END - Term MTP Evaluation Report

**Demand forecasting of freight commodities in Indian Railways using forecasting models and time series analysis**

**Submitted by**

Rupesh Garg – 19IM30019

**Under the guidance of**

Prof. Goutam Sen



Department of Industrial and Systems Engineering

Indian Institute of Technology, Kharagpur

March 2024

**Declaration**

I certify that

(a) The work contained in this report has been done by me under the guidance of

my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Con- duct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

**Date: April 29, 2024 (Rupesh Garg)**

**Place: Kharagpur (19IM30019)**

Department of Industrial and Systems Engineering

Indian Institute of Technology, Kharagpur



**Certificate**

This is to certify that the project report entitled " Demand forecasting of freight commodities in Indian Railways using forecasting models and time series analysis” submitted by Rupesh Garg (Roll No. 19IM30019) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Master of Technology in Industrial and Systems Engineering is a record of bona fide work carried out by him under our supervision and guidance during Spring Semester, 2023-24.

**Date: April 29, 2024 Prof. Goutam Sen**

**Place: Kharagpur** Industrial and Systems Engineering

Indian Institute of Technology Kharagpur

**1. Introduction**

Indian Railways (IR) boasts the distinction of possessing the world's third-largest railway network. Given its vast expanse and intricate structure, there arises an imperative need to scrutinize the various operations that wield substantial influence over the organization's revenue streams and customer outreach. It is incumbent upon the industry to usher in novel algorithms and cutting-edge technologies, thereby amplifying the efficiency and advantages inherent in these operations. The panorama of Indian industry has witnessed an exponential surge in freight traffic over the preceding decade. With projections indicating that India is poised to commandeer a substantial 40 percent of the global rail activity share by the year 2050, the pivotal role of freight in augmenting the revenue of Indian Railways comes into sharp focus. Yet, the arena of freight transport is currently witnessing an intensification of competition, ushering in a wave of specific opportunities and challenges for the railway freight industry. The ability to accurately forecast the volume of railway freight transport assumes paramount importance. This forecast serves as the bedrock upon which railway enterprises devise their transport strategies and compile meticulous transport plans. It is, in essence, the linchpin for the future progression of railway enterprises. Forecasting railway freight volume with scientific rigor and precision holds immense significance. This precision aids in the optimal allocation of railway freight resources, the refinement of the industrial structure within the railway domain, the elevation of the railway freight management landscape, and the enhancement of the overall customer experience within the ambit of railway freight services. One of the main concerns is the lack of information of future demand of wagons at different spatial locations present in Indian railway due to which a lot of empty rake movements take place.

**2. Literature Review**

The study notes that prior research on COVID-19 forecasting was restricted to particular nations or areas and that only a few numbers of forecasting techniques were applied. For comprehensiveness, this study, however, uses widely used forecasting methods and covers several European nations. Particularly, they note out that their research, which is rare in the literature, compares and evaluates the efficacy of the ARIMA, Prophet, and Holt-Winters exponential smoothing models. The paper does not give a full analysis of all prior COVID-19 predicting research. It primarily highlights the originality and significance of their own research, which compares the exponential smoothing models of ARIMA, Prophet and Holt-Winters

The research on parking availability prediction covers two basic approaches: determining parameters for an underlying parking process model and combining observed data with machine learning and statistical techniques to estimate future occupancy. Parking occupancy has been predicted by a variety of strategies, including wavelets neural networks, neural autoregressive models (NAR), chaotic time-series analysis, ARIMA models, simple regressions, multivariate spatio-temporal regression, clustering, and deep learning methods like CNN and LSTM networks. Using a neural network approach, prediction of the number of available parking space, prediction of parking lot, analysis of the impact of freight deliveries on traffic congestion, and modeling of stoppage times were implemented.

Traditional methods of forecasting demand in this literature include econometric models such as the ARIMA method and parametric smoothing. Autocorrelation in the data is covered by ARIMA, while autoregressive smoothing deals with trend and seasonality. Fifteen distinct models have been created using exponential smoothing techniques, which are categorized according to trend and seasonality.

For time series forecasting, neural networks—particularly recurrent neural networks (RNNs)—have also been employed extensively. To solve the vanishing gradient issue, long short-term memory cells (LSTM), a unique class of RNNs, were created. LSTMs are appropriate for time series forecasting. Partial demand information can be incorporated for future time steps by using LSTM's ability to selectively retain or forget information based on cell states. Nonlinear patterns and dependencies between data that are not amenable to modeling by econometric models or basic artificial neural networks are captured by LSTM.

**3. Research Gap**

The gap in demand forecasting for freight transportation in the Indian Railways is tried to be filled using traditional methods such as econometric models and exponential smoothing but they have limitations in capturing nonlinear patterns and dependencies in the data.

In addressing such problems, ARIMA and LSTM models are commonly employed. However, these models typically operate on a per-class basis, necessitating the creation of distinct models for each location and storing each model in JobLib or Pickle. Consequently, a time series must be established for each location, detailing its historical demand over a specific timeframe, to facilitate the accurate prediction of future demand for that particular location. Employing Holt Winter's Exponential Smoothing enables us to predict demand for all locations collectively, considering both the demand and their corresponding time, within a unified model. This approach eliminates the need for individual models per location, providing a more efficient and streamlined prediction framework.

**4. Data Description**

4.1 Data Collection

Following data sets were collected for two years (2021-2023) for modelling the demand of rakes for different commodities:

4.1.1 Demand Placement data:

This data comprises of details of demand placed at each spatial location of Indian

Railway such as type of rake required, commodity, location from which demand is

placed, information about the consignor and consignee.

4.2 Findings from commodity wise data:

• More than thirty percent of the demand is known on the same day of allocation

• For the commodity group Container on an average ninety percent of the demand is known on the allocation day only.

• On an average 70 percent of the demand of all the commodities is either known

on the same day or a day before the allocation of demand for loading.

From above statistics, we can clearly see that, most of the demand is known only

towards the day of loading and very meagre or no information on demand is available before 4 days. Thus, the fleet manager won’t have any clear idea on where to route the empty rake during the turnaround time. This necessitates the need of prediction on the various commodities without losing the partially available information.

There were a lot of missing values in between the months which made it a poor time series data and thus required some values to be filled for every missing date.

**5. Methodology**

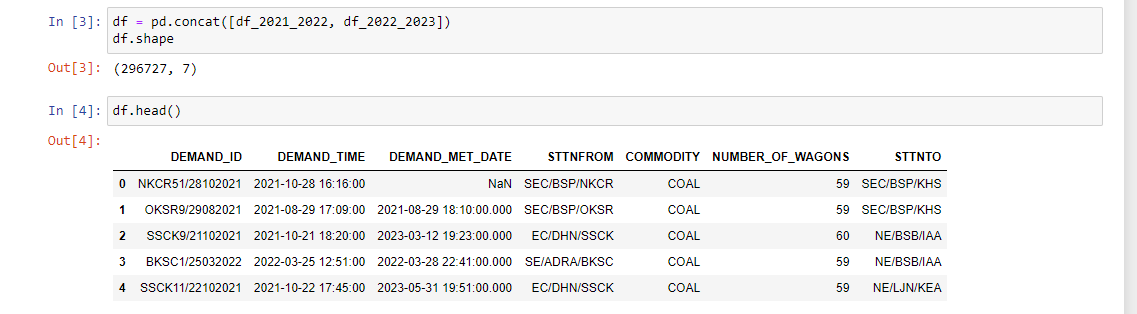
5.1 Data Preparation

This process involved cleaning of data i.e., checking for the missing data and imputing the same with appropriate values wherever required. All the data sets were verified and most of the data is found correct. The data on which demand was met is found to have missing data but it was a relevant feature in demand prediction.

5.2 Feature selection

After analysis of the data which is explained above, following features were finalized for demand prediction:

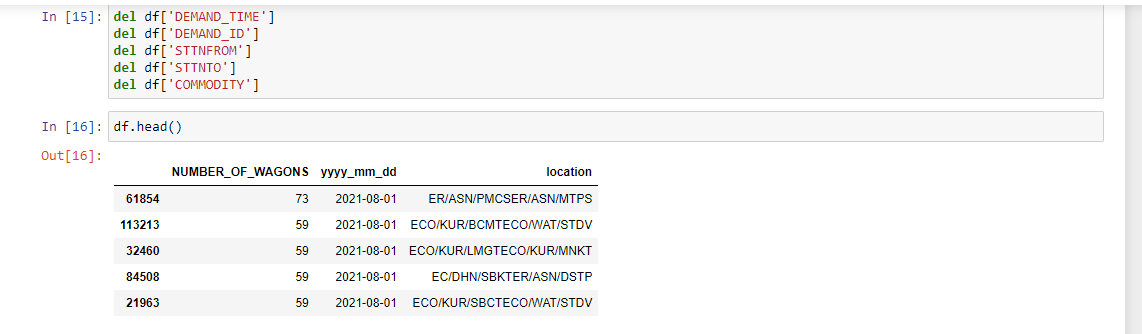
The dataset column names were changed into appropriate abbreviations. The dataset after changing looked as:



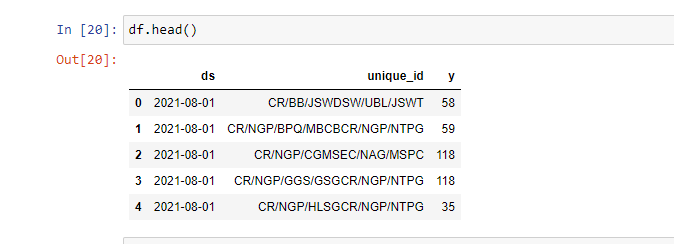
From the above 7 features, only STTNFROM (station from where train left), STTNTO (destination station of train), NUMBER\_OF\_WAGONS (from one particular station to another) and DEMAND\_TIME (as we want to predict the demand for a particular day, we need the historical data in a time series format of the day it was sent, not the day it was received) are needed.

I needed the location or path of the train as demand depends on both STTNFROM and STTNTO, so I created a new feature called ‘location’ by concatenating STTNFROM and STTNTO. Then I deleted the STTNFROM and STTNTO columns too as they became irrelevant after getting the ‘location’ feature.

Further, I am doing daily prediction of data thus I just needed the date from DEMAND\_TIME, thus I created a new column ‘yyyy\_mm\_dd’ extracting date from the timestamp DEMAND\_TIME and deleted DEMAND\_TIME column as well. The final table after making the above changes in the features looked as:



Now I had to create my dataset into a time series data format. Thus, I grouped the NUMBER\_OF\_WAGONS by ‘yyyy\_mm\_dd’ and \*‘location’ so that I can get a non-duplicate time series demand data for freight commodities for each day for every location. Also, I changed the column names as ‘y’ from ‘NUMBER\_OF\_WAGONS’, ‘unique\_id’ from ‘location’ and ‘yyyy\_mm\_dd’ as ‘ds’ (otherwise HoltWinter’s Model would give error). Final dataset ready to be trained and tested looked like:



5.3 Model Selection

Out of the number of models explained in (Section 2), we are going to adopt HoltWinter’s Exponential Smoothing and ARIMA model.

ARIMA Equation:

Ŷt = μ + ϕ1 yt-1 +…+ ϕp yt-p - θ1et-1 -…- θqet-q ∀ t = 2, ..., T

Where “e” is an error term and “c” is a constant.

ARIMA models are typically expressed like “ARIMA(p,d,q)”, with the three terms p, d, and q where

p captures “auto regressive nature” i.e., the number of prior ('lagged') Y values to be included or subtracted from Y within the model

d represents the count of differentiations needed to transform the data into a stationary signal and

q denotes the number of preceding/lagged values for the error term to be added or subtracted from Y, capturing the 'moving average' component of ARIMA

HoltWinter’s Equation:

𝑦̂𝑇+1|𝑇 = 𝛼𝑦𝑇 + 𝛼(1 − 𝛼)𝑦𝑇−1 + 𝛼(1 − 𝛼) 2𝑦𝑇−2 + ⋯

where, α signifies the model's responsiveness. The Holt–Winters model is classified as either an additive or multiplicative model, depending on the nature of the seasonal pattern.

5.4 Completing the time series:

As the data was having few missing dates, the time series was completed using the mean values separately for each location.



**6. Model Implementation Details**

I used the additive model of ‘HoltWinters’ exponential smoothing for the combined data and predicted it in a dataframe named ‘p’.

6.1 Loss Functions:

The loss functions I have considered in my work are:

Root Mean Squared Error:

RMSE = (2)\*0.5

Mean Absolute Error:

MAE =

: Actual value

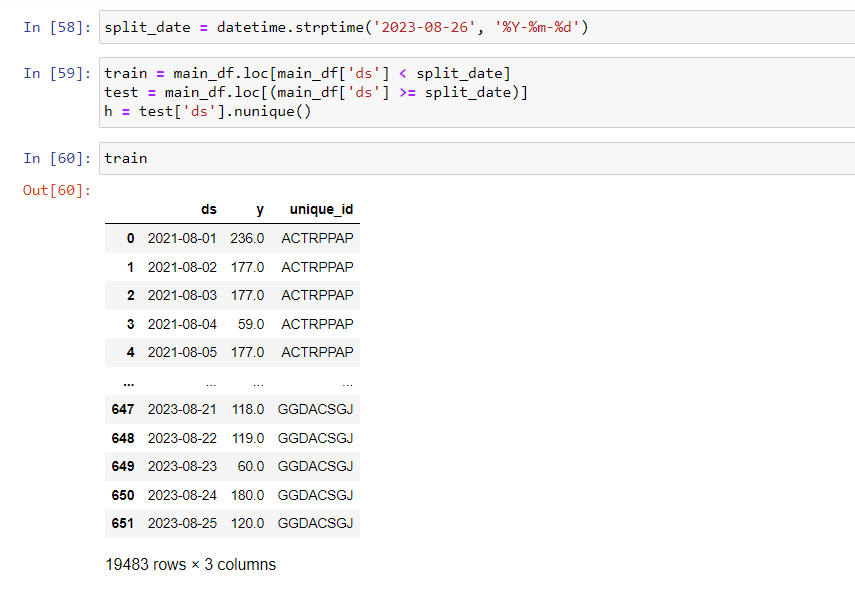
\_pred: Predicted value

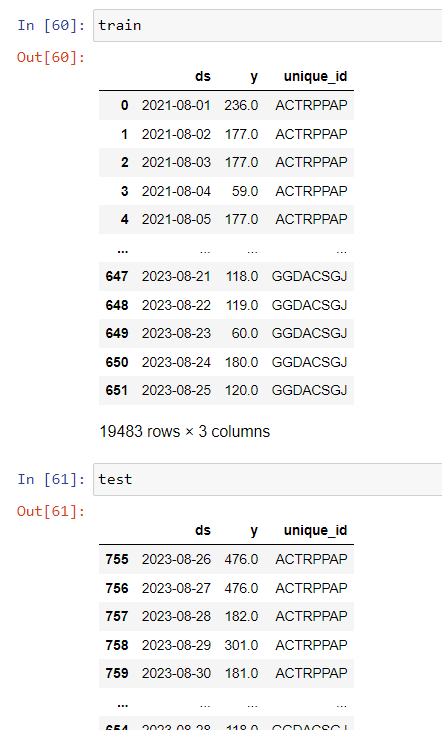
6.2 How the model worked?

As a first step data prepared in (Section 5.1) was divided into two parts training data (dataset which is used to train the model or is fitted in the model) and testing data (dataset which is used to check the accuracy of the model). The split date was set such that the dataset can be split into train data and test data.

The data was converted in following input structure before feeding it into the model:

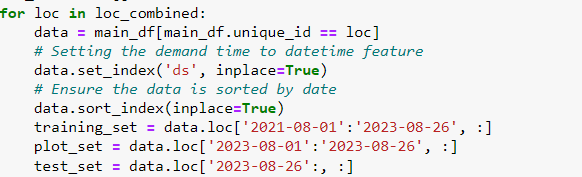
For HoltWinters:





This was the model considering all the locations into a single model. A p-matrix is calculated which contains the forecasted vs testing data for each date for each location. All the error terms are calculated accordingly for each location.

For ARIMA:

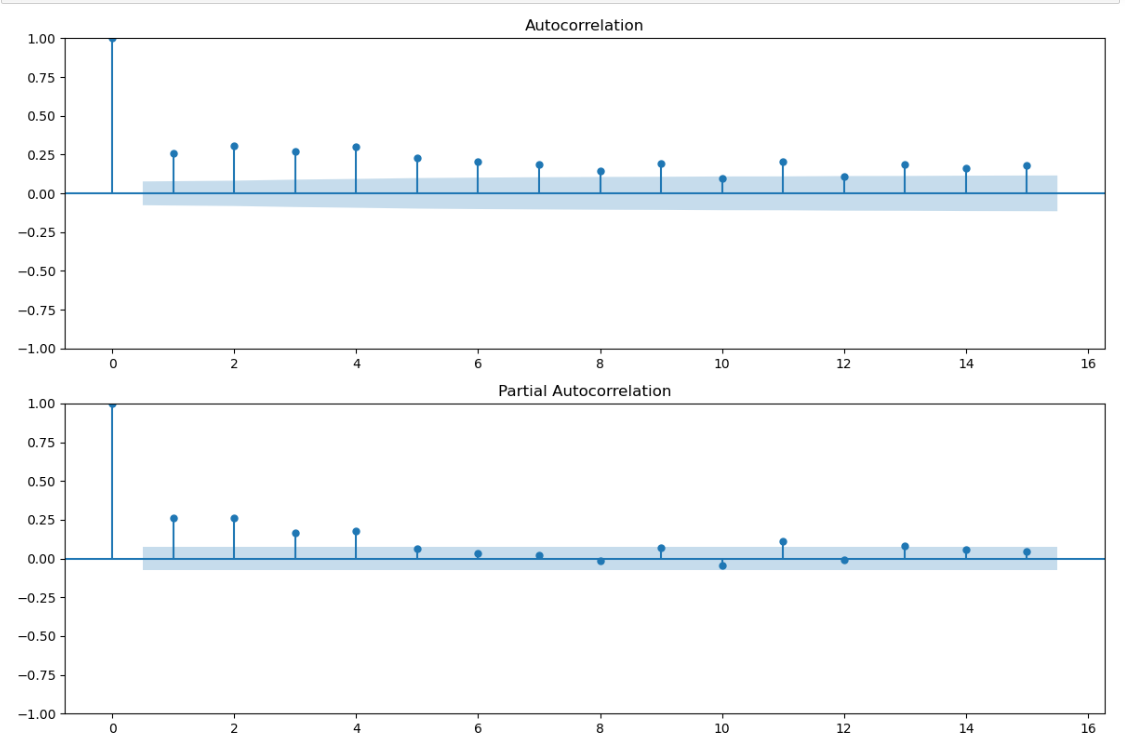


Looped over every location and models saved using JOBLIB for every location

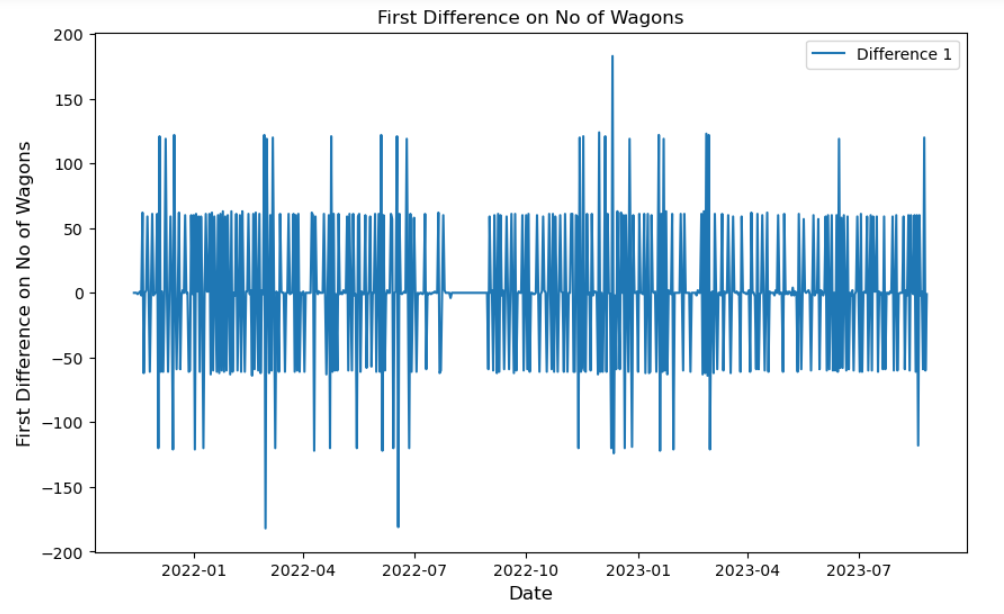
**7. Graphical Inferences, Results & Discussion**

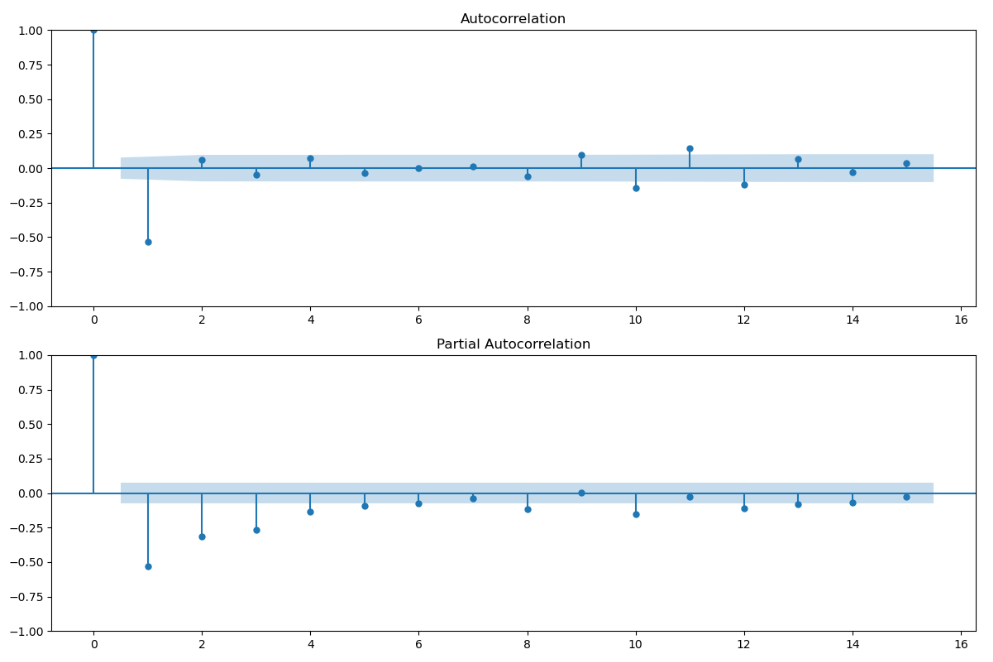
The plot of dataset, autocorrelation and partial autocorrelation before any differencing of the dataset for location - 'HLSGNTPG':



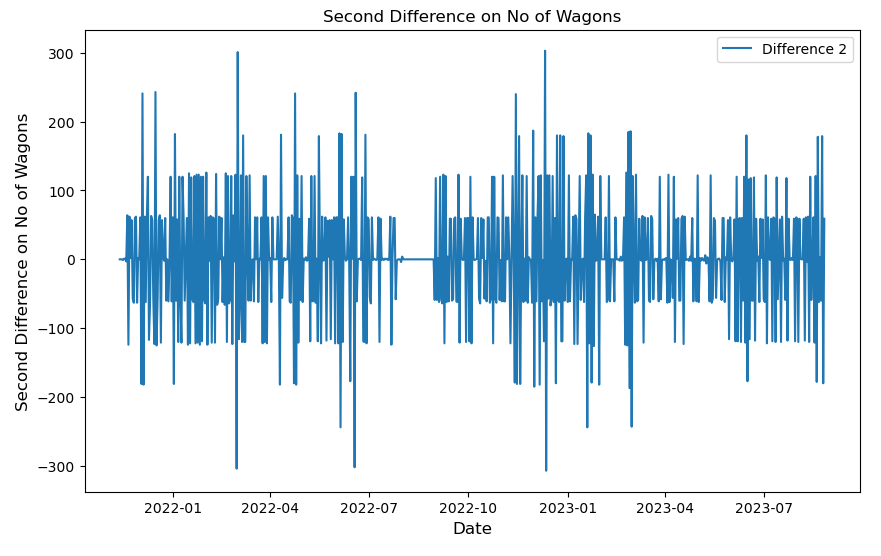


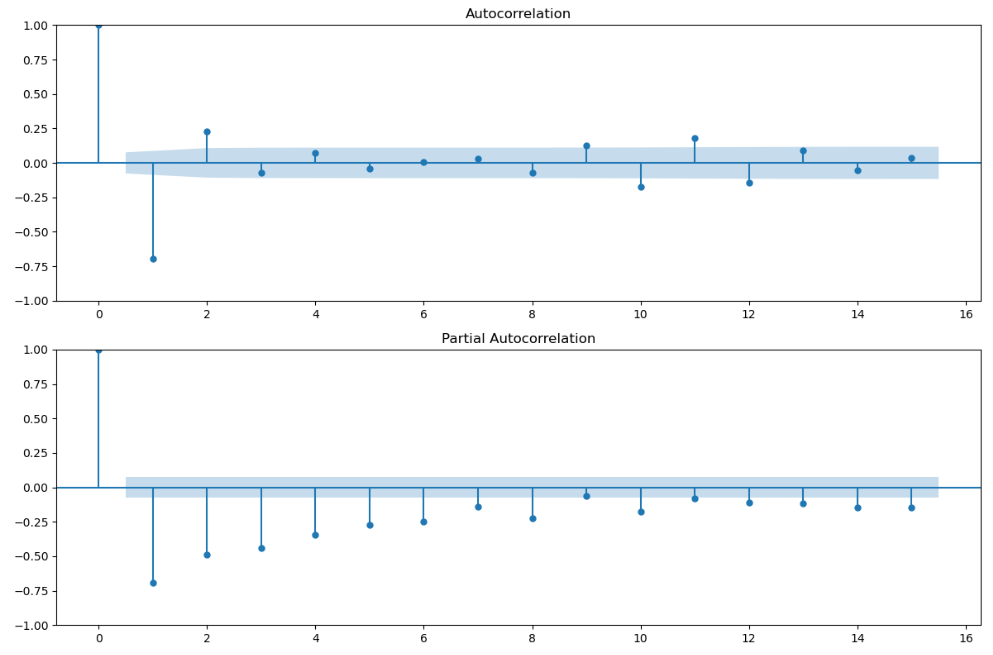
The plot of dataset, autocorrelation and partial autocorrelation after first differencing of the dataset for location - 'HLSGNTPG':



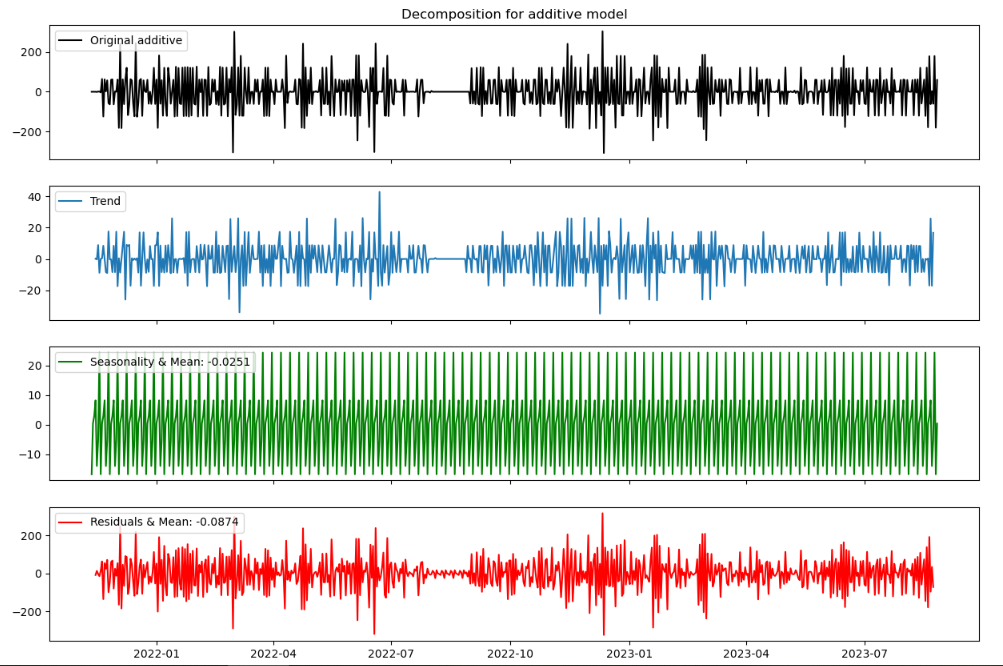


The plot of dataset, autocorrelation and partial autocorrelation after second differencing of the dataset for location - 'HLSGNTPG':

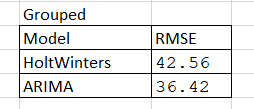




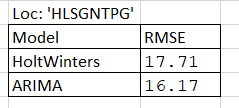
Time Series Decomposition of Final Dataset (after second differencing) for location - 'HLSGNTPG':



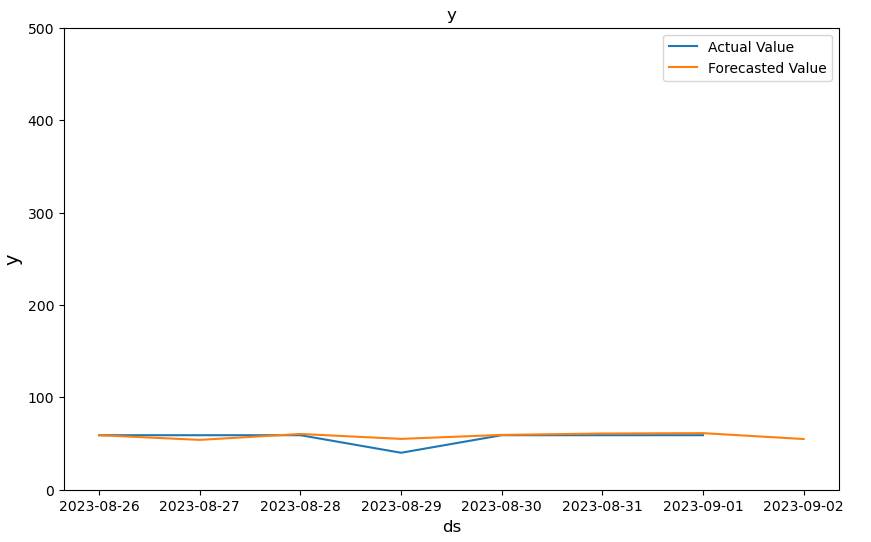
Root Mean Absolute Error curve calculated in grouped Models:



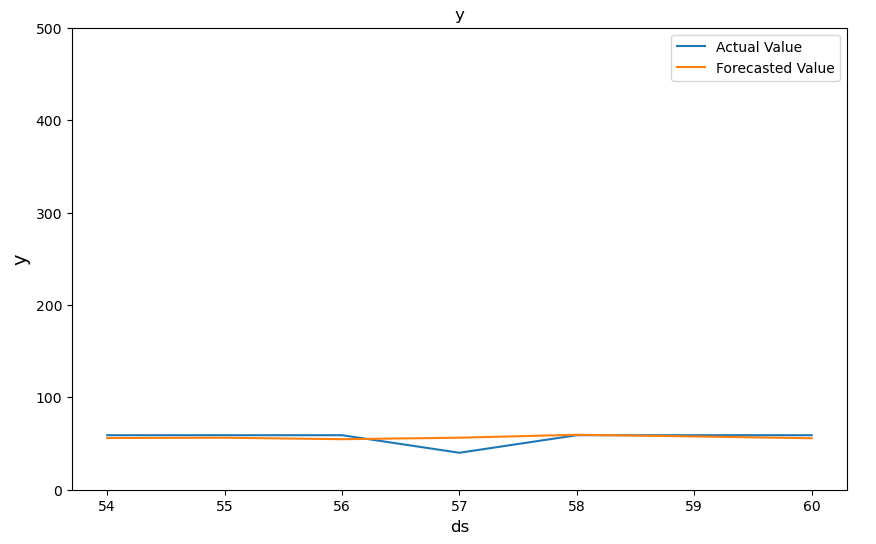
Root Mean Absolute Error curve calculated for individual location - 'HLSGNTPG':



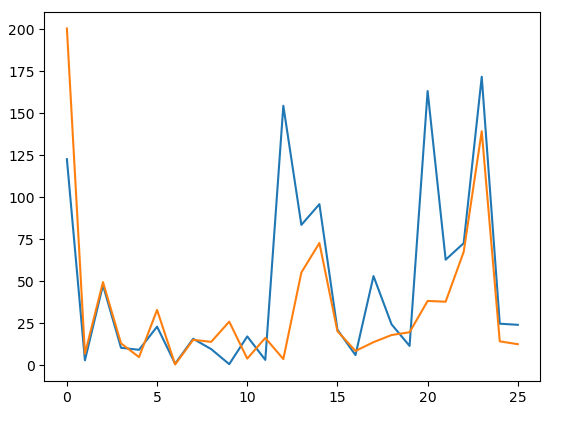
Forecasted vs actual value for ARIMA model for location - 'HLSGNTPG':



Forecasted vs actual value for HW model for location - 'HLSGNTPG':



MSE Comparison for HW and ARIMA at each location:



To ascertain the d-value for the ARIMA model, one location was selected ('HLSGNTPG' in this case) to undergo examination. The Augmented Dickey-Fuller (ADF) test was then employed to assess the stationarity of the dataset. It was noted that after applying two rounds of differencing, the data achieved stationarity, thereby determining the d-value to be 2.

ARIMA has been observed to yield a marginally better root mean square error (RMSE) value compared to HoltWinters. This enhanced accuracy, however, comes at the expense of slightly larger storage requirements, as elucidated in the research gap. Despite this disparity in storage space, both models remain viable options for time series forecasting tasks. Yet, in the current landscape where accuracy holds precedence over memory utilization, the superiority of ARIMA becomes apparent. The prioritization of accuracy aligns with contemporary trends in data-driven decision-making, emphasizing the importance of precise forecasts in optimizing resource allocation, enhancing operational efficiency, and ultimately improving overall performance metrics. Hence, in light of these considerations, ARIMA emerges as the preferred choice for time series forecasting applications where accuracy is paramount.

**8. Conclusion**

In various segments of this study, the aim was to underscore the significance of demand forecasting within the freight supply chain, particularly emphasized in Section 1. Furthermore, the necessity of examining each commodity individually was addressed, as highlighted in Section 7. This study also illustrated the efficacy of Holt-Winters in such contexts, showcasing its benefits through its equations.

After filling the time series gap using the mean filling technique, both models were applied to individual location and grouped datasets. In total, 26 ARIMA models were necessary to capture the unique characteristics of each of the 26 locations, and they were stored using JobLib. Conversely, only a single HoltWinters model was needed to forecast for all the locations. Upon comparing the accuracy using root mean square error (RMSE) of ARIMA and HoltWinters, it was observed that ARIMA outperformed HoltWinters. These findings are significant as they will contribute to providing more accurate predictions for the Indian Railways, aiding in efficient planning and operations.

**9. References:**

* Time series with trend and seasonality components by Andrius Buteikis andrius.buteikis@mif.vu.lt

<http://web.vu.lt/mif/a.buteikis/>

* Comparative Performance Analysis of ARIMA, PROPHET and Holt-Winters forecasting methods on European COVID-19 data by AUTHORS: NUR SEBNEM ERSOZ, PINAR GUNER, AYHAN AKBAS, BURCU BAKIR GUNGOR
* Using Machine Learning to Predict Freight Vehicles’ Demand for Loading Zones in Urban Environments by Andres Regal Ludowieg, Ivan Sanchez-Diaz, Lokesh Kumar Kalahasthi: SageJournals
* Demand forecasting of freight commodities with partial information revealed by shippers - a case study in Indian Railways by Ayush Sharma, K Kalaivani, Raja Gopalakrishnan