Project Report B5

Name	PGID
Ashish	61710147
Saurabh Shekhar	61710907
Rishab Choraria	61710589
Kirti Pandey	61710568
Rahul Singh	61710158
Krishna Guha	61710740

Table of Contents

Executive summary

a. Problem Description

Our client is a German drug manufacturer who owns and operates six stores in different parts of Germany. Each store has a unique layout and some stores are open for fewer days than others. Moreover, customer footfall varies by location. Our client contracts staffing personnel on a daily basis from a staffing agency for all the stores, and each contracted staff is paid on per hour basis every day. The management in our client's organization has been tasked with an objective to bring down the staffing costs by optimizing the number of store staff that is contracted every day. The number of staff required at each store depends on the number of customers who visit the store. Therefore, as a first step towards optimizing the number of staff required at each of the six stores, our client wanted to estimate customer footfall at each store.

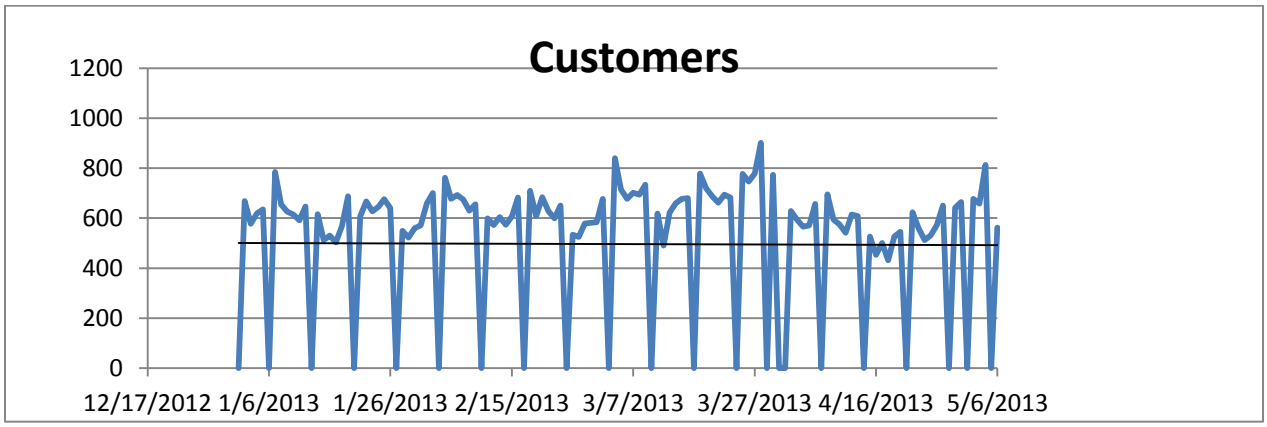
Our task was to forecast customer footfall on a daily basis at each store location over a **forecast period of six weeks** starting August 1, 2015. To accomplish this task, our client provided daily sales and customer footfall information from January 1, 2013 to July 31, 2015.

b. Brief Description of Data

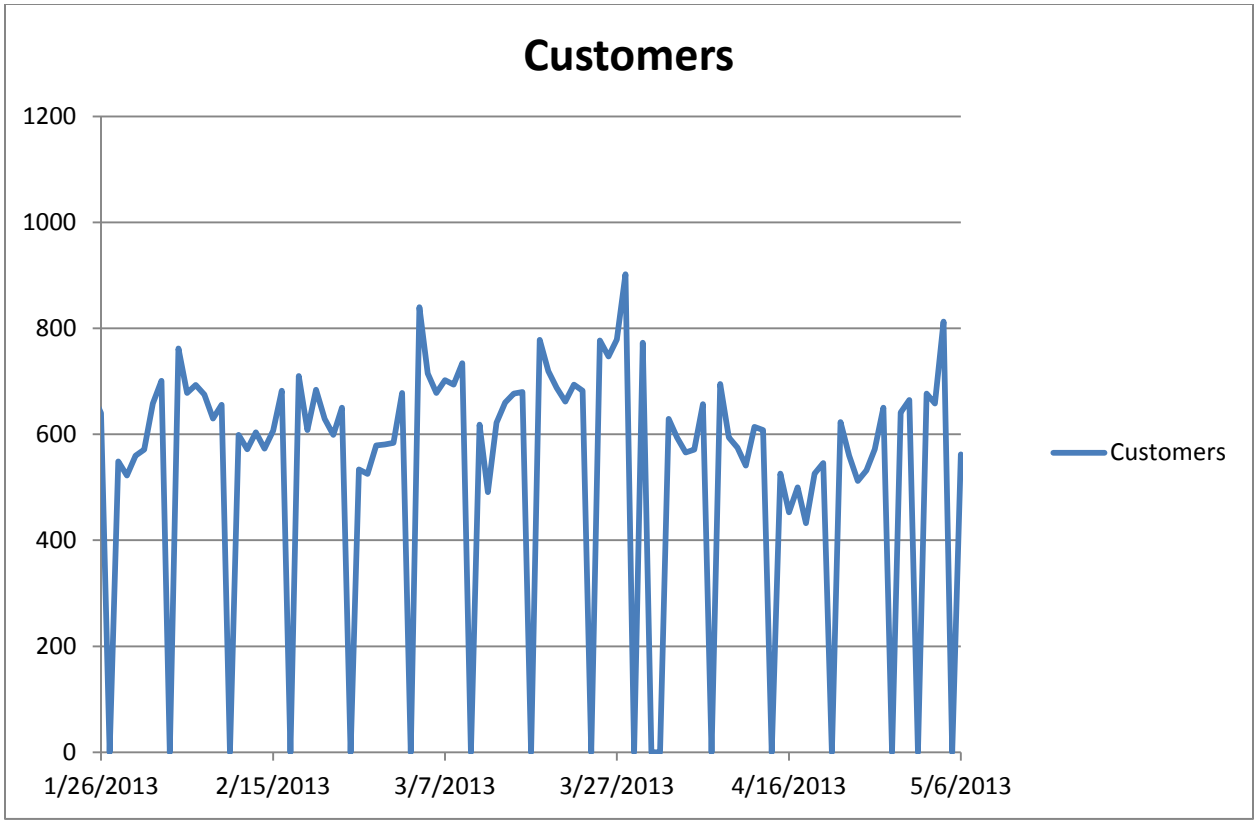
The dataset was obtained from Kaggle.com. For each store, the dataset contained daily sales and customer footfall. In addition to these two fields, there was further information provided on whether there was any sales promotion on any given day and whether a given day was a state holiday or a school holiday.

For the purpose of this analysis, only customer footfall was considered as it would directly impact the staffing requirement at the store. The customer data contained level, noise and seasonality while flat trend was observed for all the series. Seasonality was observed to be 7.

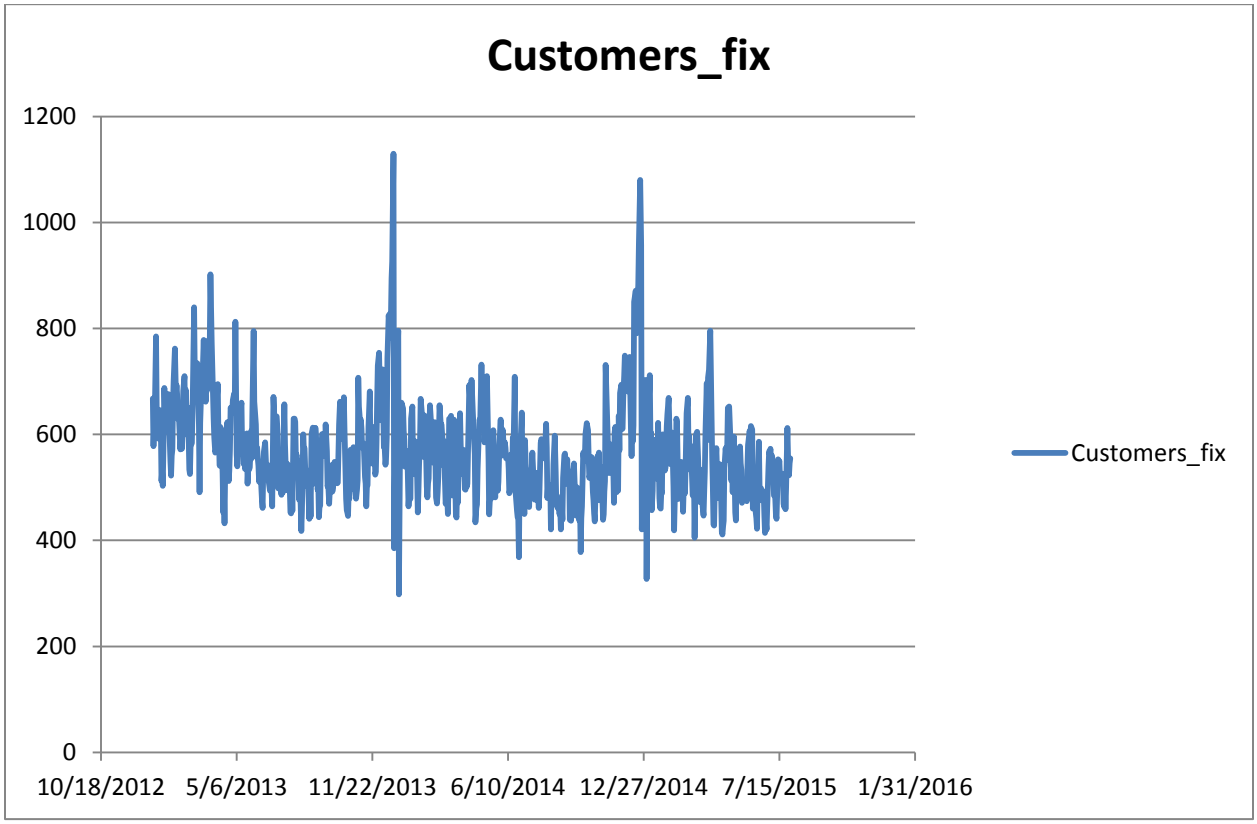
Trend



Plots below show the data before and after removing the closed day's data.

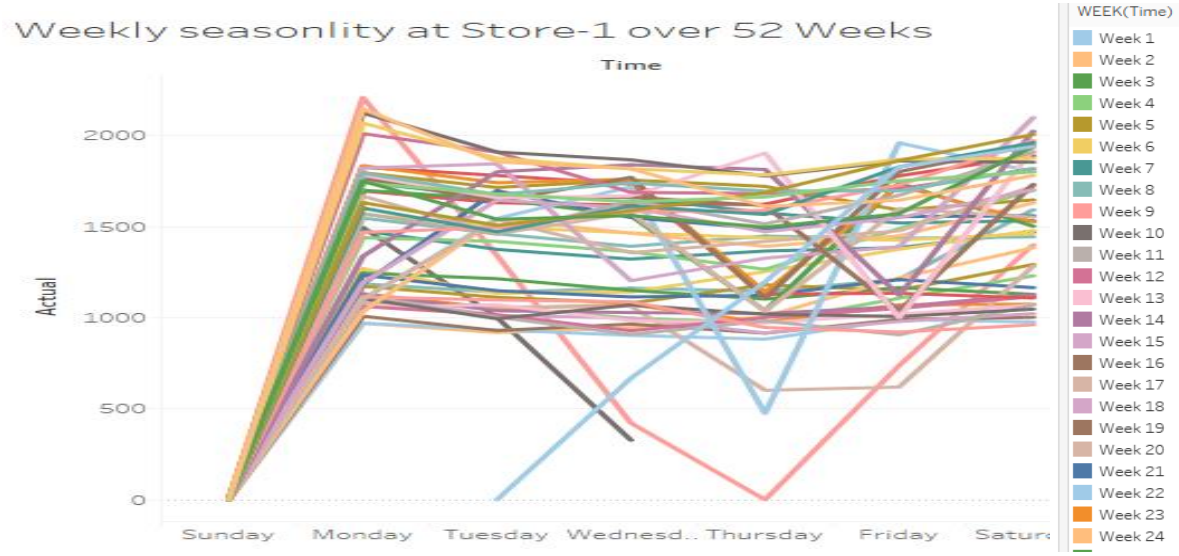


Above plot shows the customer footfall before removing the closed days.



Above plot shows the customer footfall after removing the closed days.

Day of Week Seasonality



c. High Level Description of Forecasting Methods

The following methods were used for forecasting-

- Naïve (Benchmark Prediciton)
- Holt-Winters Smoothing – Due to the presence of seasonality
- Multi Linear Regression (MLR) – Due to the presence of seasonality
- Ensemble – To evaluate if the combination of model is better than individual models

Evaluation -

The methods and results were evaluated based on Root Mean Square Error (RMSE) and based on the plots of the actual and forecasted values from various methods.

The **validation errors** for each of the methods are shown in the table below

Validation Errors (RMSE)	**Store 1**	**Store 2**	**Store 4**	**Store 7**	**Store 85**	**Store 562**
Naïve	**166.87**	**209.59**	381.33	276.59	240.60	**350.66**
Holt Winter	**55.65**	**173.04**	**147.18**	**117.92**	**90.60**	**169.13**
MLR	145.44	174.09	**400.83**	**401.84**	**246.42**	306.93

The actual and forecasted plots for each of the methods used are shown in the plot below:

Actual Vs Predicted Vs MLR

Store 1 – Residual Plot

d. Conclusions and Recommendations

- Based on the plots and errors above we conclude that **Holt Winter's** method should be used for forecasting for all stores.
- We believe that hourly data may help the store better predict the number of customer during peak hours and this could enable the store managers to plan staffing appropriately while reducing the total operational cost of the store.
- We also recommend that the store managers consider the 95% confidence band for staffing as that account for majority of the cases rather than the single value of the forecast. This would ensure that customer service levels are maintained across the stores.

Technical Summary

Data Preparation- The objective of data preparation is to make the available data suitable for forecasting using different methods. We followed the following steps for data preparation:

a) Sorting of Data- Data was in descending order, so sorting was done to get in ascending order
b) Time Indexing- For regression model, time index has to be created to capture Trend
c) Created Dummy Variables- It has to be created to capture Seasonality. Dummy variable for Thursday was dropped while running the regression calculation.
d) Data Partition- We used Tableau to plot our data to observe Seasonality, Trends, Level and Noise. We realized after the plot that we have a seasonality of 7 and our forecast period was 42. Since our forecast period is greater than the seasonality number, thus we chose forecast period as our validation period.

Issues with Data Preparation-

Forecasting Methods used- we used the following four methods for forecasting:-

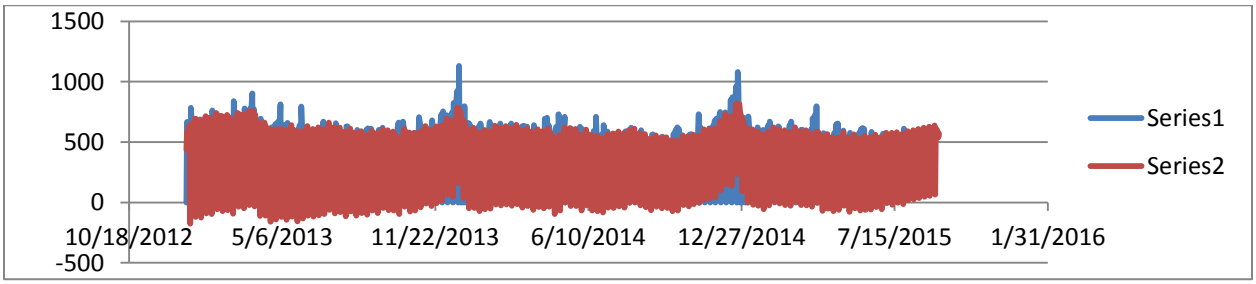
a) Naïve Method- Naïve method with seasonality 7 was chosen as a benchmark method against which other methods will be compared. We got RMSE of 166.87 by Naïve method
b) Holt Winter's Method- The data had seasonality and among all other methods available only Holt Winter's Method can handle seasonality. Thus this method was chosen. The output of this method can be interpreted by the following results:-

Training Error Measures

Mean Absolute Percentage Error (MAPE)	11.06220593
Mean Absolute Deviation (MAD)	78.09891152
Mean Square Error (MSE)	16581.65466
Tracking Signal Error (TSE)	-0.455215284
Cumulative Forecast Error (CFE)	-35.55181822
Mean Forecast Error (MFE)	-0.03950202

Validation Error Measures

Mean Absolute Percentage Error (MAPE)	10.01386915
Mean Absolute Deviation (MAD)	45.02638189
Mean Square Error (MSE)	3096.914479
Tracking Signal Error (TSE)	-2.276971193
Cumulative Forecast Error (CFE)	-102.5237745
Mean Forecast Error (MFE)	-2.44104225



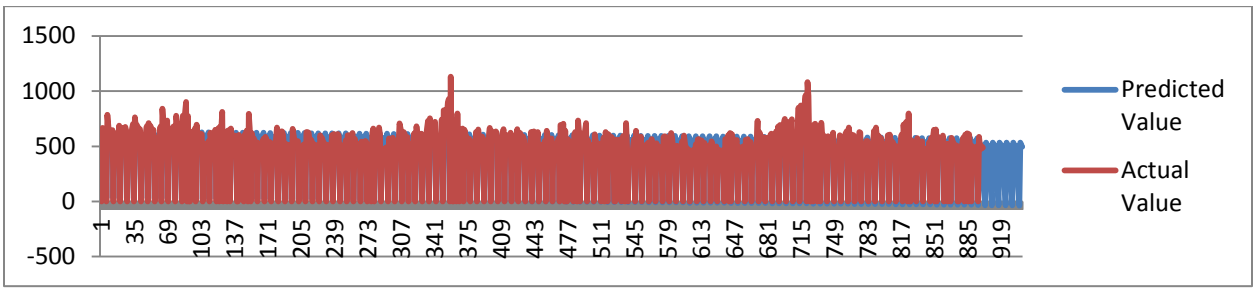
c) Multi- Regression Method- This method is also used to handle data with trends and seasonality. Further, it has an added advantage that it can handle missing values. Also, on top of that it can be used to identify the key variables affecting the dependent variables. We used this method on our data and results are as shown below:-

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
13852259	124.0621	9.06337E-13

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
888363.2	145.4356	61.97586993



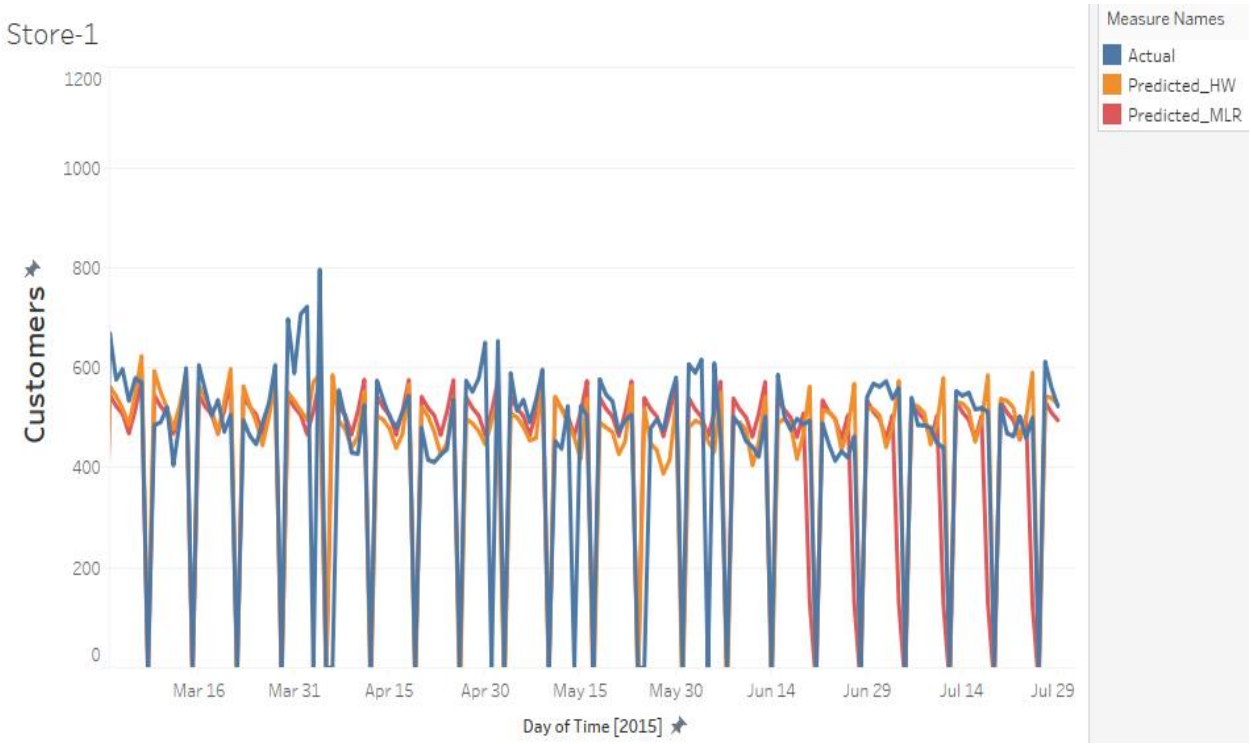
d) Ensemble- To capture good qualities of different models, Ensemble is created. We have taken the average value of the results obtained by Naïve, Holt Winter's and MLR methods and got our Ensemble output. The final results is captured in the following chart:-

Methods	RMSE
Naïve	166.87
Holt Winter's	55.65
Multi- Linear Regression	145.44
Ensemble	63.31

Conclusion and Recommendations – Holt Winter's smoothing is giving the least RMSE, so this method will be used to forecast future sales at the stores. Also, Ensemble method gave a higher error and its costly to implement Ensemble method so we suggest not to use it. Further, we could have looked further for auto-regressive model beyond this.

The detailed charts and graphs (plots of Actuals and Residuals), obtained from Tableau have been included in the Exhibits following the report. **Exhibit Part- A** contains training and validation data for the two methods (Holt Winter's and MLR) and **Exhibit Part- B** contains forecast plots for the 6 stores based on Holt Winter's Method

Exhibit Part- A



..

Exhibit Part- A

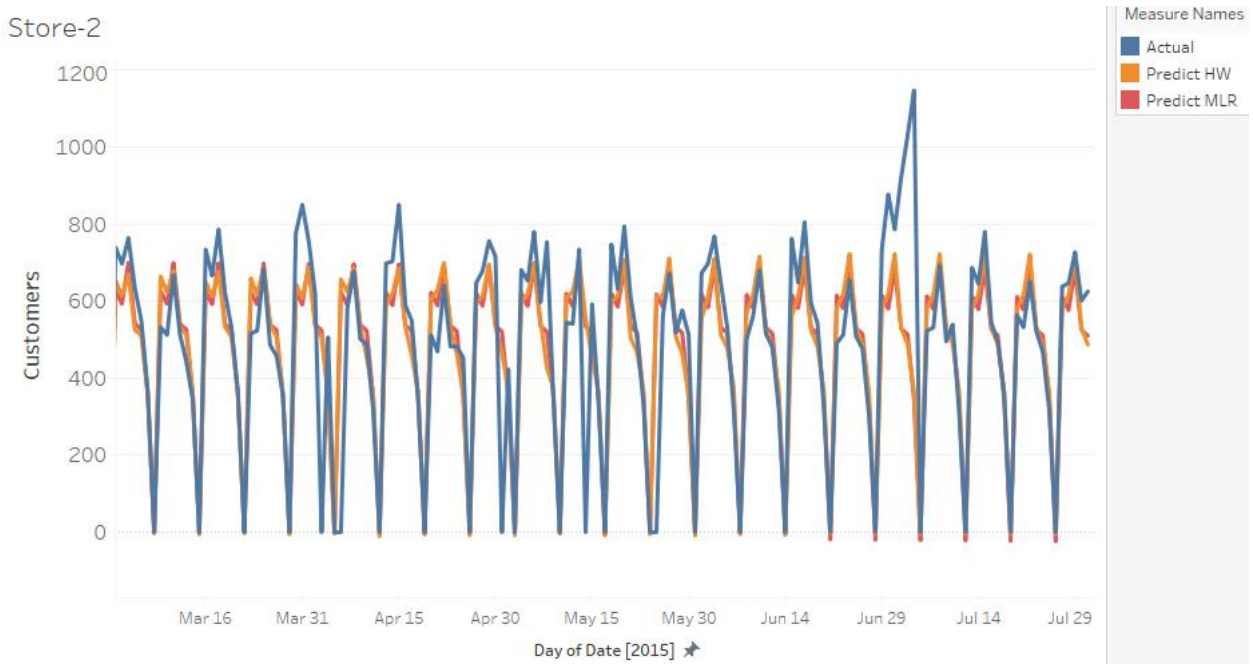


Exhibit Part- A

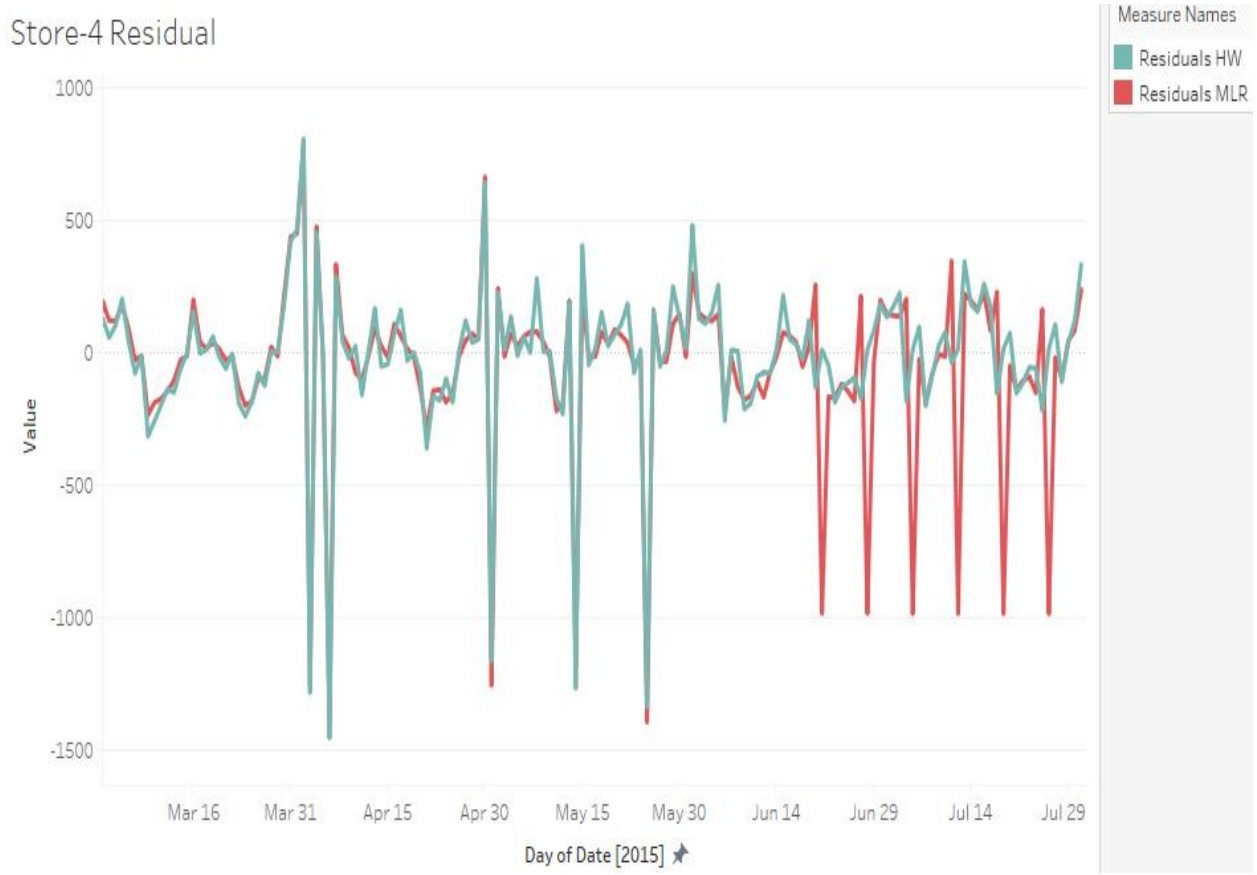
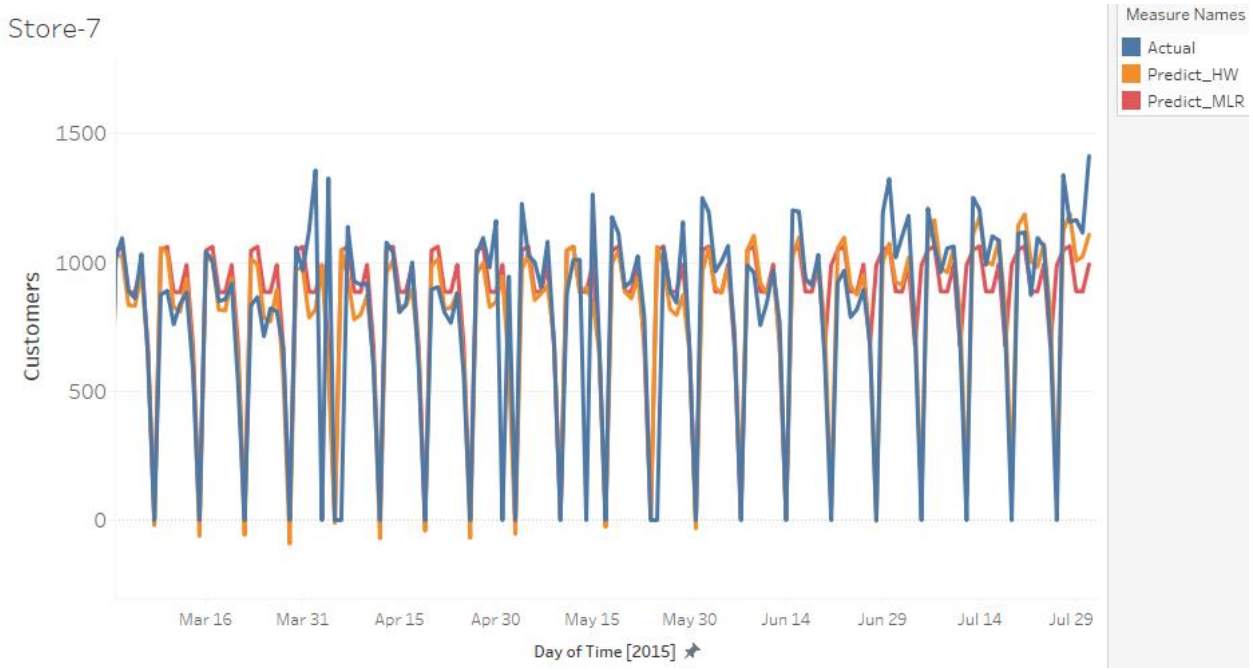


Exhibit Part- A



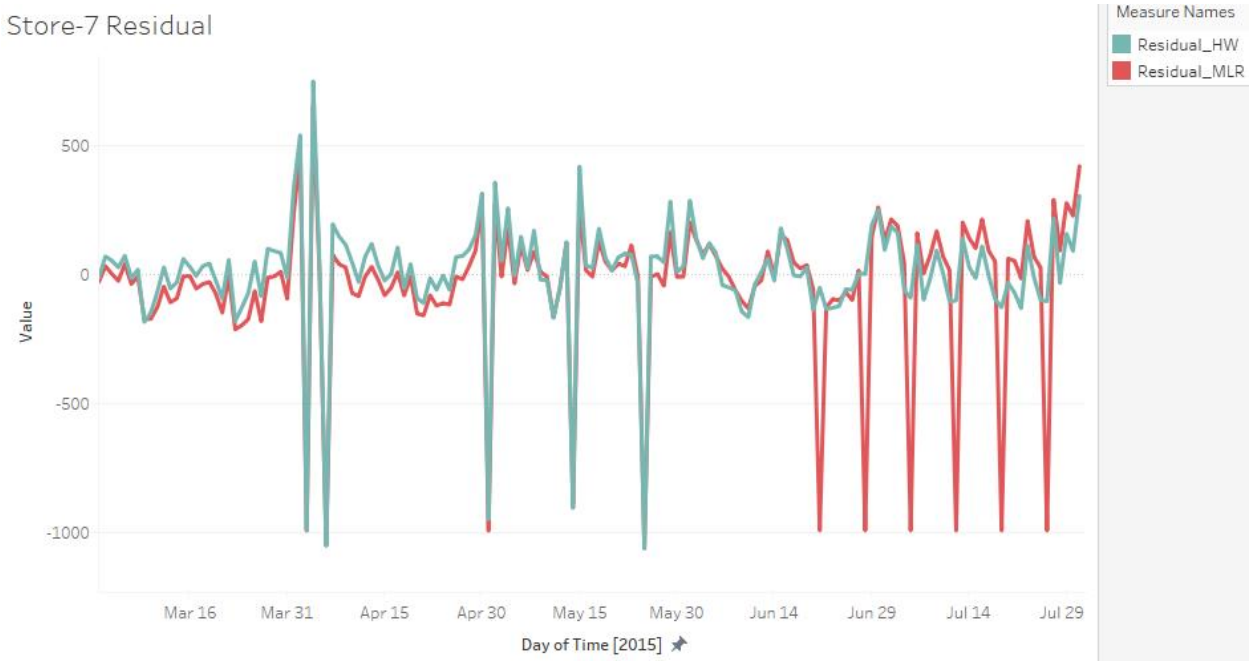


Exhibit Part- A

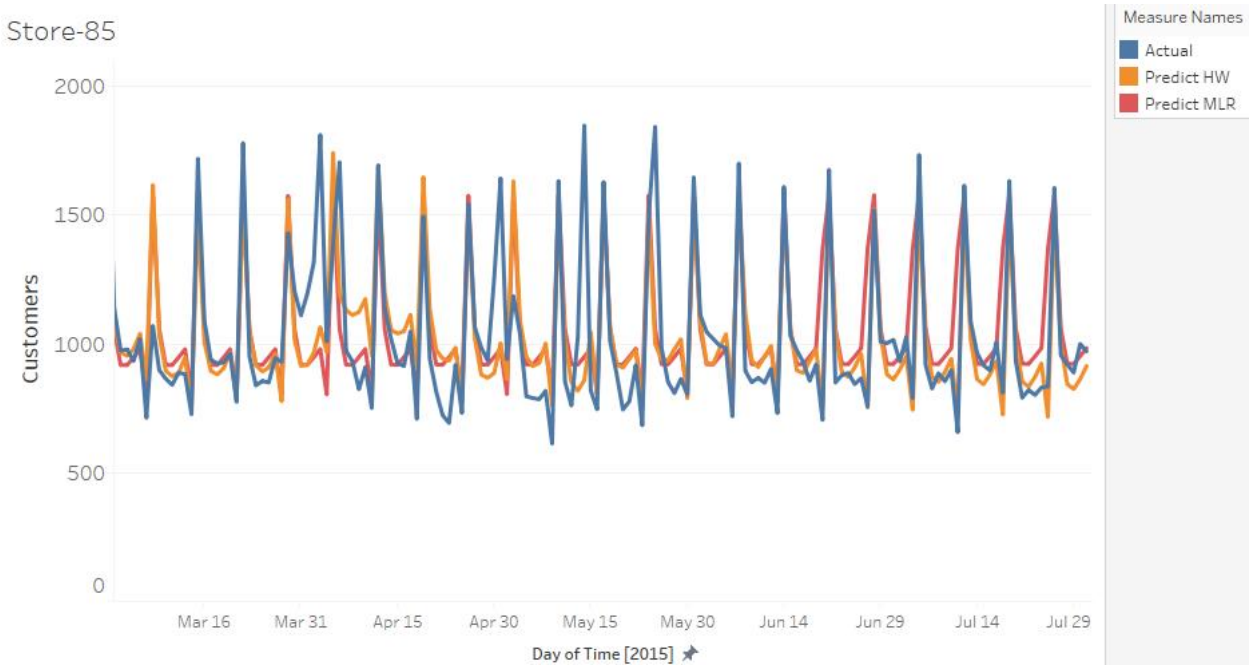
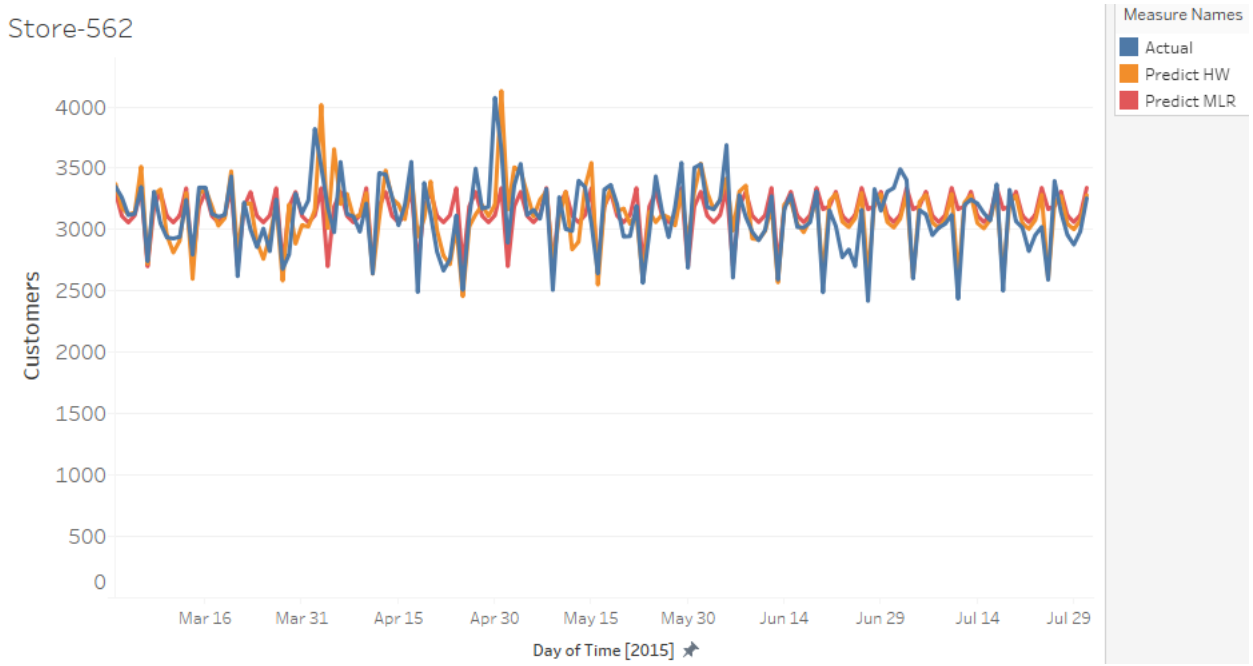


Exhibit Part- A



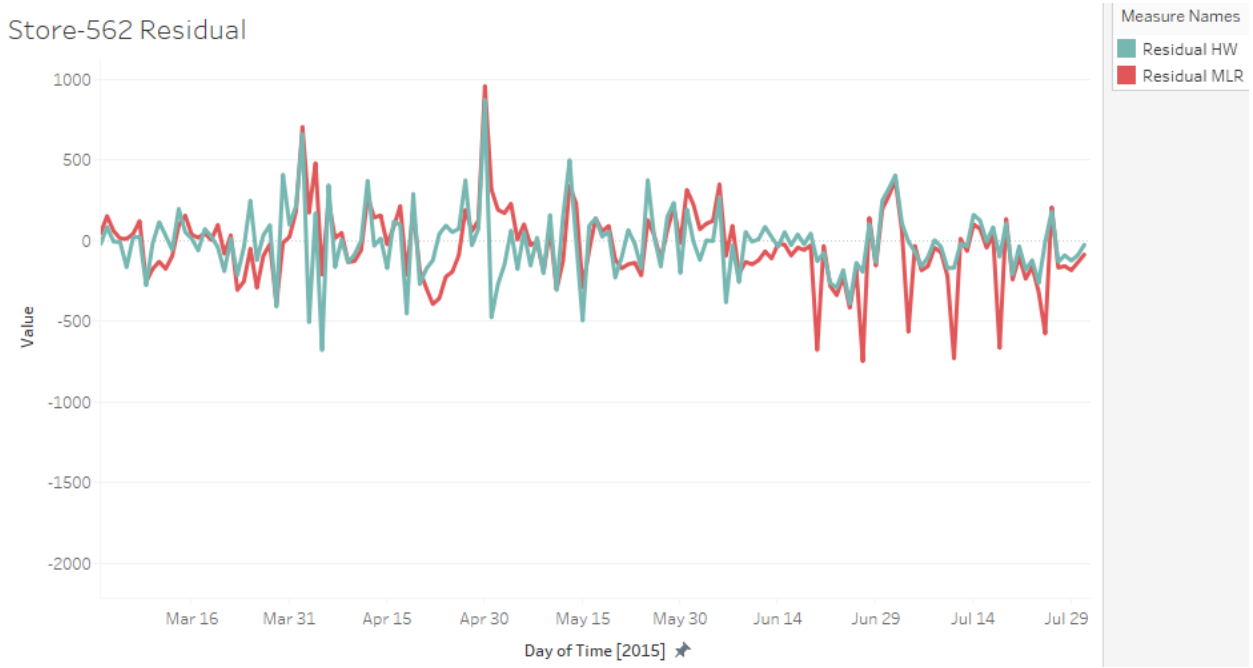
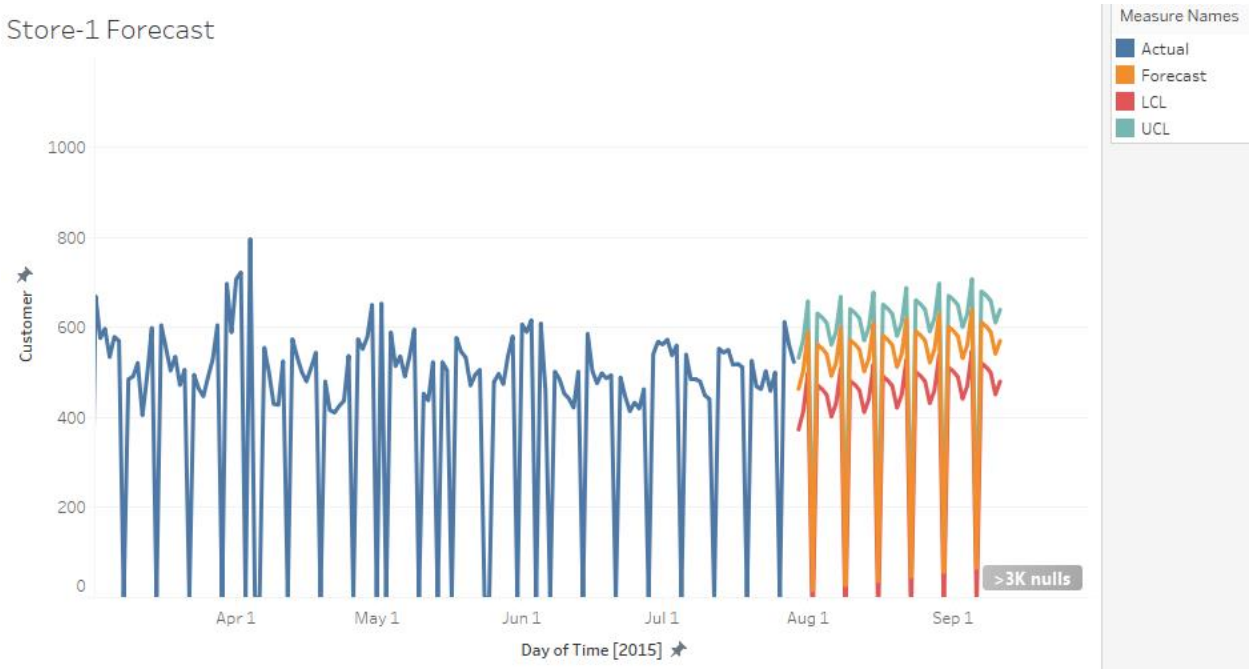


Exhibit- Part- B



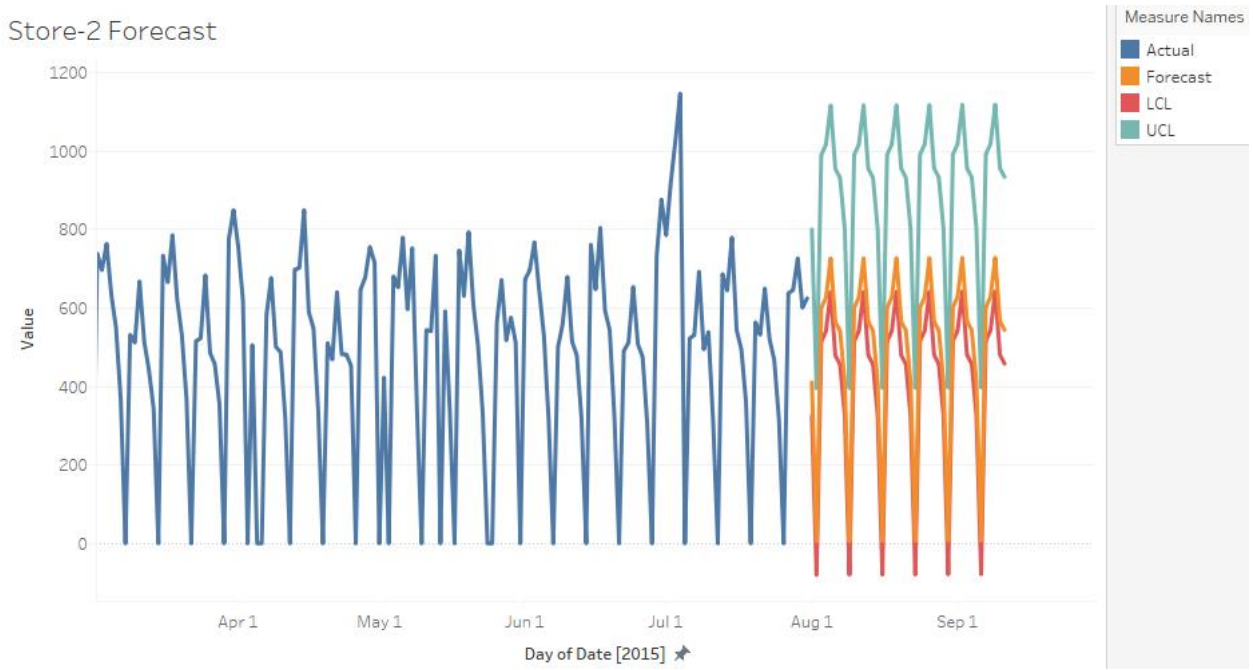


Exhibit- Part- B

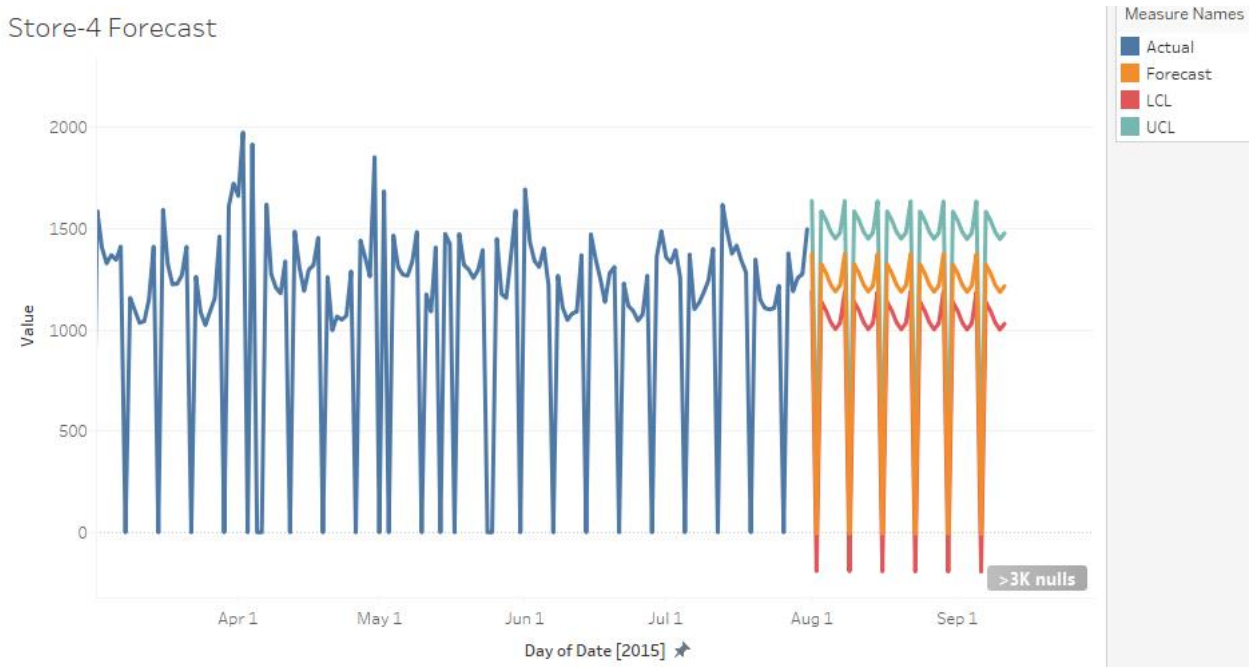


Exhibit- Part- B