

Advanced Statistics CourseWork project

By- W.A.C Fernando (26545)

M.A Pasindu Malinda (26033)

R.T Dulmina (27999)

H.K.I Dhananjaya (27595)

W.G.M.M.A Mansoor (26775)

Contents

[Abstract 3](#_Toc197373568)

[Introduction 3](#_Toc197373569)

[Hypothesis Testing 4](#_Toc197373570)

[1st scenario based on the health\_data.xlsx 4](#_Toc197373571)

[Introduction 4](#_Toc197373572)

[Methodology 4](#_Toc197373573)

[Results 5](#_Toc197373574)

[Conclusion 5](#_Toc197373575)

[2nd Scenario based on the - employee\_performance.xlsx 6](#_Toc197373576)

[Introduction 6](#_Toc197373577)

[Methodology 6](#_Toc197373578)

[Results 7](#_Toc197373579)

[Conclusion 7](#_Toc197373580)

[Linear Regression Analysis 7](#_Toc197373581)

[1st scenario based on the house\_prices.xlsx 7](#_Toc197373582)

[Introduction 7](#_Toc197373583)

[Methodology 8](#_Toc197373584)

[Results 9](#_Toc197373585)

[Conclusion 14](#_Toc197373586)

[2nd scenario based on the employee\_salaries.xlsx 14](#_Toc197373587)

[Introduction 14](#_Toc197373588)

[Methodology 14](#_Toc197373589)

[Results 15](#_Toc197373590)

[Conclusion 18](#_Toc197373591)

[Time Series Analysis 19](#_Toc197373592)

[1st scenario based on the monthly\_retail\_sales (1).xlsx 19](#_Toc197373593)

[Introduction 19](#_Toc197373594)

[Methodology 19](#_Toc197373595)

[Results 24](#_Toc197373596)

[Conclusion 27](#_Toc197373597)

[2nd scenario based on the weather\_data.xlsx 28](#_Toc197373598)

[Introduction 28](#_Toc197373599)

[Methodology 28](#_Toc197373600)

[Results 33](#_Toc197373601)

[Comparison between ARIMA and SARIMA 36](#_Toc197373602)

[Challenges, limitations and practical implications faced 37](#_Toc197373603)

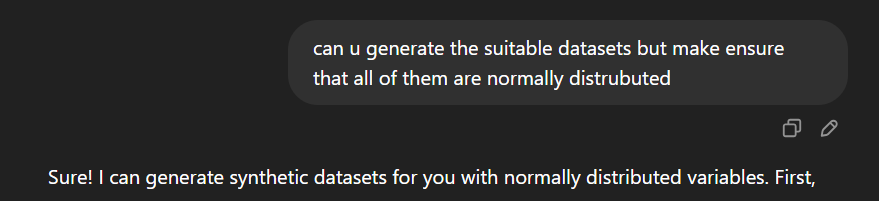
[Final Conclusion for the project 37](#_Toc197373604)

# Abstract

In this project, several datasets were generated, and various statistical tests such as Hypothesis testing, Linear Regression Analysis and Time series analysis were implemented on them to explore various statistical methods and analysis, and it shows how they can give various insights into how certain statistical algorithms should be implemented and how they can impact certain decisions of people.

# Introduction

**The following datasets were used for this project (due to unforeseen calculation issues, the datasets had to be generated unfortunately by using AI and the datasets were all generated in way that all were normally distributed as shown below)**

****

* [health\_data.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EdufUtz1DOZGpWLAve_ok4kBSuhZzhl1T_yvWrsC4kb3Fg?e=WubZw9) - here the parametric population mean test was implemented to see if the mean blood pressure was equal or different among diabetics and non-diabetics.
* [employee\_performance.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EXI9hIp6jtJNm0c0WayGx7oBGgc5P9krt3rxJBIRwfzu5Q?e=1AlSzL) – Mann Whiteney U test was performed to see if the distribution of performance scores is equal/the same for both groups of employees.
* [house\_prices.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/Ebb13NXhp1lKg1PC3hsID0oBQoidigGHkHXjik6DltBcug?e=X97aqX) – Linear regression analysis was done to see the impact of certain factors impact the price of a house.
* [employee\_salaries.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EQ3_ycXqiPRMt9jvN8f8HpUBO66_O3WPh5bYsugfqLjnCw?e=gpBr7z) – Linear regression analysis was done to analyze how certain factors effect the salary of an employee.
* [monthly\_retail\_sales (1).xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EX5cbkz87OdAlKaiH7GMe9wB8psCPD-LNb5bLXrOEaA1Zw?e=nOGwXw) - SARIMA and ARIMA Algorithms were implemented on this to identify certain patterns and predictions.

* [weather\_data.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EesNX1NHSgFLkTCADuST6h0BDkSVblQJm3FUmGueteJoyg?e=1l3QFR) – SARIMA and ARIMA Algorithms were implemented on this to identify certain patterns and predictions.

The link to the code - [StatProject.ipynb](https://nsbm365-my.sharepoint.com/:u:/g/personal/wacfernando_students_nsbm_ac_lk/EbKeJbl44JdNk_c7QNlmdVIBJSo68Jx8NoMzJx_XW-neTg?e=faUAd2) , [forcasting.ipynb](https://nsbm365-my.sharepoint.com/:u:/g/personal/wacfernando_students_nsbm_ac_lk/EQgyURTQpv5NoPlIzfPWkDIBAeA-NNzaMDwNRUncsLaUkg?e=ciRZiH)

# Hypothesis Testing

## 1st scenario based on the [health\_data.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EdufUtz1DOZGpWLAve_ok4kBSuhZzhl1T_yvWrsC4kb3Fg?e=WubZw9)

### Introduction

Certain health factors of 500 people are covered in this dataset such as,

* + Age - basically from 20-80
  + BMI - normally distributed
  + Blood pressure (systolic) – normally distributed
  + Diabetes Status - shown in binary where 0= non-diabetic and 1= diabetic
  + Smoker status - shown again in binary where 0 = nonsmoker and 1 = smoker

### Methodology

1. Firstly, had to make sure that the dataset was clean which It was since it was a synthetic dataset.
2. Then proceeded onto defining the null and alternative hypothesis.

**H0 = The mean blood pressure between diabetics and non-diabetics is the same (µ1= µ2)**

(µ1=mean blood pressure for diabetics

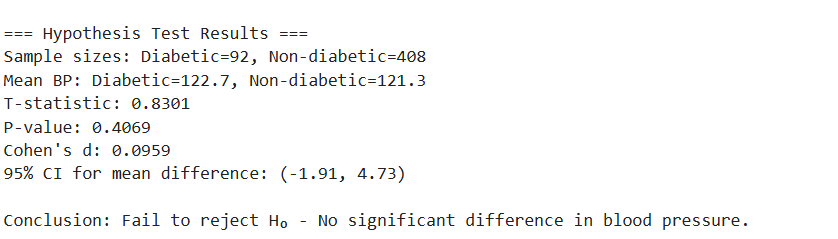
µ2=mean blood pressure for non-diabetics)

**H1= The mean blood pressure between diabetics and non-diabetics is not the same (µ1≠ µ2)**

Since the hypothesis is based on equality this is a two-tailed test.

Since what is being done here is basically comparing 2 independent groups (diabetics and non-diabetics) and since data is normally distributed, **the T-Test for population mean (which is parametric) was used.** The reason being both H0 and H1 is based on the mean BMI of the population.

### Results



* Based on the above output, the sample means of diabetics and non-diabetics are 122.7 and 121.3, meaning diabetics mean value is slightly higher).
* Since P-value is greater than the alpha value which is 0.05 it basically means that it’s not statistically significant and in terms of the confidence interval, the true population difference would probably be from, non-diabetics having 1.91mmHg higher BP (this is the negative bound) to diabetics having 4.73 mmHg BP (positive bound).
* Therefore, since the p-value which is 0.4069 is greater than 0.05, the only choice left is to accept the null Hypothesis(H0).

### Conclusion

Therefore, in conclusion, this means that there is not enough evidence to support that there is a difference of the mean blood pressure between people who have diabetes and non-diabetics. (based on this dataset).

## 2nd Scenario based on the - [employee\_performance.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EXI9hIp6jtJNm0c0WayGx7oBGgc5P9krt3rxJBIRwfzu5Q?e=1AlSzL)

### Introduction

This dataset was also generated to simulate employee work hours and employee performance scores of a company. There are 500 employees mentioned in this and they are categorized into 2 main groups based on the hours they are working.

The following variables are mentioned in the dataset

* EmployeeID
* Group (To classify the employees as either Group a or Group B)
* Work Hours
* Performance Score (represents a standardized performance rating which is scaled from 0-100 and is normally distributed)

### Methodology

1. The dataset was cleaned first.
2. Here the Mann-Whitney U Test (which is a non-parametric test was used to compare performance scores of people who work few hours and people who work more hours. Which means each employee belongs to only one group hence making them independent making the Wilcoxon Signed Rank Test not applicable in this case and making the Mann Whitney U Test the better choice.
3. Moving on to defining the hypothesis

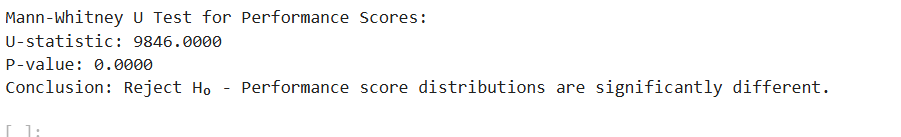
**H0 (Null hypothesis) = The distribution of performance scores is equal/the same for both groups**

**H1(Alternative hypothesis) = The distribution of performance is not equal for both groups.**

1. Then of course the Mann-Whitney U test was performed
2. The U statistics were calculated, and the p-value was also obtained.

And finally, the p-value was compared to the alpha value which is 0.05 to decide whether to accept the H0 or to reject it.

### Results



* So, what the U-value shows is, times, as in the counts as to how many times the scores from one group outperformed the other group when all pairs are compared. Therefore, in this case for instance Group B is the higher performing group, then its scores dominate Group A in ~9846 pairwise comparisons.
* As the above screenshot shows, it is clear that since the p-value is indeed less than 0.05 (which also indicates in turn that it’s highly statistically significant, we can reject the H0 (null hypothesis).

### Conclusion

Which in conclusion means that there is sufficient evidence to accept that people’s working hours do indeed significantly impact their performance scores. In other words, this means that people who work longer hours (which is Group B) will tend to have higher performance ratings to those who work shorter hours (group A).

# Linear Regression Analysis

## 1st scenario based on the [house\_prices.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/Ebb13NXhp1lKg1PC3hsID0oBQoidigGHkHXjik6DltBcug?e=X97aqX)

### Introduction

This data set shows the relationship of how the price of a house can change depending on the square feet, number of bedrooms, bathrooms etc. Since this dataset has variables such as,

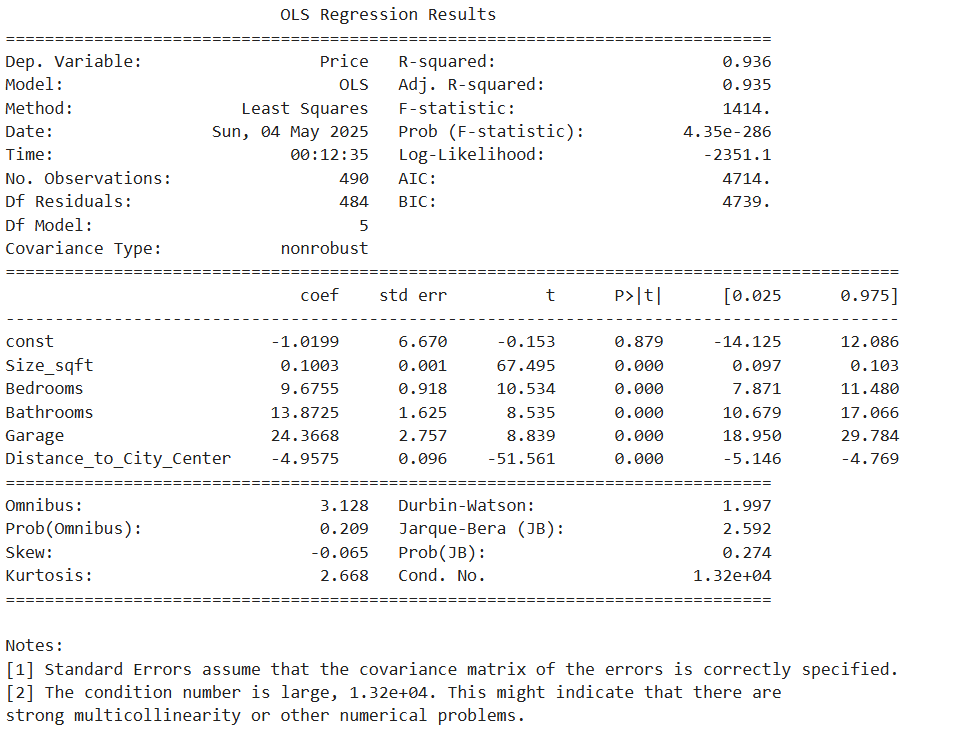
* Garage (which is in binary, as in 0= no garage and 1 = has garage)
* Size\_sqft- Size of the house in square feet
* Bedrooms - as in the number of bedrooms
* Bathrooms - as in the number of bathrooms
* Distance to city center - which is in miles
* And of course, the price of the house itself (which is in dollars),

one can analyze how certain housing features can impact the price of the house itself.

### Methodology

1. The dataset was generated and cleaned and preprocessed as well before moving on to the linear regression analysis.
2. In this scenario the dependent variable (Y) is price, and all the other variables are basically all independent variables(X). The reason being price of any house will always depend /change on its size and its facilities and of course its location etc.
3. Then the regression analysis was done, and its values were also tested with ANOVA tables.
4. Checking the residual errors.
5. Q-Q plot
6. Residuals vs Fitted values plot

### Results



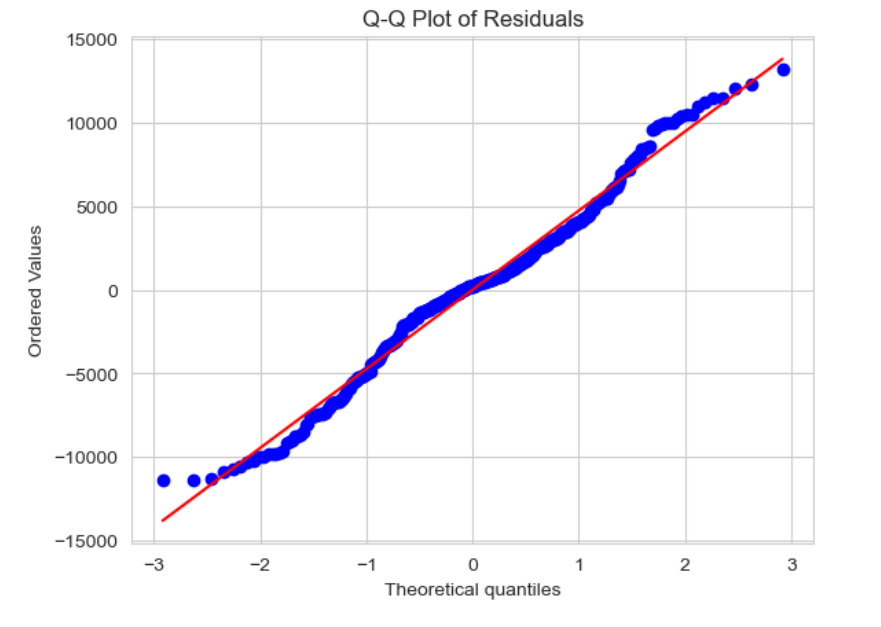
According to the above image, the coefficients for each variable can be seen first.

* In terms of Size\_sqft= since its coefficient is 0.1003, it simply means that if the square feet of the house were to increase by 1sqft then that will raise the price of the house by basically 100$.
* In terms of bedrooms, since its coefficient is 9.6755, it means that each bedroom adds almost 9675$ to the price of the house.
* In terms of the Bathrooms also since its coefficient is 13.8725, it simply means that each bathroom will increase the price of the house by 13872$.
* In terms of the garage, since its coefficient is 24.3668, it simply means that houses that have a garage are worth 24,369 more than the houses without a garage.
* In terms of distance to the city center, since its coefficient is -4.9575, that means that the farther the house is from the city by each mile, the price of the house also decreases by 4,958 which in turn shows that urban locations are indeed more valuable when it comes to selling houses.
* The p-value for all the variables here is 0.0000. What that means is that every variable here can impact the price of a house in a very big way.
* In terms of applying the above coefficients from that output into the multiple linear regression formula it is,

**Price = -1.0199 + 0.1003(size\_sqft) + 9.6755(bedrooms) + 13.8725(bathrooms) + 24.3668(garage)-4.9575(distance\_to\_city\_center)**

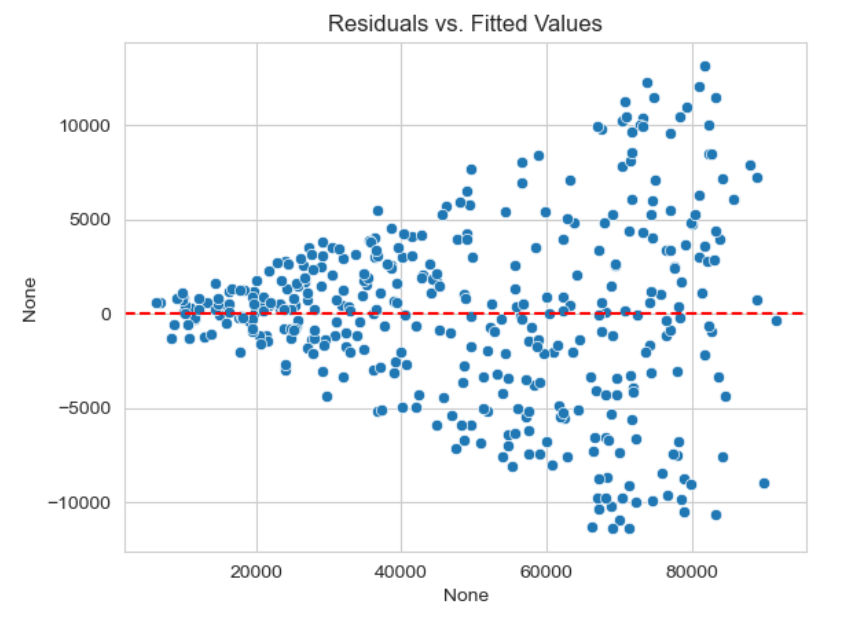
**And, in terms of the model quality, for R-squared, the value is 0.936 which means that the model explains about 93.6% of the variation in property prices and also proves that the model does indeed fit the data well. Also since there is a high F-statistic (1414), that means that the model as a whole is significant.**

* Checking the regression model’s errors(residuals) follow a normal distribution as shown below.



According to the above graph the middle points of the line, (specifically from -1 to +1), since the dots are fairly close to the straight line, that means that the errors are normally distributed for the most part. However, on the tails (the 2 corners of the line though), they curve away a bit, which in a way means that this model makes some larger errors than it was expected at first for some data points. So, the residuals are not perfectly normal.

In order to further check for any other issues, the following plot was also created.

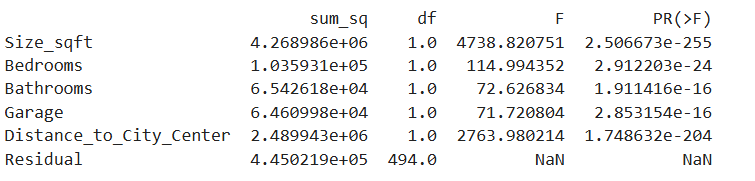


So, in the above graph the X axis represents the fitted values, and the Y axis represents he residuals.

According to the above graph one can see that the residuals are randomly scattered around zero for the most part, which indeed is mostly good. But the issue here is that as the predictions increase the residuals are fanning out (this is also known as heteroscedasticity. So, in order to resolve this the data will have to be checked to see if there are any unforeseen errors unfortunately.

#### Regarding the ANOVA Test,

The results are shown below.

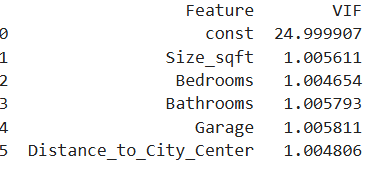


According to the above output the factor that has the biggest potential of impacting the price of a house is indeed the size (Size\_sqft). Second, its distance and the rest of the factors also still have strong effects on the price. The PR>F column (the p-value column) there at each row there’s a minus and some digits (-255,-24…etc.). That means the decimal should be moved 255 or 14 or 24 etc. places to the left which essentially means all of them are basically 0 (almost). Which in tun shows that all of those factors are indeed very significant.

#### VIF Analysis (Variance Inflation Factor)

This was done to measure how much one predictor variable is explained by the others in the model.

Outputs are shown below.



So as shown in the above output, there are no multicollinearity issues as all of them were close to 1 which in turn means that all the predictors here are independent as well except for “intercept”.

### Conclusion

Therefore, in conclusion it is evident that when it comes to the price of a house, certain factors such as the size of the home, its facilities, and location etc. will always have a major impact on the price. The bottom line is according to these outputs, especially for people who are worried about their finances, when it comes to buying houses, they should try to buy smaller houses closer to the city. The worst move one could potentially make is to buy a big house very far away. Also, when it comes to buyers and the sellers of houses, factors such as square footage and location will play a major role according to the results obtained above.

## 2nd scenario based on the [employee\_salaries.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EQ3_ycXqiPRMt9jvN8f8HpUBO66_O3WPh5bYsugfqLjnCw?e=gpBr7z)

### Introduction

This dataset shows some career related factors/variables related to 400 employees of a company.

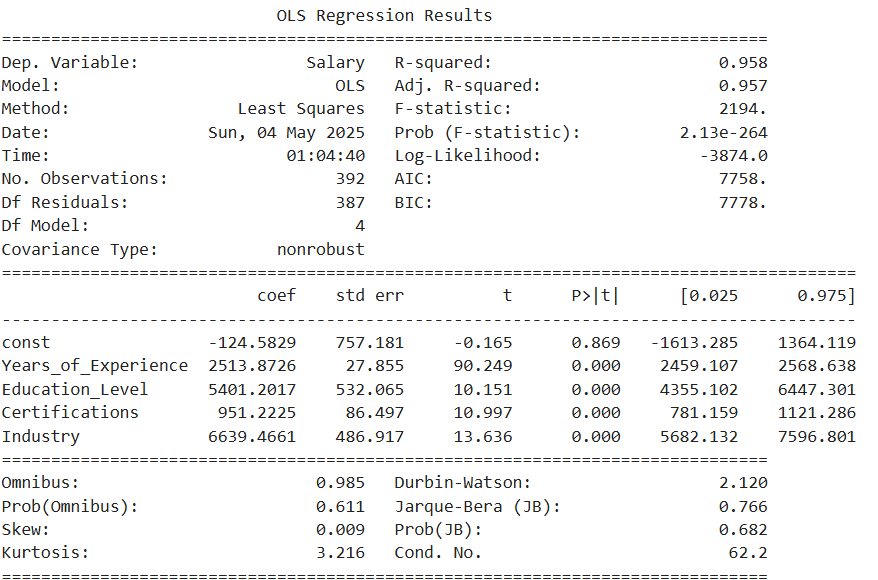
* Years of experience - as in the work experience shown in years
* Education level - shown in binary as in 0=noon graduate and 1 = graduate
* Certifications - basically shows the number of certifications
* Industry - again shown in binary, 0 = non-tech and 1 = tech
* Salary

Here the dependent variable is the salary, and all the other variables are independent variables. The reason being, in any case, the salary of anyone will always depend on that person’s professional experience, qualifications etc.

### Methodology

1. The dataset was cleaned
2. The linear regression analysis was performed.
3. Anova tables were also implemented.
4. Residual normality checking
5. Homoscedasticity Check plot

### Results



* Now as shown in the above image, the regression coefficients of each independent variable can be seen.
* When it comes to Years\_of\_Experience = Since its coefficient is 2513.87 and its p=0.000 that simply means each extra year of experience will add basically 2513$ to the salary and the p-value being 0.00 shows that it’s highly significant.
* When it comes to Edcuation\_Level = Since its coefficient is 5401.20 and its p-value is again 0.00 it simply means that employees with higher education earns 5401$ more than people with no higher education.
* When it comes to the certifications, since its coefficient is 951.22 and p=0.00, it means that certifications help increase the salary for people by 951$ each.
* Then when it comes to industry since its coefficient is 6639.46 and p=0.000, it means that employees in the tech industry earns 6639$ more than the people in non-tech industries and this is also the largest effect after education, which in turn shows that the industry people choose to be in also matters a lot when it comes to being paid well.

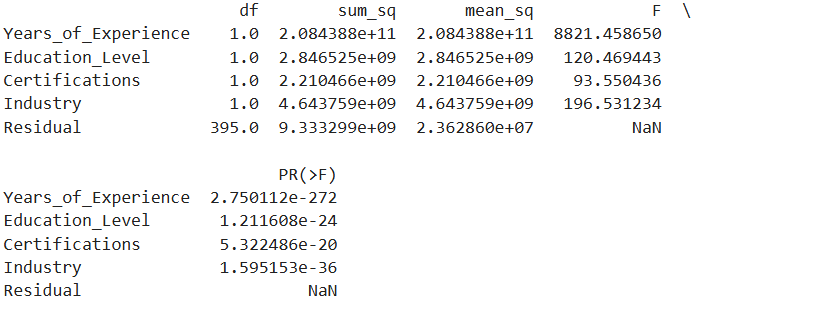
So, based on the above output the following multiple linear regression formula can be created.

Salary = -124.5829 + 2513.8726(Years\_of\_experience) + 5401.2017(education\_level) + 951.2225(certifications) + 6639.4661(industry)

In terms of the model quality, R-sqaured is 0.958 which means that the model explains about 95.8% of the variation in the salary data and that the model does fit the data well. Here also there is a high F-statistic value which is 2194 which means with a low p-value, thus confirming that the model as a whole is indeed statistically significant as well.

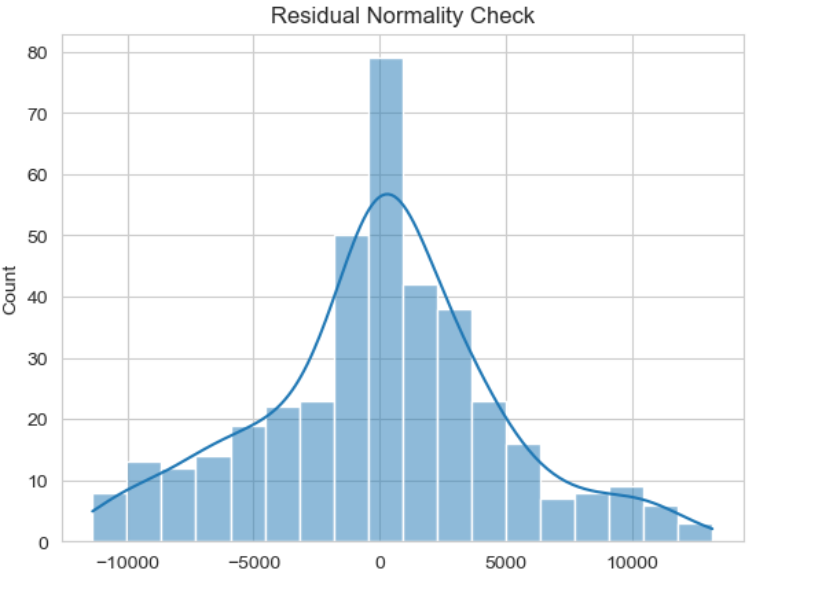
#### Interms of the ANOVA table,

The results are shown below.



According to the above output, it is clear the years of experience have had a big impact on the salary. Since all the p=values are basically 0 here as well it indicates that all of these factors are indeed significant when it comes to affecting the salary.

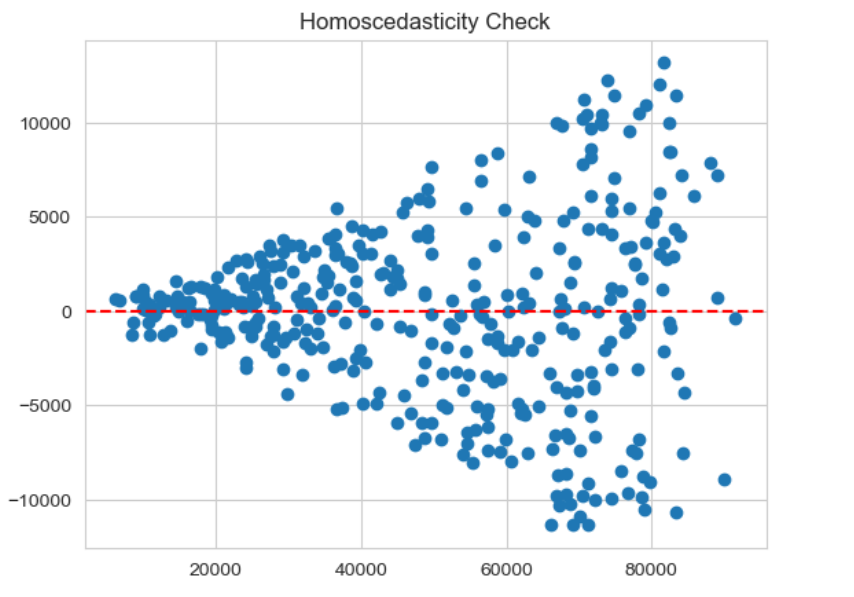
#### Residual normality check



According to the above graph, since the residuals are basically bell-shaped and are centered at 0 it is indeed a good sign because that means that the errors in this model are normally distributed indeed. Which means in general this model is trustworthy for inference as well.

#### Homoscedasticity Check plot

The plot is shown below



Unfortunately, here also the spread increases as the fitted values also increase which means that the residuals are not evenly spread which unfortunately violates the homoscedasticity assumption. Therefore, this model is not that reliable at higher values. So, in order to resolve these certain changes would have to be applied.

### Conclusion

Therefore, in conclusion, it is very evident that, certain factors such as education, experience, certifications and industry does indeed dictate one’s earning potential. Also, when it comes to one’s career, ensure prioritizing job experience before anything else and when entering an industry it is better to aim for a tech company and education still has its own place in this world.

# Time Series Analysis

## 1st scenario based on the [monthly\_retail\_sales (1).xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EX5cbkz87OdAlKaiH7GMe9wB8psCPD-LNb5bLXrOEaA1Zw?e=nOGwXw)

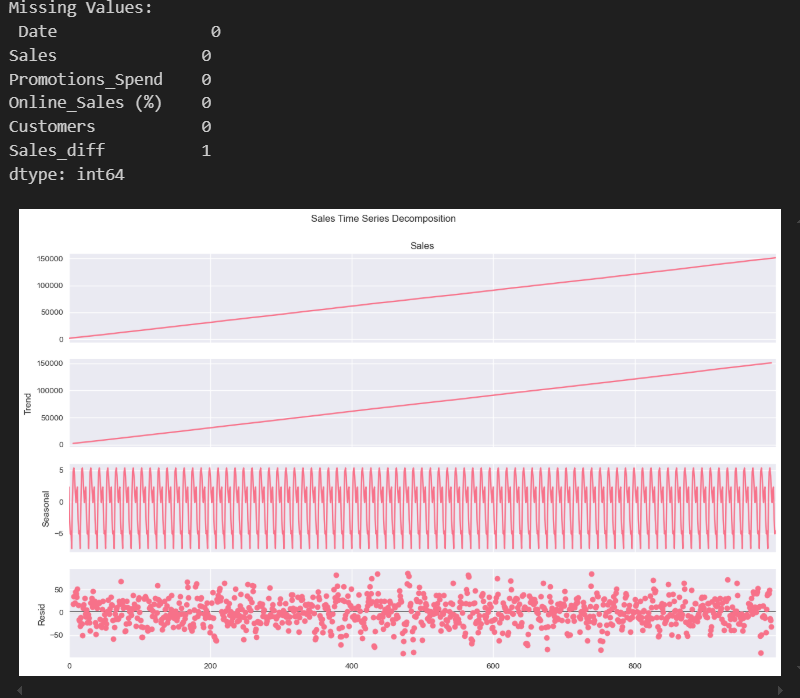
### Introduction

The monthly sales data of a retail store (from January 2015 to December 2098) is shown in this dataset where it covers certain factors/aspects such as,

* Date
* Sales – total sales amount
* Promotions\_Spend – amount spent on promotions
* Online\_Sales – percentage of sales that came from online
* Customers – number of customers

### Methodology

1. The dataset was first cleaned, preprocessed.
2. Stationarity testing was also done to ensure that the statistical properties such as mean and variance etc. have not changed over time (The ADF Test was applied where the Dickey-Fuller Test was run to check if the data was stationary or not.)
3. Then we also thought of including the computing first order differencing as well to make the series even more stationary.



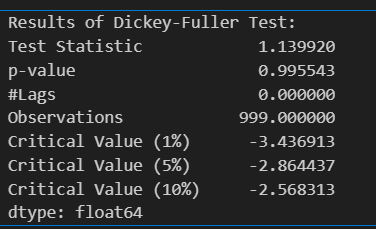
1. So here both the 1st and the 2nd graphs are rising steadily, which basically shows that sales are increasing consistently over time. The 3rd graph basically captures the repeating patterns which in this case indicate that the data has strong seasonality. The last graph shows the leftover fluctuations after removing trend and seasonality.

So, in summary this shows that the data has a strong upward trend in general, has clear seasonal patterns and has random residuals.

A graph on a screen

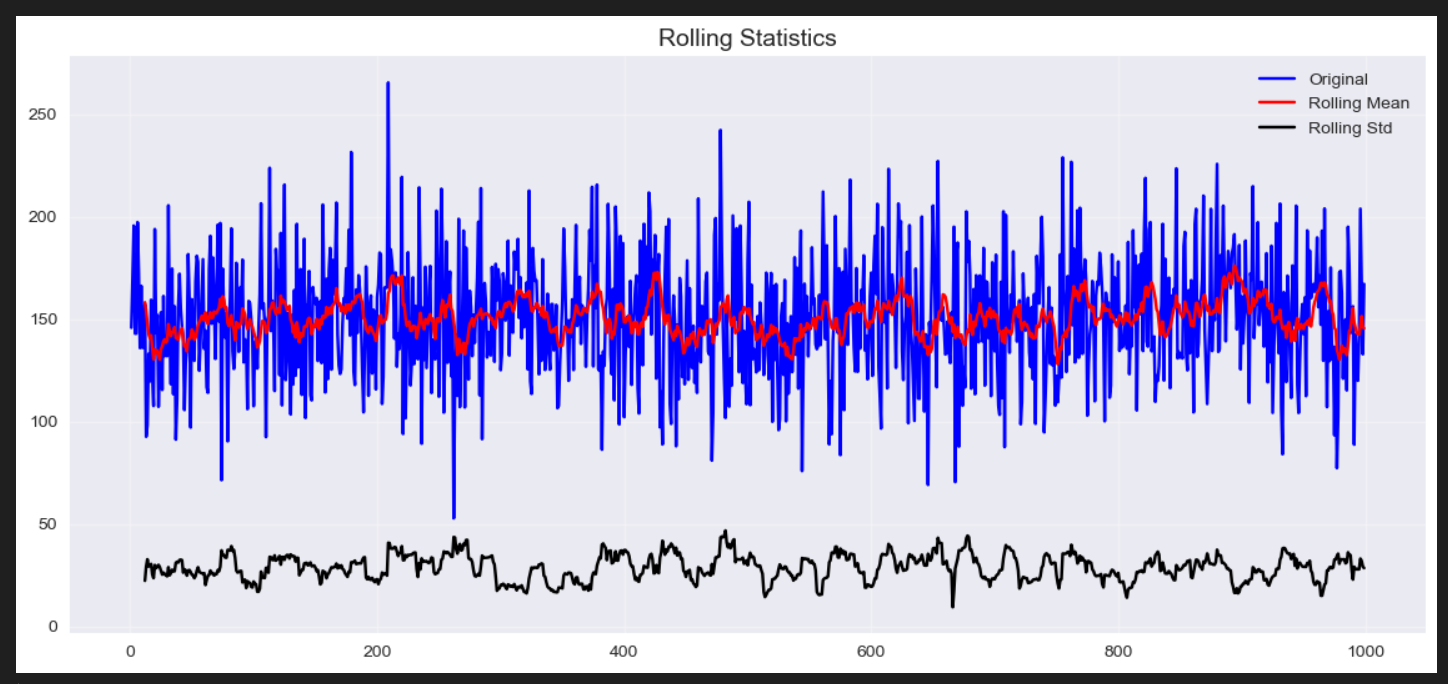
AI-generated content may be incorrect.

1. So here, the blue line shows the raw series sales data and it is indeed increasing consistently over time. The red line is also following the blue line very closely, which suggests that the average sales over a 21-period window also increase in a similar pattern. The black line on the other hand remains quite low, which shows that the variability in sales within the rolling window does not fluctuate much. Therefore, since the rolling mean increases overtime, that means the data is probably nonstationary. Below are the outputs of this test.

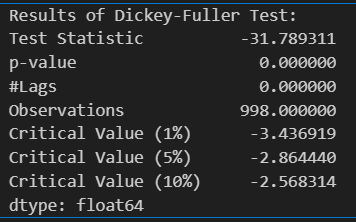


1. Here, since the test statistics are much higher than the critical values, that means the null hypothesis can’t be rejected , it must be accepted of the Dickey-Fuller test.

Since the p-value is almost near to 1 that means the data is indeed non-stationary. This can be seen in the graph below as well.



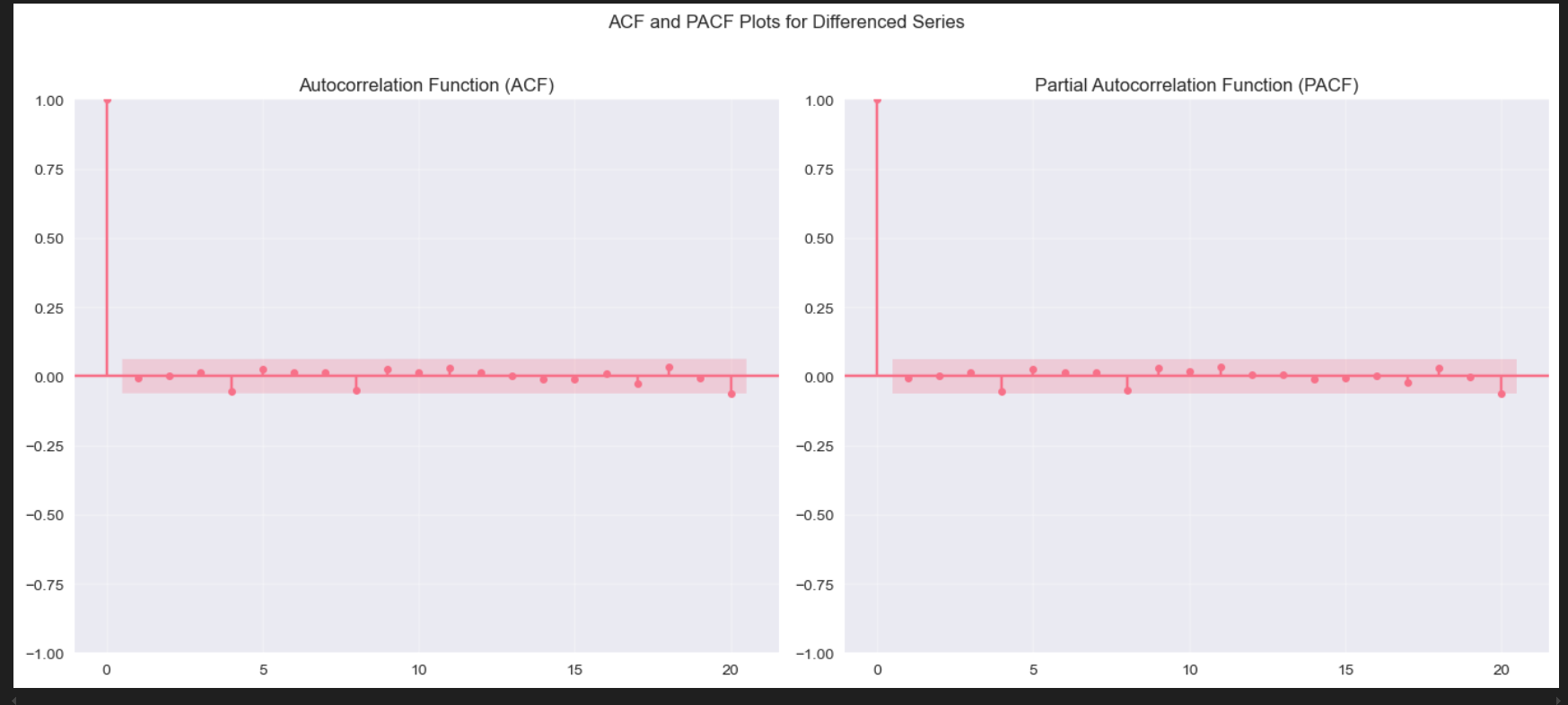
1. Since the average value of the series changes over time and the rolling standard deviation also fluctuates slightly the data is indeed not stationary as was proved earlier as well.
2. Below are the results after differencing



Here the test statistic is far more lower now than all the critical values which in turn now means that we can reject the null hypothesis. P-value is also 0, meaning the data is indeed stationary.

#### ACF and PACF implementation

The graph is shown below



1. Afterwards, before moving onto the ARIMA model it is crucial to identify the parameters for that, which can be identified above. Therefore, in summary according to the above 2 plots since, both show only the 1st lag as significant, that suggests that the ARIMA model should be (p=0, d=1, q=0) or (p=1, d=1, q=1). Either is fine. Theres also no significant correlation at higher lags which indicates that additional AR (p) or MA (q) terms would not be needed.
2. Afterwards we moved onto the Arima model implementation
3. Once that was done, we moved onto the SARIMA implementation

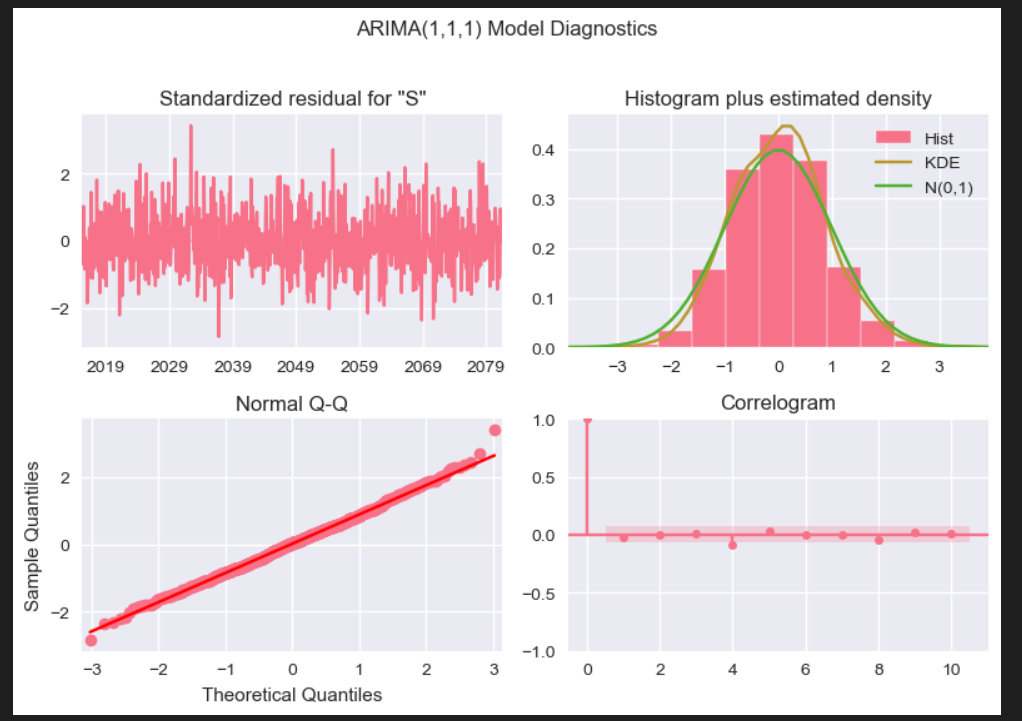
### Results

#### Results of the ARIMA

A screenshot of a computer

AI-generated content may be incorrect.

As shown above, the model is an ARIMA (1,1,1) model which means that p=1 (1 autoregressive lag), d=1 which means 1 differencing step to achieve stationarity and q=1 q=1 which means 1 moving average lag. The AIC and BIC can be used to compare models as well. The ar. L1 = 1.000 indicates a strong persistence in past values and high autocorrelation which means past values mostly dictate the future ones. The ma.L1 indicates that past errors can strongly impact the present and that immediate past errors almost completely correct themselves. And sigma2 measures the error variability. In summary this model fits well for the most part with stationary residuals and normal error distribution. But the AR and MA terms are indeed near perfect which could mean overfitting unfortunately.



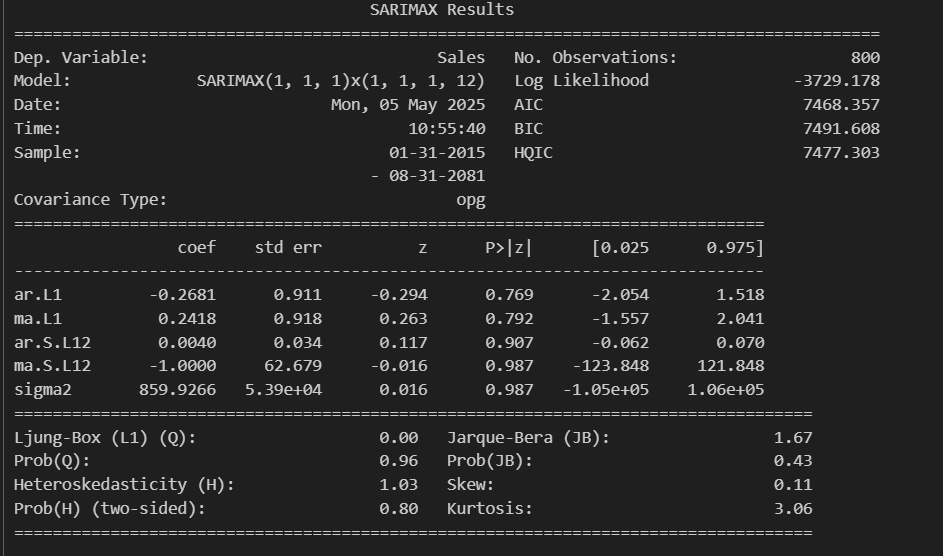
The above diagnostic plots were also done for the Arima model to help in assessing its performance and validity. As shown in the graph above in summary, there are no significant autocorrelation in residuals which means the model fits time dependencies well. The residuals are approximately normal as well which means model errors are well behaved and there are no trends or patterns that can be seen in residuals as well which show that the model captures the data structure.

A screen shot of a graph

AI-generated content may be incorrect.

So here the ARIMA forecast, which is shown with the green line, tracks very closely with the actual values which are shown with the orange line and the training data in pink. Here the lines overlap with each other almost perfectly, which shows that the model fits the data well and the predictions it makes are highly accurate as well.

#### Results of the SARIMA

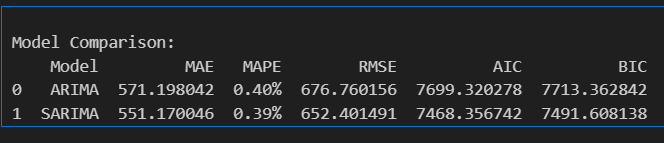


According to the above output, in summary it is clear that, this model also fits the sales data well in general. The reason being the residuals are random and normally distributed as well. Their variance is also stable. The AR and MA terms (both seasonal and non-seasonal) proved to be not that significant here, probably due to the data being very smooth in general and to also having a strong trend captured by differencing. Regarding the leftover patterns there aren’t any in the residuals which indeed confirms that the model has captured the time series structure very well.

### Conclusion

So, in conclusion the ARIMA model captures the sales pattern with exceptional accuracy and in the long run, this model shows that the sales data does indeed support long-term forecasting, which suggests that there is very little change in the relevant market/business. It can also be concluded that there is a certain product or service where its sales are showing a steadily increasing pattern which explains that product / service is capable of satisfying a lot of people.

#### Comparison between the ARIMA and the SARIMA



In terms of the Accuracy metrics (MAE, MAPE, RMSE) SARIMA has the lower values which means several things.

Regarding the actual sales values, SARIMA’s predictions are much closer to them than ARIMA’s predictions are.

With regards to the percentage of errors, even though both have low percentages, SARIMA is just slightly better than ARIMA in this case.

## 2nd scenario based on the [weather\_data.xlsx](https://nsbm365-my.sharepoint.com/:x:/g/personal/wacfernando_students_nsbm_ac_lk/EesNX1NHSgFLkTCADuST6h0BDkSVblQJm3FUmGueteJoyg?e=1l3QFR)

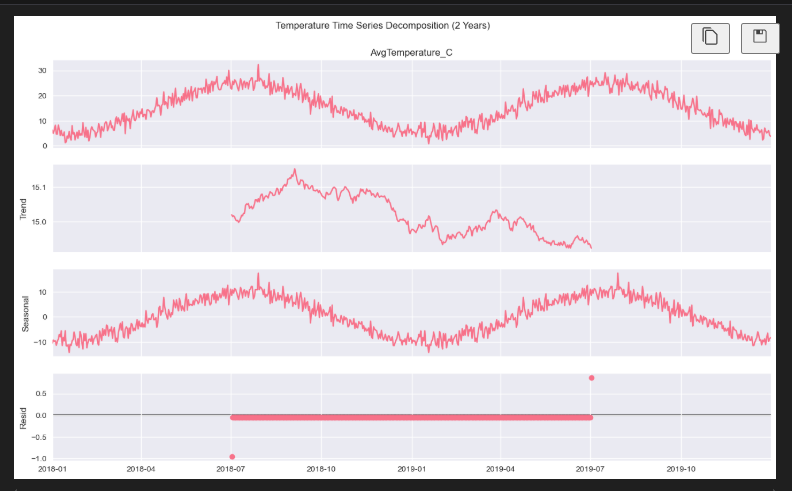
### Introduction

This dataset was also generated, and it contains some observations related to the daily weather over several months in 2018. It covers some factors such as,

* Date- normal calendar dates
* AvgTermperature\_C – the average temperature for the day
* Rainfall\_mm – the average rainfall for the day
* WindSpeed\_kmph – the average wind speed for the day
* Humidity\_% - the average humidity for the day

### Methodology

1. The dataset was first of all cleaned and preprocessed like before.
2. Afterwards proceeded on to the time series decomposition and stationarity testing.
3. Here, firstly seasonal decomposition was done to the time series – broke it into trend, seasonal and residual components. (shown in the graph below)



1. Then the rolling statistics were performed on the series as well as an ADF Test to check if its stationery or not.
2. Then proceeded onto testing the original temperature series for stationarity and if it wasn’t stationary then the first order differencing was applied to it and also if the dataset has 730 days, then seasonal differing was also applied to remove the seasonal effects as well.

Before differencing was applied the relative test values were as below.

A screenshot of a computer

AI-generated content may be incorrect.

Below is how the series looked after differencing.

A screen shot of a graph

AI-generated content may be incorrect.

1. Here the time series appears more stationary, and the rolling mean and standard deviation are both stable as well and there’s seems to be no autocorrelations as well.
2. The below image shows the test values to prove that the first differencing step worked perfectly to make the temperature data stationary.

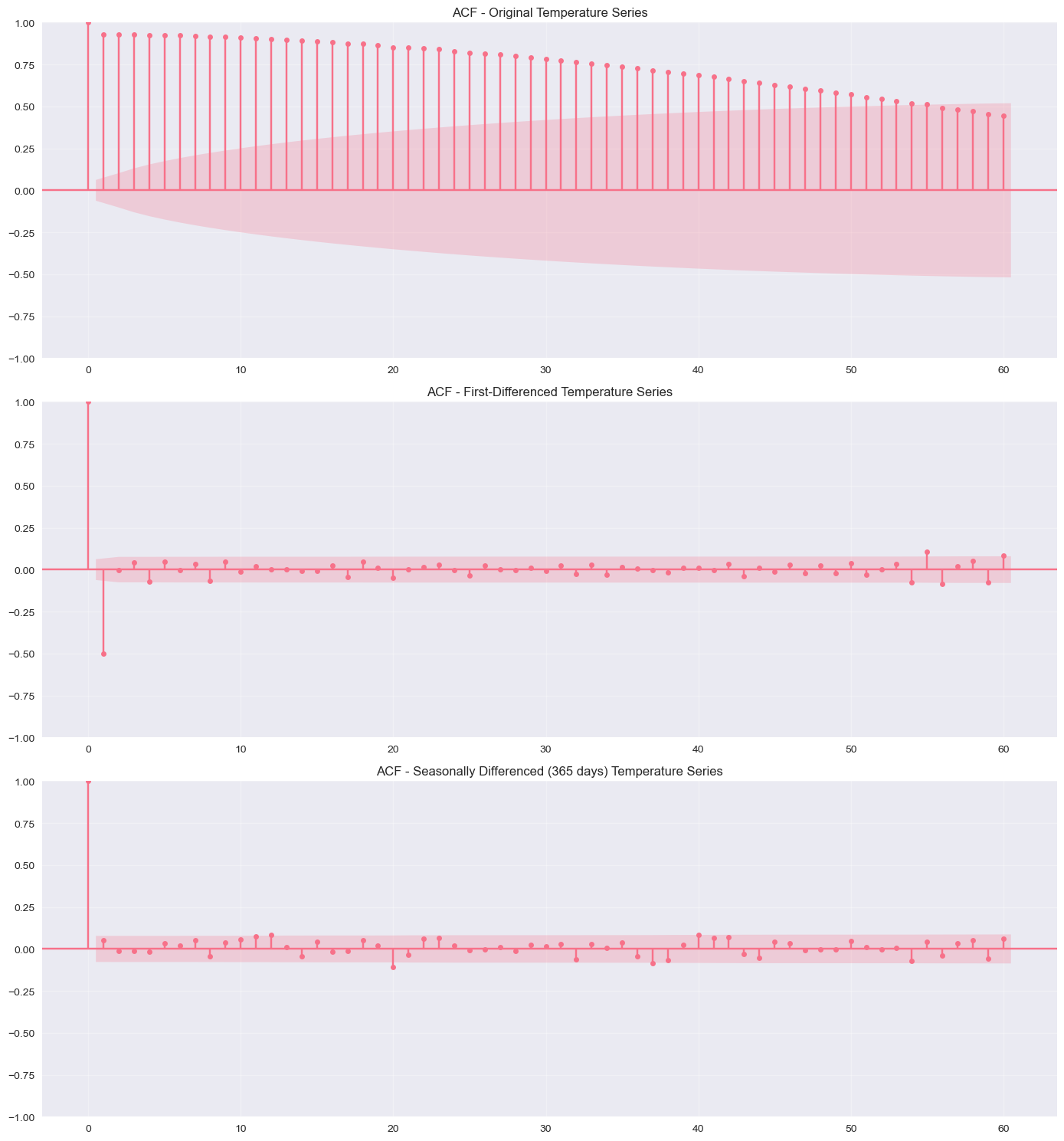
A screenshot of a computer

AI-generated content may be incorrect.

1. Here the large negative test statistics prove that there is now a very strong rejection of non-stationarity.

#### Model identification (ACF and PACF)

##### ACF

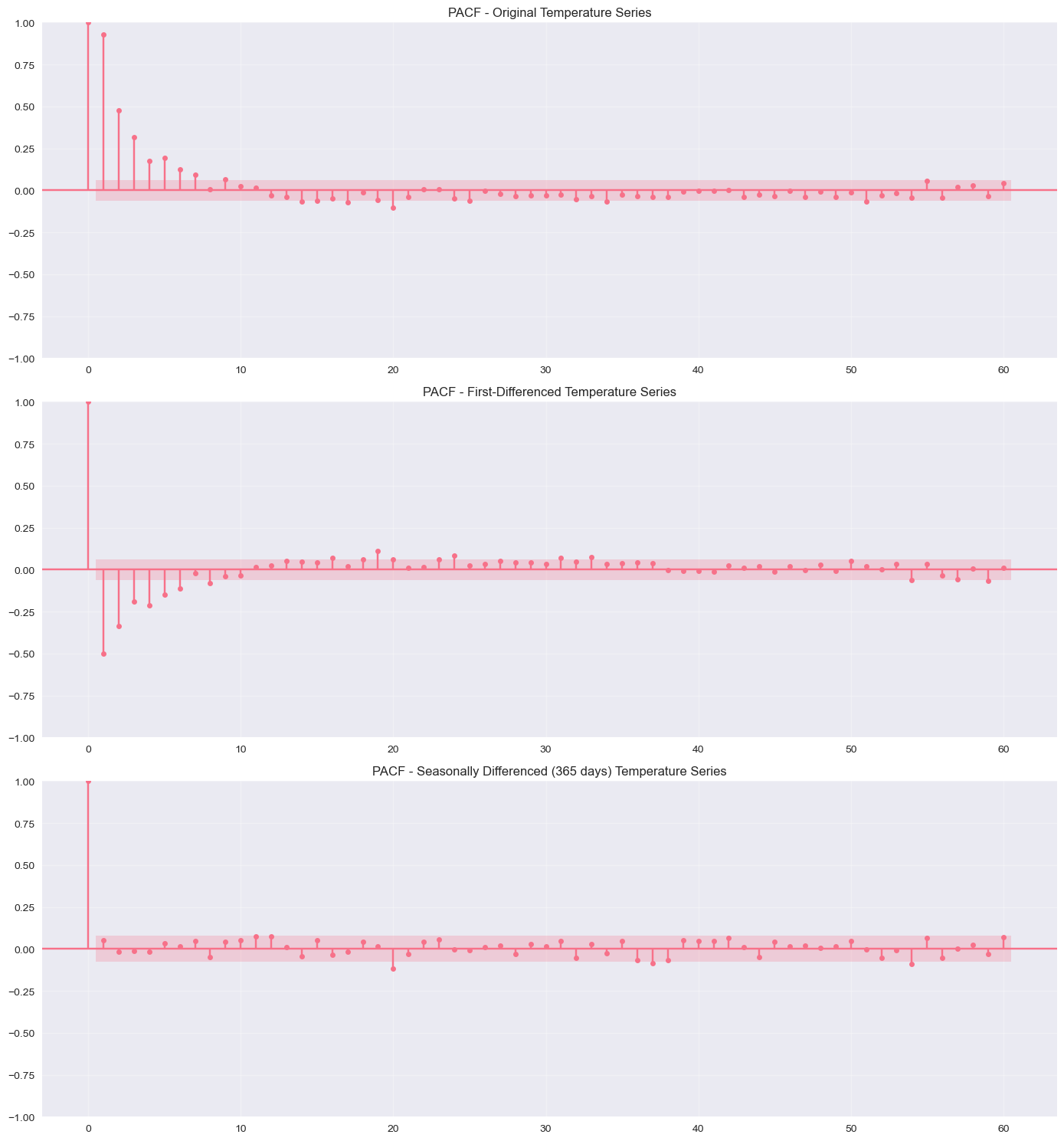


In the above image, the very top plot is the original temperature series plot which basically shows that the series shows persistent patterns and is not suitable for ARIMA modeling unless it gets transformed.

Then the middle plot is the first differenced temperature series plot which shows that the series now mainly stationary with no autocorrelations.

Then the bottom plot is the seasonality differenced temperature series plot which shows that after the seasonal differencing the series is now indeed well suited for time series modeling.

##### PACF



Here also the 1st plot is the original series, and it is not ready for modeling.

Then the 2nd plot, it’s mostly ready for ARIMA but only if there is no strong seasonality.

Then the 3rd plot is the ideal one for SARIMA modeling.

1. Once all of this was done, we moved onto the ARIMA model implementation.
2. Then SARIMA implementation.

### Results

#### ARIMA implementation

A screenshot of a computer screen

AI-generated content may be incorrect.

According to the above results one can see that the ARIMA model (2,1,2) is statistically perfect for forecasting average daily temperature for this dataset because it is capable of capturing the main patterns and also can give reliable predictions as shown in the above output values.

Then we also created some diagnostic plots to check if the residuals from the ARIMA model meet the assumptions that are required in order to do reliable forecasting as shown below.

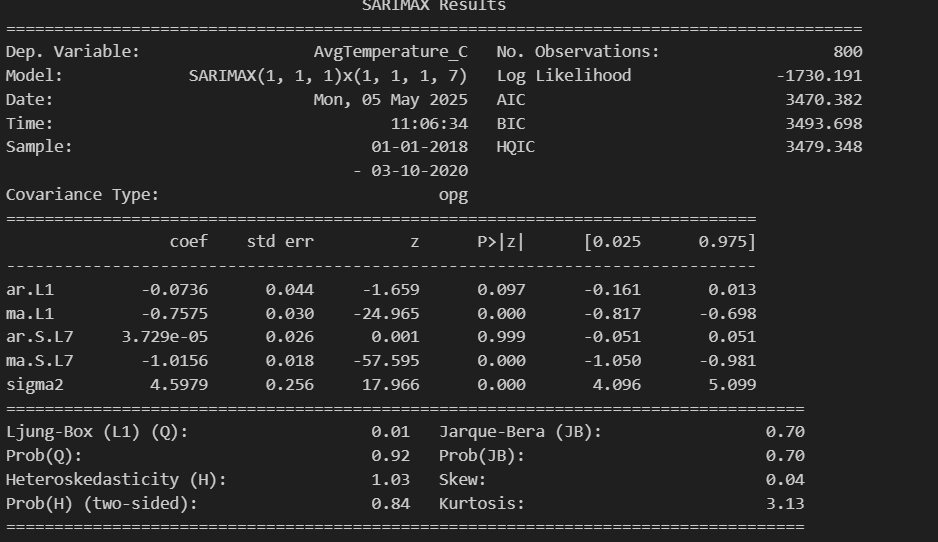
A collage of graphs

AI-generated content may be incorrect.

According to the above output, it is clear that the ARIMA model fits the data well and that the residuals are random and normally distributed and uncorrelated. There appears to be no evidence of model misspecifications or anything such as well.

#### SARIMA implementation

Below are the output values.



In summary, according to the above output shows that the model for the average temperature is well specified and passes all the key diagnostic checks. It proves that the model is suitable for forecasting temperature patterns.

Then we proceed onto some diagnostic plots for the SARIMA model as shown below.

A screenshot of a graph

AI-generated content may be incorrect.

### Comparison between ARIMA and SARIMA

A screenshot of a computer

AI-generated content may be incorrect.

With regards to the accuracy metrics (MAE, MAPE, RMSE) since ARIMA has slightly lower values than SARIMA, for this weather dataset, ARIMA has the capability to produce slightly more accurate forecasts than SARIMA.

In terms of the model selection criteria (AIC and BIC), since SARIMA has lower values than ARIMA it is clear that SARIMA is more statistically favored.

So, to summarize ARIMA has slightly better lower forecast errors for this dataset while SARIMA is indeed more statistically preferred due to lower AIC/BIC which in turn indicates better balance.

# Challenges, limitations and practical implications faced

There was a small issue when it came to running the ANOVA test on the regression models as there was a small mismatch with the model, we created vs how the function anova\_lm() expects it. For the model we created it by using sm.OLS() which is basically a dataframe method but the anova function expects a formula model. So in order to fix that we had to switch models essentially where we ended up using model = smf.ols("Price ~ Size\_sqft + Bedrooms + ...", data=df).fit() instead. Reason being this version explicitly tells the stats models how variables connect, so the anova function was able to compute properly without any errors.

Also, when it came to finding datasets suitable for the above tests to be done it proved to be quite tricky as in some datasets the calculations did not work properly.

# Final Conclusion for the project

Since the datasets were generated, it doesn’t necessarily resemble the real world applications properly but the datasets here and by performing the statistical calculations and analysis on them, it still showed just how much statics can impact on various aspects and it also highlights the pros and cons of the and ARIMA , Mann Whiteny U test and Wilcoxon test, Population mean T test - when they are suitable to use , how to verify them etc.