Forecasting Total Sales & EDA

By : Moch Ardhi Kurniawan

Table of Content

OBJECTIVE & DATA UNDERSTANDING

Objective

What Is Forecasting?
Forecasting is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends.

Why Forecasting ?
Forecasting is valuable because it gives the ability to make informed business decisions and develop data-driven strategies, also helping to eliminate potential failures or losses before they happen.

Data Understanding

Date <date>	Day <int>	Month <chr>	Units <dbl>	Product_Name <chr>	Product_Category <chr>
2017-01-01	1	January	1	Chutes & Ladders	Games
2017-01-01	1	January	1	Action Figure	Toys
2017-01-01	1	January	1	Deck Of Cards	Games

Product_Cost <dbl>	Product_Price <dbl>	Total_Cost <dbl>	Total_Price <dbl>
9.99	12.99	9.99	12.99
9.99	15.99	9.99	15.99
3.99	6.99	3.99	6.99

Store_Name <chr>	Store_City <chr>	Store_Location <chr>
Maven Toys Aguascalientes 1	Aguascalientes	Downtown
Maven Toys Puebla 2	Puebla	Downtown
Maven Toys Mexicali 1	Mexicali	Commercial

This data is taken from Kaggle Mexico Toy Sales

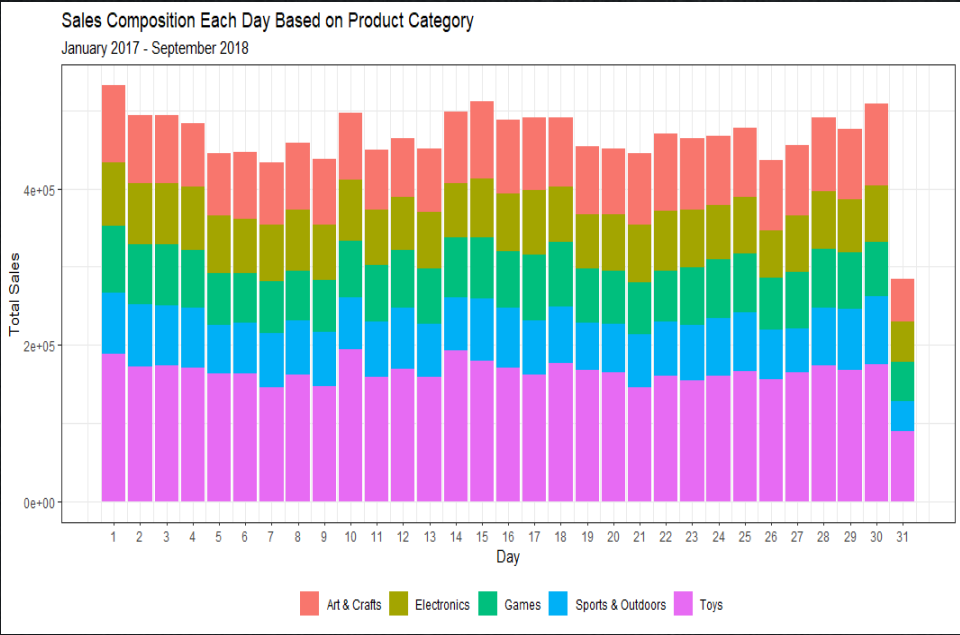
The data that I take consist of 3 files: products ; sales ; stores

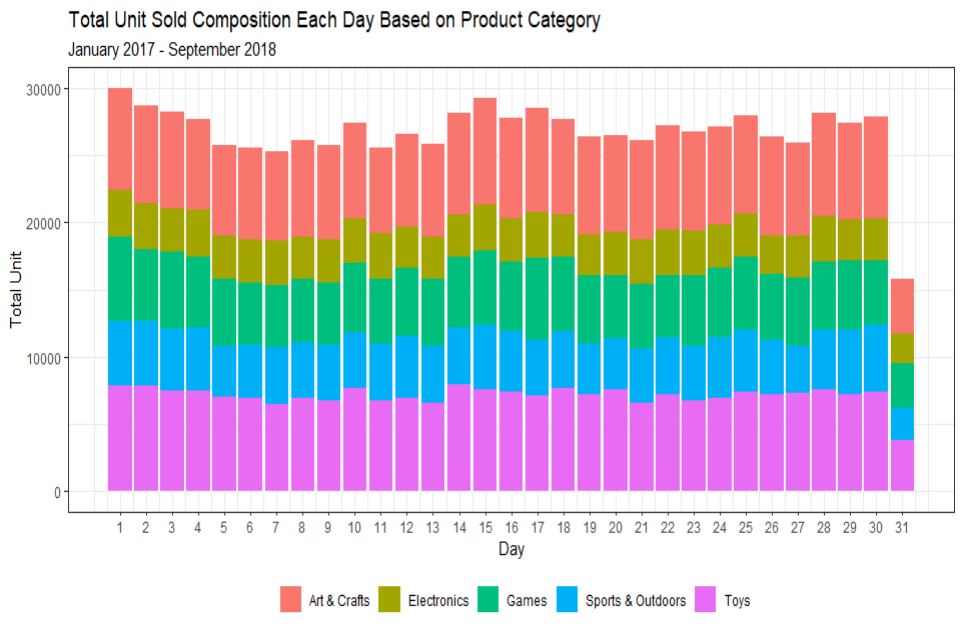
This is daily sales data from 01 January 2017 – 30 September 2018

After some combine and processing, the data consists of 829,262 rows and 13 columns

EXPLORATORY DATA ANALYSIS (EDA)

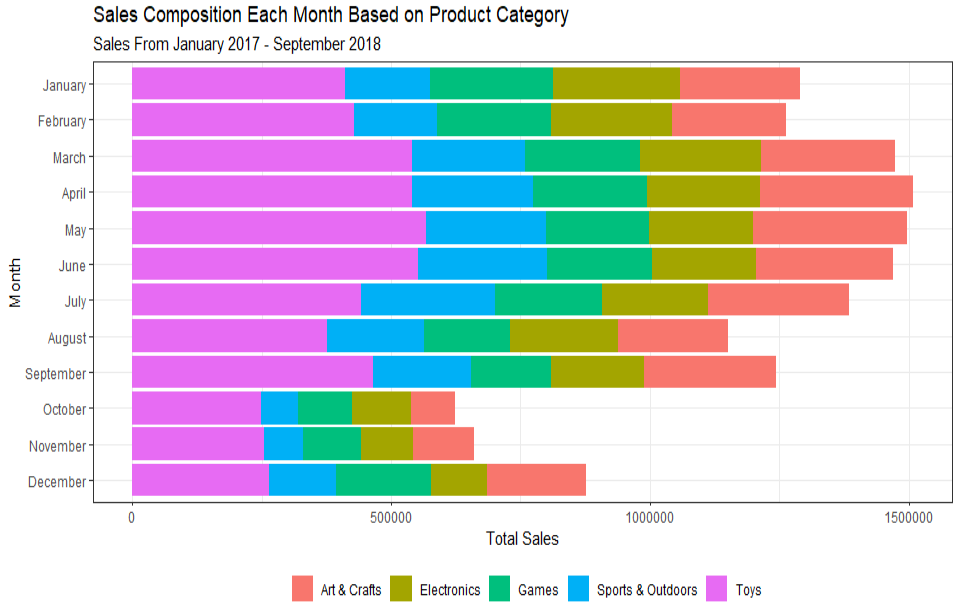
Exploratory Data Analysis (EDA)

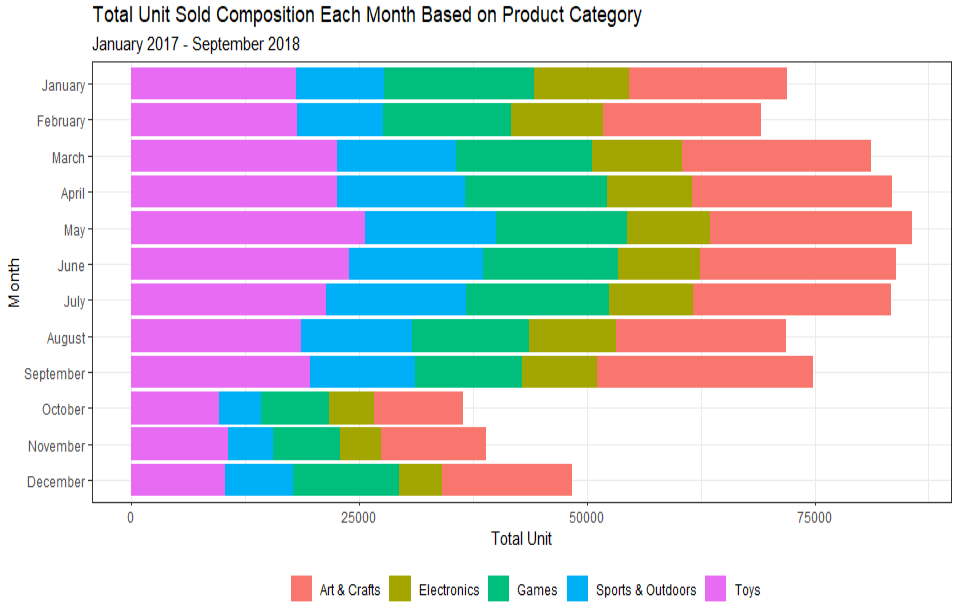




1. In every date 1 and 15 are the highest
2. Highest sales category is toys and the lowest is sports&outdoors
3. Highest unit sold category is art&crafts and the lowest is electronics

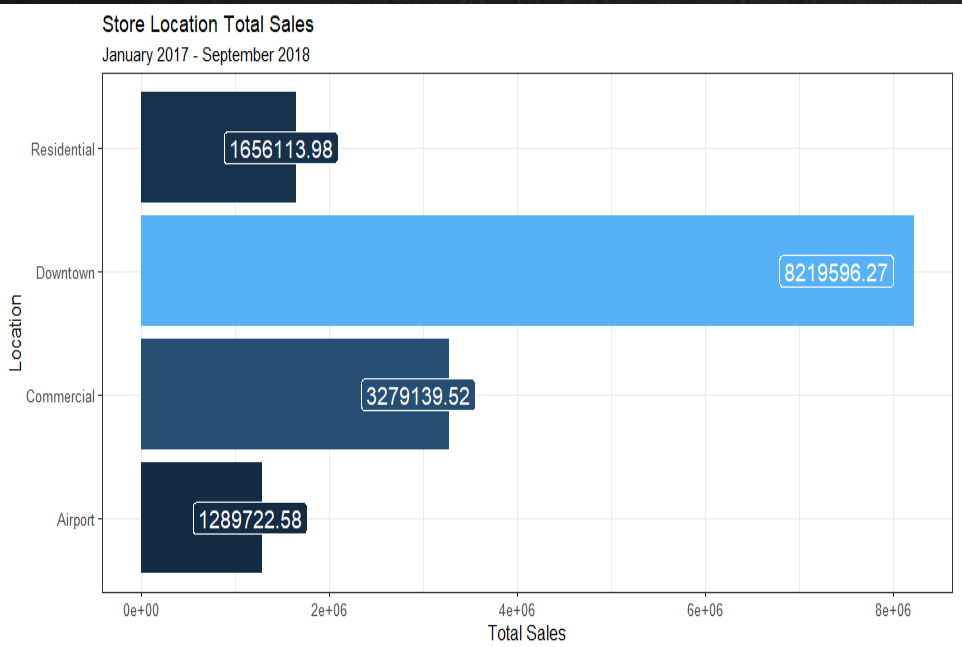
Exploratory Data Analysis (EDA)

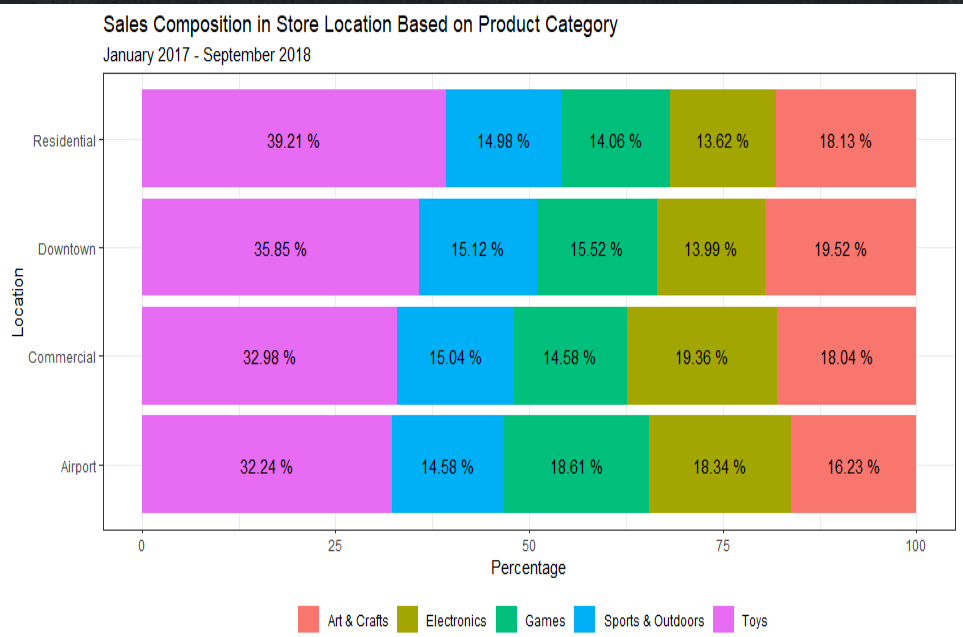




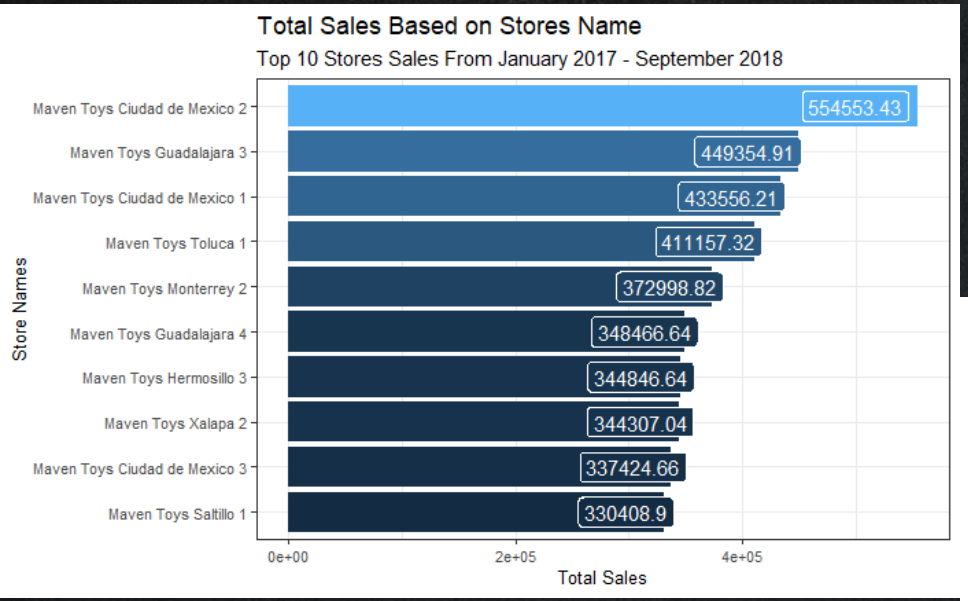
1. April is the highest is sales, May is the highest is unit sold
2. Q2 is the highest in both sales and unit sold
3. Category of sales and unit sold same with the day analysis

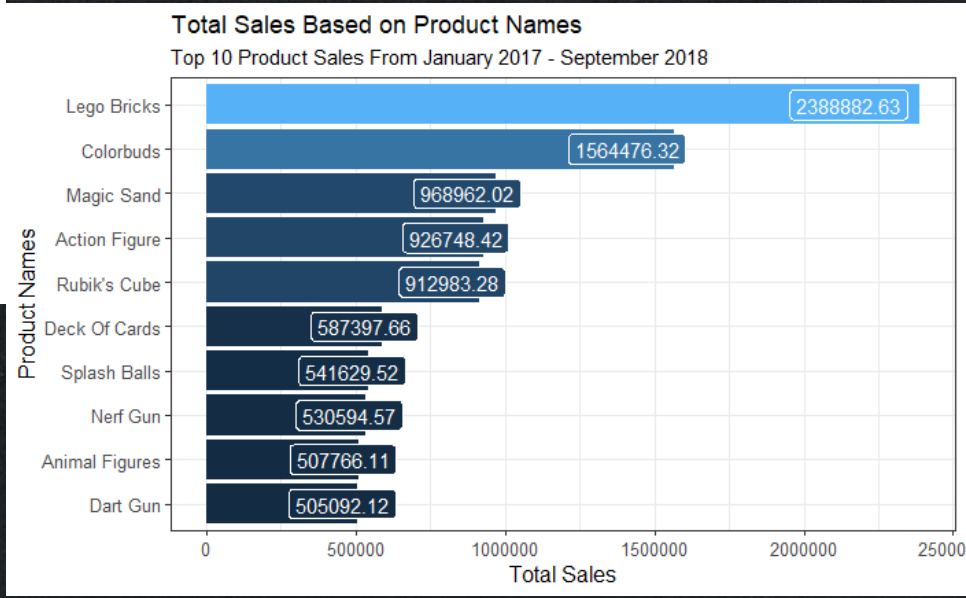
Exploratory Data Analysis (EDA)



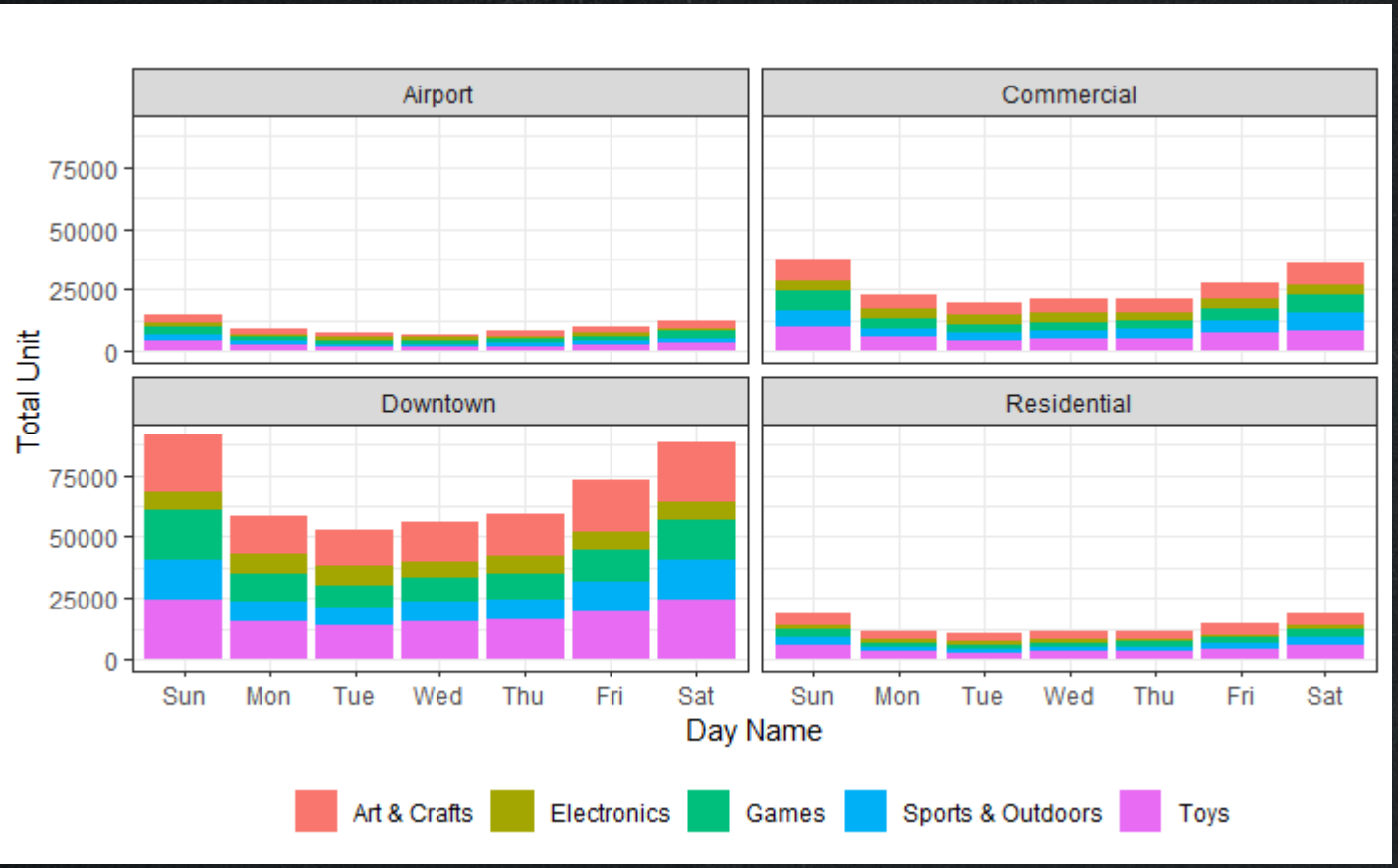


Exploratory Data Analysis (EDA)





Exploratory Data Analysis (EDA)



1

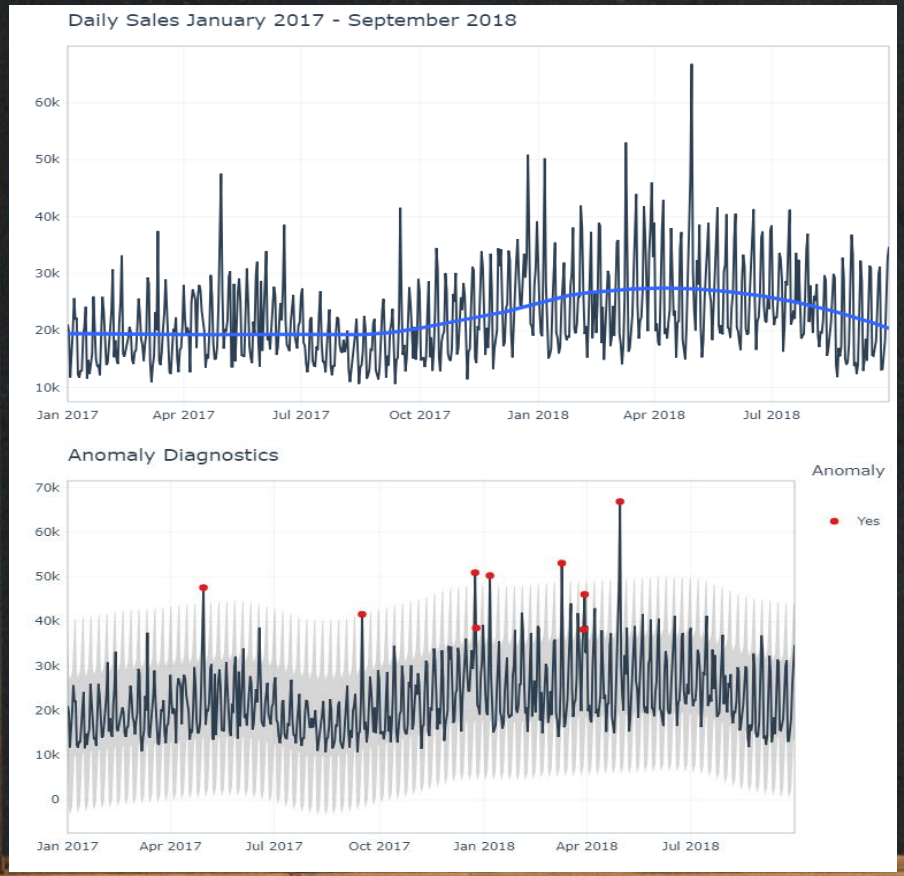
Saturday and Sunday are the highest in every store location

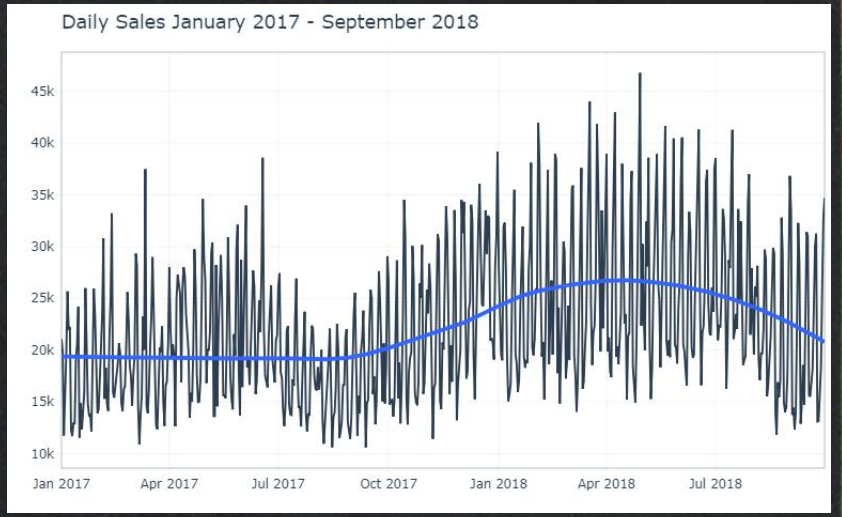
2

Every store location has same pattern, After sunday decrease then slowly increase

DATA PREPROCESSING

Data Preprocessing



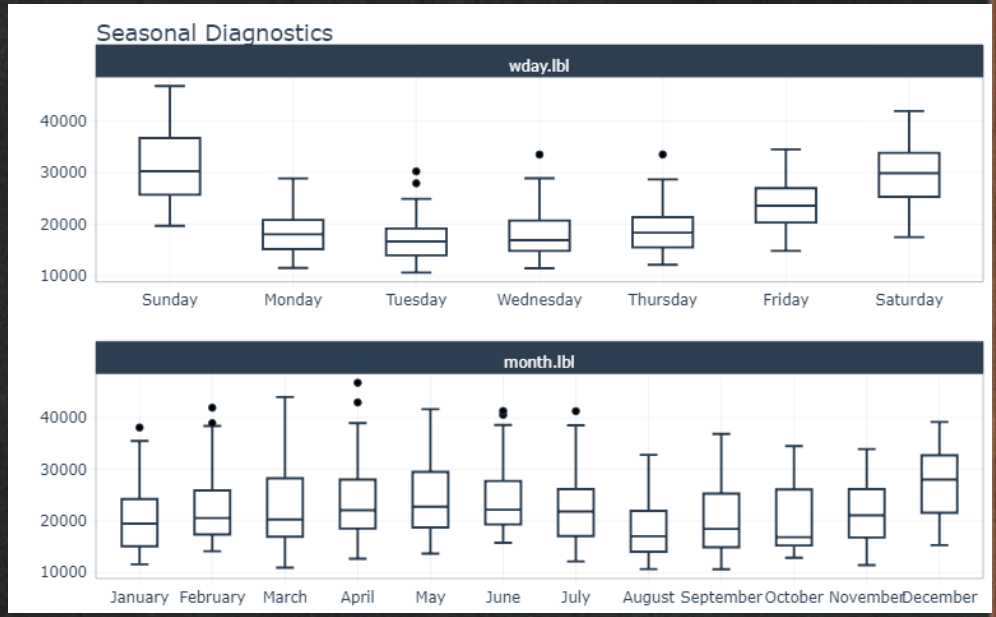


I need to remove the anomaly to reducing forecasting error. There is some seasonality in the data.

Data Preprocessing

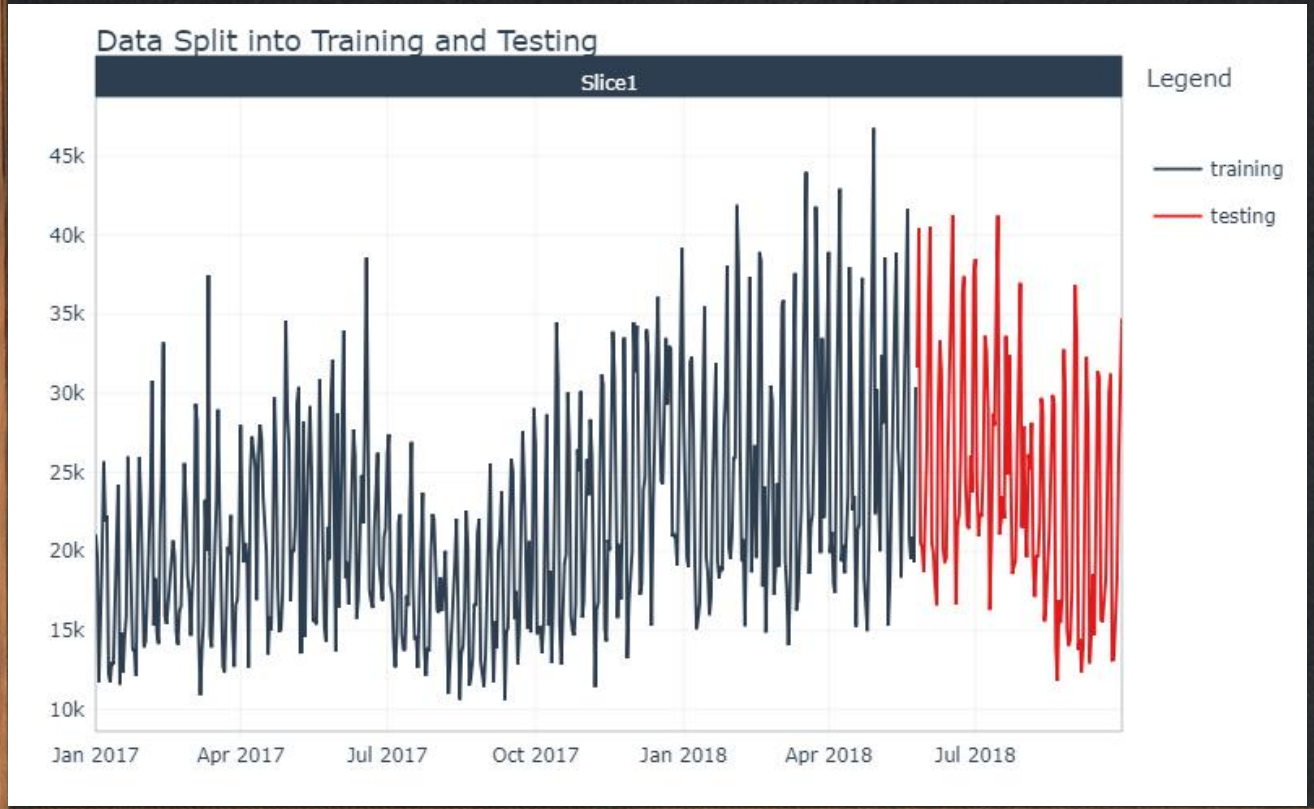
```
        Augmented Dickey-Fuller Test

data:  sales_daily_test
Dickey-Fuller = -4.1858, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary


        Phillips-Perron Unit Root Test

data:  sales_daily_test
Dickey-Fuller Z(alpha) = -214.28, Truncation lag parameter = 6, p-value =
0.01
alternative hypothesis: stationary
```

Since p-value of adf and pp test less than 0.05,
data is stationary, so we can continue to use it



The data has seasonality In Sunday and Saturday. After Sunday, the sales decreasing, and slowly increase after Tuesday.

Data Preprocessing



The data split into 80% Training ; 20% Testing

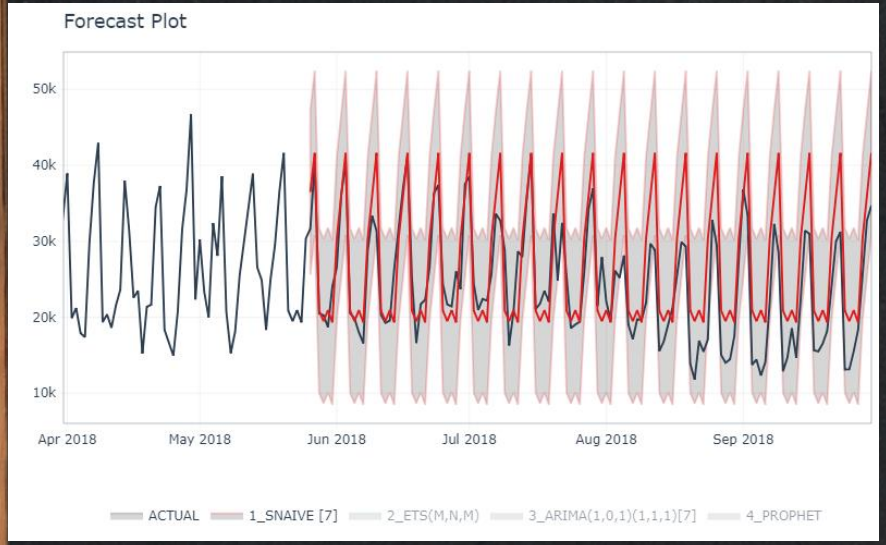
MACHINE LEARNING MODELLING

Machine Learning Modelling

.model_desc <chr>	.type <chr>	mae <dbl>	mape <dbl>	mase <dbl>	smape <dbl>	rmse <dbl>	rsq <dbl>
SNAIVE [7]	Test	4348.039	19.79304	0.7778153	17.56274	5522.421	0.7003501
ETS(M,N,M)	Test	3869.720	19.27055	0.6922496	16.75440	4689.019	0.7213572
ARIMA(1,0,1)(1,1,1)[7]	Test	3964.644	19.19044	0.7092303	16.76355	4916.771	0.7120835
PROPHET	Test	3171.751	15.05140	0.5673907	14.04289	3953.066	0.7333770

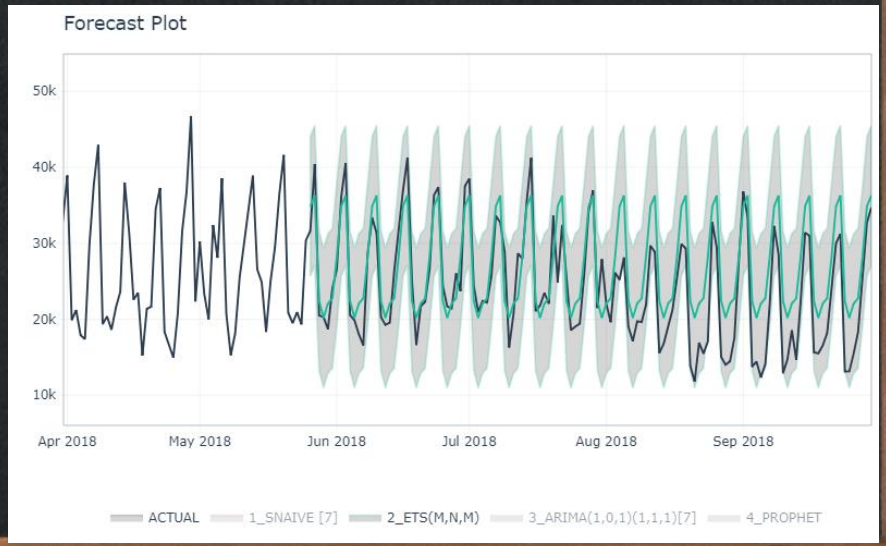
This is the metric used to evaluate the model's performance.The common way is to look at the MAPE (Mean Absolute Percentage Error). The lower the MAPE, the better.

Machine Learning Modelling

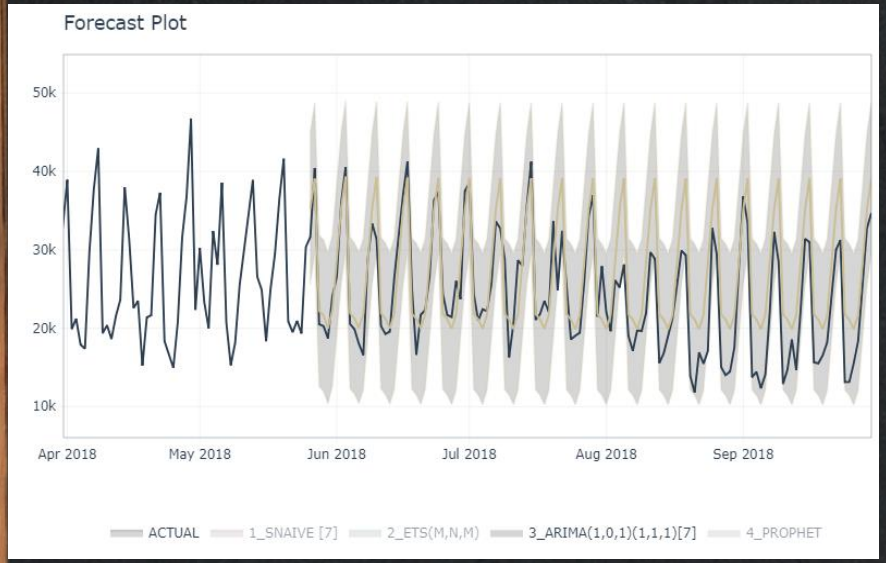


SNAIVE: This is a bad model, because the predicted value in the red colour exceeds the actual value in the black colour, and the shade colour is the interval of the predicted value

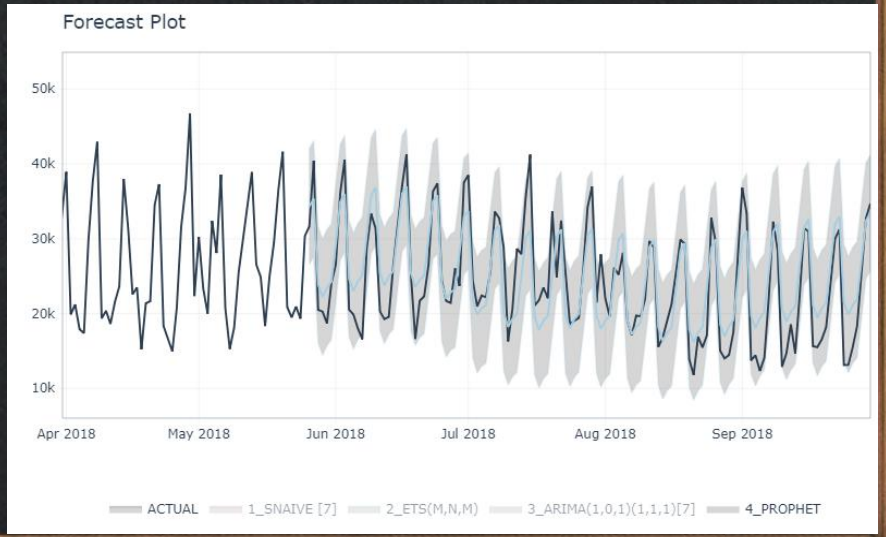
ETS: The predicted value is slightly better than the SNAIVE, but there isn't trend in the interval (shade colour)



Machine Learning Modelling



ARIMA: Same with the ETS model, but the predicted value is slightly better, There are several values that are the same in both the predicted and actual values in the beginning



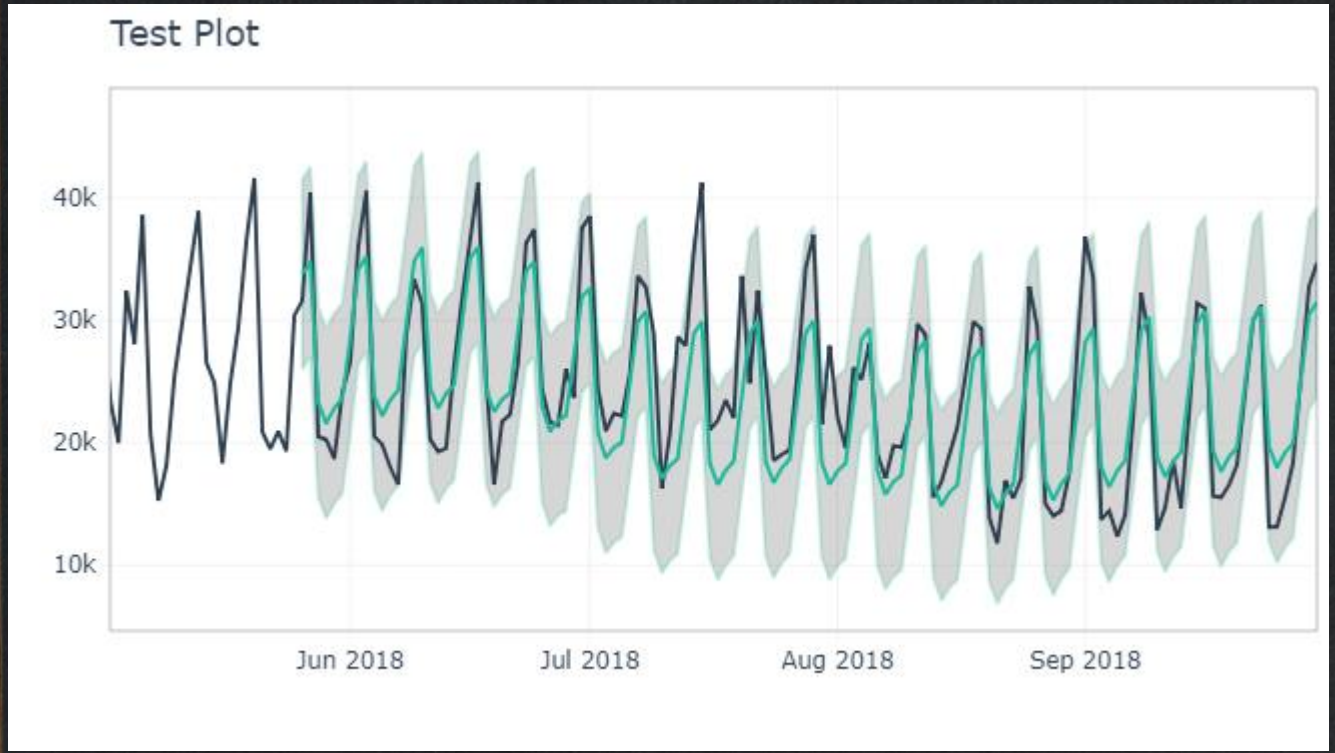
PROPHET: best model out of all of them. The predicted and the interval

Machine Learning Modelling

.model_desc <chr>	.type <chr>	mae <dbl>	mape <dbl>	mase <dbl>	smape <dbl>	rmse <dbl>	rsq <dbl>
PROPHET	Test	3171.751	15.05140	0.5673907	14.04289	3953.066	0.7333770
PROPHET	Test	3184.847	13.97567	0.5697333	13.92089	3956.564	0.7374106

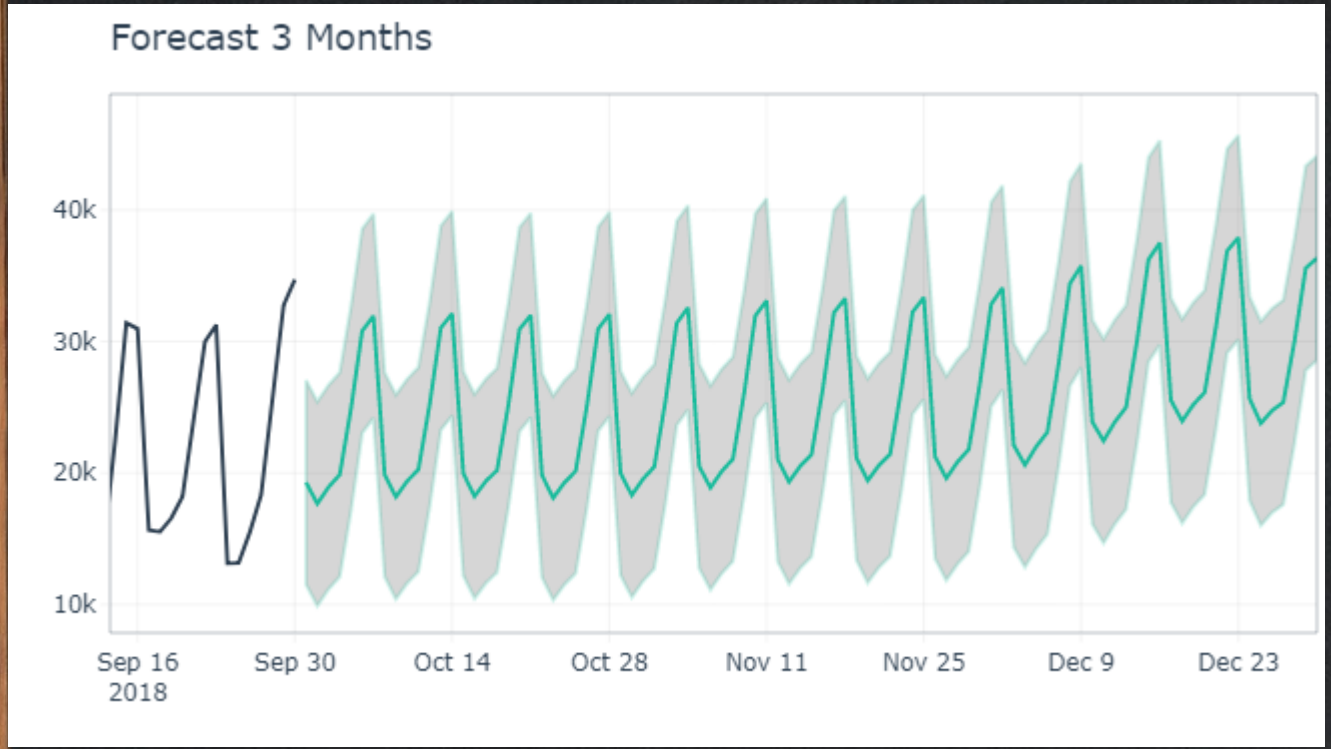
After some little change on its features of Prophet model, the MAPE is decrease

Machine Learning Modelling



The predicted value can match the actual value and the interval range is not too far

Machine Learning Modelling



The forecasting it shows that the trend is increasing until the end of the year

RESULT SUMMARY

Result Summary

- We can see from the forecast for the next 3 months that total sales are increasing and there is seasonality in Saturday and Sunday, and we can make some promotions on these days
- We can make some promotions based on its category on April or May
- In every day 1 and 15 our sales and unit sold are the highest we can make something in this day
- We can combine the day and the month for promos to attract new customers
- We can give our attention on airport and residential locations to make our sales more higher

Thank You


085156296689


mochardhik@gmail.com


www.linkedin.com/in/mochardhikurniawan


https://github.com/MArdhiK