Retail Analytics

1. Introduction and background of the project.

Retail industry is a vast and major sector of the economy. In today's world, many businesses are starting retail stores. It is important for the stores to run successfully and control the loss of profit. Many businesses and retailers are still using the old forecasting techniques and gut instinct.[Rod Daugherty,2019]. Retailers need to practice predictive analysis to improve the growth of sales, understand the product management, optimize the supplies. This way the business can operate dynamic and have competitive business. According to research companies with sales forecast are 10% more likely to grow their revenue year over year.

A toy store company named Maven toys located in Mexico wants to manage inventory efficiently to improve sales. To achieve this, I will be using machine learning techniques on the company data and study their products and sales. The dataset is provided by Kaggle. The link to the dataset is – Mexico Toy Sales | Kaggle. The dataset contains inventory, products, sales, and stores data. The inventory dataset has store ID, productId, stocks_on_hand. Product dataset has Product Id, product_name, product_category, product_cost, product_price. Sales dataset has sale_id, date. Product_id, product_cost, Product_price. Store dataset has store_id, store_name, store_city, store_location, store_open_date. The data has the sales record for the year 2017 and 2018 sep. This project is important as I get to analysis the store sales, inventory, and their profits. I expect to learn the way to perform exploratory data analysis, feature selection and data modelling. It is important to know which features are required for the data model prediction and their algorithms. Apply the machine learning models like regression, random forest, time series analysis , helps me to improve my skills and make me expertise as data scientist. There are many machine learning techniques provided for retail analytics however many retail companies are not using it yet. My work would help to find the more accurate and consistent model in the analysis.

2. Statement of the project problem.

In retail business, goal is to predict their company sales which helps to predict the future growth and help to analyze their performance and revenue. This analysis will also help the company to optimize the inventory to meet the demand and supply of products. Retail stores can maximize the expected profit and decrease the storage space required for inventory.

3. Describe related work such as publications and systems

Paper1: How To Apply Machine Learning To Demand Forecasting

Reference: Liudmyla Taranenko **,** Machine Learning To Demand Forecasting, Jul 7, 2021

Summary:

1.1 Purposes of the Study: This article helps to build demand forecasting models for retail products and helps to grow their sales. The focus here is to apply the models for both stable environment and crisis.

1.2 **Research Design or Strategy:** Demand forecasting is to predict the demand of the products and help the store to load the required inventory. Here the article focuses on supplier relationship management which helps to understand the need of new supply chains or reduce the suppliers, customer relationship management is to gain the customer satisfaction by provide the product on time , order fulfillment and logistics is to order the required product and save the space by not ordering the products not in demand , Marketing campaigns is to adjust the ads and influence the products, Manufacturing flow management is to predict how many goods will be sold.

Demand forecasting predictions is now using the upgraded Machine learning techniques. The models are ARIMA/SARIMA , Linear regression, XGBoost, K-Nearest Neighbors Regression,Random forest, long short-term memory(LSTM). The data is evaluated based on Consistency, Accuracy, Validity, Relevance, Accessibility, Completeness, Detalization. The forecast model differs for small company , large company , offiline and online business. The model needs to be reframed accordingly.

1.3 Conclusions: This article helped with the concepts of demand forecasting, supplier, customer relationship and marketing campaigns.

1.4 **Contributions:** Demand forecasting in retail field products.

<u>Section 2: CRITICAL ANALYSIS</u>

2.1 Overall Assessment: This article helps me to understand the requirements of demand forecasting and the overall business flow and machine learning techniques which are applied to build.

2.2 Research Methodology: The article well stated covering each topic in detail.

2.3 Future Research: The article can further analyze the models for pandemic situation in detail.

2.4 New Knowledge Learned: Yes, learnt the overall flow of demand forecasting**.**

Paper2: 10 Ways Retail Predictive Analytics Can Transform Your Business

Reference: Rohini Mandiwal , 10 Ways Retail Predictive Analytics Can Transform Your Business

1.1 Purposes of the Study: This article provides the 10 ways for retailers to practice predictive analysis to improve the growth of sales, understand the product management, optimize the supplies. This way the business can operate dynamic and have competitive business.

1.2 **Research Design or Strategy**: Understand Customer Behavior: The customers are interactive about their brands, products and store and collections in social media and websites. Hence using this data to build predictive models helps to increase revenue, campaign ads, improve products and gain high-value customers, strategizing to up-sell and cross-sell, and minimizing acquisition cost. Improve Customer Service and Experience: The customer satisfaction is very required to improve the customer's service. The customers express their opinions by rating and review, The retails can gather these data and also by sensors, cameras used to understand in-store behavior and improving selling strategies. Enhance Customer Segmentation**:** The customers are targeted by analyzing there online and offline behavior. By this analysis retail can created target-based campaigns and create customer's lifetime value and reduce customer churn rate. Enhance Inventory Management: Inventory analysis enables to achieve the demand supply and increase sales. Unavailability of product can impact sales and customer loss. Insights for Store Expansion: Retailers can use their current store data and predictive the new store location based on customer reach, product demand, potential revenue. Optimize Trade Promotions: consumer data helps to predict trade promotions and can improve the ROI. Improve Pricing: Pricing needs to alter and predicted based on the weather, inventory load, products selling and real time sales. **Forecast Revenue**: To forecast revenue, predictive analytics can be used on special events, holidays, customer purchasing trend. Give Personalized Recommendations: In this e-commerce life, retailers should recommend the products based on customer preferences and provide similar ads . Enhance Marketing Campaign Targeting: Predictive analytics needs to find to optimize targeting and find new market places, reduce expenses.

1.3 **Conclusions:** This article covered the essentials ways to accomplish the analysis for retail business improvement.

1.4 Contributions: Predictive analysis required in retail business.

Section 2: CRITICAL ANALYSIS

2.1 Overall Assessment: This article provided the best ways to apply predictive analytics for retailers

2.2 Research Methodology: The paper clearly mentioned the topics and ways to help business.

2.3 Future Research: Researchers can dig detail into each way and provide the use case.

2.4 New Knowledge Learned: Learnt a lot ways in retail predictive analysis.

Paper3: How Data Can Optimize Inventory Forecasting

Reference: Rod Daugherty, How Data Can Optimize Inventory Forecasting , August 6, 2019

1.1 Purposes of the Study: This article states that many businesses and retailers are still using the old forecasting techniques and gut instinct. He explains how the advance machine learning techniques can reduce the forecasting errors and improve the inventory and sales.

1.2 Research Design or Strategy: A survey showed Demand forecasting is top challenge face by the retailers. By this business are losing their revenue by buying excess stocks and not meeting the sales. This loss of business can be saved by used the modern machine learning techniques. The data information like sales, product, inventory history, customer details can help to predict results with more accuracy. The paper further discusses about the tools need to use for managing seasonal or slow-moving demand, promotion related changes, demand by location and to understand how price changes affect sales. Managing seasonal or slow-moving products is challenging, by using advanced data techniques like top-down forecasting, sales can be predicted more accurately. The promotion related changes have three consequences i.e., halo – increase in demand for non-promoted product, cannibalization – demand reduced for items similar to promoted items, post-promotion dip- sales decreased after promotion as customer have stocked up.

1.3 Conclusions: This paper mentioned the importance of using model tools and techniques for business.

1.4 **Contibutions**: Brief introduction of required data variables and machine learning techniques.

Section 2: CRITICAL ANALYSIS

2.1 Overall Assessment: It is good to read to know about how the companies are focusing on machine learning technologies.

2.2 Research Methodology: This article mentions the importance of using the modern machine learning technologies in companies.

2.3 Future Research: Research to be carried out about the techniques and there accuracy.

2.4 New Knowledge Learned: how the companies are focusing on machine learning technologies.

Paper4: Retail Analytics

Reference: James Collins, Lee Gates, Bulent Kasman, Vicky Nguyen, Mark Taylor, John Yamartino : Retail Analytics

http://scet.berkeley.edu/wp-content/uploads/UCBSCETRetailAnalyticsReport.pdf

Summary: This article describes briefly about the retail analytics and why we need it and how the companies are using it.

1.1 Research Design or Strategy: This article states retail analytics solutions and approaches to build their business. The approaches are categorized into categorical model which covers the 5 categories of company ecosystem and visual landscape and inventory for retail analytics. The retail model can be divided into forecast for ordering, stocking, selling, pricing. The retail analytics can be divided into strategy and planning, store operations, marketing, supply chain management, and merchandising. Machine learning techniques and tools help to analyze the business goals such as supplier management, revenue growth, inventory optimization, customer management. This strategy sets the success of the business.
The paper also mentions the company analysis who has performed well using retail analytics. The information was collected from Gartner paper about top 6 company who performed well and listed the company rankings based on business model categories.

1.2 Conclusions: This article mentions the retail analytics and it is used to build the business model of a company.

1.3 Contributions: Retail analytics, business model and companies using these strategies.

Section 2: CRITICAL ANALYSIS

2.1 Overall Assessment: Useful article provide with overall business knowledge.

2.2 Research Methodology: This article is high level write up of companies using the defined business model

2.3 Future Research: Research can be further carried with more companies and provide the analysis

2.4 New Knowledge Learned: Learned about the companies and models.

Paper 5: Maturing supply chain analytics for optimal inventory management

Summary:

Reference: Consultancy.eu, Maturing supply chain analytics for optimal inventory management, 19 July 2021

1.1 **Research Design or Strategy**: This article mentions how the descriptive, predictive, and prescriptive analytics can help the supply chain to manage the inventory and business. Descriptive explains how the inventory levels are working and tries to answer the question like is inventory overloaded, how to manage it effectively. Predictive helps to forecasts the issues and provide accurate results. By following the forecast, business can improve their sales.

1.2 Conclusions: This article briefly describes the all the techniques step by step.

Section 2: CRITICAL ANALYSIS

2.1 Overall Assessment: Good article to learn about all the analytics used.

2.2 Research Methodology: It was easy to understand the research**.**

2.3 Future Research: To focus on more about prescriptive analytics**.**

2.4 New Knowledge Learned: Learned about descriptive and prescriptive analytics.

4. **Objectives of the study.**

The objective is to estimate the future sales of the retail store which helps to plan and prepare to have a logistics and inventory ready.

5. **Data Collection**

The dataset is provided by Kaggle. The link to the dataset is – Mexico Toy Sales | Kaggle.

6. **Research design and methodology.**

In today’s world, many businesses are starting retail stores. For the stores to run successfully and control the loss of profit is important. I have chosen to research this area and collected the data from Kaggle. Performing exploratory data analysis and feature selection, I would predict the data by applying data model linear regression, random forest, and ARIMA model. I will be calculating the absolute error and percentage error to calculate the accuracy. I will use python language to perform my analysis.

7. Data analysis and result

Linear regression is the classic statistical way to predict an output. It tries to find the best fit line between dependent and other independent variables with the least squared error. Here in my dataset the dependent variable is sales and independent variables are costs and units, based on this I have developed linear regression model.

```
#############Importing packages##############
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn import metrics
```

```
x = sales_weekly[['Units','cost']]
```

```
y = sales_weekly['sales']
```

```
x_train1, x_test1, y_train1, y_test1 = train_test_split(x, y, test_size = 0.2)
```

```
model1 = LinearRegression()
```

```
model1.fit(x_train1, y_train1)
```

```
LinearRegression()
```

```
print(model1.coef_)
```

```
[2.8235433  1.07522447]
```

```
print(model1.intercept_)
```

```
2.820248966837994
```

```
pd.DataFrame(model1.coef_, x.columns, columns = ['Coeff'])
```

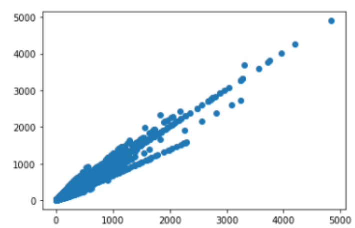
	Coeff
Units	2.823543
cost	1.075224

```
model1.score(x_train1, y_train1)
```

```
0.9559412261276342
```

```
plt.scatter(y_test1, y_pred1 )
```

```
<matplotlib.collections.PathCollection at 0x1e783bed5b0>
```



```
plt.hist(y_test1 - y_pred1)
```

```
(array([3.000e+00, 1.800e+01, 1.660e+02, 3.672e+03, 6.788e+03, 3.700e+02,
        7.700e+01, 3.800e+01, 1.300e+01, 8.000e+00]),
 array([-495.0113689 , -374.63750055, -254.26363221, -133.88976386,
         -13.51589552,  106.85797283,  227.23184117,  347.60570952,
         467.97957786,  588.35344621,  708.72731455]),
 <a list of 10 Patch objects>)
```

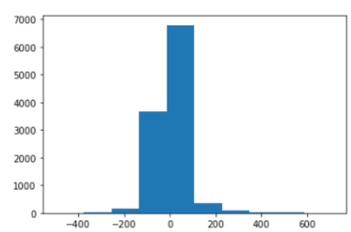


Fig 1a **Fig 1b**

```
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test1, y_pred1))
print('Mean Absolute Error:', metrics.mean_squared_error(y_test1, y_pred1))
print("The mean absolute % error is {}%".format(round(np.mean(np.abs(y_pred1- y_test1.values)/np.abs(y_test1.values))*100,2)))
```

```
Mean Absolute Error: 37.324363729396865
Mean Absolute Error: 4495.3459376202445
The mean absolute % error is 19.55%
```

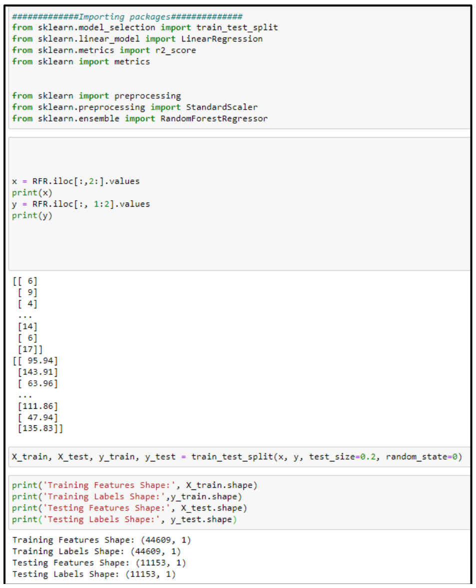
```
r2_score(df.Actual, df.Predicted)
```

```
0.9557761534070274
```

Fig 1c

The above model has the intercept 2.82 and r2_score is 0.95 which is good. However, when we check for assumptions of the linear regression model, we can see the residual have a patterns, that is variance of residuals grows with dependent variable. The autocorrelation would exist in a time series while it is 0 in Linear Regression. Hence, we do not use this model on this dataset.

Next, I have used Random Forest regressor model to perform the analysis. Here number of decision trees are made in which each tree is created from different bootstrap samples of training dataset. The dataset is divided into training and testing sets and data set is scaled by using standardscaler. The performance is measured mean absolute error, mean squared error, and root mean squared error. The model mean absolute error is 121.737 and mean baseline error is 255.91. The lesser mean absolute error than mean baseline error, the performance of model has better results.



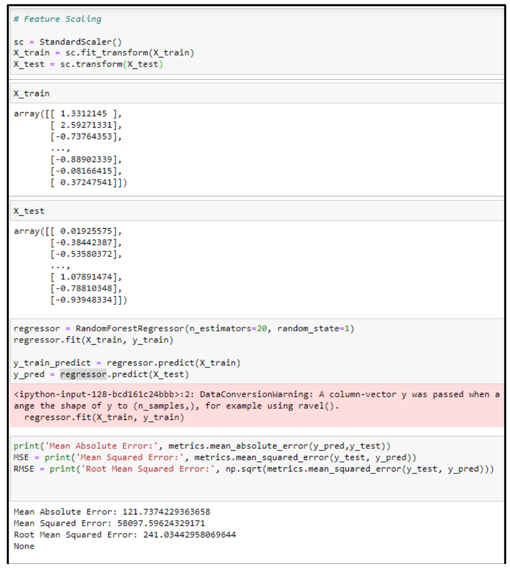
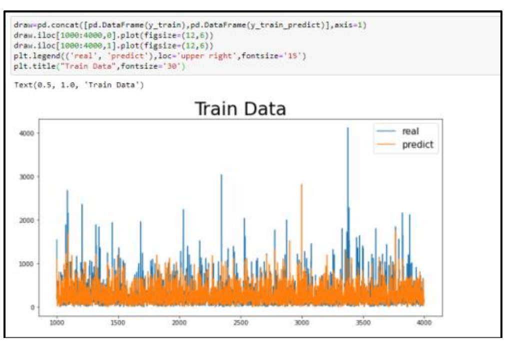


Fig 2a **Fig 2b**



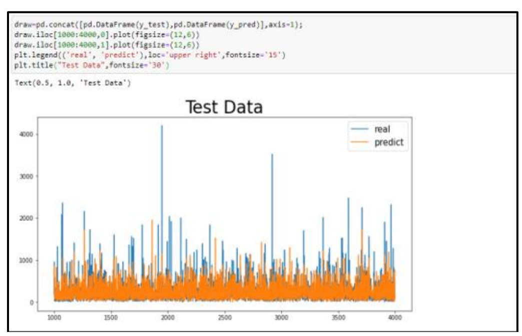


Fig 2c **Fig 2d**

Further I have used Time series analysis to continue with forecast and prediction of the goal. The important criteria to be followed is that time series formulation is stationary. Time series can have patterns like trend, seasonality, and cycles. I have used ARIMA (Auto-Regressive Integrated Moving Average) model, this takes lag terms and forecast errors in account for predicting.

Hypothesis to check if time series is stationary using dicky fuller test for significance level of 0.05

Null: The sales data is not stationary
Alternate: The sales data is stationary

```
from statsmodels.tsa.stattools import adfuller
from numpy import log
result = adfuller(weekly_df.sales)
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
```

```
ADF Statistic: -2.283518
p-value: 0.177339
```

Fig 3

In the above test, p value is greater then 0.05 and hence we reject Alternate hypothesis and sales data is not stationary and we need to difference the series. This is achieved by providing the difference parameter 1 in the build model. I have also checked the lag through correlation and given the value 5 for lag parameter in the model. Below shows the forecast output with mean error percent of 34.

```
# PACF plot of 1st differenced series
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
plt.rcParams.update({'figure.figsize':(9,3), 'figure.dpi':120})


# fig, axes = plt.subplots(1, 2, sharex=True)
# axes[0].plot(weekly_df.total_amt.diff()); axes[0].set_title('1st Differencing')
# # axes[1].set(ylim=(0,5))
plot_pacf(weekly_df.sales)
plt.xlabel('AR Lag')
plt.ylabel('Strength/Correlation')
plot_acf(weekly_df.sales)
plt.xlabel('MA Lag')
plt.ylabel('Strength/Correlation')


plt.show()
```

Fig 3a

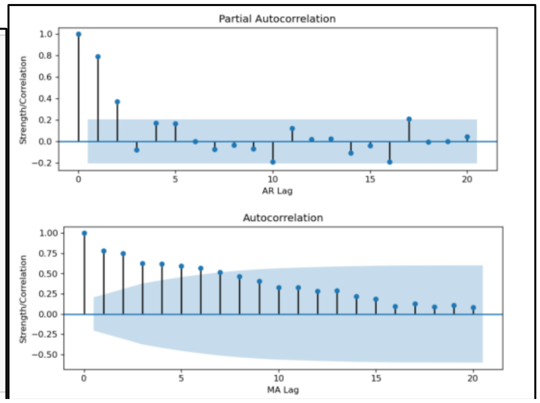


Fig 3b

```
from statsmodels.tsa.arima_model import ARIMA

# 1,1,2 ARIMA Model
model = ARIMA(weekly_df.sales, order=(5,1,0))
model_fit = model.fit(disp=0)
print(model_fit.summary())
```

```
C:\Users\Sruja\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:213: ValueWarning: An
ed and will be ignored when e.g. forecasting.
  warnings.warn('An unsupported index was provided and will be'
C:\Users\Sruja\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:213: ValueWarning: An
ed and will be ignored when e.g. forecasting.
  warnings.warn('An unsupported index was provided and will be'
                              ARIMA Model Results
==============================================================================
Dep. Variable:                D.sales   No. Observations:                   90
Model:                 ARIMA(5, 1, 0)   Log Likelihood                -998.905
Method:                       css-mle   S.D. of innovations          15952.324
Date:                Thu, 02 Dec 2021   AIC                           2011.810
Time:                        22:52:25   BIC                           2029.309
Sample:                             1   HQIC                          2018.867

================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const          249.5541    816.546      0.306      0.760   -1350.846    1849.954
ar.L1.D.sales   -0.4826      0.105     -4.580      0.000      -0.689      -0.276
ar.L2.D.sales   -0.1145      0.116     -0.988      0.323      -0.342       0.113
ar.L3.D.sales   -0.2889      0.112     -2.588      0.010      -0.508      -0.070
ar.L4.D.sales   -0.1807      0.116     -1.562      0.118      -0.407       0.046
ar.L5.D.sales   -0.0199      0.106     -0.188      0.850      -0.227       0.187
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            0.7268           -1.2613j            1.4558           -0.1668
AR.2            0.7268           +1.2613j            1.4558            0.1668
AR.3           -1.7161           -0.6283j            1.8275           -0.4441
AR.4           -1.7161           +0.6283j            1.8275            0.4441
AR.5           -7.1025           -0.0000j            7.1025           -0.5000
-----------------------------------------------------------------------------
```

Fig 3c

```
# Build Model

model = ARIMA(X_train, order=(5,1,0))
fitted = model.fit(disp=-1)
print(fitted.summary())

# Forecast
fc, se, conf = fitted.forecast(26, alpha=0.1)  # 95% conf

# Make as pandas series
fc_series = pd.Series(fc, index=X_test.index)
lower_series = pd.Series(conf[:, 0], index=X_test.index)
upper_series = pd.Series(conf[:, 1], index=X_test.index)

# Plot
plt.figure(figsize=(12,5), dpi=100)
plt.plot(X_train, label='training')
plt.plot(X_test, label='actual')
plt.plot(fc_series, label='forecast')
plt.fill_between(lower_series.index, lower_series, upper_series,
                 color='k', alpha=.1)
plt.title('Forecast vs Actuals')
plt.legend(loc='upper left', fontsize=8)
plt.show()
```

```
C:\Users\Sruja\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:213: ValueWarning: An
ed and will be ignored when e.g. forecasting.
  warnings.warn('An unsupported index was provided and will be'
C:\Users\Sruja\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:213: ValueWarning: An
ed and will be ignored when e.g. forecasting.
  warnings.warn('An unsupported index was provided and will be'
                              ARIMA Model Results
==============================================================================
Dep. Variable:                D.sales   No. Observations:                   64
Model:                 ARIMA(5, 1, 0)   Log Likelihood                -706.718
Method:                       css-mle   S.D. of innovations          15074.520
Date:                Thu, 02 Dec 2021   AIC                           1427.436
Time:                        22:52:03   BIC                           1442.548
Sample:                             1   HQIC                          1433.389

================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         1331.6440   1025.146      1.299      0.194    -677.604    3340.892
ar.L1.D.sales   -0.3887      0.135     -2.880      0.004      -0.653      -0.124
ar.L2.D.sales   -0.0351      0.143     -0.245      0.806      -0.316       0.246
ar.L3.D.sales   -0.2255      0.138     -1.632      0.103      -0.496       0.045
ar.L4.D.sales   -0.1573      0.143     -1.102      0.271      -0.437       0.122
ar.L5.D.sales   -0.0720      0.134     -0.538      0.591      -0.334       0.190
                                    Roots
```

Fig 3d

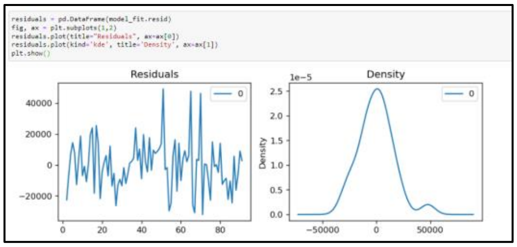


Fig 3e

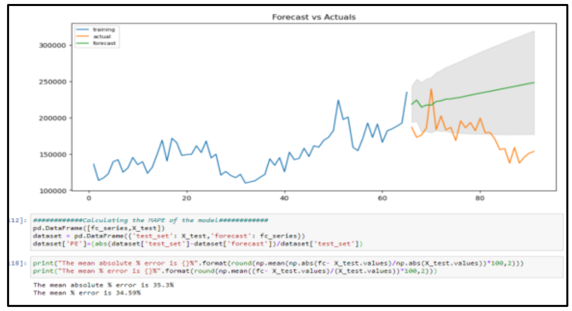


fig 3f

Seasoned ARIMA (SARIMA)

I have also used SARIMA model which considers seasonal trends in the forecast. This result less mean error which is 17% compared to ARIMA. Hence, SARIMA give best results for retail sales data and forecast is shown below.

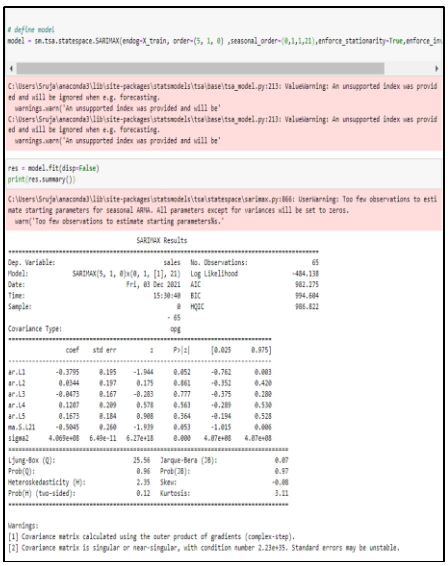


Fig 4a

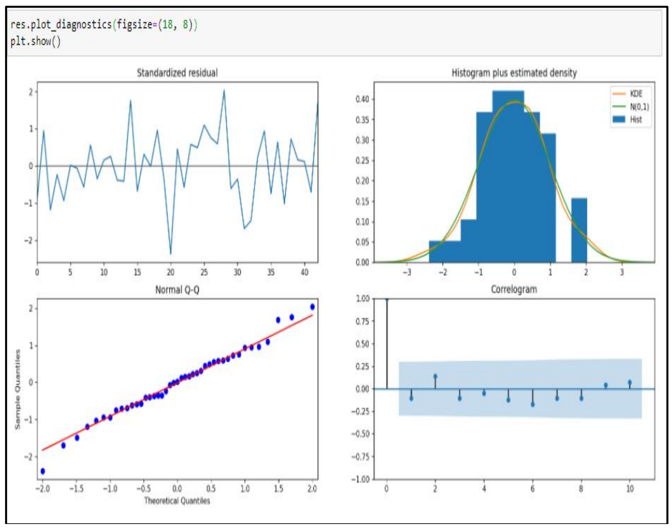


Fig 4b

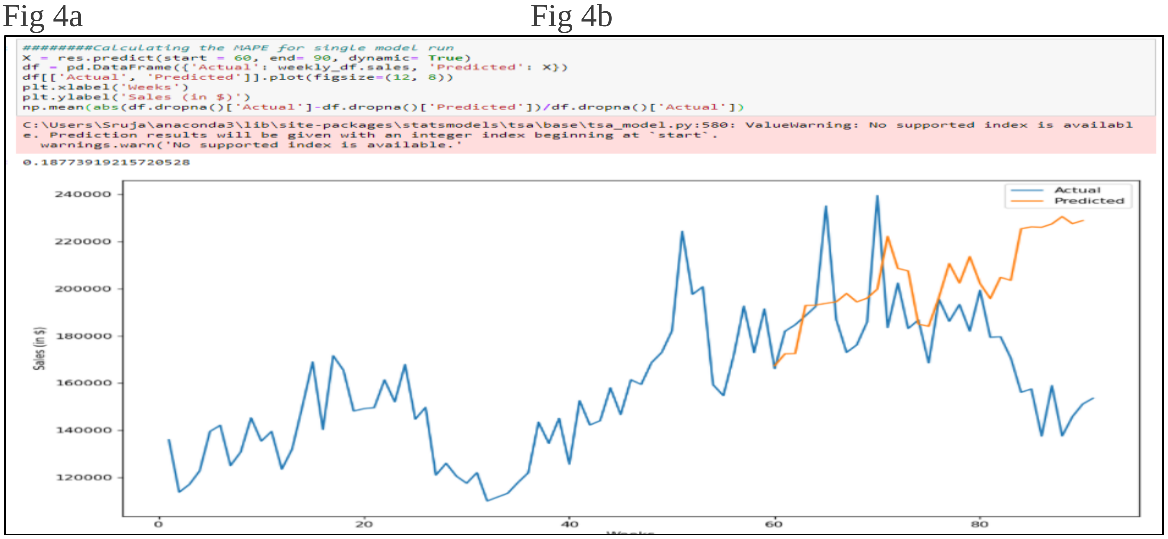


Fig 4c

8. Data Visualization and Results Report

I have performed exploratory data analysis to find the patterns in the given dataset. In this data set there is no null values and have checked for outliers and removed it before building the model. I have used Tableau for visualization charts. Below charts fig 5 shows the product category which are toys, arts crafts, electronics, sports and outdoors and games. It also shows the sales of the products and highest sales are from toys category. Fig 6 shows the products in stock and highest stored are arts and craft. Fig 7 shows the number of units sold according to the product price and highest sold units are around 15.99$ to 19.99$. Fig 8 and fig 9 shows the sales across locations and city. Fig 10, fig 11, fig 12 shows the monthly sales, monthly sales across locations and weekly sales.

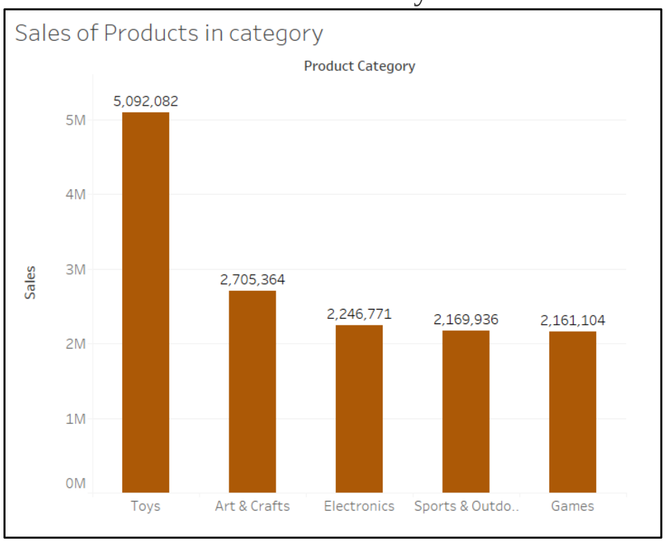


Fig5:Sales of Products in category

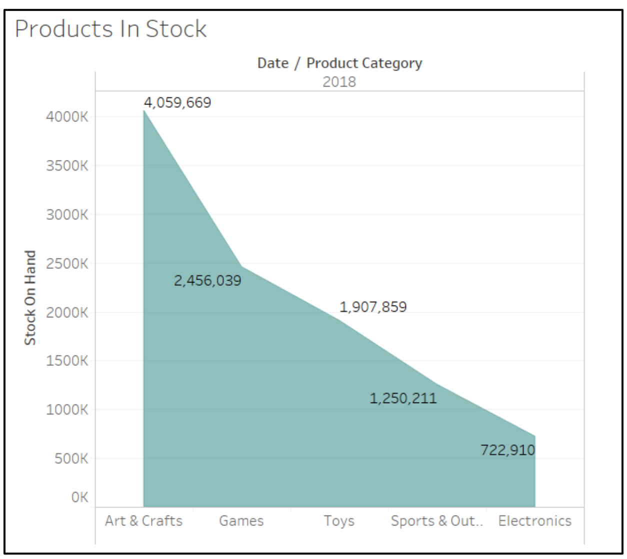


Fig 6: Products In-Stock

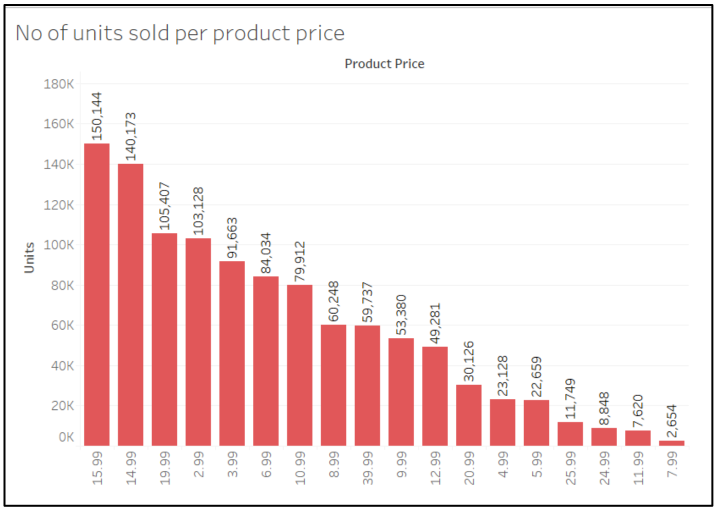


Fig 7:no of units sold per product price

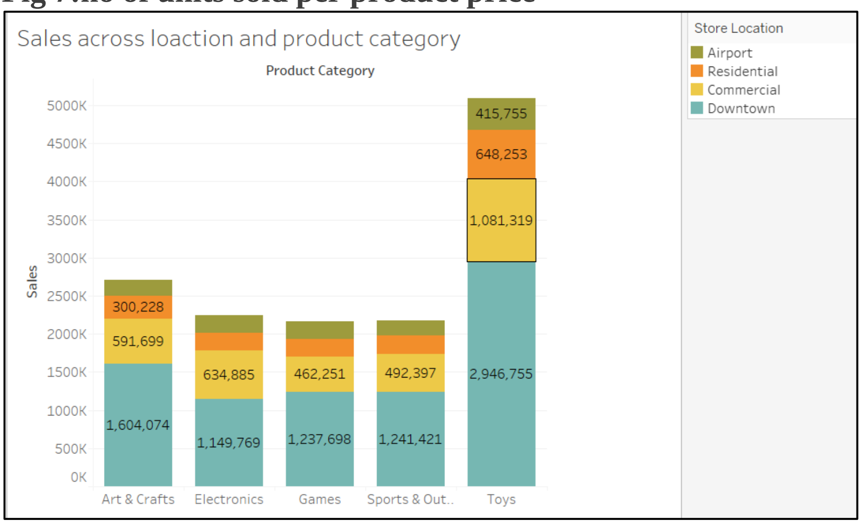


Fig 8:sales across location and product category

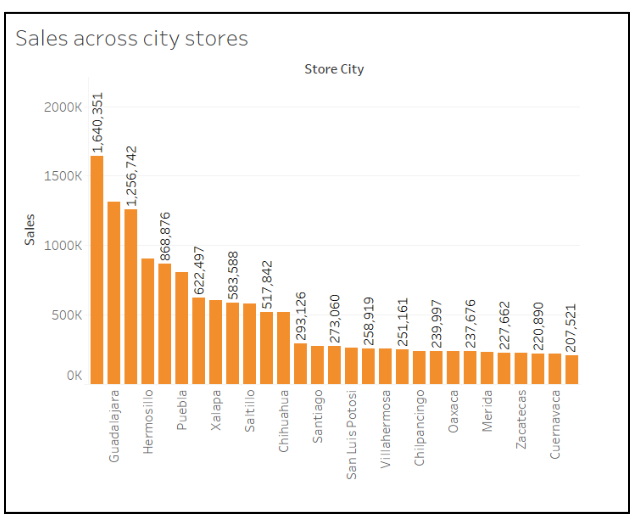


Fig 9: sales across city stores

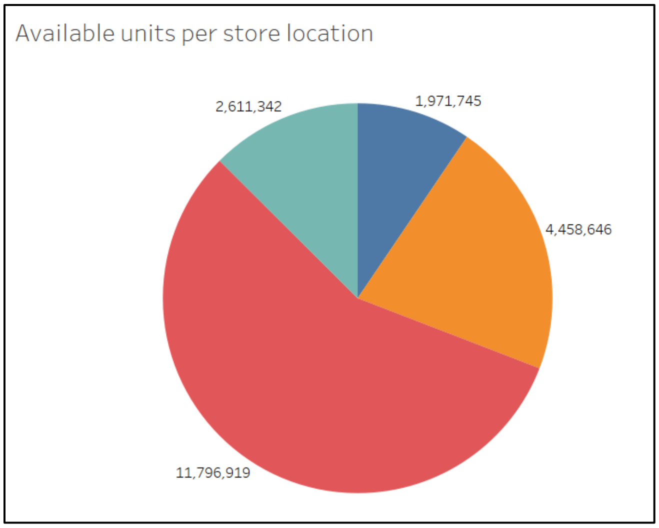


Fig10: Available units per store location

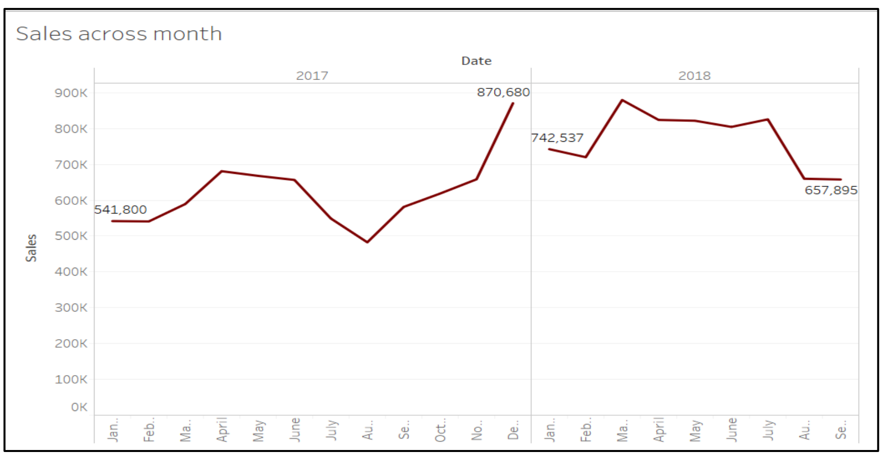


Fig 11: sales across month

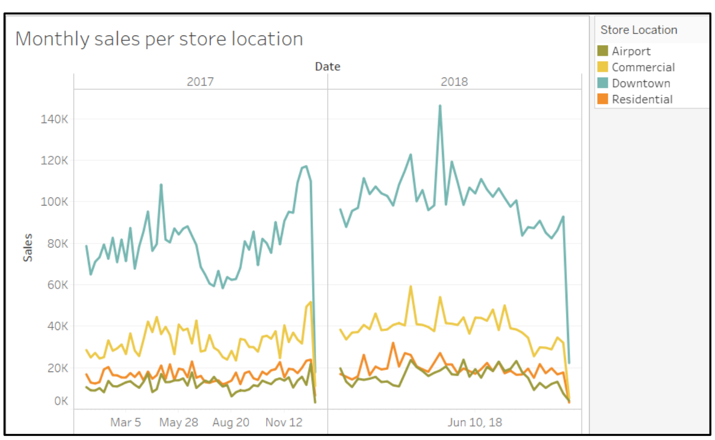


Fig 12: monthly sales per store location

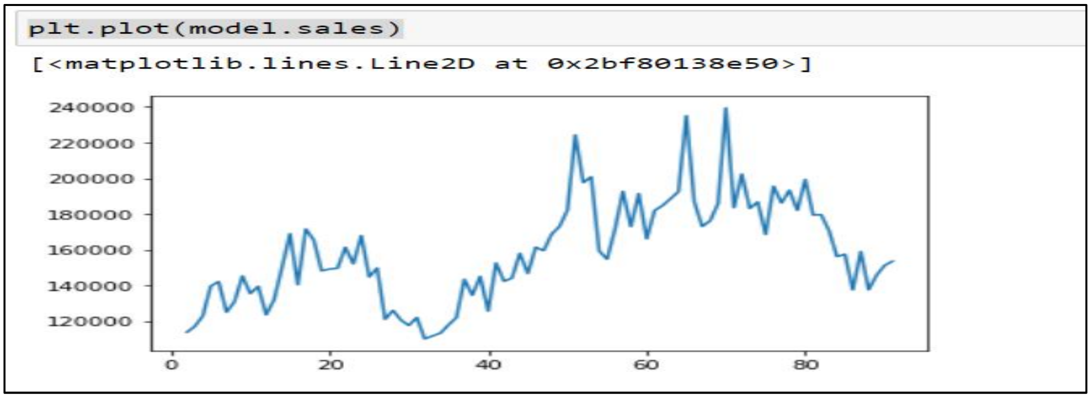


Fig 13: Sales Across Week

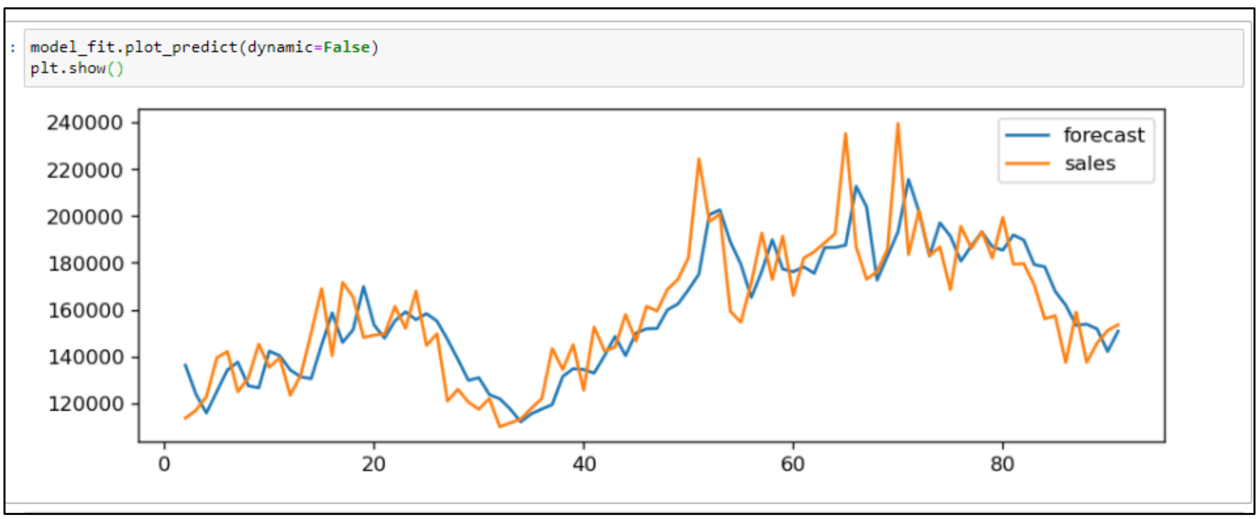


Fig 14: Sales prediction

9. **Explain and usage of your results.**
   In retail industry, with developed technology applying analytics helps the businesses to make a key decision. Analytics allows retailers to make a standard process to explore product categories, product in-stock, store city and location to predict the future sales and boost their revenue. The data model helps to predict the product sales for future year based on the previous year data trends and patterns. The model result show that Jan to Apr the sales increases constantly and Apr to Jun sales are steady and Jun to Aug sales are decreased and from Sep to Dec it gradually increases. This helps retailers to prepare for inflow of customers and manage the inventory.

10. **Other issues such as security and privacy. fairness and ethics issues.**
    I didn't have any issues regarding security and privacy. I have completed project on my own following the UNT policies.

11. List the datasets to be used.

- Inventory.csv
- Products.csv
- Sales.csv
- Stores.csv

The above-mentioned datasets are sufficient to complete my research.

12. Conclusion.

Predictive analytics applied for retail business helps to optimize the inventory and to predict the product demand, improve sales and helps to manage supply and customer relationship. This analysis helps business improve their demand and sale forecast accurately.

13. Selected bibliography and list them in APA format. You need well cite each reference in your proposal.

- Liudmyla Taranenko **,** Machine Learning To Demand Forecasting, Jul 7, 2021 https://mobidev.biz/blog/machine-learning-methods-demand-forecasting-retail.

- Rohini Mandiwal , 10 Ways Retail Predictive Analytics Can Transform Your https://www.intellicus.com/10-ways-retail-predictive-analytics-can-transform-your-business/

- Rod Daugherty, How Data Can Optimize Inventory Forecasting , August 6, 2019 https://multichannelmerchant.com/operations/practical-machine-learning-how-data-optimize-inventory-forecasting/

- James Collins, Lee Gates, Bulent Kasman, Vicky Nguyen, Mark Taylor, John Yamartino : Retail Analytics http://scet.berkeley.edu/wp-content/uploads/UCBSCETRetailAnalyticsReport.pdf

- Consultancy.eu, Maturing supply chain analytics for optimal inventory management, 19 July 2021 https://www.consultancy.eu/news/6510/maturing-supply-chain-analytics-for-optimal-inventory-management