**ANALYSIS OF WISCONSIN PERSONAL INCOME DATASET**

**USING TIME SERIES**

**(XL MINER)**

**ISM 6136 – DATA MINING AND PREDICTIVE ANALYSIS**

**CLASS ASSIGNMENT 10**

**MIEBAKA TOBY**

**Z23472301**

**INTRODUCTION:**

The dataset looks at the average personal income for several states per year from 1929 -1999. 5 states were selected to perform time series analysis and forecasting using xlminer. The states are California (CA), District of Columbia (DC), New York (NY), California (CA) and Florida (FL)

**DATASET:**

The dataset was split into 60:40 ratio for training and test data respectively. There were no missing data in the dataset.

**EXPERIMENT SETUP:**

Each state had three models as:

Model 1: p = 2, q = 1, d= 0

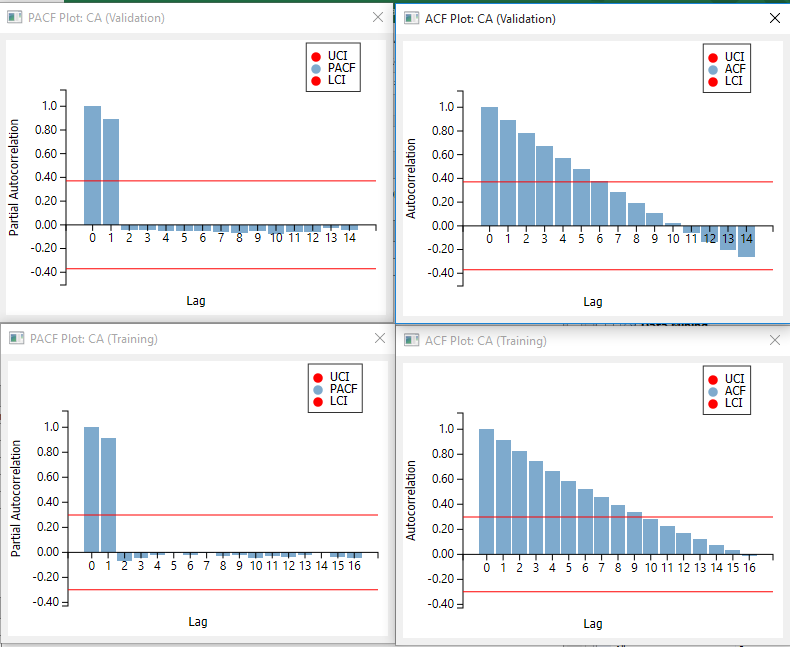
Model 2: p =2, q= 2, d= 1

Model 3: p =2, q=2, d=2

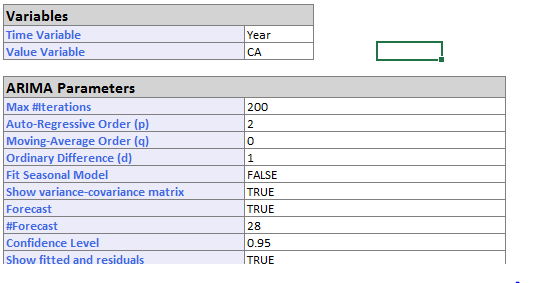
The best model was selected and a comparison was then done amongst the top model for each state

CALIFORNIA (CA)

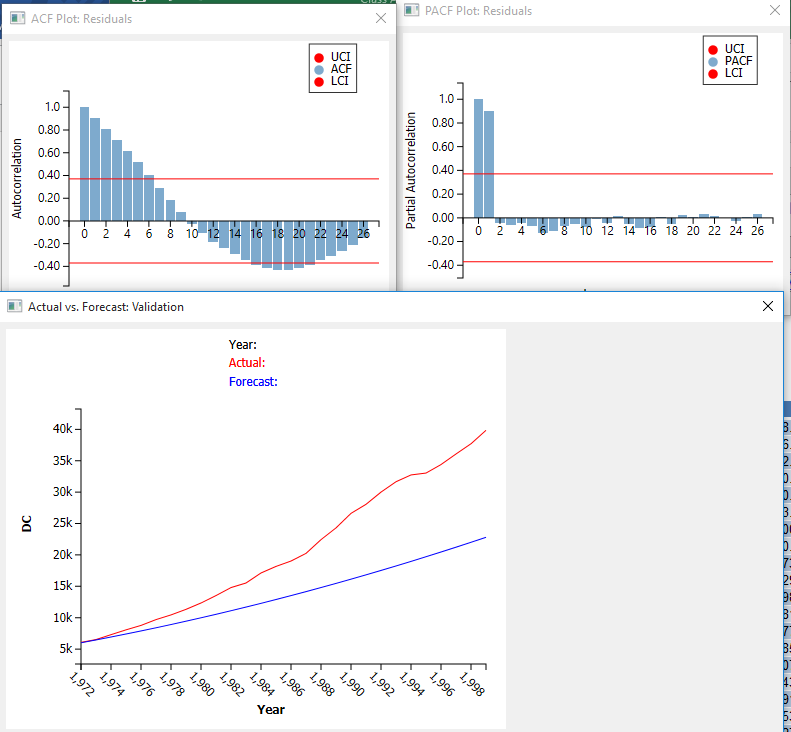
**Lagging Analysis:**



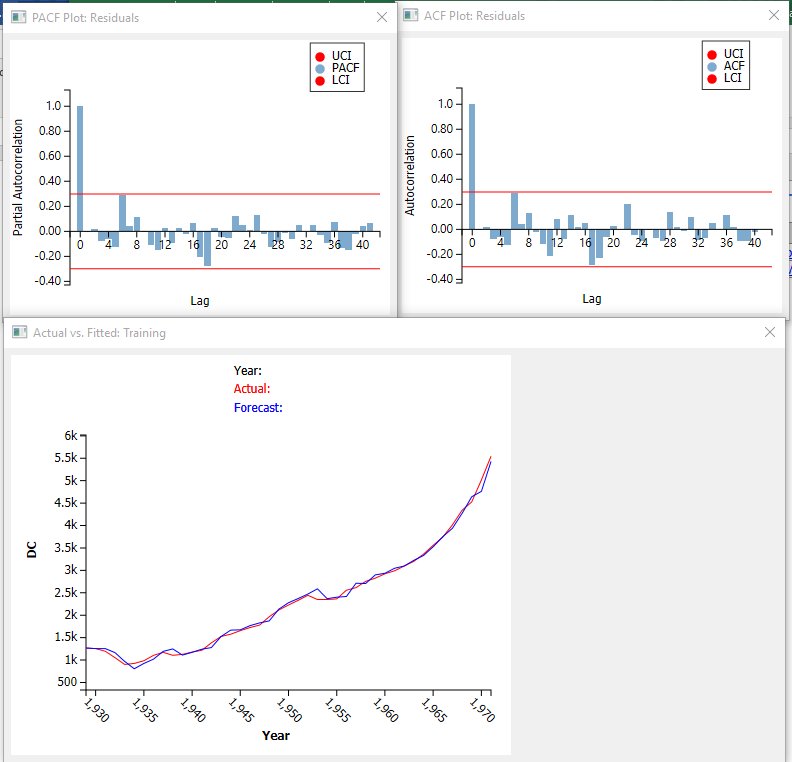
**Arima Analysis**



**Actual Arima Analysis**



**Forecasted Arima Analysis**



**Arima Performance Metrics**

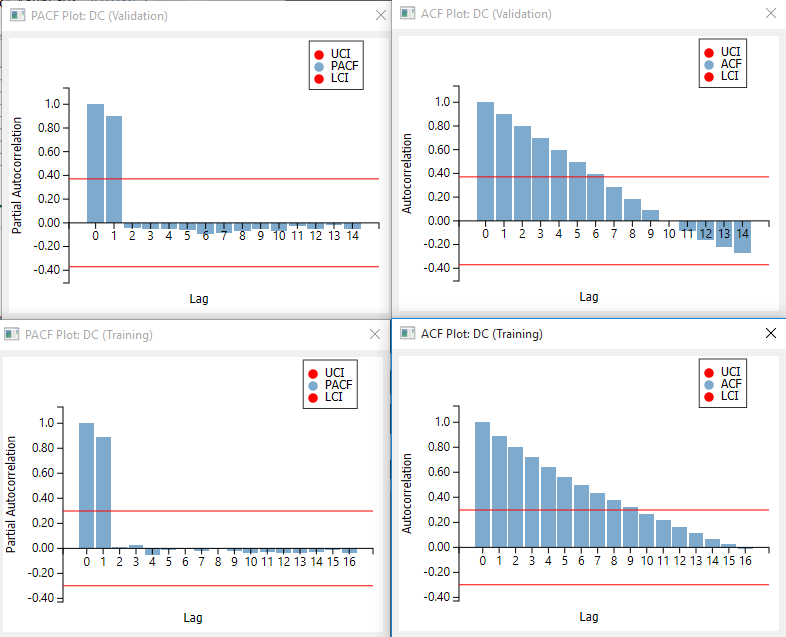
|  |  |
| --- | --- |
|  |  |

Result:

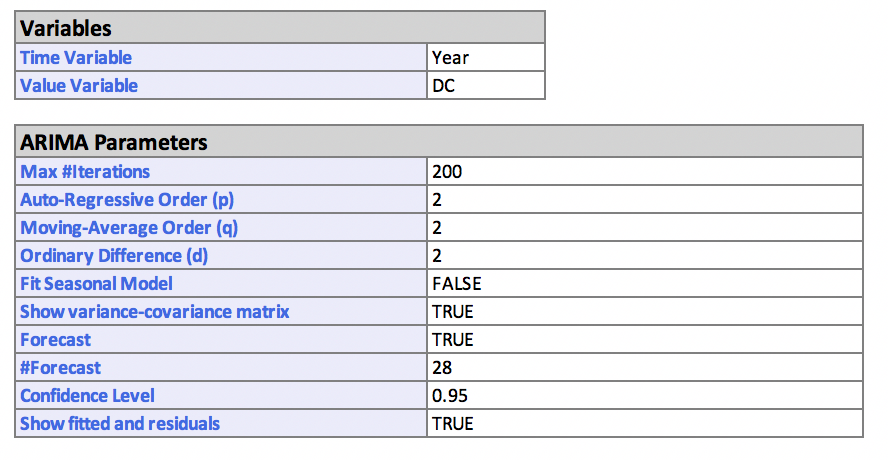
The lag graph shows there is a correlation in the data. There isn’t a similarity between the validation and training data and therefore the model may not fit the training data. There is a downward trend as the lags increase. The arima shows correlation between the series and the lag. However, the forcast was not good.

DISTRICT OF COLUMBIA (DC)

**Lagging Analysis:**



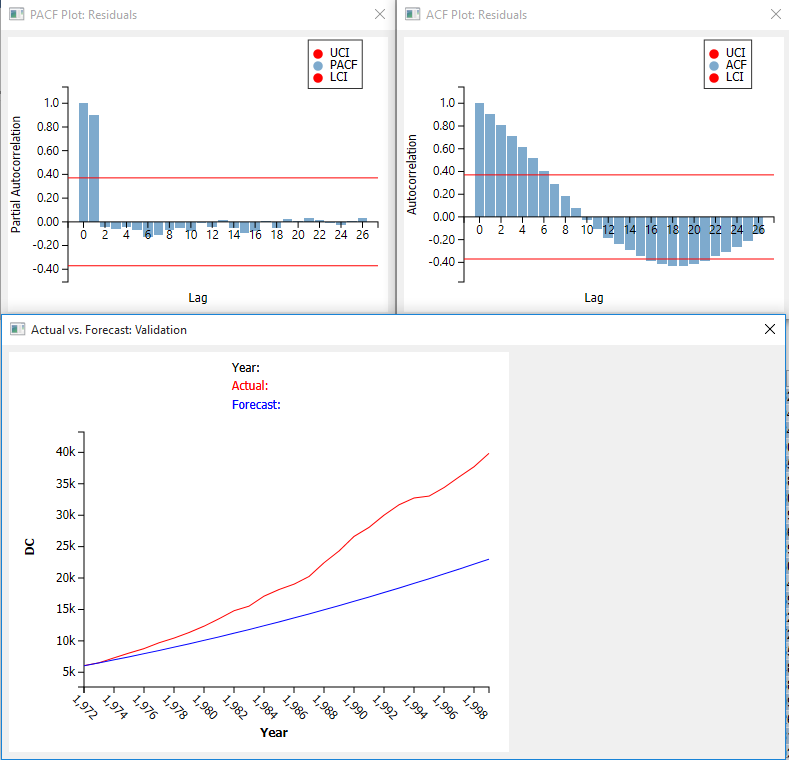
**Arima Model**



**Actual Arima**



Forecasted Arima



**Arima Performance Metrics**

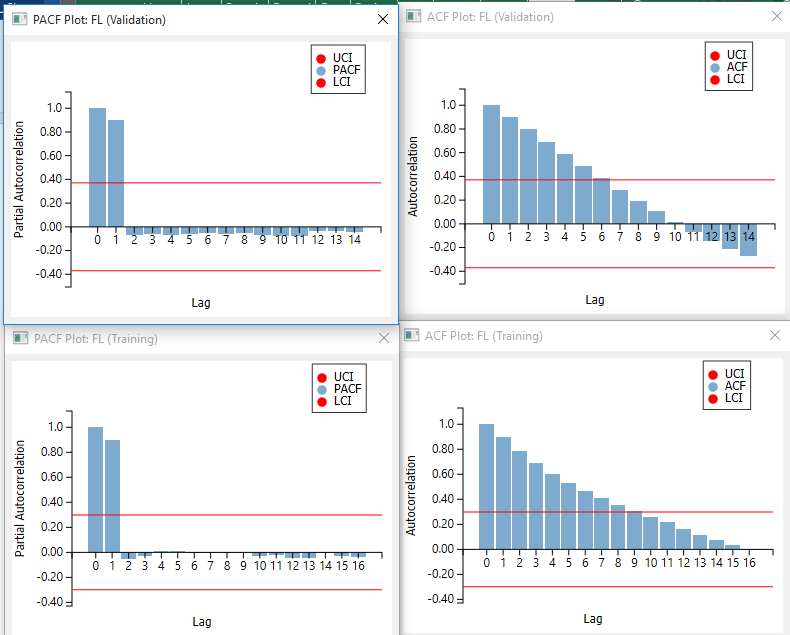
|  |  |
| --- | --- |
|  |  |

Result:

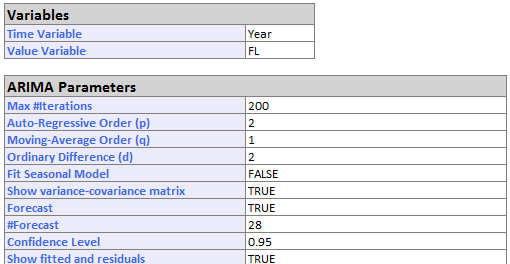
Model 3 was the best out of the three models. The lag graph shows there is a correlation in the data. There isn’t a similarity between the validation and training data and therefore the model may not fit the training data. There is a downward trend as the lags increase. The arima shows correlation between the series and the lag. However, the forcast was not good.

FLORIDA (FL)

**Lagging Analysis:**



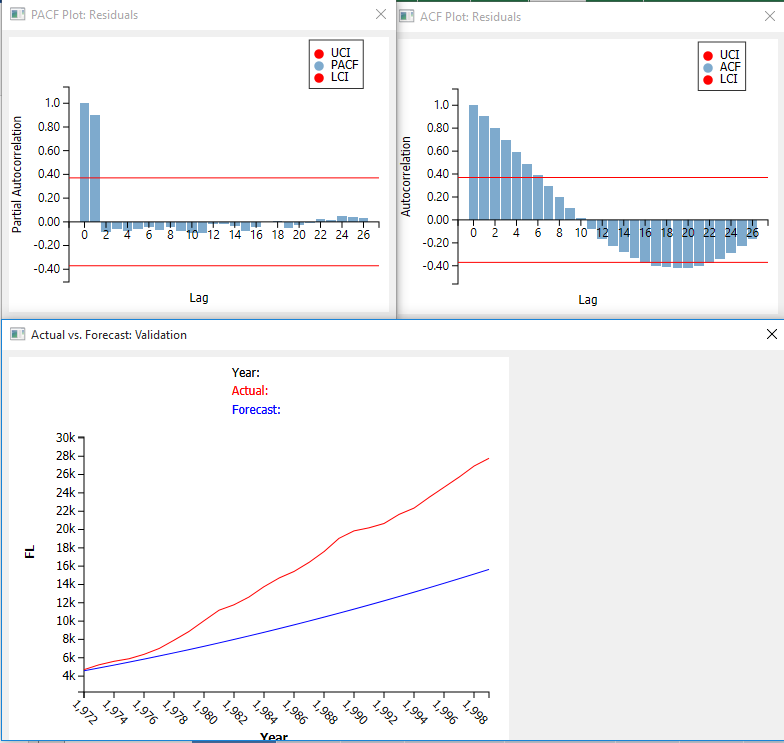
**Arima Model**



**Actual Arima**



**Forecasted Arima**



**Arima Performance Metrics**

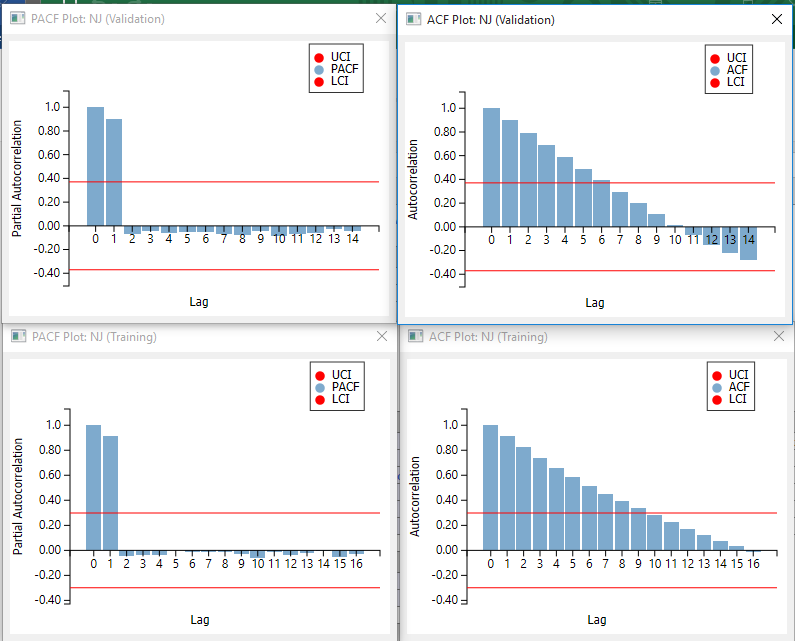
|  |  |
| --- | --- |
|  |  |

Result:

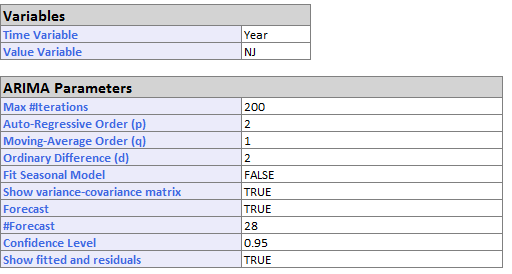
Model 2 was the best model from the three models. The lag graph shows there is a correlation in the data. There isn’t a similarity between the validation and training data and therefore the model may not fit the training data. There is a downward trend as the lags increase. The arima shows correlation between the series and the lag. However, the forcast was not good.

NEW JERSEY (NJ)

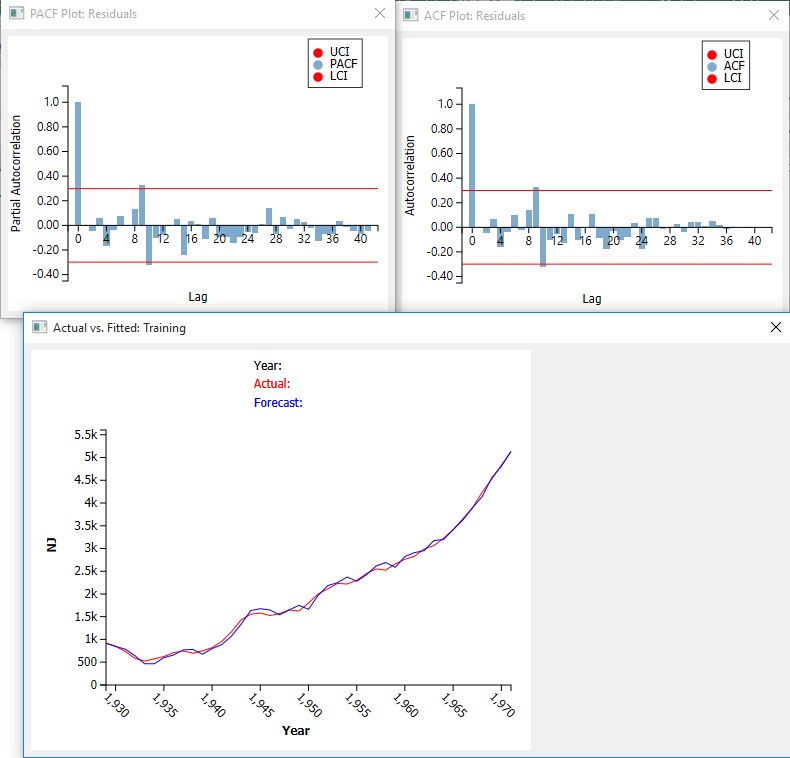
**Lagging Analysis**



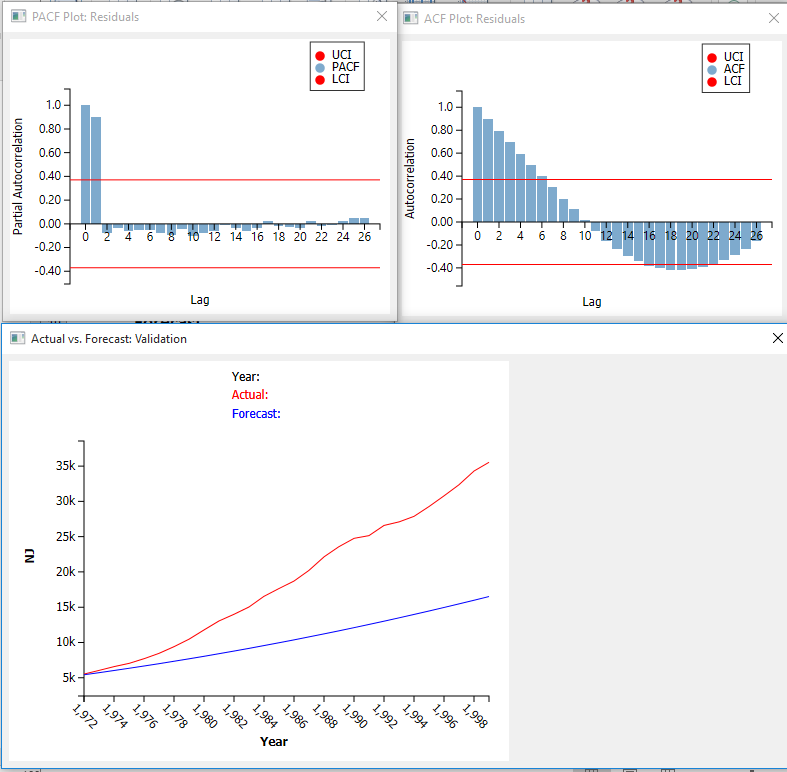
**Arima Model**



**Actual Arima**



**Forecasted Arima**



**Arima Performance Metrics**

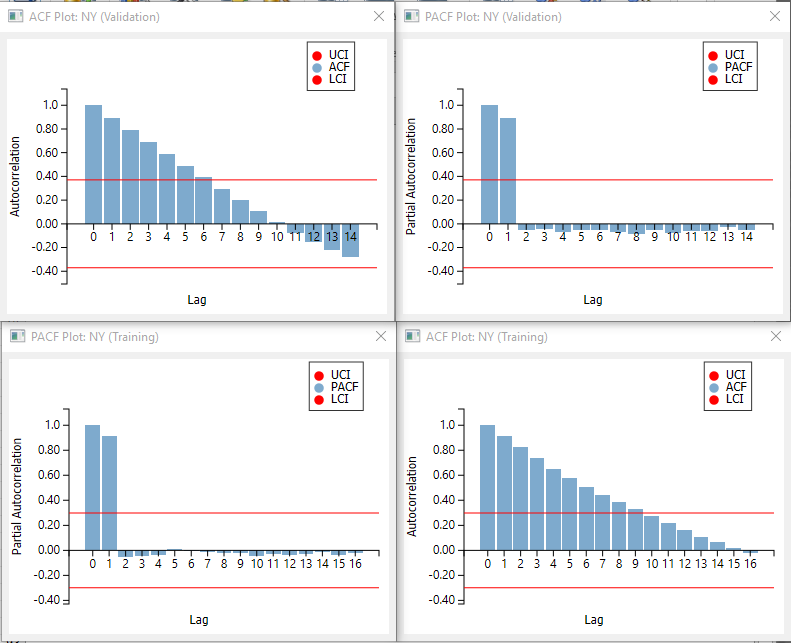
|  |  |
| --- | --- |
|  |  |

Result:

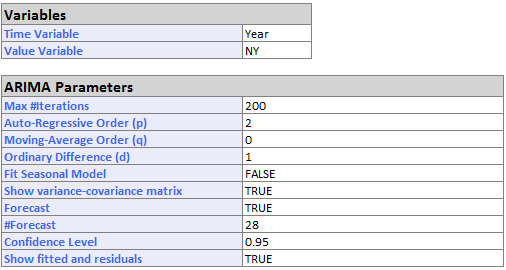
Model 2 was the best model out of the three models. The lag graph shows there is a correlation in the data. There isn’t a similarity between the validation and training data and therefore the model may not fit the training data. There is a downward trend as the lags increase. The arima shows correlation between the series and the lag. However, the forcast was not good.

NEW YORK (NY)

**Lagging Analysis**



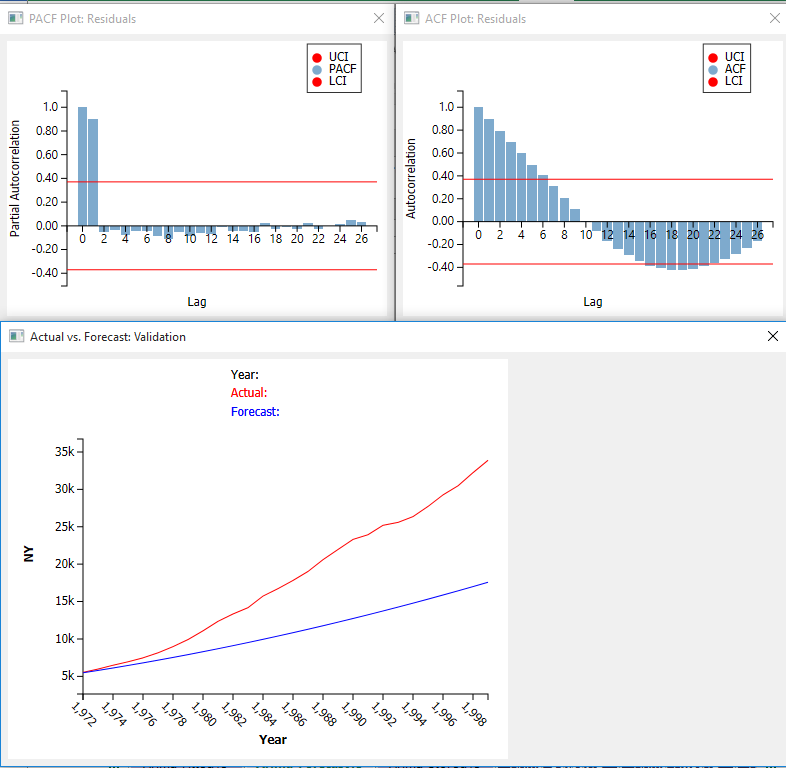
**Arima Analysis**



**Actual Arima**



**Forecasted Arima**



**Arima Performance Metrics**

|  |  |
| --- | --- |
|  |  |

Result

NY: Model 2

The lag graph shows there is a correlation in the data. There isn’t a similarity between the validation and training data and therefore the model may not fit the training data. There is a downward trend as the lags increase. The arima shows correlation between the series and the lag. However, the forcast was not good.

Comparison between forecasted models for 1999

The forcast was far from the actual value . This was identified in the lag where the graphs for validation was not similar for training showing the model could be good for the training dataset but would not perform well on the test dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STATE | YEAR | ACTUAL | FORCAST | STD DEV | LCI | UCI | RESIDUAL |
| CA | 1999 | 29910 | 15291.44095 | 82.95417512 | 15128.85 | 15454.03 | 14618.5591 |
| DC | 1999 | 39858 | 23008.66819 | 88.215878 | 22835.77 | 23181.57 | 16849.3318 |
| FL | 1999 | 27780 | 15650.66034 | 59.79370528 | 15533.47 | 15767.85 | 12129.3397 |
| NJ | 1999 | 35551 | 16444.94115 | 81.60369215 | 16285 | 16604.88 | 19106.0588 |
| NY | 1999 | 33890 | 17585.21421 | 66.68791482 | 17454.51 | 17715.92 | 16304.7858 |