**DSCI 5340-001, 003: Predictive Analytics and Business Forecasting**

Project 2 Report

Predicting Residential Power Usage

Nishith Reddy Mannuru: 11467417

**PREDICTING RESIDENTIAL POWER USEAGE**

I identified this dataset in Kaggle. [Link](https://www.kaggle.com/srinuti/residential-power-usage-3years-data-timeseries)

Contents

[INTRODUCTION AND BUSINESS OBJECTIVE 3](#_Toc70681878)

[DATA DICTIONARY 3](#_Toc70681879)

[LITERATURE REVIEW: 4](#_Toc70681880)

[HYPOTHESIS GENERATION WITH RESPECT TO PROBLEM STATEMENT 5](#_Toc70681881)

[5](#_Toc70681882)

[CLEANING UP THE DATASET 5](#_Toc70681883)

[METHODOLOGY: 7](#_Toc70681884)

[RESULTS and DISCUSSIONS: 8](#_Toc70681885)

[Step 1: Importing cleaned dataset to SAS and plotting the data. 8](#_Toc70681886)

[Step 2: Plotting the data 9](#_Toc70681887)

[Step 3: Perform different time series methods to forecast. 9](#_Toc70681888)

[ADF test: 13](#_Toc70681889)

[Auto Arima in Python: 17](#_Toc70681890)

[Step 1: importing data. 17](#_Toc70681891)

[Step 2: Checking if the data is stationary using ADF test in Python. 18](#_Toc70681892)

[Step 3: Performing differencing. 18](#_Toc70681893)

[Python code: 19](#_Toc70681894)

[Step 4: Checking ACF and PACF before moving to auto arima. 19](#_Toc70681895)

[Step 6: Auto arima model run. 22](#_Toc70681896)

# INTRODUCTION AND BUSINESS OBJECTIVE

The main aim is to build a predictive model which can predict the power usage of 2 storied houses in Houston, Texas. In this project I used 3 years of data i.e. 2017, 2018, 2019. I are trying to predict the future power consumption by residents of Houston. This can help the government to see how the graph of consumption of power is going towards future and accordingly the generation of power can be maintained.

# DATA DICTIONARY

The original data set consists of data from January 6th, 2016 to August 2020. But to maintain consistency I assumed only from January 1st, 2017 to 31st December 2019, so that exactly 3 years of data is chosen. Two different datasets are combined and cleaned to get the final dataset. More on that explanation is in the preparation of dataset section.

Dataset Link: <https://www.kaggle.com/srinuti/residential-power-usage-3years-data-timeseries>

This final data set contains a set of **1095 records (365\*3 years)** under **2 attributes**.

The attributes being as follows:

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Day | No. of Days |
| Value | Power usage in KWH |

# LITERATURE REVIEW:

* In this paper the author, talks about the clustering of residential electrical customers using load time series. The authors made, few little arrangements, demonstrated boundaries and have utilized them to sum up the extensive load time series. The gathering methodology is reasonably vivacious against a limited sort of mixed noise. The author accepted that this could help in diminishing the noise, and inconsistency in the time-series likewise can help in predicting the more precise residential power usage. I additionally attempted to follow the same comparative methodology. I utilized smoothing procedures to diminish the noise and also to normalize the inconsistency in the time-series, which vastly improved our expectations.
* In this paper the author H. M. Al-Hamadi was forecasting long term electric power load using fuzzy linear regression technique. Anticipating electric load assumes a significant part in foreseeing the future electric load and pinnacle load demand. H. M. Al-Hamadi accepted if components like trend, irregularity, seasonality, and abnormality were to be distinguished independently, at that point it would help in making exact predictions. I have attempted to follow the comparable methodology, in this undertaking. I distinguished the occasional factors and made precise predictions dependent on that.
* In this paper, the authors talk about the dynamic clustering of residential electricity consumption time series data based on Hausdroff distance. The author has thought, noticing the unique advancement of energy at every hour of the day or routinely at steady timeframe can help in resulting much accurate predictions. With the information at constant timeframe, he accepted to execute moving average model, which can help in knowing how today's information was influenced by the previous days. I implemented the same moving average model in our project, so that our predictions would be much accurate and also to know how much of an influence my previous day data is having on todays.
* The paper talks about the well-established holts’ winters method and compares with another model called Prophet. In this paper the authors are forecasting electric loads of Kuwait. Accuracy and feasibility is checked in this paper. According to authors, the later model is showing more accuracy than holts winters method. Both models are being judged based on criteria like MAPE, MAE, Rsq. And the same approach was used while performing analysis in our project. Many other models were used and checked with its MAPE results.
* In this paper the author talks about the financial time series forecasting learning. The main objective of the author was to use time series models such as ARIMA and machine learning models as LSTM and also few other deep learning techniques to help forecast the financial aspects. He has created a various number of models and has identified the best model on the basis of minimum possible error. I implemented the same strategy as well. I implemented a lot of different models in our approach and concluded the best model based on the least amount of error.
* In this paper the author, describes about the time series forecasting of petroleum production using deep LSTM recurrent networks. He accepted that the time arrangement estimating is the undertaking of anticipating future qualities with authentic information. He referenced that no specific technique is ideal for determining something. Furthermore, he advised to execute the thought from numerous points of view and has advised to choose the best model based on that. Similar methodologies were followed in our project as well. I conducted and led many strategies as possible to wind up with one amazing model. The ideal model was determined based on the least possible error.

# HYPOTHESIS GENERATION WITH RESPECT TO PROBLEM STATEMENT

* The **NULL HYPOTHESIS** in our problem statement is that the future forecast values does not increase with time period.
* **ALTERNATE HYPOTHESIS** is that the future forecast values does increase with the time period.

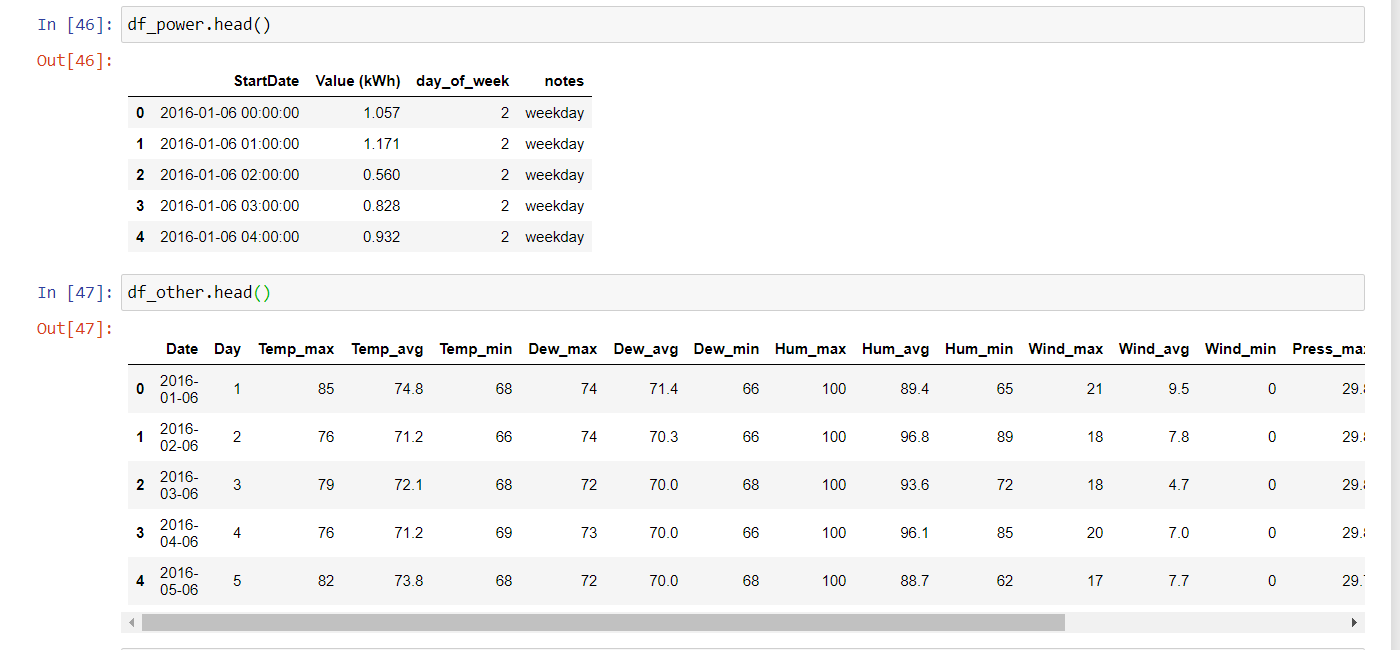
# CLEANING UP THE DATASET

Since I only need 2 attributes like time and actual values. I cleaned rest of the dataset by using python. There were many missing data in 2016 and in 2020 there was data only till August. So, I planned to remove 2016- and 2020-years’ data because of its inconsistency and then I also removed other attributes which were not necessary.

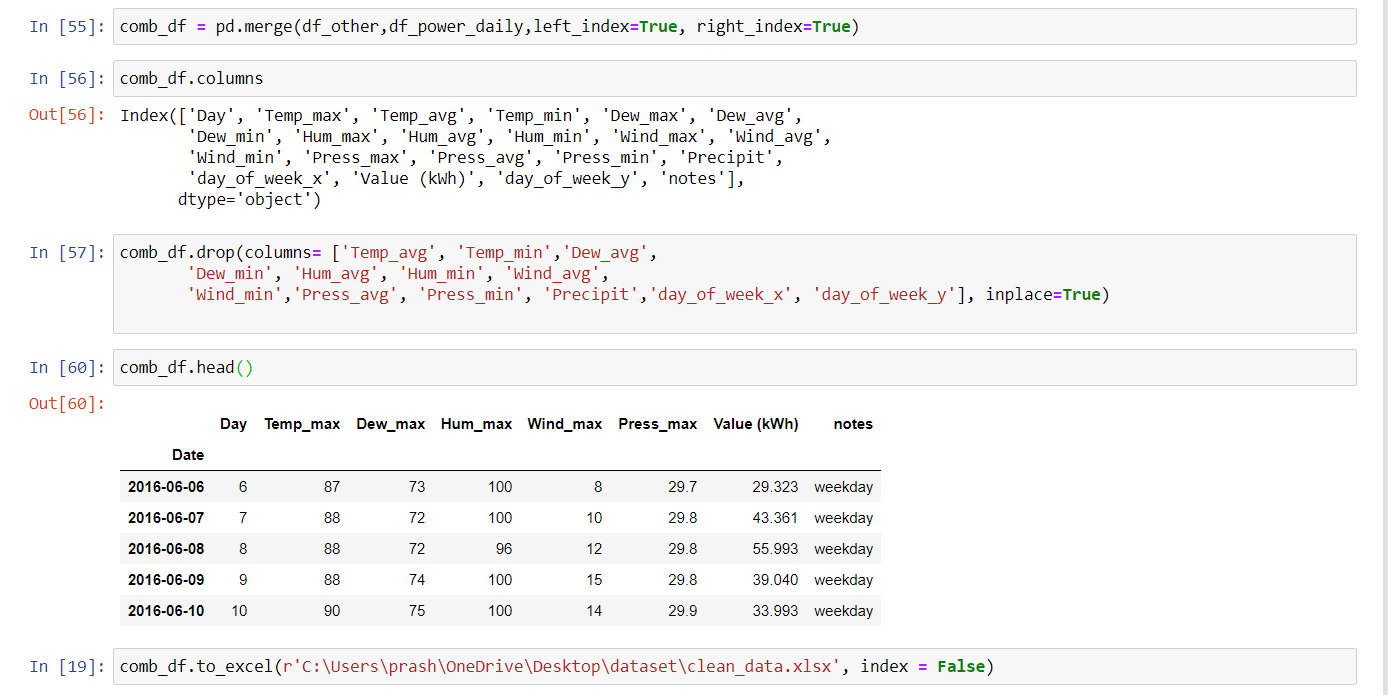
Refer to Python file Clean\_data.ipyb to see how I cleaned using two different datasets.

Below are few of the screenshots while cleaning the data.

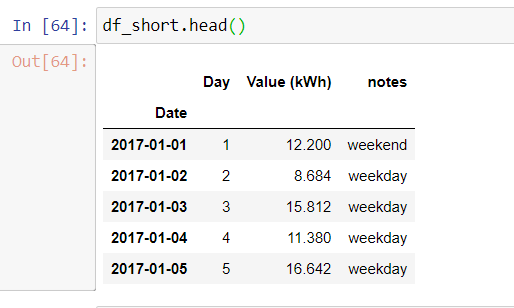
Initial two datasets before combining:



I have also updated the date format and made sure that actuals are recorded daily instead of hourly.



Final dataset:



Now that the data is ready. I used SAS to look at the plot.

First step while performing time series is to plot the data and check for trend and seasonality.

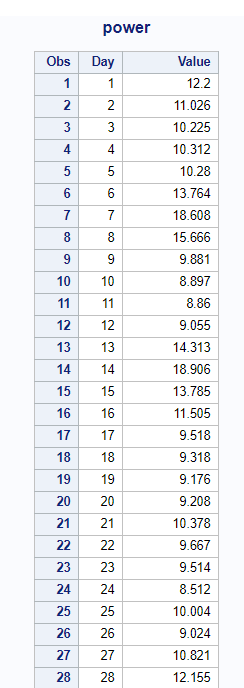
I have also made couple of changes like removing notes column and also changing attribute Value (kWh) to just ‘Value’ for easy usage while performing analysis.

# METHODOLOGY:

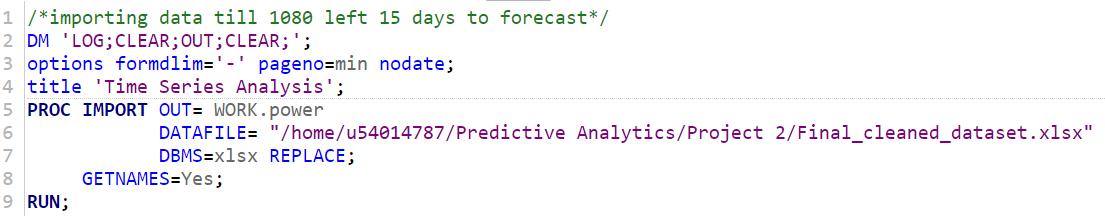
Based on the plot. Looking at the Trend and seasonality and also by performing different stationary tests like ADF tests. I have used Additive Holts winters method to perform predictions. Also, Manual and Auto ARIMA method was used.

# RESULTS and DISCUSSIONS:

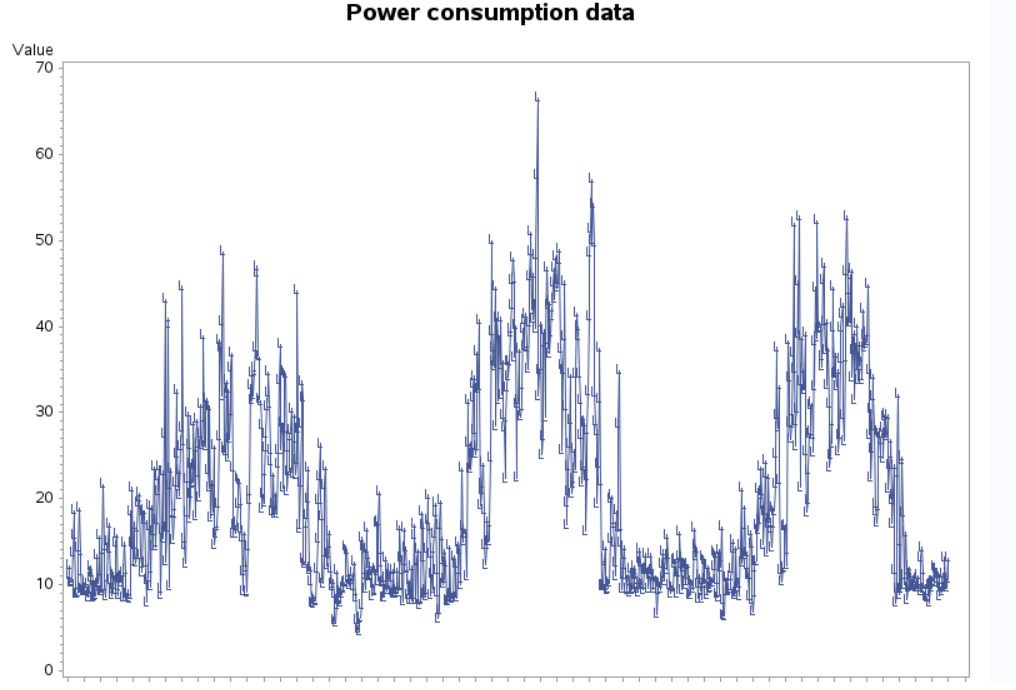
## Step 1: Importing cleaned dataset to SAS and plotting the data.



SAS code:



## Step 2: Plotting the data

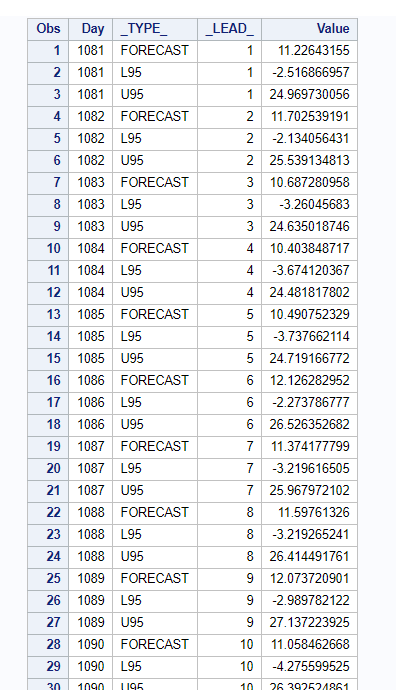


This is a daily data. As you can see that there is no trend but there is constant seasonal pattern being observed every 365 days.

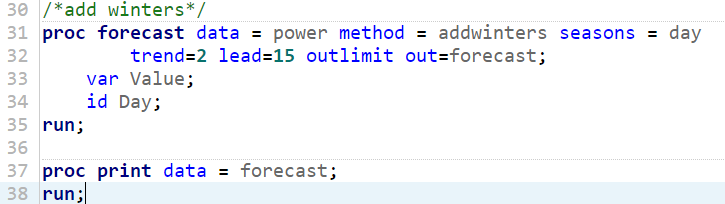
## Step 3: Perform different time series methods to forecast.

**Model 1**: *Using holt additive winters method.*

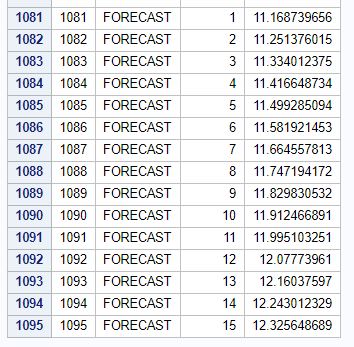
Additive because seasonality remains constant throughout.



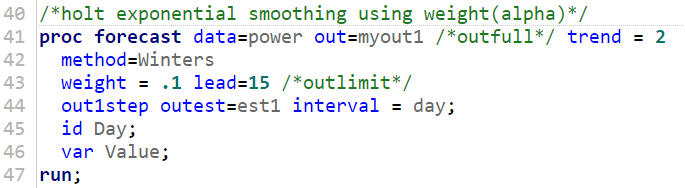
SAS code:



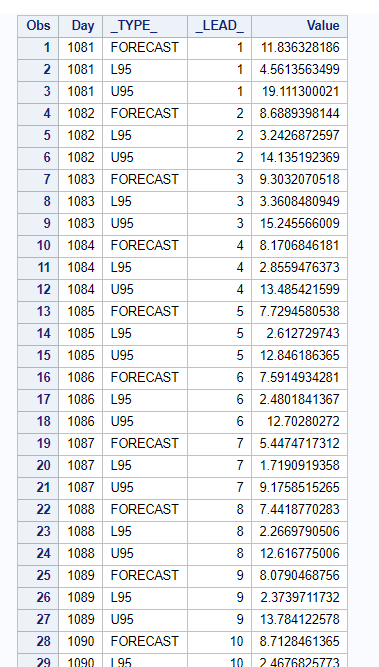
**Model 2:***Using holt exponential method with weight 0.1.*



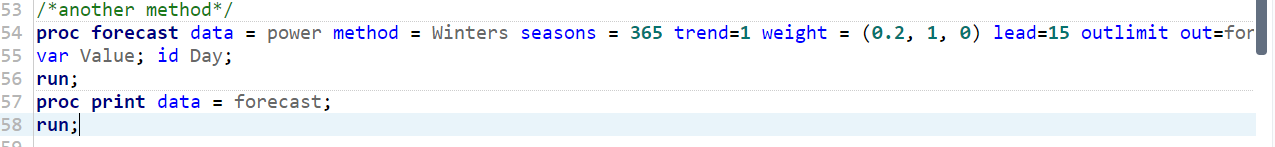
SAS code:



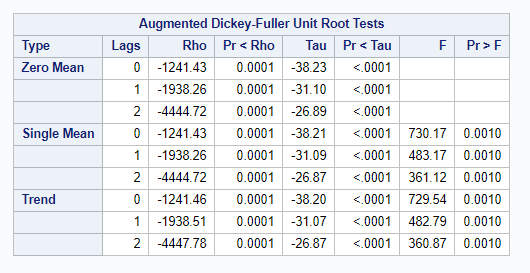
**Model 3:** *Same additive method using (0.2,1,0) values another model was run*



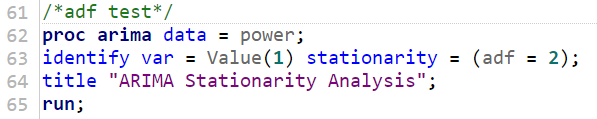
SAS code:



# ADF test:

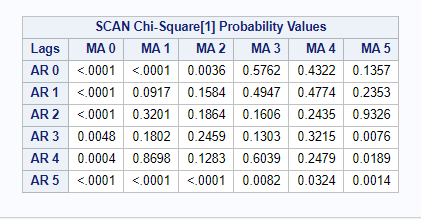


SAS code:

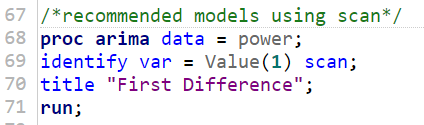


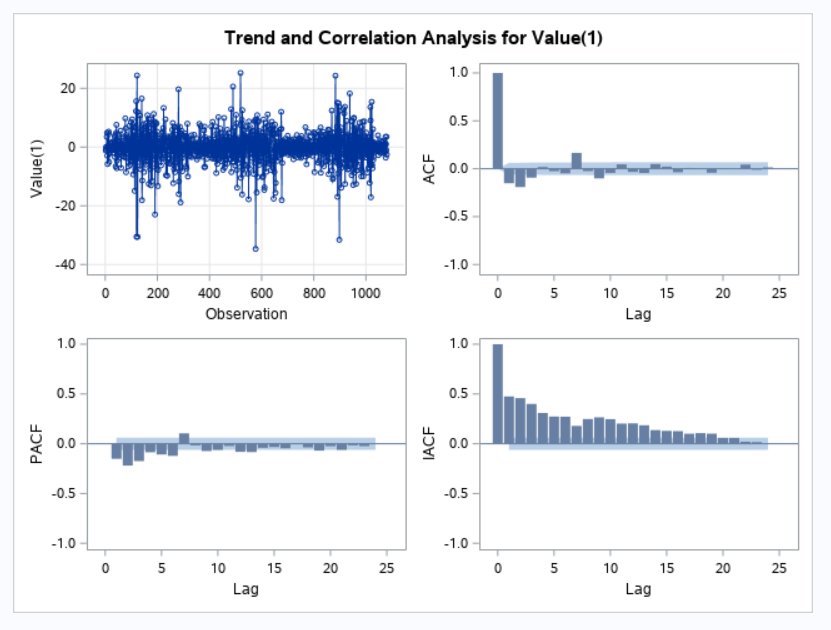
Inference: looking at the ADF test, I can see that Rho and Tau values are significant. The same ADF test is also performed using Python. **There also it stated that the data is stationary.**

**Recommended models using SCAN:**

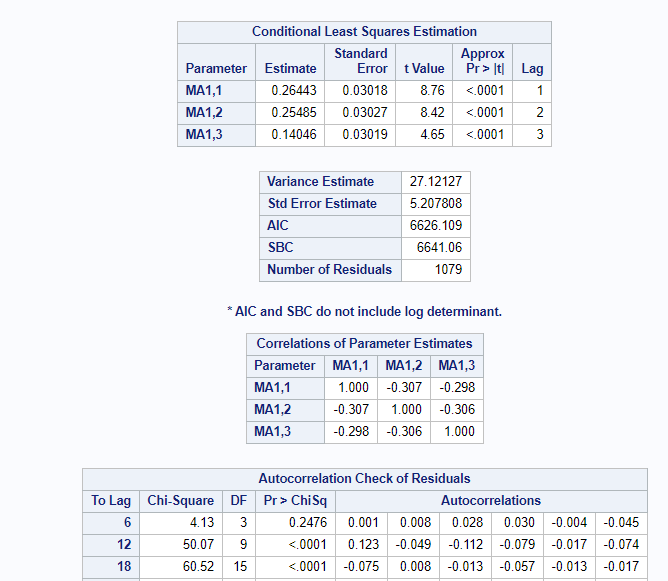


SAS code:

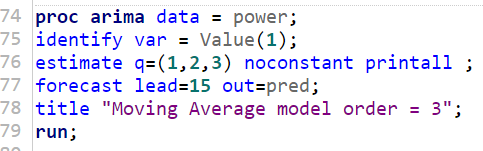




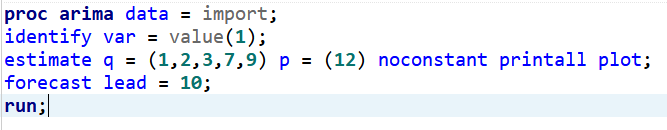
Another model:

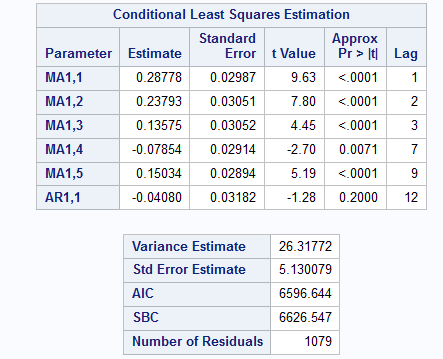


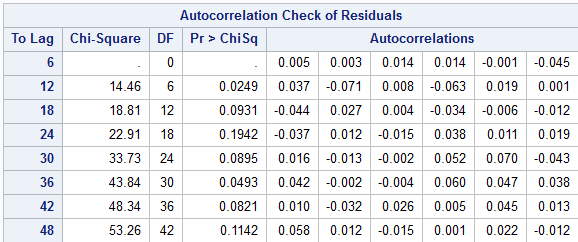
SAS Code:

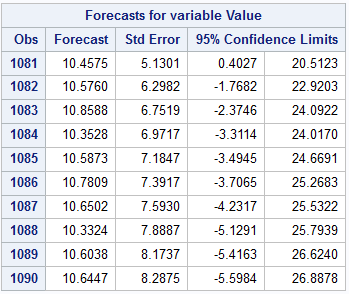


Model 4: Manual ARIMA Model by looking at the SAC and SPAC spikes







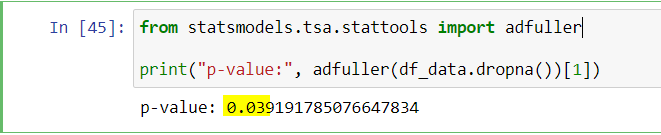


# Auto Arima in Python:

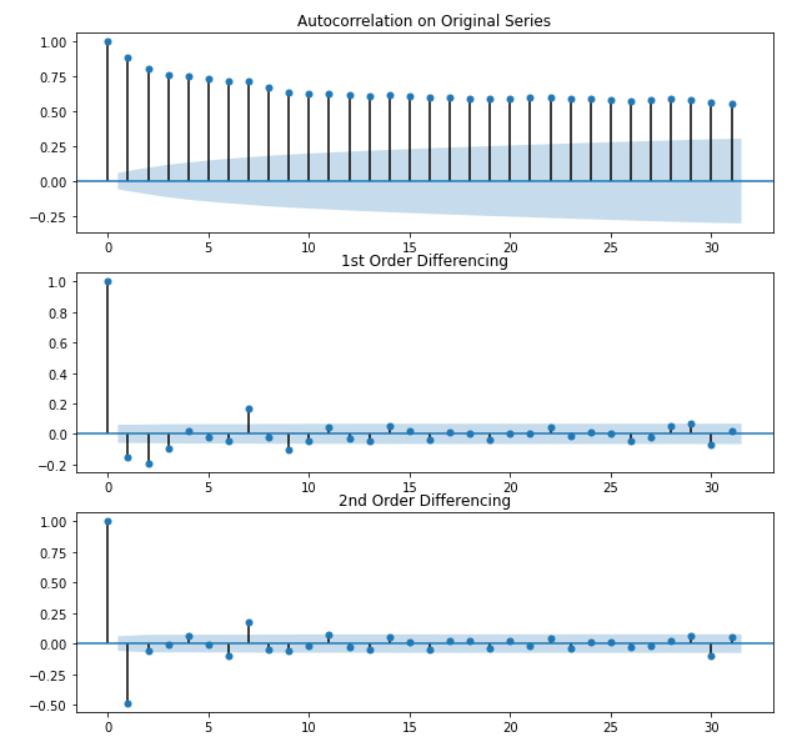
## Step 1: importing data.



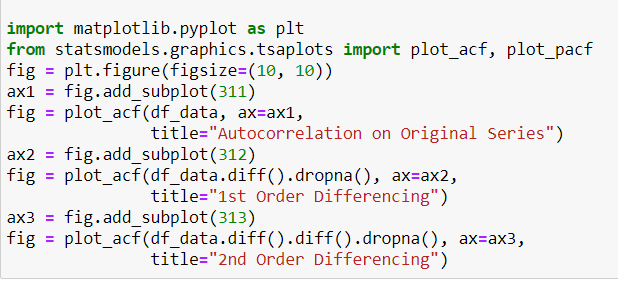
## Step 2: Checking if the data is stationary using ADF test in Python.



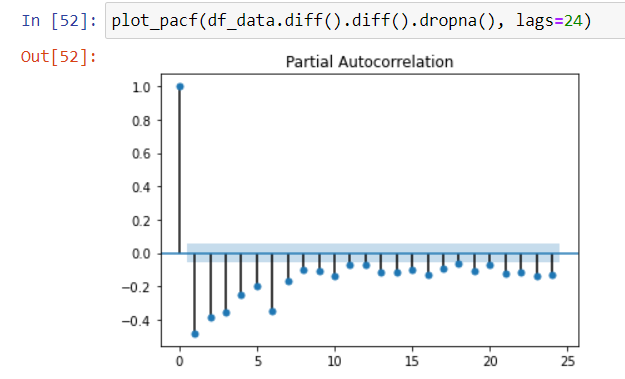
## Step 3: Performing differencing.

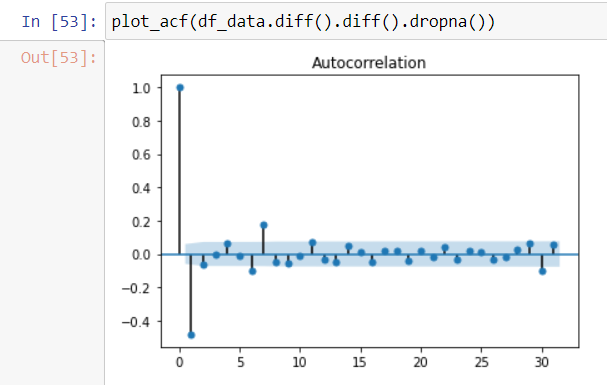


## Python code:

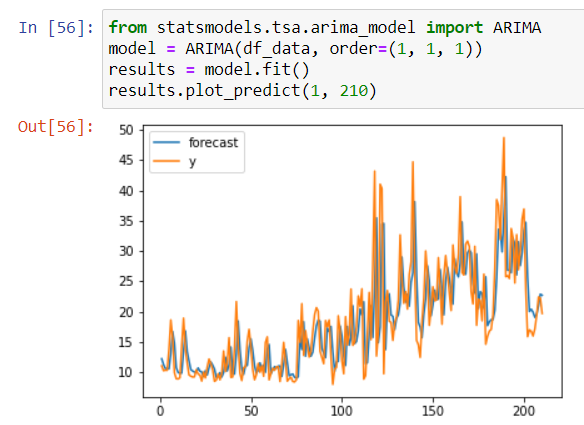


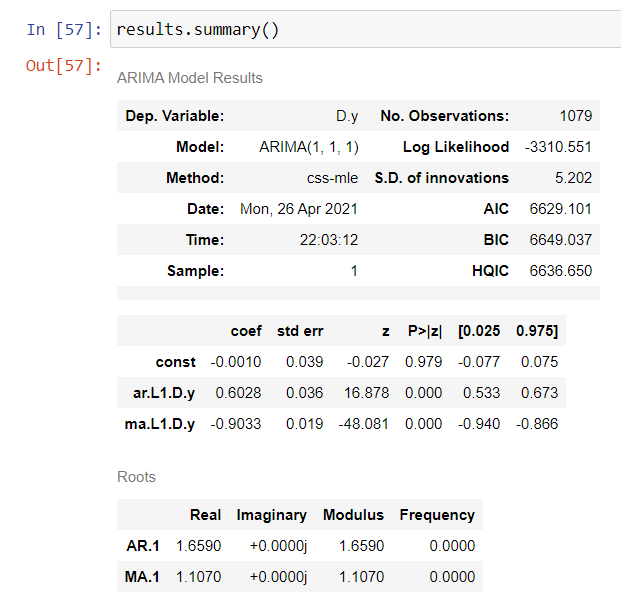
## Step 4: Checking ACF and PACF before moving to auto arima.



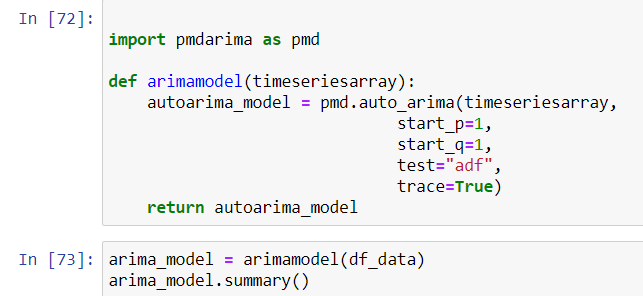


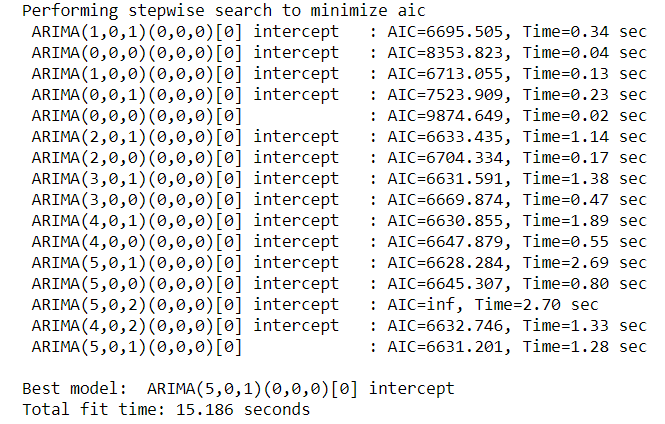
Step 5: Manual arima model run:

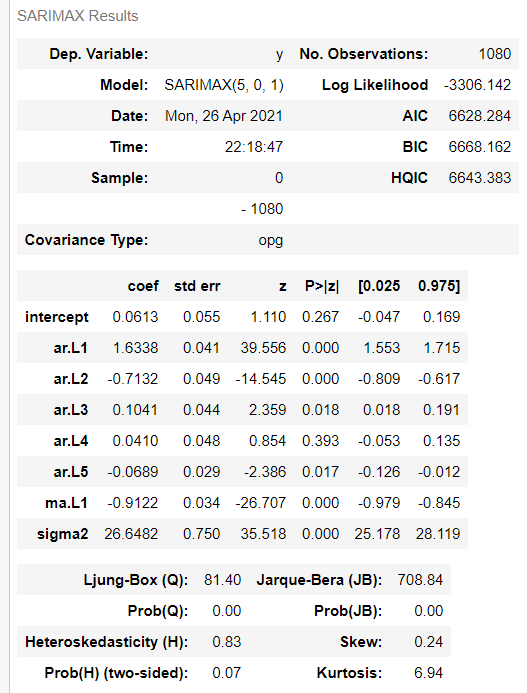




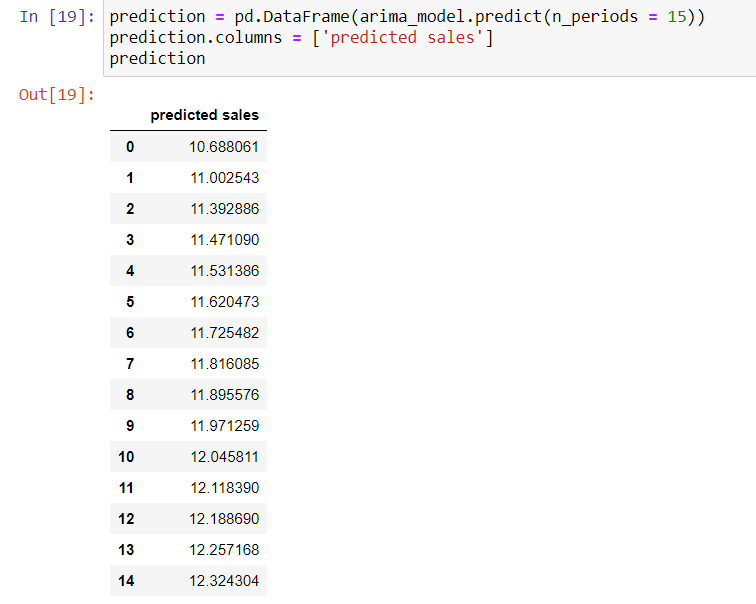
## Step 6: Auto arima model run.

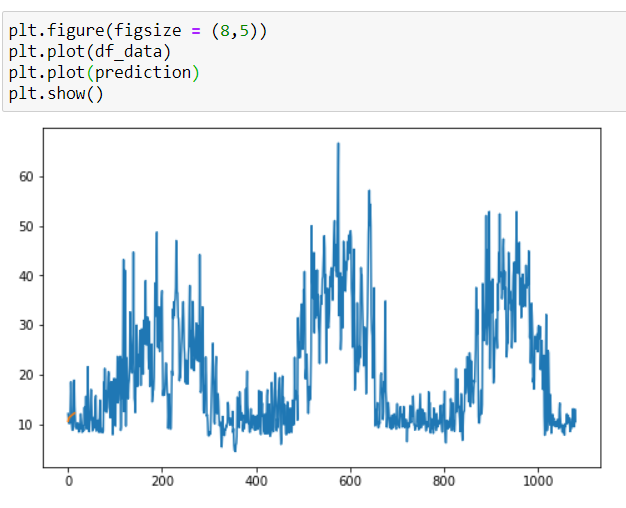






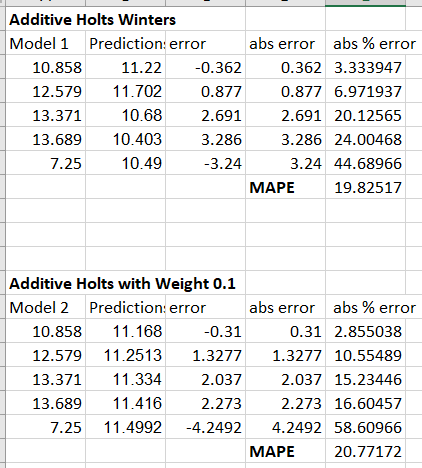
Final predicted values

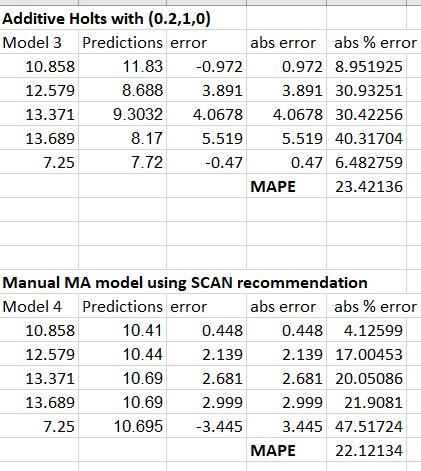


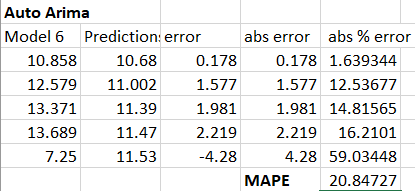


**CONCLUSION and LIMITATIONS:**

Finding MAPE for all the 5 models:







From the above all MAPE’s for all the models, I can say that Additive holts’ winters model and also Auto Arima model works very well for predicting the values for this kind of data.

Also, looking at the initial predicted values for the 4th year. I can reject null hypothesis and conclude that yes, alternate hypothesis is true. That is the future forecast values does increase with the time period.

**References:**

* Omid Motlagh, Adam Berry, Lachlan O'Neil, Clustering of residential electricity customers using load time series, Applied Energy, Volume 237, 2019, Pages 11-24, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2018.12.063>.
* H. M. Al-Hamadi, "Long-term electric power load forecasting using fuzzy linear regression technique," 2011 IEEE Power Engineering and Automation Conference, 2011, pp. 96-99, doi: 10.1109/PEAM.2011.6135023.
* Ignacio Benítez, José-Luis Díez, Alfredo Quijano, Ignacio Delgado, Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance,

Electric Power Systems Research, Volume 140, 2016, Pages 517-526, ISSN 03787796,

<https://doi.org/10.1016/j.epsr.2016.05.023>.

* Almazrouee AI, Almeshal AM, Almutairi AS, Alenezi MR, Alhajeri SN. Long-Term Forecasting of Electrical Loads in Kuwait Using Prophet and Holt–Winters Models. Applied Sciences. 2020; 10(16): 5627, https://doi.org/10.3390/app10165627
* Omer Berat Sezer, Mehmet Ugur Gudelek, Ahmet Murat Ozbayoglu, Financial time series forecasting with deep learning: A systematic literature review: 2005–2019, Applied Soft Computing, Volume 90, 2020, 106181, ISSN 1568-4946, <https://doi.org/10.1016/j.asoc.2020.106181>.
* Alaa Sagheer, Mostafa Kotb, Time series forecasting of petroleum production using deep LSTM recurrent networks, Neurocomputing, Volume 323, 2019, Pages 203-213, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2018.09.082>.
* Kaggle. *Residential Power Usage 3years data – Time Series*. <https://www.kaggle.com/srinuti/residential-power-usage-3years-data-timeseries>