Capstone Project

Employment by Sector Forecast

By: Divendra Chand Raj (Data Scientist/Data Analyst/Business Intelligence Analyst)

Date: 15/04/2023

Table of Content

Business Background

Government, some relevant business/companies, and some career building employee relay on how individual sector is performing example Primary Industry, Manufacturing Industry and Financial sector.

They are involved in either investment or diverting their career to earn more money.

The movement of workers can be a huge impact on the industry as whole. Some sectors can really struggle to find workers as some would be overwhelmed. Recently government are opening boards to bring in workers from overseas to meet the demand for workers in various sector.

If these predictions are done in time than it could have prevented lots of problems, and even start planning to face the challenges by either allocating or diverting more funds in the sector to resolve the issue. Also, for the business to avoid where they will struggle to find workers and save from potential losses.

In New Zealand employment rate is 69.30 percent (2.85 million) and unemployment rate is 3.40 percent (over 90,000) as of Dec 2022. (Sourced from Stats New Zealand)

The government can save millions of dollars if they can engage unemployed people to such areas instead of just paying social benefit if any cannot find suitable jobs. Business would save millions of dollars investing in setting up business where workers are scarce.

Business Main Objective

In this occasion, stats New Zealand collects data and using the data I must build a precise model that can forecast movement of workers for the different business sectors and save millions of dollars.

Data Summary

To build my model, the data was sourced from Stats New Zealand which gave me monthly employment figures from various sectors. These data were actual and were based on employment figures from May 2019 to Feb 2023. The dataset was big however from the data only employment by individual sector were extracted to build the model.

Original data contained lots of information example employment by age, region, sex, full time paid, part time paid, contract and by Industry.

The dataset is publicly available with all kinds of information who wish or have interest in building model of various related fields.

Data cleaning

The data was clean and did not contain any null or missing values which was used to build the model.

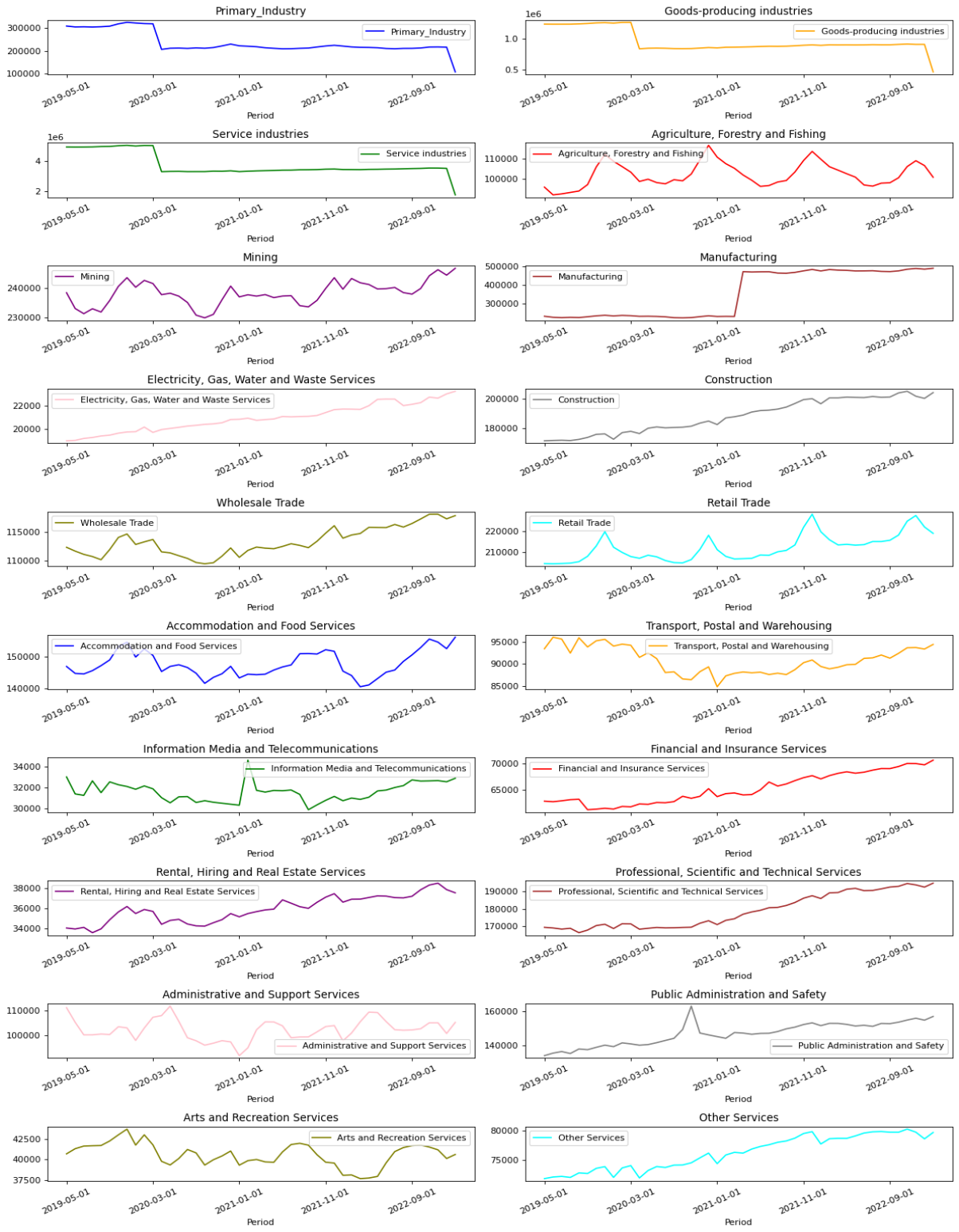
Exploratory Data Analysis

The dataset was very clean and did not contain any null values but since all the data were compiled together data extraction was bit of challenge.

Data Analysis

Finally, got a nice line graph which explains the performance of each sector. From the analysis we can see first 4 sectors employment numbers is falling at a high rate when compared with others.

Of course, that can be an alarming but let's build a model to find out.



Building Model

New Zealand has 4 different seasons so building a Times Series Model using ARIMA/SARIMA would be the best option to see if season has any effect.

Randomly used Mining sector data which has minor effect of seasons from the 20 sectors.

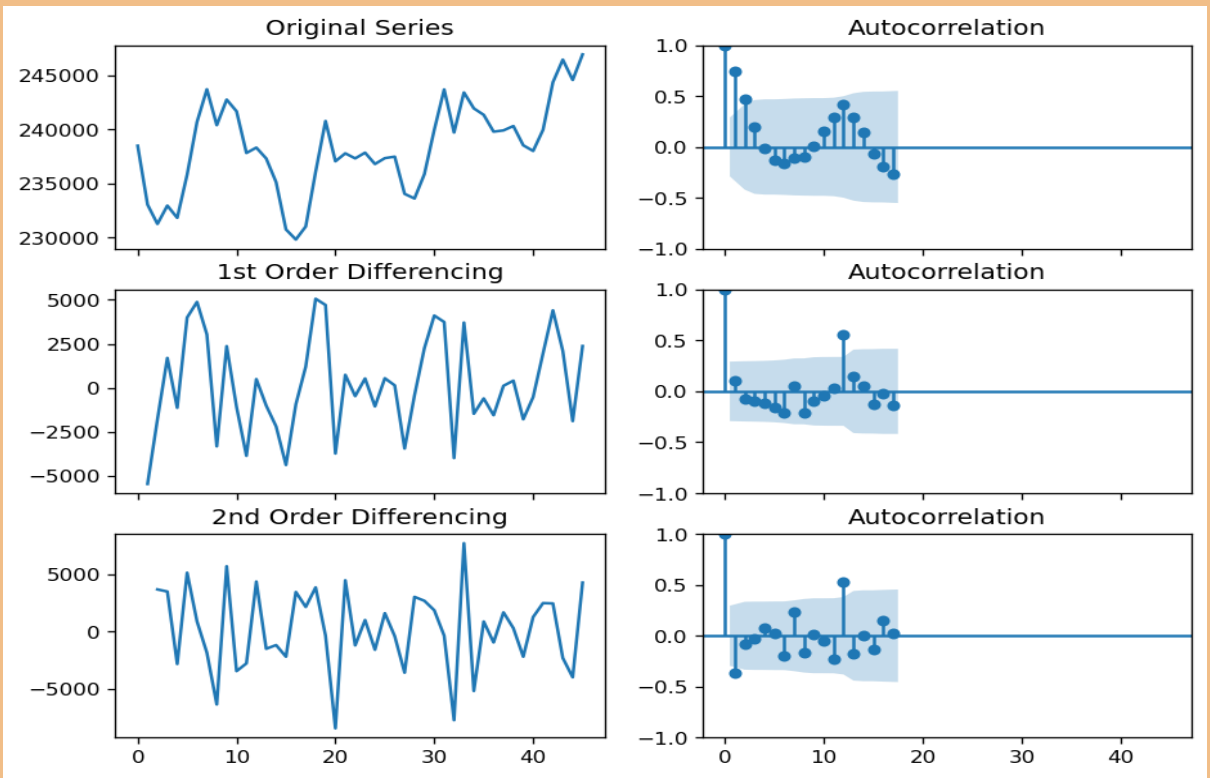
Checking the stationarity of the data

First, I checked the stationary of the data using adfuller. The result was not great, the p-value was great than 0.05 hence the data was not stationary.

```
ADF Statistic: %f : -1.7815040802192887
p-value: %f : 0.3896822937055069
#lags Used : 0
Number of Observation Used : 45
Weak evidence, indicating it is non-staionary
```

```
From the above result P- Value is greater than 0.05 we can say that the data is not stationary hence we going to do order of difference
```

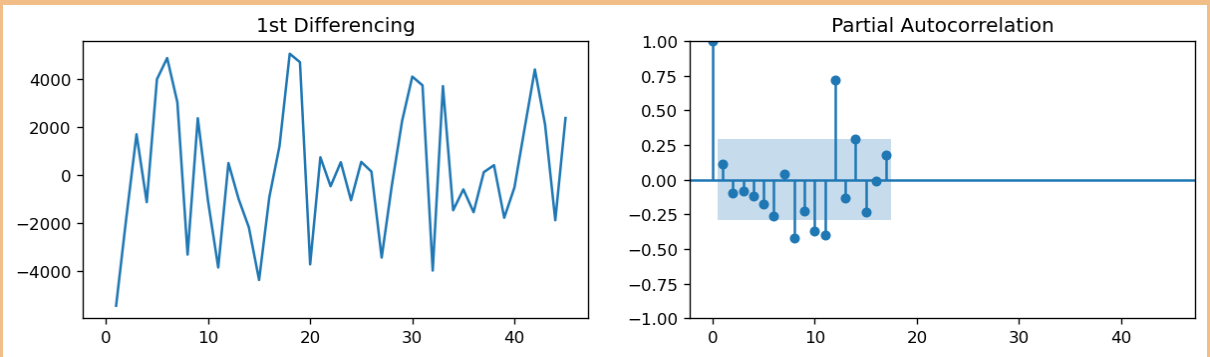
To stationary the data I use the order of differencing tactic and the outcome was what I was looking for. First order of differencing did the work.



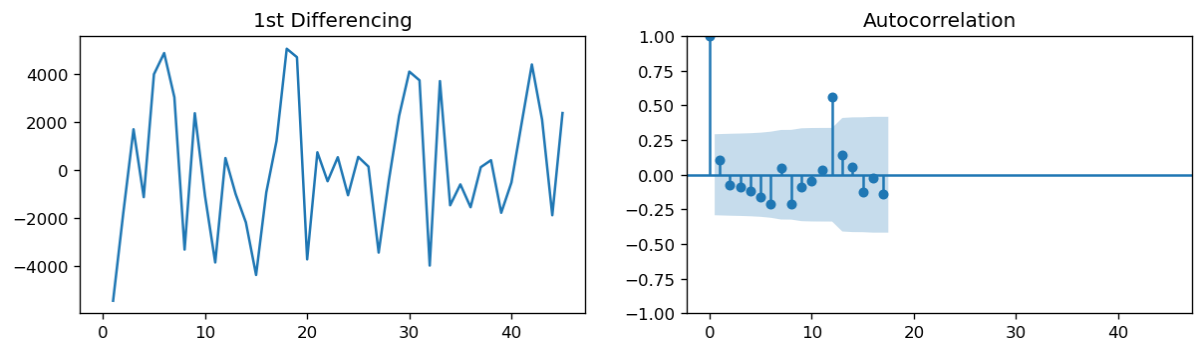
Finding p, d, and q

After finding p, d, q which was (1,1,1), I was ready to build the ARIMA model.

Plotting (PACF) plot to see AR term(p)



Plotting (ACF) plot to see AR term(q)



Also used itertool to get the p, d and q orders.

```
ARIMA(1, 0, 1)x(0, 1, 0, 12)12 - AIC:575.4989239829151
ARIMA(1, 0, 1)x(0, 1, 1, 12)12 - AIC:345.1608413919955
ARIMA(1, 0, 1)x(1, 0, 0, 12)12 - AIC:600.6411062774661
ARIMA(1, 0, 1)x(1, 0, 1, 12)12 - AIC:584.7898378005713
ARIMA(1, 0, 1)x(1, 1, 0, 12)12 - AIC:365.82376322045343
ARIMA(1, 0, 1)x(1, 1, 1, 12)12 - AIC:348.24256693251453
ARIMA(1, 1, 0)x(0, 0, 0, 12)12 - AIC:821.040983830139
ARIMA(1, 1, 0)x(0, 0, 1, 12)12 - AIC:1252.6997333978984
ARIMA(1, 1, 0)x(0, 1, 0, 12)12 - AIC:575.544202842756
ARIMA(1, 1, 0)x(0, 1, 1, 12)12 - AIC:347.52933792077226
ARIMA(1, 1, 0)x(1, 0, 0, 12)12 - AIC:570.0400763483449
ARIMA(1, 1, 0)x(1, 0, 1, 12)12 - AIC:2617.9674028028217
ARIMA(1, 1, 0)x(1, 1, 0, 12)12 - AIC:345.2974442185203
ARIMA(1, 1, 0)x(1, 1, 1, 12)12 - AIC:346.8631934789594
ARIMA(1, 1, 1)x(0, 0, 0, 12)12 - AIC:804.9776173693803
ARIMA(1, 1, 1)x(0, 0, 1, 12)12 - AIC:1392.6875766477024
ARIMA(1, 1, 1)x(0, 1, 0, 12)12 - AIC:548.7398239204156
ARIMA(1, 1, 1)x(0, 1, 1, 12)12 - AIC:328.56124144363844
ARIMA(1, 1, 1)x(1, 0, 0, 12)12 - AIC:581.798686335605
ARIMA(1, 1, 1)x(1, 0, 1, 12)12 - AIC:1574.917510046107
ARIMA(1, 1, 1)x(1, 1, 0, 12)12 - AIC:344.25669759620524
ARIMA(1, 1, 1)x(1, 1, 1, 12)12 - AIC:330.1604751668159
```

Once data was stationary and got the orders it was time to build the ARIMA model and fit the raw data. The result was good. P-Value was less than 0.05

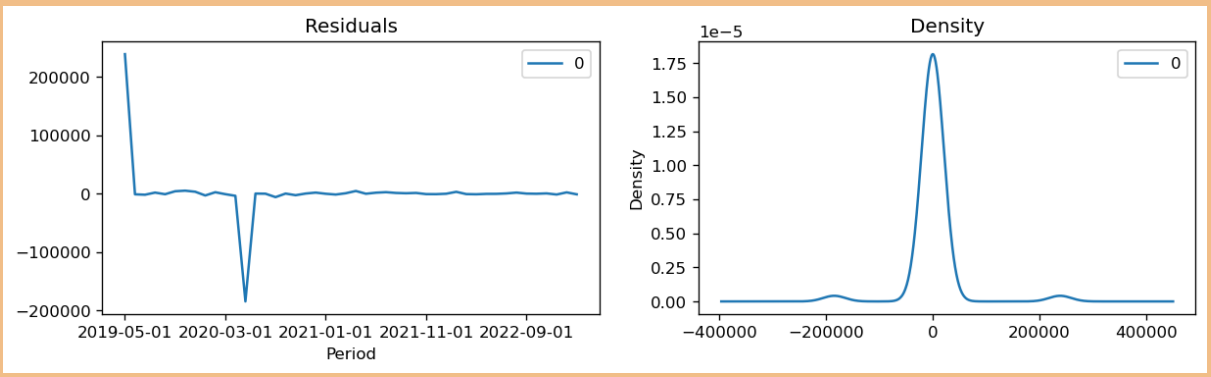
SARIMAX Results

Dep. Variable:	Mining	No. Observations:	46
Model:	ARIMA(1, 1, 1)	Log Likelihood	-417.639
Date:	Thu, 13 Apr 2023	AIC	841.278
Time:	23:05:49	BIC	846.698
Sample:	05-01-2019	HQIC	843.299
	- 02-01-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2040	1.448	0.141	0.888	-2.635	3.043
ma.L1	-0.2297	1.377	-0.167	0.868	-2.929	2.470
sigma2	7.124e+06	2.15e+06	3.318	0.001	2.92e+06	1.13e+07

Ljung-Box (L1) (Q):	0.58	Jarque-Bera (JB):	1.40
Prob(Q):	0.44	Prob(JB):	0.50
Heteroskedasticity (H):	0.78	Skew:	0.16
Prob(H) (two-sided):	0.64	Kurtosis:	2.20

The residual errors were fine with mostly near zero and uniform variance.



Next, it was time to split the data and use training data to see the outcome.

```
# Create Training and Test
train = df.Mining.values[:25]
test = df.Mining.values[25:]
```

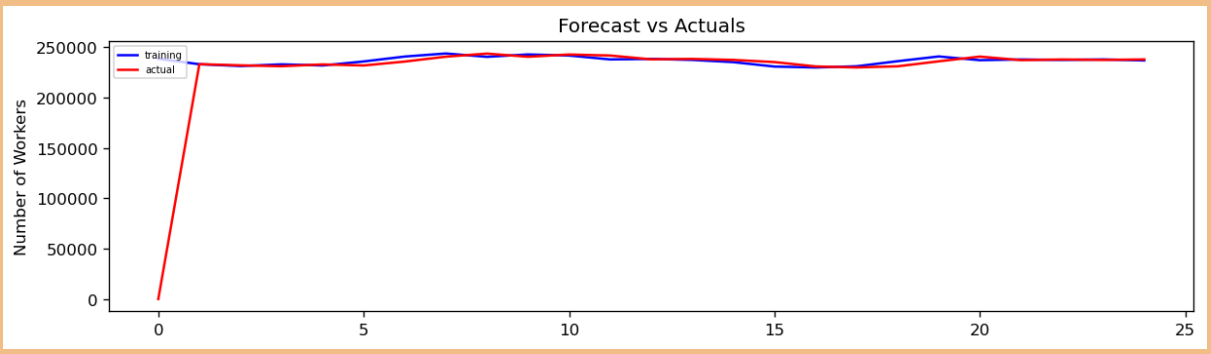
Model Result and Interpretation

The model was built on a training dataset 25 months and the result was great.

```
Plotting ARIMA Model
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   25
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -224.065
Date:                Thu, 13 Apr 2023   AIC                            454.130
Time:                        22:48:06   BIC                            457.664
Sample:                             0   HQIC                           455.067
                                  - 25
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1994      1.784      0.112      0.911      -3.297       3.695
ma.L1         -0.2247      1.685     -0.133      0.894      -3.527       3.078
sigma2      8.298e+06   4.99e-07   1.66e+13      0.000     8.3e+06     8.3e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.98   Jarque-Bera (JB):                 0.85
Prob(Q):                              0.32   Prob(JB):                         0.65
Heteroskedasticity (H):               0.99   Skew:                             0.26
Prob(H) (two-sided):                  0.98   Kurtosis:                         2.23
===================================================================================
```

Plotting on a line graph – Actual Vs Forecast

The actuals were plotted vs the forecast. The result was positive, and the goal can be achieved.



Model Review

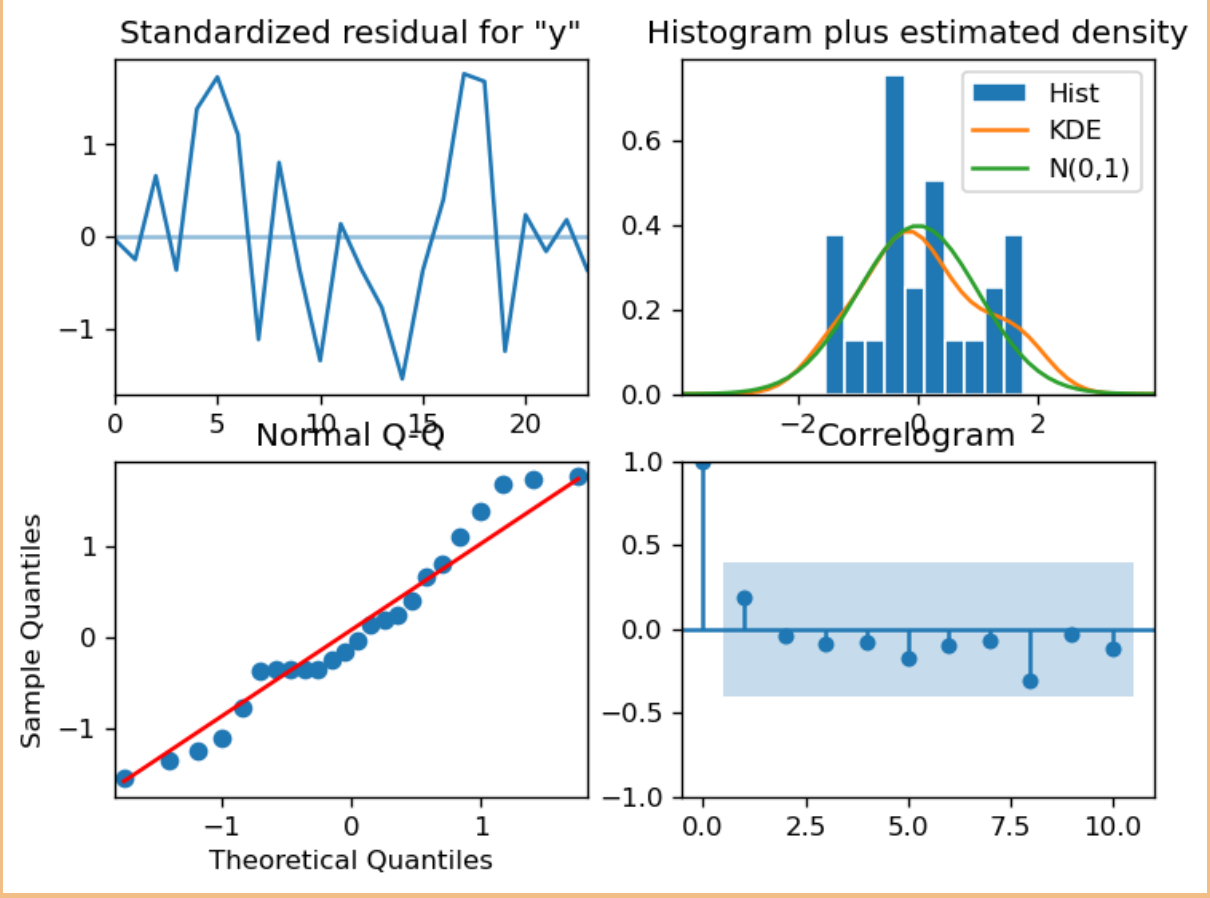
Top left: The residual errors seem to fluctuate around a mean of zero and have a uniform variance.

Top Right: The density plot suggests normal distribution with mean zero.

Bottom left: All the dots should fall perfectly in line with the red line. Any significant deviations would imply the distribution is skewed.

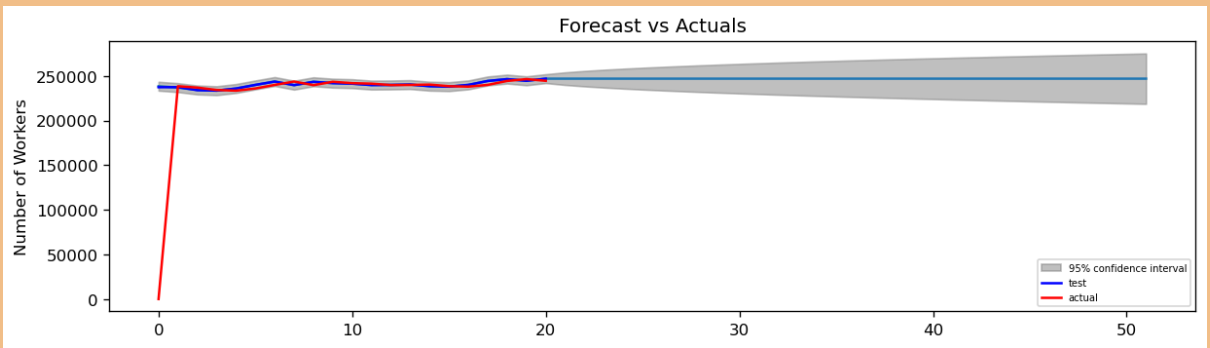
Bottom Right: The Correlogram, aka, ACF plot shows the residual errors are not autocorrelated. Any autocorrelation would imply that there is some pattern in the residual errors which are not explained in the model. So, you will need to look for more X's (predictors) for the model.

Overall, it seems to be a good fit. Let's forecast.



Model Forecast

Prediction was done for next 6 months and the model did a nice forecast.



Next Step

Despite satisfying model performance there are other tools which can be used to see if better model can be build.

Conclusion

Time Series Analysis are widely used worldwide and has been a great help in solving lots of problems. It is also good for my career as with my financial background understanding data and company requirements can be achieved easily.

Reference

- Stats New Zealand

- Authors: Prabhanshu Attri, Yashika Sharma, Kristi Takach, Falak Shah

https://colab.research.google.com/github/keras-team/keras-io/blob/master/examples/timeseries/ipynb/timeseries_weather_forecasting.ipynb#scrollTo=aAOKbLOf7M1_

- Selva Prabhakaran

https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python

- Anandhu H

https://www.kaggle.com/code/anandhuh/time-series-analysis-covid19-cases-arima-model

And thanks to Amin/ Ricky/ Sebastian

Thank You