**FORECAST THE TOTAL AMOUNT OF PRODUCTS SOLD IN EVERY SHOP**

Contents

Introduction 1-1

Problem Statement

2.1 Objective and Scope of Project 2-1

Data Collection

3.1 Source 3-1

3.2 Data Pre-Processing 3-2

3.3 Tools and Techniques 3-3

3.3.1 Tableau 3.3.1

3.3.2 Python 3.3.2

3.3.3 R 3.3.3

3.4 High-level Approach 3-4

3.5 Constraints 3-5

Exploratory Data Analysis

4.1 Data Dictionary Table 4-1

4.2 Graphs and Observations 4-1

4.3 Overall EDA Conclusions 4-1

Model Development

5.1 Model Development – ARIMA 5-1

5.1.1 Atrium 5.1.1

5.1.2 Central Shopping Mall 5.1.2

5.1.3 Kimki 5.1.3

5.1.4 Mega Teply Stan 5.1.4

5.1.5 Semonovisky 5.1.5

5.2 Model Development – Prophet 5-2

5.2.1 Semonovisky 5.2.1

Forecasting the Sales

6.2 ARIMA 6-2

6.3 Prophet 6-3

Appendices

*Appendix A* Code snippets to Data preprocessing

*Appendix B* Data Cleaning (Python)

*Appendix C* ARIMA (Python)

*Appendix D* Prophet (R)

1. **INTRODUCTION**

Sales are the lifeblood of a business. It is what helps to pay employees, cover operating expenses, buy more inventory, market new products and attract more investors. Sales forecasting is a crucial part of the financial planning of a business. Sales forecasting is  a self-assessment tool that uses past and current sales statistics to intelligently predict future performance.

With an accurate sales forecast in hand, we can plan for the future. If our sales forecast says that during December you make 30 percent of our yearly sales, then we need to ramp up manufacturing in September to prepare for the rush. It might also be smart to invest more in seasonal sales, people and start a targeted marketing campaign right after Thanksgiving. One simple sales forecast can inform every other aspect of your business.

Sales forecasts are also an important part of starting a new business. Almost all new businesses need loans or start-up capital to purchase everything necessary to get off the ground: office space, equipment, inventory, employee salaries and marketing. We cannot just walk into a bank with a bright idea and lots of enthusiasm. We need to show them numbers that prove your business is viable. In other words, you need a business plan

1. **PROBLEM STATEMENT:**

Sales forecast is the backbone of a business plan. People measure a business and its growth by sales, and the sales forecast sets the standard for expenses, profits and growth. The sales forecast is almost always going to be the first set of numbers we’ll track for plan vs. actual use, even if we do no other numbers.

The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

* 1. **OBJECTIVE AND SCOPE OF THE PROJECT:**

**1.2.1 Objective:**

The Primary objectives of the study are:

* Study the sales of various items across different stores of Russia in a given interval and forecast the sales.
* Identify the various factors that affect the sales of items across stores.
* Explore the possibility of developing a Time Series ARIMA model to forecast sales.
* Based on ARIMA model built features, performing a Multiple Linear Regression, Checking cross-validation score, and compare results with Auto ARIMA models.

**1.2.2 Scope:**

* The scope of the study covers sales of Electronic items across different shops of Russia
* The study covers three-year Data starting 1stJanuary’13 to 31st December’ 15. This is done to ensure seasonality factors are covered.
* Since the size of the file is too large to handle we have considered only five shops and five items in our study
* The Study’s focus to forecast number of products sold in each shop in a day

1. **DATA SOURCE:**

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends. Accurate data collection is essential to maintaining the integrity of research, making informed business decisions and ensuring quality assurance. Hence, we chose dataset consisting of daily sales data, provided by one of the largest Russian software firms - 1C Company.

There are 3000000 observations and 11 attributes. The dataset contains daily sales of 22170 different items in different 60 shops of Russia in the interval of January 2013 to October 2015.

**3.2 DATA PRE-PROCESSING**

Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data pre-processing transforms the data into a format that will be more easily and effectively processed for the purpose of the user -- for example, in a neural network. There are a number of different tools and methods used for pre-processing.

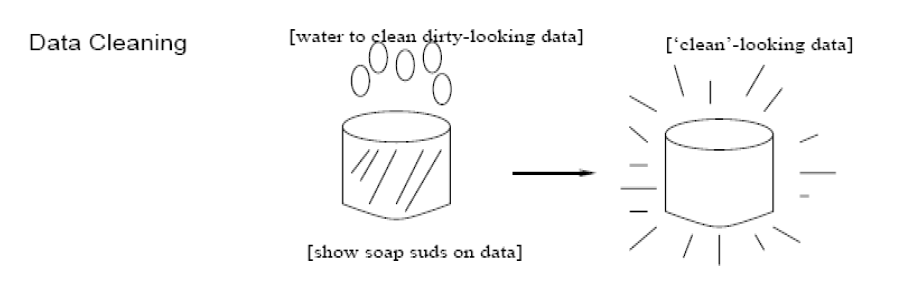
Steps in Data Pre-Processing-

Data preprocessing methods are divided into following categories

* Data Cleaning
* Data Integration
* Data Transformation
* Data Reduction

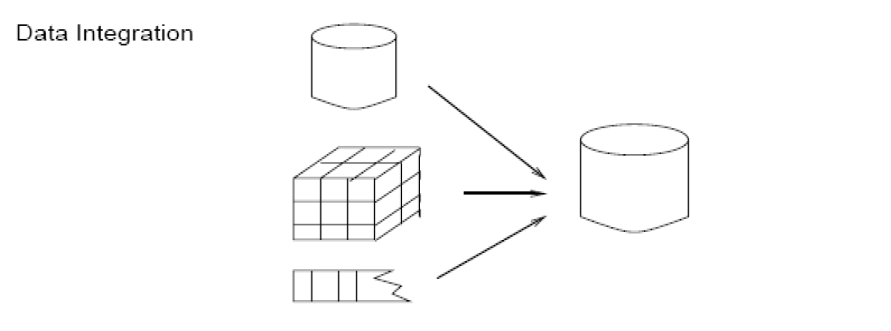
**3.2.1 DATA CLEANING**

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After analysing we came to know there are no missing, corrupt or inaccurate records in the dataset



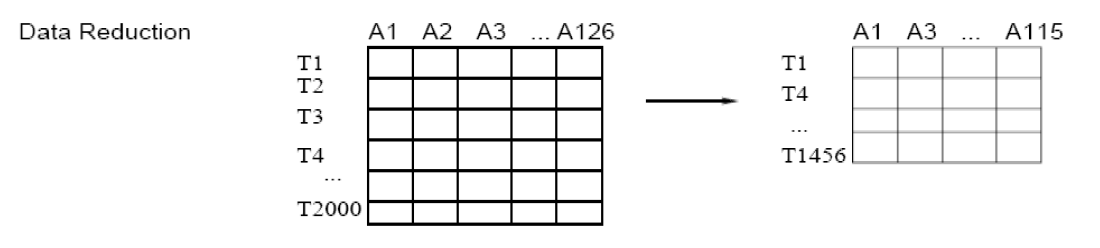
**3.2.2 DATA INTEGRATION**

Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information. A complete data integration solution delivers trusted data from various sources. We had combined sales test and train data with items, item categories, shops to obtain a clear information about the sales of products.



**3.2.3 DATA REDUCTION**

Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts. Since we had sales information about 60 stores, we had to filter top five shops based on sales.



**3.2.4 DATA TRANSFORMATION**

Feature engineering is an essential part of building any intelligent system. Features can be of two major types based on the dataset. Inherent raw features are obtained directly from the dataset with no extra data manipulation or engineering. Derived features are usually obtained from feature engineering, where we extract features from existing data attributes.

**Derived features :**

We derived the following features from the date column

* Name of the week day
* Day number of the weekday
* Whether day is weekday or weekend
* Day is holiday or not
* Festival day or not
* Season of the current date i.e., Winter, Summer, Autumn, Spring

**3.3 TOOLS & TECHNIQUES**

We have used the following Analytical techniques/Methodology for analysing the Data

1. Using Graphs to visually represent them

2. Tools used: Python, R Programming, Tableau & Excel

* **Tableau:** Tableau allows for instantaneous insight by transforming data into visually appealing, interactive visualizations called dashboards. This process takes only seconds or minutes rather than months or years, and is achieved through the use of an easy to use drag-and-drop interface. We have used it for EDA on our data
* **Python:** Python is a multi-paradigm programming language: a sort of Swiss Army knife for the coding world. It supports object-oriented programming, structured programming, and functional programming patterns, among others. We have used it for building model on forecasting.
* **R Programming:** R is a language used for statistical computations, data analysis and graphical representation of data. R is the second most popular language in data science. This shows how popular R programming is in data science. We have used it for building model on forecasting.

3. Techniques: Tableau dash boarding, ARIMA, Prophet Forecasting, MLR.

**3.4 HIGH LEVEL APPROACH**

The Analytical Approach will involve the following (not necessarily in the order) activities:

* Data extraction from Primary Data source as well as secondary data sources
* Data quality check
* Data cleaning and data preparation
* Study each of the variables by exploring the data
* Study the variables for its relevance for the study
* Adding additional features required for the data
* Running a ARIMA algorithm on the data to forecast the sales

We used the following Seven Step Analytical Approach to the Project:

Understand the Domain, Identify business problem

Data Collection

Data Cleaning, Preparation

Data reduction and transformation

Develop Model

Validate Model

Assessment and Review with stakeholders

**Figure: High Level Process Functional Architecture**

**3.5 CONSTRAINTS**

There are few limitations that this study has w.r.t data and the methodology that can be used.

* Due to size of data, we could not deploy whole dataset for the study so we considered chose randomly five stores and five items and forecast their sales.

## EXPLORATORY DATA ANALYSIS

In statistics, exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. The Exploratory Data Analysis (EDA) is done using Tableau.

## 4.1 DATA DICTIONARY TABLE

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

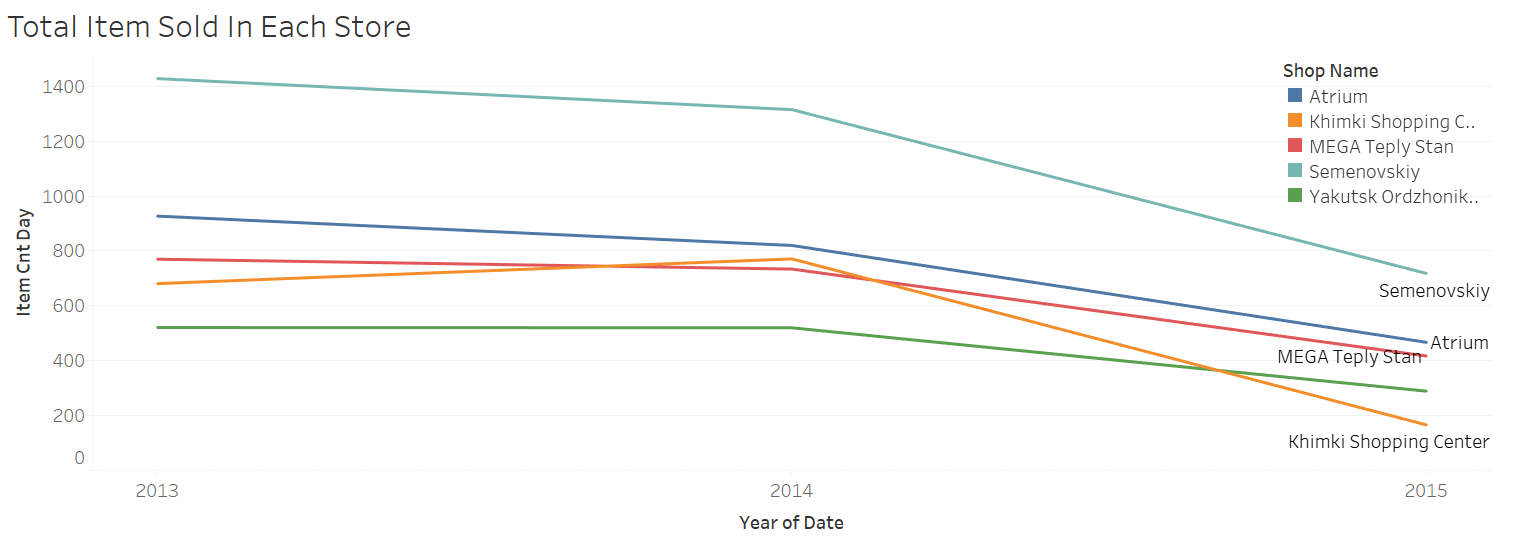
# 4.2 GRAPHS AND OBSERVATIONS:

The Exploratory Data Analysis can be classified into seven parts:

* Analysing the Total number of items sold in each store
* Analysing the sale of individual items in each store
* Analysing the monthly sales of individual items in each store
* Analysing the daily sales of individual items
* Analysing the effect of seasons in sales of individual items
* Analysing the sales of items over a weekday and weekend
* Analysing the sales of items over a holiday and a non-holiday
* Forecasting Using Tableau

**4.2.1 ANALYSING THE TOTAL NUMBER OF ITEMS SOLD IN EACH STORE**

We used simple line Graph to plot the sales across the interval. The year is plotted on X– axis and the total item count were plotted on the Y-axis and each colour representing an individual store.



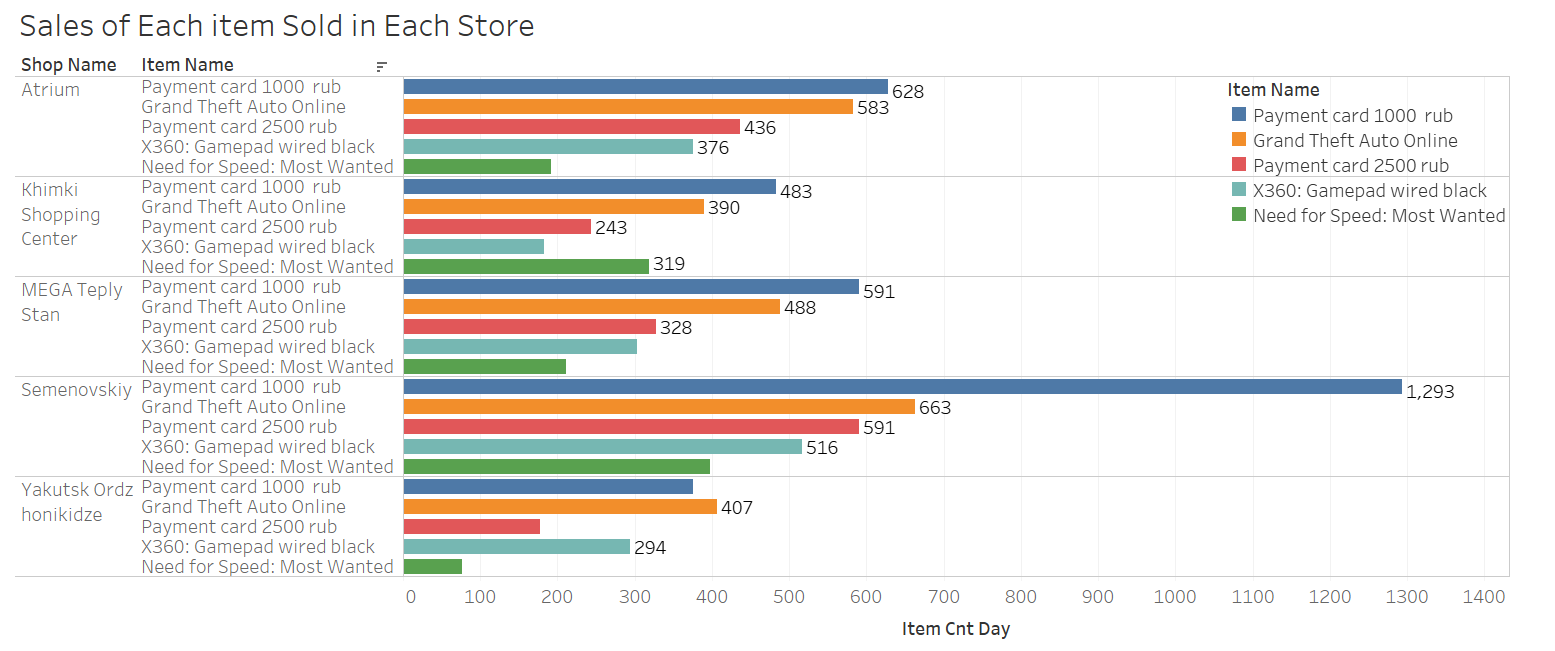
**Figure : Total Item Sold In Each Store**

**OBSERVATIONS:**

* In all the five stores the sale is dropping from 2013 to 2015
* The sales of Semenovskiy and Khimki shopping centre were drastically decreasing from 2014 to 2015
* Total number of items sold Semenovskiy is higher than all other stores in the time interval from 2013 to 2015
* Yakustsk store has the least number of items sold in the time interval from 2013 to 2015

* + 1. **ANALYSING THE SALE OF INDIVIDUAL ITEMS IN EACH STORE:**

We used simple line Graph with dual axis to plot the sales of individual items in each store across the interval. The year and item name is plotted on X– axis and the total item count were plotted on the Y-axis and colour representing each store.



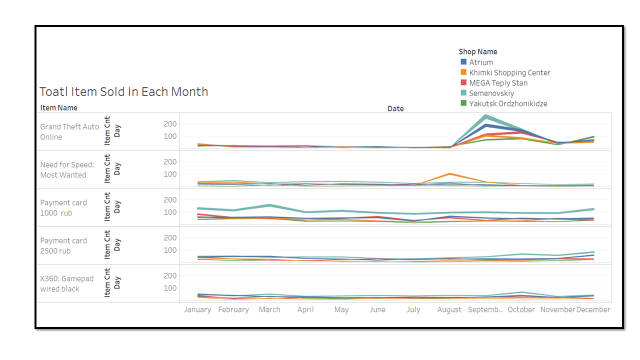
**Figure : Sales Of Each item Sold in Each Store**

**OBSERVATIONS:**

* Sales of all the 5 items high in Semenovskiy followed by Atrium
* Compared to all the products ‘Payment card 1000’ had been brought most in 2014
* Need for speed is the least bought game in the time interval from 2013 to 2015
* Grand Theft Auto was the most popular game during the time interval 2013 to 2015

**4.2.3 ANALYSING THE MONTHLY SALES OF INDIVIDUAL ITEMS IN EACH STORE**

We used simple line Graph to plot the monthly sales of individual items in each store across the interval. The month is plotted on X– axis and the total item count and item name were plotted on the Y-axis and colour representing each store.



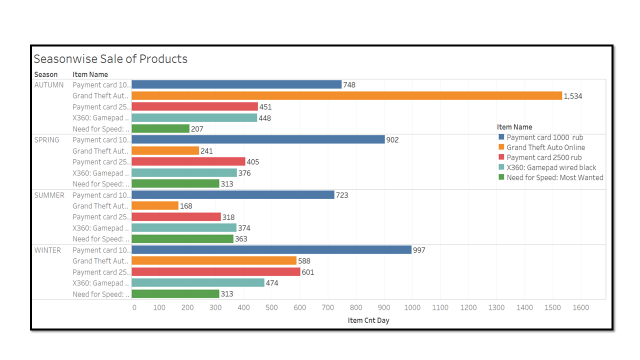
**Figure : Total Item Sold in Each Month**

**OBSERVATIONS:**

In all the three years, September has high sales and we are going deeper to understand why there is a hike in September.

**4.2.4 ANALYSING THE EFFECT OF SEASONS IN SALES OF INDIVIDUAL ITEMS:**

We used simple line Graph with dual axis to plot the sales of individual items in each store across seasons. The date and season is plotted on X– axis and the total item count were plotted on the Y-axis and colour representing the name of the item.



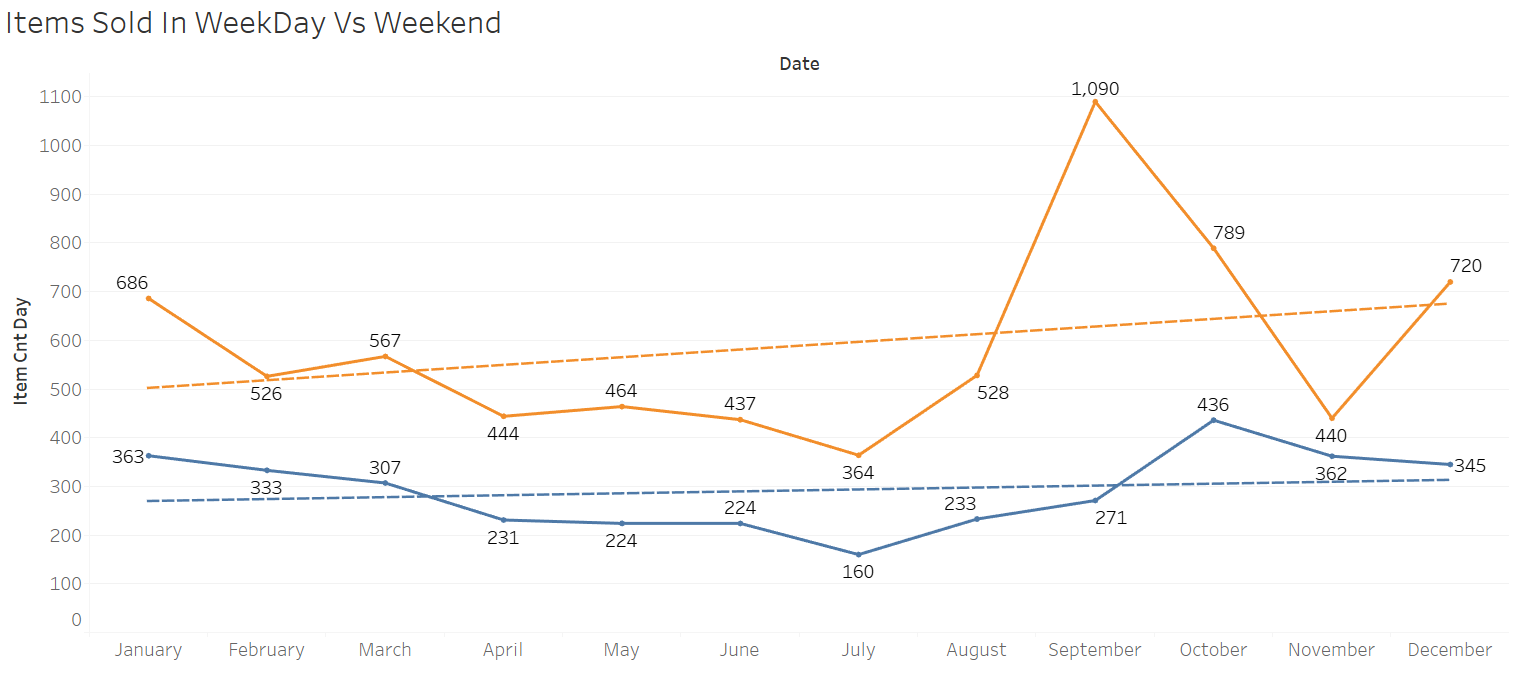
**Figure : Seasonwise sale of product**

**OBSERVATIONS:**

* Autumn (Sep-Oct-Nov) has a greater number of sales than the other seasons
* GTA was sold mostly in autumn season
* Sales of all the items is less in summer season

**4.2.5 ANALYSING THE SALES OF ITEMS OVER A WEEKDAY AND WEEKEND:**

We used simple line Graph with dual axis to plot the sales of individual items in each store across seasons. The year and item name is plotted on X– axis and the total item count were plotted on the Y-axis and colour representing whether the date is a weekend or a week day (Saturday and Sunday are weekend else weekday)



**Figure : Item Sold in Weekday and Weekend**

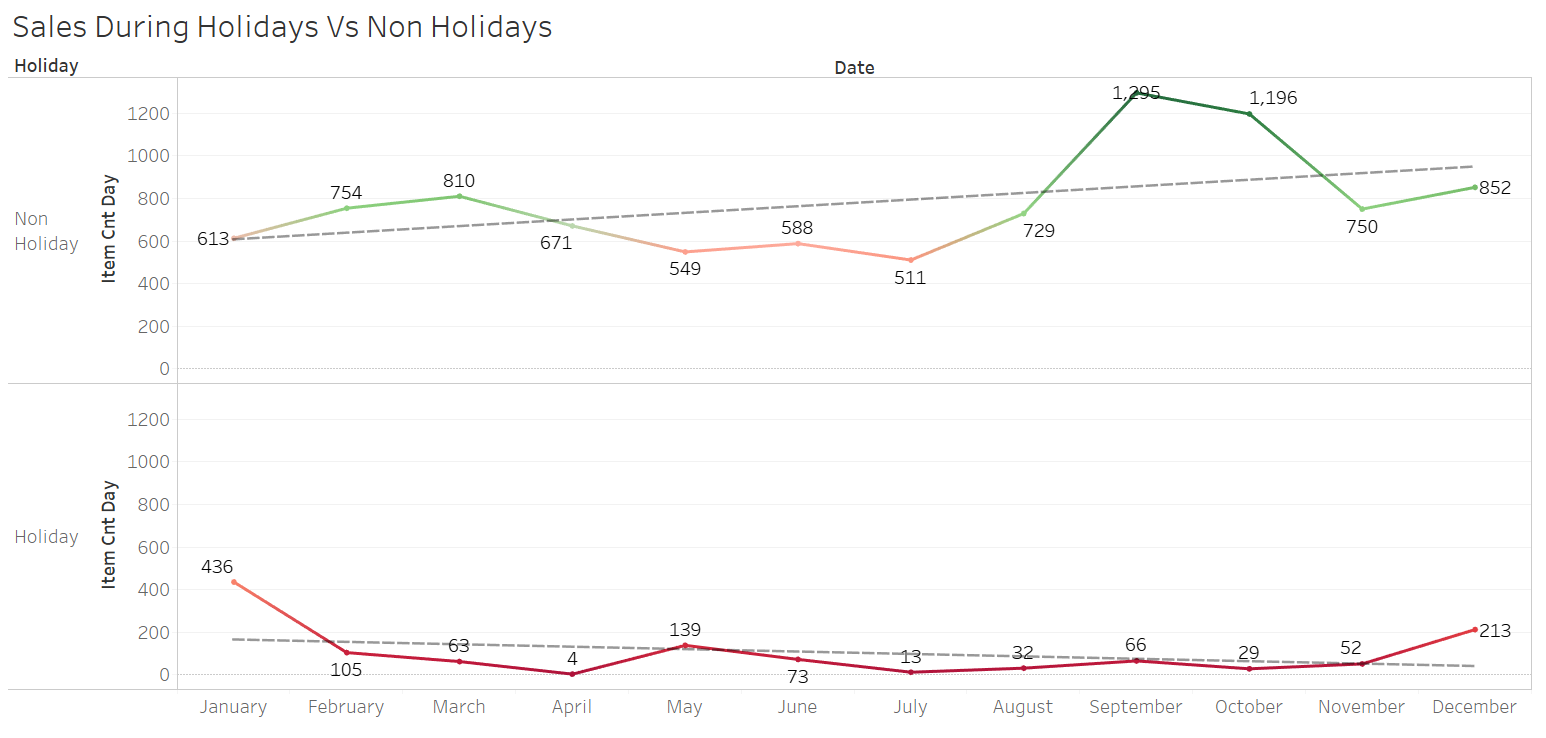
**OBSERVATIONS:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Day Labels** | **Sum Of Item\_Cnt\_Day** | **Count Of Week Day\_Week Age** | **Proportion** |
| **weekday** | 7055 | 4085 | 0.58 |
| **weekend** | 3489 | 2116 | 0.61 |
| **Grand Total** | 10544 | 6201 |  |

From the proportion, we could see the sales on weekend is little high than in weekdays.

**4.2.6 ANALYSING THE SALES OF ITEMS OVER A HOLIDAY AND A NON-HOLIDAY:**

We used simple line Graph with dual axis to plot the sales of individual items in each store across seasons. The year and item name is plotted on X– axis and the total item count were plotted on the Y-axis and the dual axis is made in which one is representing the holiday and other one a non-holiday.



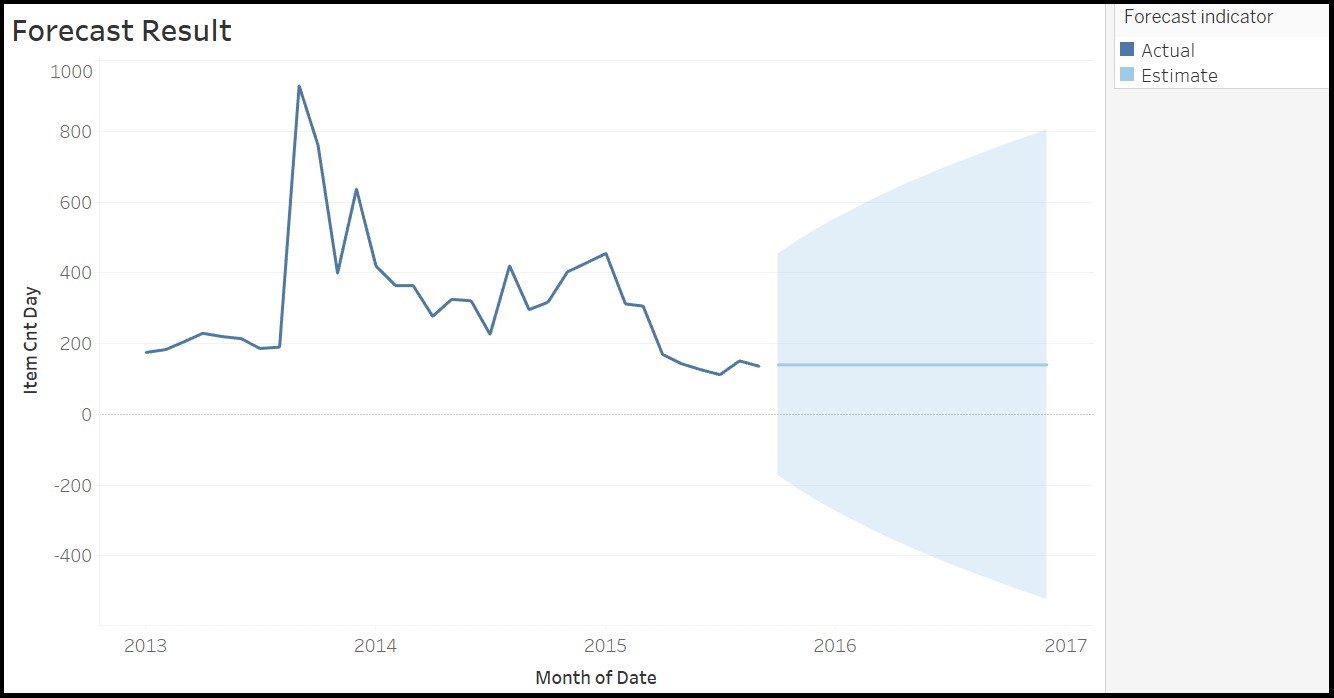
**Figure : Sales During Holiday vs Non-Holiday**

**OBSERVATIONS:**

From this, it is clearly visible that, sale over non-holiday is higher than holiday. It is most over the months of September and October.

**4.2.8 FORECASTING USING TABLEAU**

We have done a simple forecasting using Tableau before taking it into Python. We have used simple Moving Average method with 95% confidence interval



**Figure : Forecast**

## FORECASTING MODEL DEVELOPMENT

* 1. ARIMA

We are using ARIMA, A popular and widely used statistical method for time series forecasting. ARIMA is an acronym that stands for Autoregressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data.

The parameters of the ARIMA model are defined as follows:

* p: The number of lag observations included in the model, also called the lag order.
* d: The number of times that the raw observations are differenced, also called the degree of differencing.
* q: The size of the moving average window, also called the order of moving average. An ARIMA model can be created using the statsmodels library as follows:

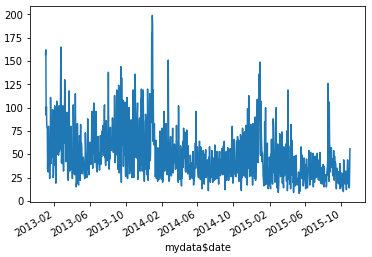
* Define the model by calling ARIMA () and passing in the p, d, and q parameters.
* The model is prepared on the training data by calling the fit () function.
* Predictions can be made by calling the predict () function and specifying the index of the time or times to be predict

In ARIMA model, as the dates are ought to be continuous, from the whole dataset, the particular five shop ID’s which include ‘Atrium’, ‘mega teply stan’, ‘Central Shopping’, ‘Kimki’, ‘Seminovisky’ are taken completely with all the products in it.

# CENTRAL SHOPPING MALL

# Time-series:

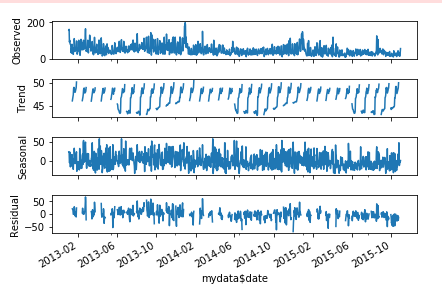
The first graph is a basic time-series graph which shows the spread of our data through a time period



**Figure:** Time series graph

From the time series graph, it can be said with prima facie that there is no trend and seasonality in that. For further analysis, decomposition has to be done and analysed.

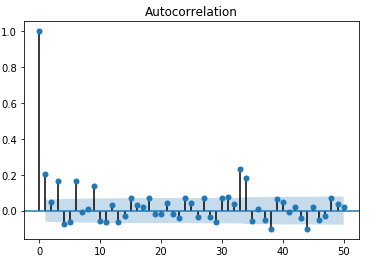
# Decomposition of Time Series



### **Figure:** Decomposition of time series

The decomposition of time-series splits the time series data into trend, seasonality, residual and the observed value. Since there are no such trend or seasonality in our data, the pattern is random.

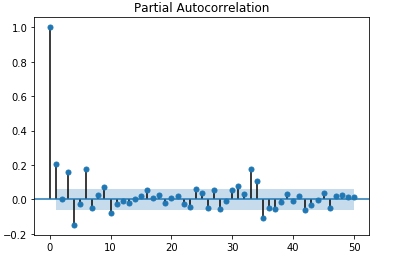
# Auto-Correlation Function Plot (ACF):



**Figure:** ACF plot

From this plot, the ‘q’ value which is to be passed in ‘ARIMA’ function is taken.

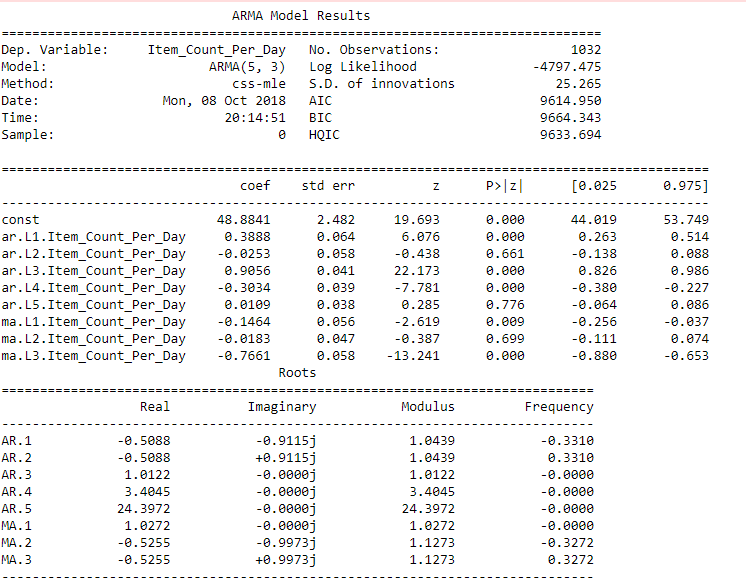
# Partial Auto-Correlation Function Plot (PACF):

****

**Figure:** PACF plot

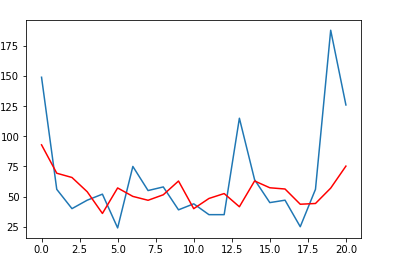
From this plot the ‘p’ value which is to be passed in ‘ARIMA’ function is taken.

1. **Fitting ARIMA model:**



**Figure:** Fit Arima to the time series

1. **Forecast:**



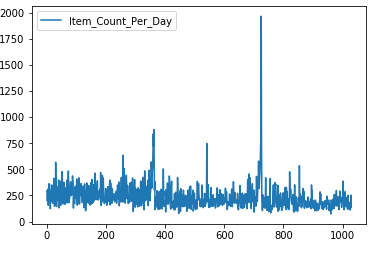
**Figure : Forecast For Central Shopping**

### **ATRIUM:**

1. **Time-series:**

The first graph is a basic time-series graph which shows the spread of our data through a time period.

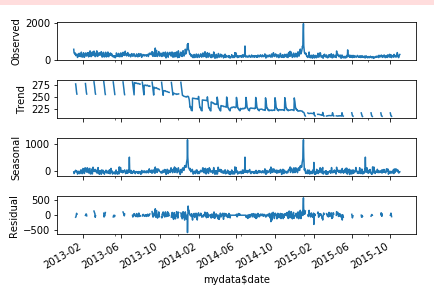
### 



### **Figure:** Time series graph

From the time series graph, it can be said with prima facie that there is no trend and seasonality in that. For further analysis, decomposition has to be done and analysed.

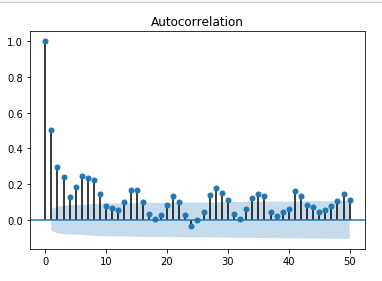
1. **Decomposition of time-series:**



### **Figure:** Decomposition of time series

The decomposition of time-series splits the time series data into trend, seasonality, residual and the observed value. Since there are no such trend or seasonality in our data, the pattern is random.

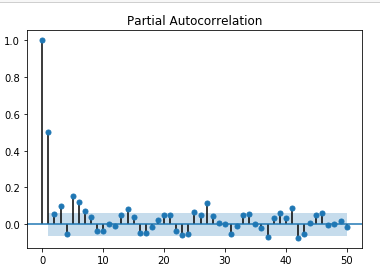
1. **Auto-Correlation Function Plot (ACF):**



**Figure:** ACF plot

From this plot, the ‘q’ value which is to be passed in ‘ARIMA’ function is taken.

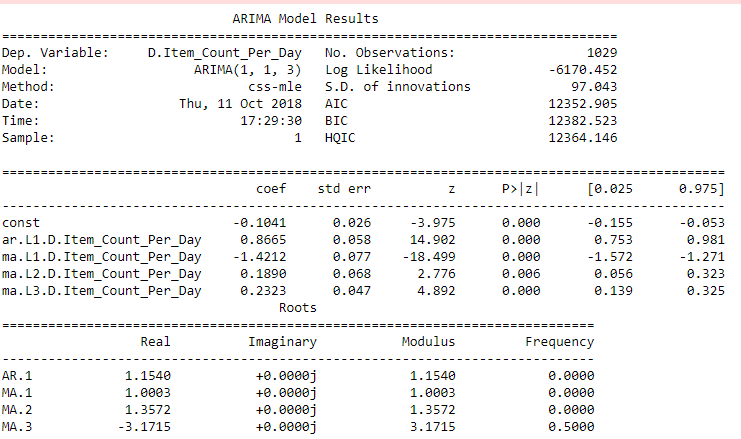
1. **Partial Auto-Correlation Function Plot (PACF):**

****

**Figure:** PACF plot

From this plot the ‘p’ value which is to be passed in ‘ARIMA’ function is taken.

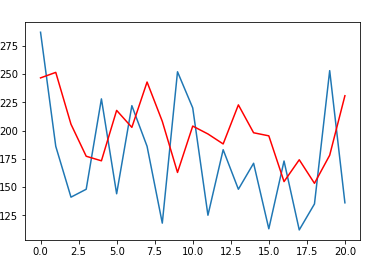
1. **Fitting ARIMA model:**



### **Figure:** Fit Arima to the time series

With the help of ‘p’, ‘d’, ‘q’ values obtained we fit the ARIMA model

1. **Forecast:**

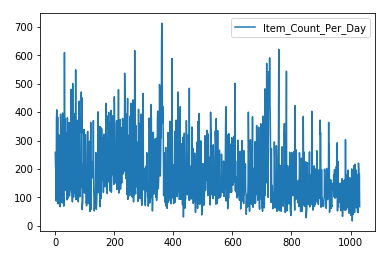


**Figure : Forecast for Atrium Store**

### **MEGA TEPLY STAN**

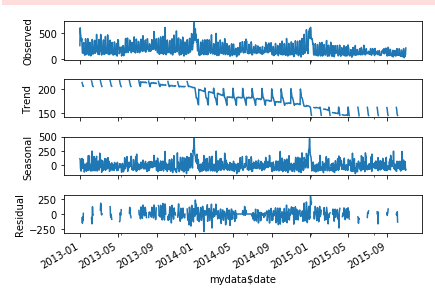
1. **Time-series:**

The first graph is a basic time-series graph which shows the spread of our data through a time period.



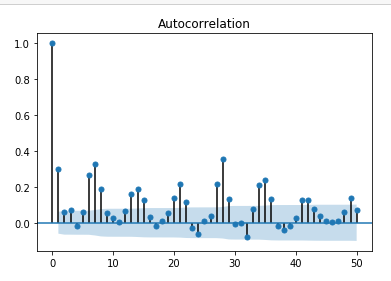
### **Figure:** Time series graph

1. **Decomposition of time-series:**



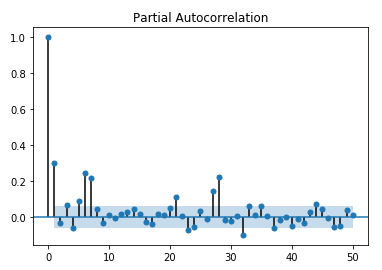
**Figure: Decomposition of time series**

1. **Auto-Correlation Function Plot (ACF):**



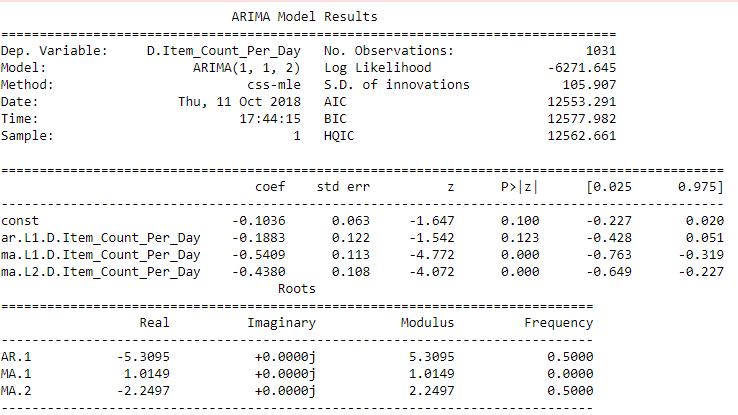
**Figure:** ACF plot

1. **Partial Auto-Correlation Function Plot (PACF):**

****

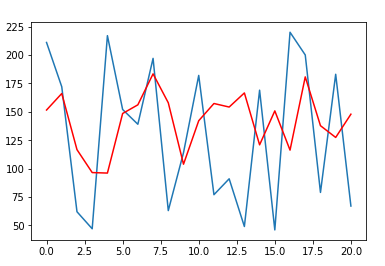
**Figure:** PACF plot

1. **Fitting ARIMA model:**



### **Figure:** Fit Arima to the time series

# Forecast:

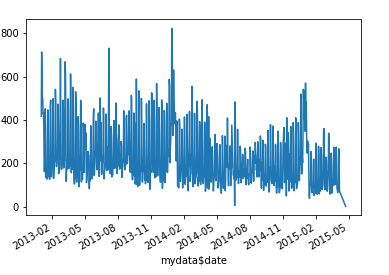


**Figure : Forecast for Mega Teply Stan Store**

### **KIMKI**

1. **Time Series**

### 

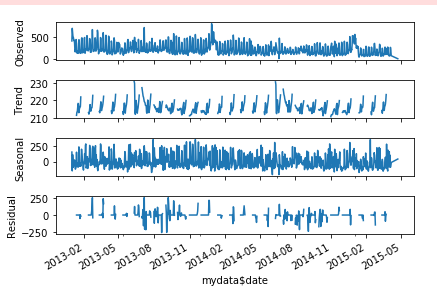


### **Figure:** Time series graph

### 

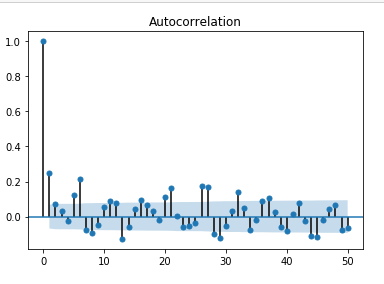
### **Decomposition of time-series:**

### 



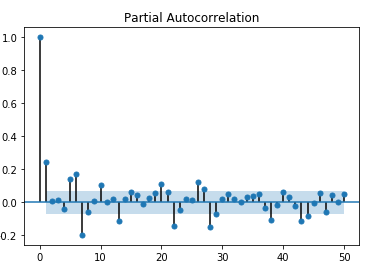
### **Figure:** Decomposition of time series

1. **Auto-Correlation Function Plot (ACF):**

****

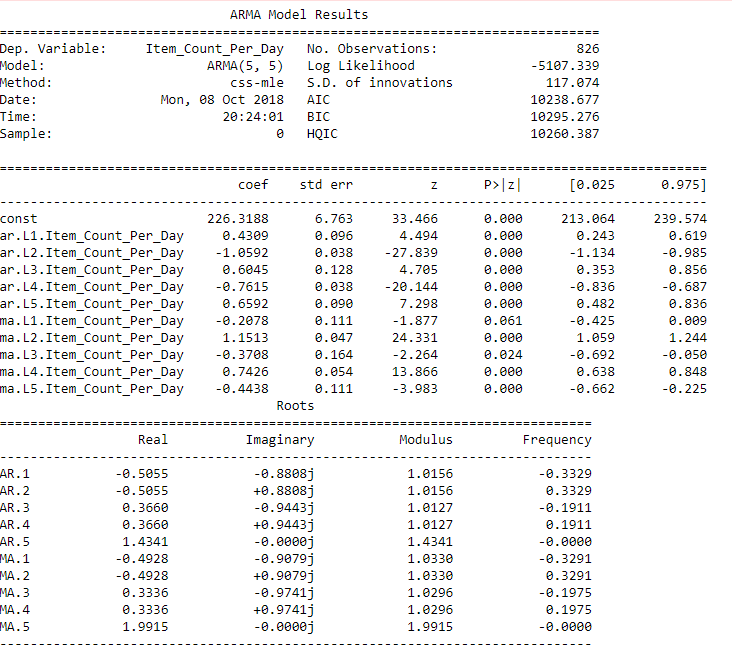
**Figure:** ACF plot

1. **Partial Auto-Correlation Function Plot (PACF):**

****

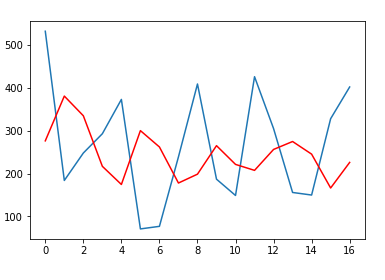
**Figure:** PACF plot

1. **Fitting ARIMA model:**



### **Figure:** Fit Arima to the time series

# Forecast:

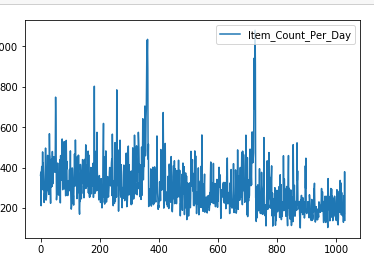


**Figure : Forecast for Kimki**

### **SEMONOVISKY**

### 

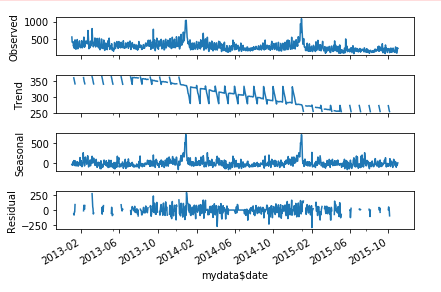
1. **Time Series**



### **Figure:** Time series graph

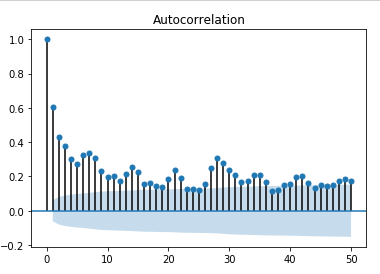
### **Decomposition of time-series:**

### 



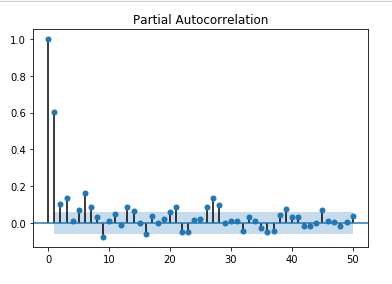
### **Figure: Decomposition of time series**

1. **Auto-Correlation Function Plot (ACF):**

****

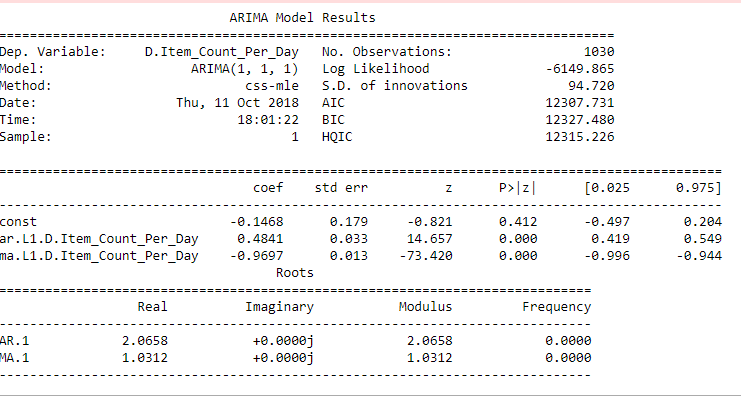
**Figure:** ACF plot

1. **Partial Auto-Correlation Function Plot (PACF):**

****

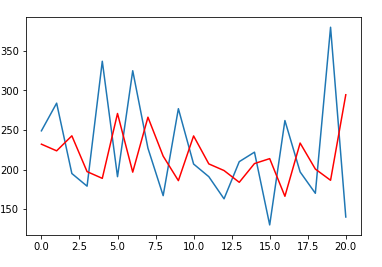
**Figure:** PACF plot

1. **Fitting ARIMA model:**



### **Figure:** Fit Arima to the time series

# Forecast:

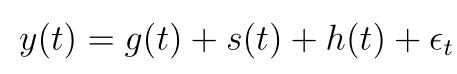


**Figure: Forecast for Semenivisky Store**

**5.2 PROPHET**

Prophet is an open source library published by Facebook that is based on **decomposable (trend + seasonality + holidays) models.** It provides us with the ability to make time series predictions with good accuracy using simple intuitive parameters and has support for including impact of custom seasonality and holidays!

Prophet is a decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in the following equation:



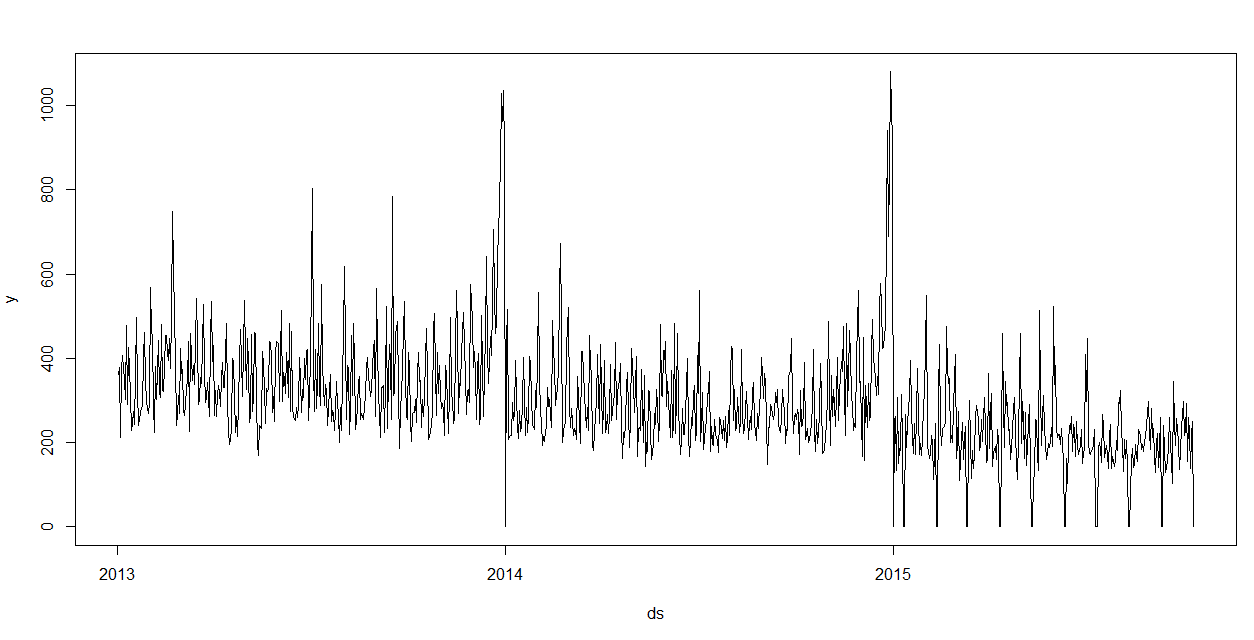
* **g(t)**: piecewise linear or logistic growth curve for modelling non-periodic changes in time series
* **s(t)**: periodic changes (e.g. weekly/yearly seasonality)
* **h(t)**: effects of holidays (user provided) with irregular schedules
* **εt**: error term accounts for any unusual changes not accommodated by the model

Using time as a regressor, Prophet is trying to fit several linear and non linear functions of time as components. Modeling seasonality as an additive component is the same approach taken by exponential smoothing in [Holt-Winters technique](https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/) . We are, in effect, framing the forecasting problem as a curve-fitting exercise rather than looking explicitly at the time based dependence of each observation within a time series.

### **SEMONOVISKY**

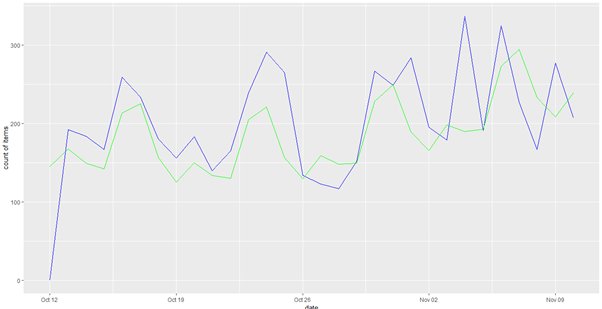
1. **Time Series**

We have use FB prophet model of forecasting for forecasting sales in store semonovisky in R programming environment.



**Figure : Time Series Graph**

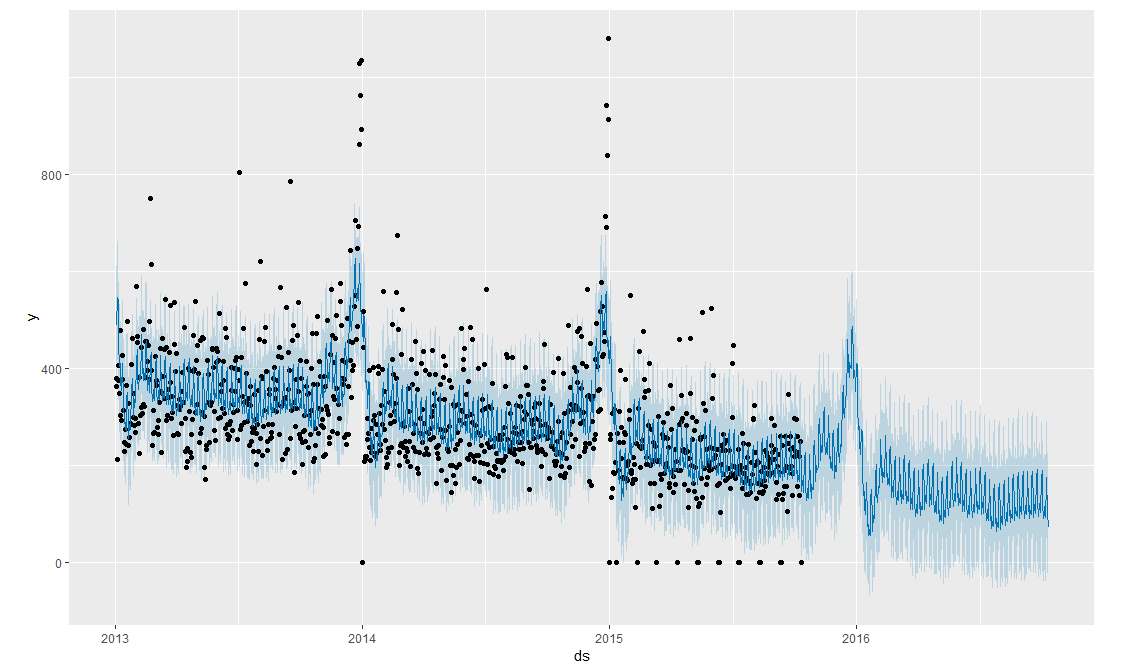
1. **Forecast on test Data**



**Figure : Forecast For Test Data**

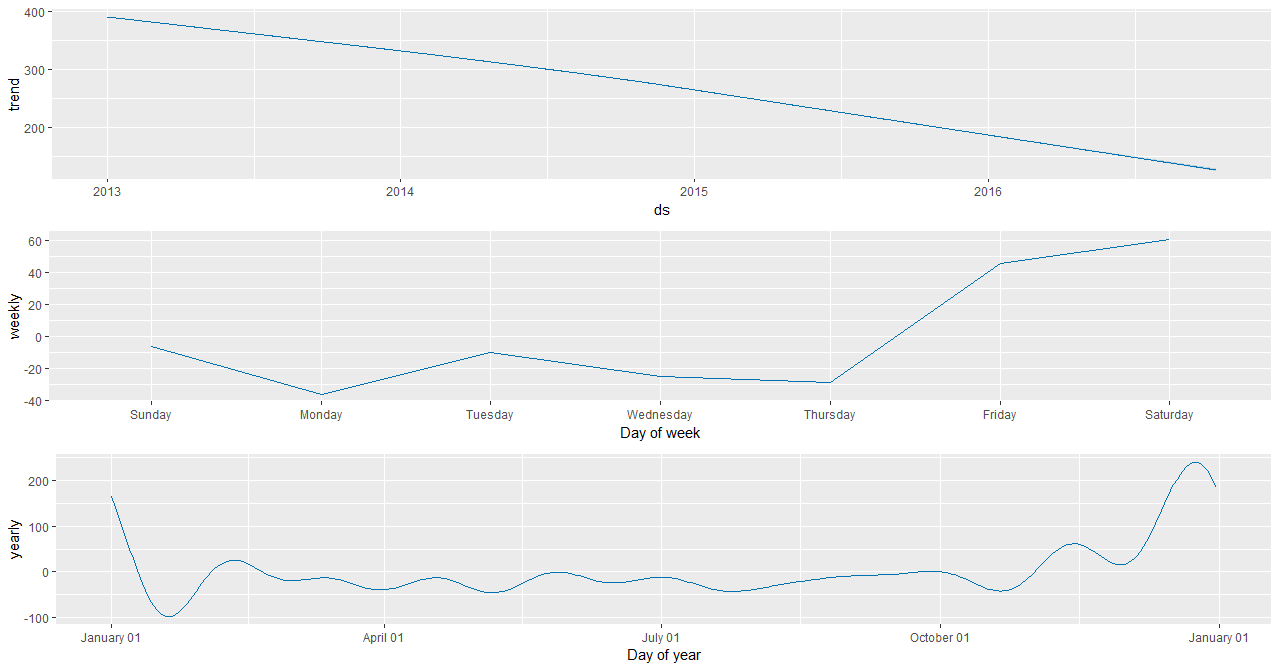
|  |
| --- |
|  |

1. **Forecast using Fb Prophet for 365 Days:**



**Figure : Forecast using Fb Prophet for 365 Days**

1. **Forecast Components:**



1. **Model Evaluation of Prophet:**

|  |
| --- |
|  |

**Observations:**

* From the trend graph, Prophet did a good job by fitting the trend. The Sales show a decreasing trend
* The graph of weekly seasonality leads to the conclusion that usually there are less sales on Mondays, Wednesdays, and Thursdays than on the other days of the week.
* In the yearly seasonality graph, there is a prominent dip after New Year vacation. But Sales are high towards the last quarter of the year

## REFERNCES

* G.E.P. Box, G. Jenkins, “Time Series Analysis, Forecasting and Control”, HoldenDay, San Francisco, CA, 1970.
* C. Chatfield, “Model uncertainty and forecast accuracy”, J. Forecasting 15 (1996)  G.P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model”, Neuro computing 50 (2003)
* J. Lee, “Univariate time series modelling and forecasting (Box-Jenkins Method)”, Econ 413
* https://machinelearningmastery.com/time-series-forecasting/
* <https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/>
* <https://rpubs.com/riazakhan94/arima_with_example>
* <https://robjhyndman.com/hyndsight/dailydata/>
* https://machinelearningmastery.com/time-series-forecasting-performance-measures-with-python/
* https://www.analyticsvidhya.com/blog/2018/02/timhttps://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/e-series-forecasting-methods/