observe more clearly how different are the types of items being sold by the company and understand better the forecast. After applying the 'Elbow Method', we concluded that the optimal number of clusters to group the items is k=3, see *Figure 6.a.* As seen in *Figure 6.b*, we have 3 well-defined groups of clusters meaning that the company has 3 well defined types of products.

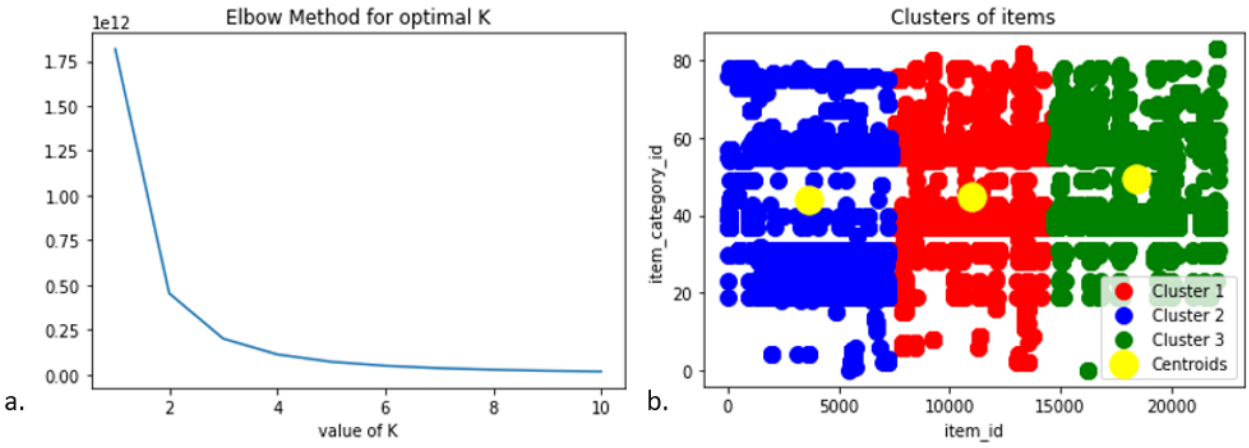

Figure 6. Elbow method and Cluster for items

Model for Time Series forecasting

This project aims to create a forecast of the number of items that will be sold by 1C Company during the months November 2015 to April 2016. To do this, we created a model based on the historical time-stamped data of the items sold by the company from January 2013 to October 2015.

First, we used the seasonal decompose function from the 'statmodels' library to observe the trend a seasonality of our dataset, see *Figure 7*. As observed in the graphs, we concluded that the dataset presents a Downtrend and a seasonal component that needs to be accounted when applying the model.

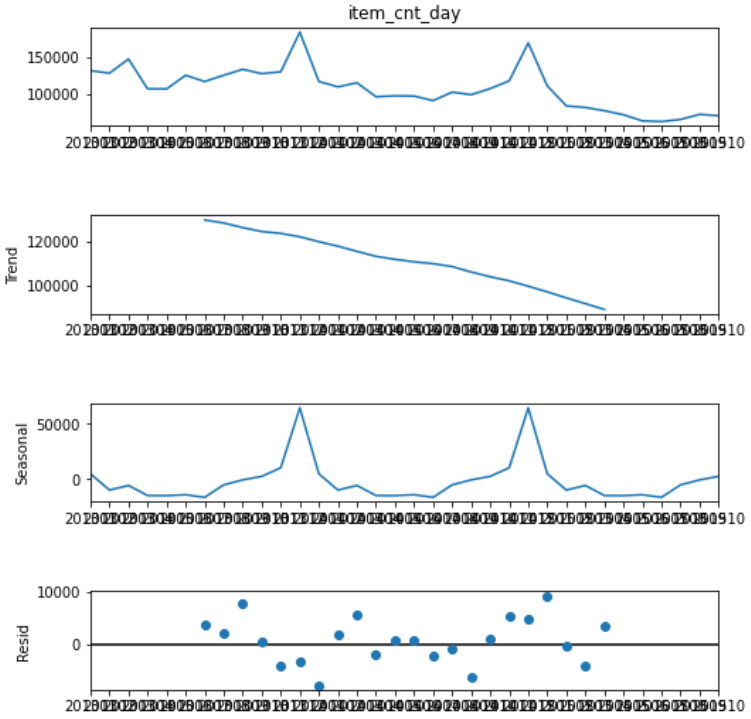

Figure 7. Seasonal decompose graph

- **Trend:** Pattern in the data that shows the movement of a series to relatively higher or lower values over a long period.
- **Seasonality:** Predictable pattern that recurs or repeats over regular intervals. Seasonality is often observed within a year or less.

- *SARIMA model*

The Seasonal AutoRegressive Integrated Moving Average (SARIMA) model is an extension of the widely used forecasting method ARIMA. The main difference with the ARIMA model is that it supports direct modeling of the seasonal component present in time series.

The parameters of the SARIMA method are specified as follows:

- **p** and seasonal **P**: indicate the number of autoregressive terms.
- **d** and seasonal **D**: indicate differencing that must be done to stationarize series.
- **q** and seasonal **Q**: indicate the number of moving average terms (lags of the forecast errors).

- **s**: indicates the seasonal length in the data.

Stationarity

Stationarity is a property of time-series data sets that needs to be evaluated before applying any model of forecasting. This parameter describes the statistical properties of a time series, looking at the way the data changes over time and making sure that the way it changes, does not changes itself over time.

As described above and observed in *Figure 7*, we have already confirmed that the model presented a trend and a seasonal component, thus, it is very likely that the dataset is non-stationary, meaning that the way the data changes over time is not stationary. We will applied the Augmented Dickey-Fuller test to confirm this hypothesis.

- *H0*: The time series is non-stationary.
- *HA*: The time series is stationary.

```
Results of Dickey-Fuller Test:
Test Statistic   -2.392387
p-value           0.143897
dtype: float64
```

Figure 8. Results of Dickey Fuller Test

Since the p-value we obtained from this test is higher than 0.05, we can't reject the null hypothesis. We conclude then that our dataset is **non-stationary.**

Differencing Method

In order to prepare our predictive model, we stabilized the dataset by applying the differencing method. This method can help stabilize the mean of the time series by subtracting a previous observation from the current observation, and so eliminating (or reducing) trend and seasonality.

```
Results of Dickey-Fuller Test:
Test Statistic   -3.262794
p-value           0.016626
dtype: float64
```

Figure 9. Results of Dickey Fuller Test after Differencing method

After applying the Differencing Method and re-applying the Augmented Dickey-Fuller test to the new dataset, we obtained a p-value lower than 0.05. We can confirm that our dataset is now **stationary**.

Autocorrelation (ACF) and Partial Autocorrelation (PACF)

Finally, we looked at the auto-correlation and partial autocorrelation of our dataset to determine the degree of similarity between the time series and the lag over successive time periods, see *Figure 10*. These two graphs can give an intuitive approach at estimating the terms for seasonal and non-seasonal parameters (p,d,q)x(P,D,Q,s) of the SARIMA model. We used the function 'auto_arima' to optimize the selection of these parameters by applying Grid Search to the model.

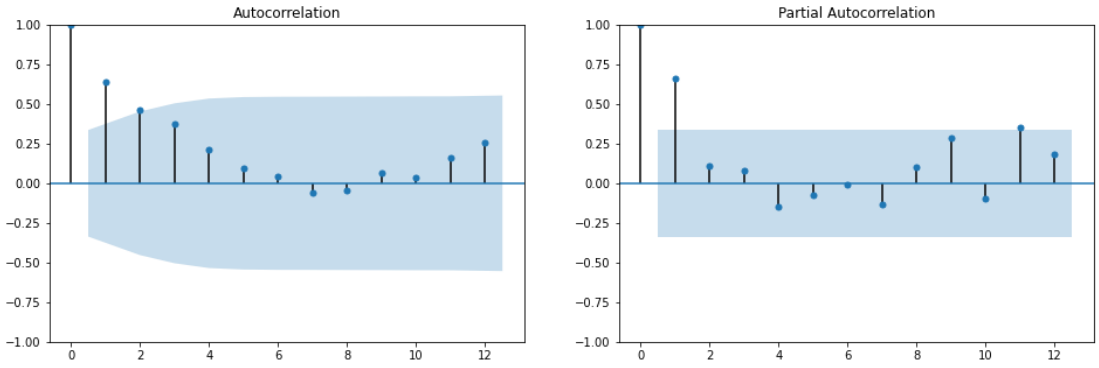

Figure 10. ACF and PACF curves

We obtained that the most optimized combination of parameters for the model was:

p = 0, d = 2, q = 1, P = 0, D = 1, and Q = 1, see *Figure 11.*

```
Best model:  ARIMA(0,2,1)(0,1,1)[12]
Total fit time: 4.195 seconds
```

Figure 11. SARIMA parameters

Evaluation of the Model

Since the actual number of items sold by the 1C Company during the desired months is completely unavailable, we evaluated our model based on the summary results and the plot diagnostic. From the summary results of the model (see *Figure 12*) we concluded that:

- Ljung-Box p-value > 0.05: Residuals are independent.
- Heteroskedasticity p-value > 0.05: Fail to reject the null hypothesis of Homoscedasticity.
- Jarque Bera p-value > 0.05: Sample data follows normal distribution.

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                   34
Model:             SARIMAX(0, 2, 1)x(0, 1, 1, 12)   Log Likelihood                -216.014
Date:                            Tue, 09 Aug 2022   AIC                            438.028
Time:                                    13:09:36   BIC                            441.016
Sample:                                         0   HQIC                           438.611
                                              - 34
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9858      0.299     -3.301      0.001      -1.571      -0.401
ma.S.L12      -0.0853      0.167     -0.511      0.610      -0.413       0.242
sigma2      2.049e+08   6.07e-10   3.37e+17      0.000    2.05e+08    2.05e+08
===================================================================================
Ljung-Box (L1) (Q):                   0.74   Jarque-Bera (JB):                 0.88
Prob(Q):                              0.39   Prob(JB):                         0.64
Heteroskedasticity (H):               0.80   Skew:                             0.24
Prob(H) (two-sided):                  0.77   Kurtosis:                         2.09
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.79e+33. Standard errors may be unstable.
```

Figure 12. SARIMA model summary results

From the plot diagnostics (see *Figure 13*) we observed that:

- The Standardized errors seem to fluctuate around a mean of zero.
- The Histogram of the model suggests that the data follows a normal distribution slightly skewed to the left.
- As we see from the Sample Quantities plot, not all the dots fall perfectly in the red line, nevertheless they are not far away from it either. We can imply the distribution is skewed as we saw in the histogram.
- The correlogram plot shows that the residual errors are not autocorrelated.

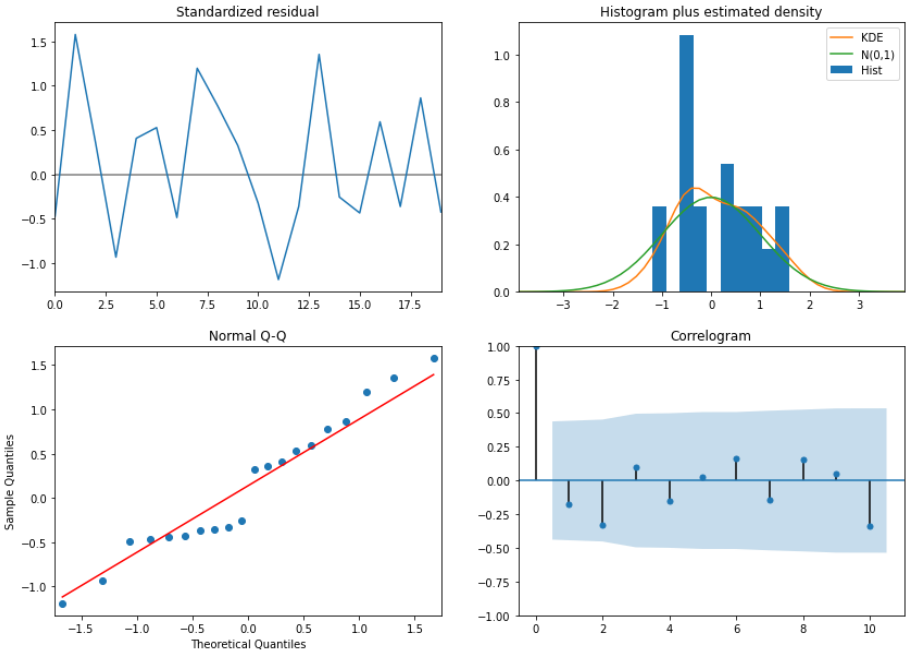

Figure 13. SARIMA model plot diagnostic

Overall, the model seems to have a good fit for the dataset and can produce a forecast for the future.

- *Prediction*

As observed in Table 2 and Figure 14, when using the SARIMA model to do a prediction for the months of November 2015 to April 2016, we obtained a significant increase in sales for the winter of 2015 and an important decrease in sales at the beginning of 2016. We also observed that the overall downtrend behavior of the sales is maintained through the months predicted, but we obtained a higher value for the initial items sold in October 2015.

Table 2. Predicted items sold from November 2015 to April 2016.

date	Predicted item_cnt_month
2015-10	78010.30
2015-11	125925.95
2015-12	65870.27
2016-01	37780.98
2016-02	33427.73
2016-03	24867.50
2016-04	17003.73

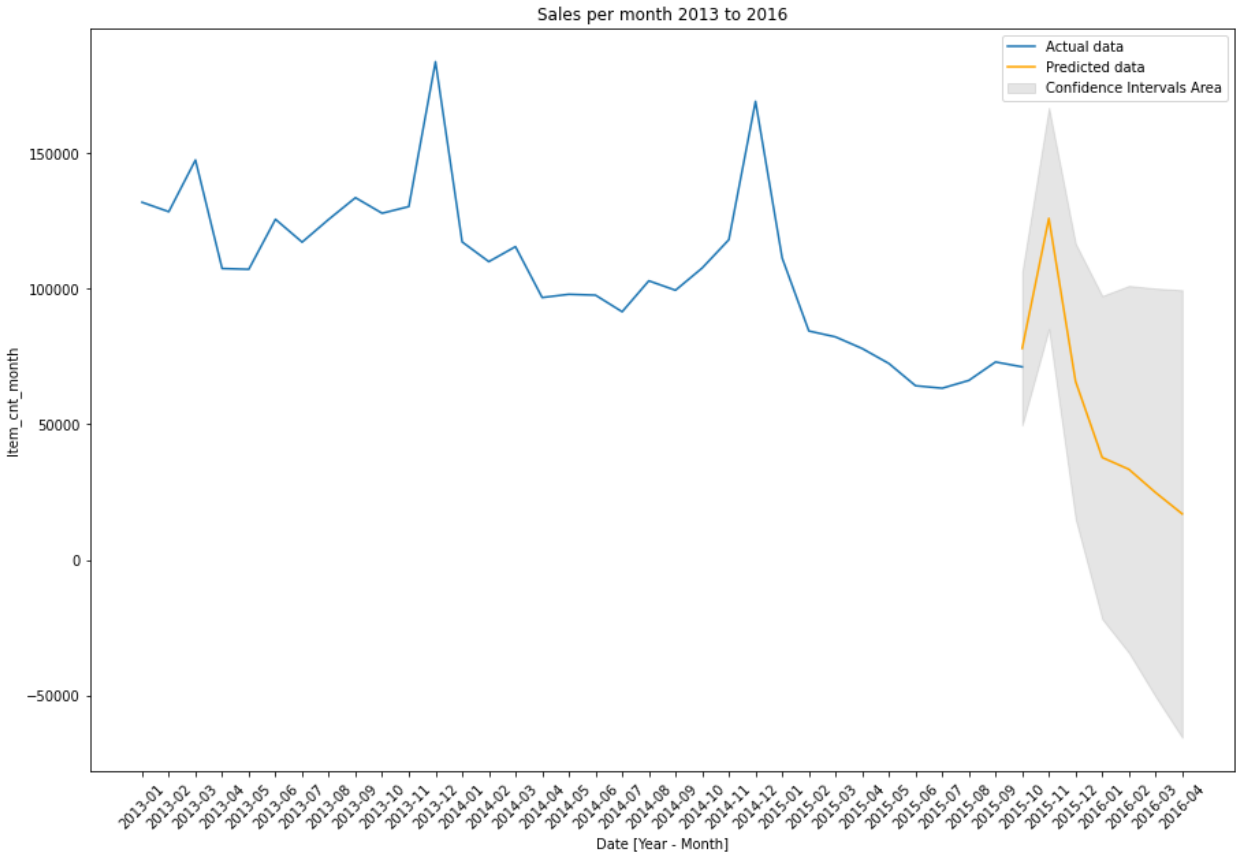

Figure 14. Sales per month from January 2015 to April 2016

Conclusion

This report showed that using a SARIMA model to predict the sales of the following six months for 1C Company, can generate a reliable and accurate forecast. Even though the model struggled to fit the initial value of the forecast (as observed in October 2015, see *Figure 14*), it quickly learnt and captured the seasonality and trending behavior within the dataste to produce a more fitted prediction for the desired months. These predictions can be incredibly useful for the 1C company as knowing the approximate sales volume will give them key info on how much stock to order, how many staff do they need for each month, project profits and other important aspects of the business.

Limitations

The SARIMA model can be handy and relatively easy to tunn to create a baseline forecasting due to is linear nature. Nevertheless, the seasonal components of the parameters used in this model require a significant amount of data to work on. It is likely that the model didn't have enough data to do an accurate estimation of the seasonal parameters, given that the dataset we used had only three years of historical data. A potential improvement that could been applied to the model is to connect it to a live dataset that updates the sales data daily. This way, the hyperparameters of the model could be optimized more precisely automatically and the model could generate a better prediction for the following six months to a year of sales data, with the most up to date and accurate information.