Business Analytics

Kamaljeet Kaur Sidhu

Student ID: 19489382

MODULE: (2022) 7BUIS024W.2

# Problem 1

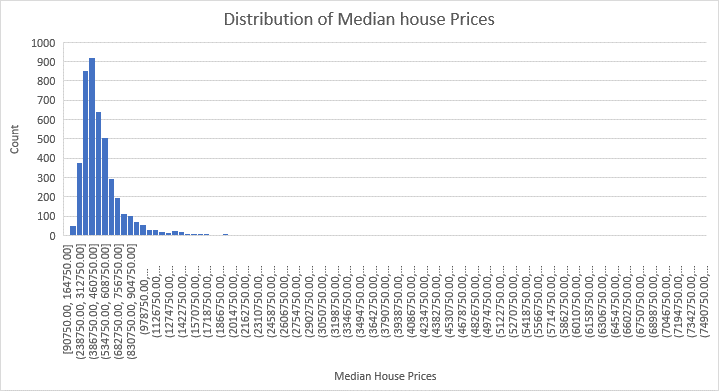
## Median House Prices:

## Suitable Summary and Distribution Chart

Table

Description automatically generated

**Fig. 1.1**



**Fig. 1.2**

It is clear from **Fig. 2;** the distribution of Median house prices is skewed to the right that means the peak amount of prices lies to the left of the centre. Histogram is the best option to display distribution.

## Statistical calculations to compare median house prices.

Timeline

Description automatically generated

**Fig. 1.3**

## Chart to compare the median house price by district.

**Fig. 1.4**

It is clear from the line graph, the most expensive house is In Westminster among others. The top five boroughs (highlighted in red) having maximum house prices are Kensington and Chelsea, Camden, Merton and Hammersmith and Fulham including Westminster. However, the least house prices are being noticed in Barking and Dagenham. Followed by, Bexley, Redbridge, Enfield and Brent. It is highlighted with green colour in Fig. 1.3.

## Confidence Intervals

Table

Description automatically generated

**Fig. 1.5**

## Significant difference in Prices between different pairs of Districts

A screenshot of a computer

Description automatically generated with low confidence

**Fig. 1.6**

After finding 95% confidence interval for three different pairs of local districts, the difference of sample mean between Westminster and Kensington and Chelsea is less as compared to two other pair of samples.

Further, P-value gives the information about the probability of data occurrence under null hypothesis.

I made assumptions about the significance level of the difference between two different boroughs.

Statistically significance means that the p-value is small enough so that the null hypothesis of the test can be rejected.

I found, the difference between Camden and Westminster (First pair) and Lewisham and Camden (Third pair) is statistically significant as the p value is < 0.01, below the alpha value of 0.05. However, the difference is not statistically significant between Westminster and Kensington and Chelsea (Second pair) as the p value is > 0.01, above the alpha value of 0.05.

# Age Group

## Percentage of the population in each group based on Local district.

A picture containing text, screenshot

Description automatically generated

**Fig. 1.7**

## Chart based on split of population as per Districts.

**Fig. 1.8**

In all the boroughs, most of the population belong to 16-64 years of age group. However, the percentage of young and elderly people is less as compared to the middle-aged population.

## Relationships between variables

## Calculate correlation coefficients.

A picture containing text, font, number, line

Description automatically generated

**Fig. 1.9**

## Explore the relationship between median house price and rest of the variables.

A scale that varies from -1 through 0 to 1 is used to measure the correlation coefficient. Based on the above matrices, I have analysed that there is no variable that has a positive relation with Median House Price except Median House Price itself. While a perfectly negative relationship is observed among four different pairs of variables highlighted in yellow. Namely, Living Environment Rank, Population aged 0-15, Population aged 16-64 and Area Square km versus Median House Price. All the other variables have a negative relation with Median House Price.

So, I found from this analysis, if the value of one variable is increasing then surely the value of another variable will decrease.

## Model(s) to forecast house price in terms of the other variables.

## Develop a model using all variables.

A screenshot of a computer

Description automatically generated

**Fig. 1.10**

**Fig. 1.11**

**Fig. 1.12**

**R-Square** is the important statistical measure used to determine the variance in the Median House Price that can be explained by the independent variables. R-Square defines the goodness of the fit. The higher R-Square value will enhance the model performance and correlation among variables. Also, there are two variables with higher **VIF** value highlighted in red. So, to improve the stability and accuracy of the model I decided to remove these variables. It will also help to overcome the multicollinearity. There are three variables with **P-value** of more than 0.05. Which indicates that these variables do not have a statistically significant association with the dependent variable. I will keep a check on P-values if there are any changes after removing higher VIF variables.

A Regression model will be considered good if **the scatterplot** is showing some randomness instead of a pattern of data. Also, clumping should not be there around the line. But this scatterplot is displaying completely opposite to the preference of a good model.

It can be assumed from **the histogram**, the data is not normally distributed. So, it is an indication that I need to work on normal distribution of the data.

### Develop an improved model using only significant variables.

A screenshot of a computer

Description automatically generated with medium confidence

**Fig. 1.13**

In this updated model, the **R-Square value** is improved by approximately eight percent by using the significant variables. In total there are just 8 significant variables.

**Fig. 1.14**

After updating the model, still there is not much improvement in the distribution of data. It is skewed to the right. So, I need to come up with a new idea to make it normally distributed.

**Fig. 1.15**

The new dataset contains 98 outliers. And, the scatter plot still shows a pattern rather than randomness. So, the model requires further improvement.

### Enhance further the model by including/excluding and transforming variables.

A screenshot of a computer

Description automatically generated with medium confidence

**Fig. 1.16**

Finally, I have transformed the data to logarithm form in hope of some development in the model.

Resultantly, I got the maximum R-Square value of approximately 40 percent as compared to previous models. The higher R Square value means the higher correlation among dependent and independent variables. Also, there are a smaller number of outliers as compared to all the models.

**Fig. 1.17**

After performing logarithmic transformation of variables, now the data is normally distributed.

**Fig. 1.18**

A randomness can be noticed in the scatter plot of Residuals vs Fit. Finally, the data is not following a pattern anymore.

Based on Residuals values, it can be assumed that this model has made good predictions. As the comparison between original and fitted value of Median House price is smaller.

## Findings:

* There are quite a few numbers of highly priced house.
* In Westminster, houses are more expensive as compared to all the boroughs.
* Cost of houses in Barking and Dagenham is low among other boroughs.
* City of London has less population, even the proportion of people of each age group are less as compared to other regions.
* Barnet has the maximum population of three age groups.
* The percentage of middle-aged population is higher in all the areas.
* Education, skills and training Rank and Income Rank, Employment Rank, Health Deprivation and Disability Rank has a strong positive relationship with Median house prices.
* Living Environment Rank, Population aged 0-15, Population aged 16-64 and Area Sq. Km. has a negative relationship with Median House Prices. **Based on d) (i)**
* **Findings based on enhanced model to forecast house prices in terms of other variables:**

A picture containing text, font, line, screenshot

Description automatically generated

* Each one unit increase in the coefficient of Education, Skills and Training Rank, Health Deprivation and Disability Rank, Barriers to Housing and Service Rank and Population aged 65+ is associated with a 0.3603711, 0.02166277, 0.04300045, 0.06169743 increase in the Median House Price in order.
* However, each unit increase in the Crime Rank, Living Environment Rank, Population aged (0-15), Population aged (16-64) is associated with 0.02166277, 0.08099527, 0.1337471, 0.09790358 unit decrease in the dependent variable.
* The association between each variable and Median House Price is also statistically significant ( P< 0.0001).
* Further I accessed the relative importance of the independent variables based on t-value.

A picture containing text, screenshot, font, number

Description automatically generated

* Even, Crime rate Rank does not reach the statistically significant level of (P<0.0001) still I need to retain in the model to control the possible confounding.
* The multiple regression model can be used to predict the Median House Price as a function of a boroughs Education, Skills and Training Rank, Health Deprivation and Disability Rank, Barriers to Housing and Service Rank, Population aged 65+, Crime Rank, Living Environment Rank, Population aged (0-15) and Population aged (16-64) value.

# Problem 2

## Representation of the problem, recommendation and expected payoff.

A picture containing text, screenshot, diagram, line

Description automatically generated

**Fig 2.1**

**Recommendation:**

Considering the maximum profit into account, I recommend London Coaches (LC) to run the service independently rather than getting into a partnership with the incumbent operator. The optimal expected payoff associated with this decision is £260,000.

**Expected Payoff:**

A screenshot of a computer

Description automatically generated with low confidence

**Fig 2.2**

## More informed decision tree (Hire a transport consultant?)

As per my understanding, a transport consultant has delivered two results ‘Fairly Low’ and ‘Fairly High’ and the results reveal new possible probabilities of service users.



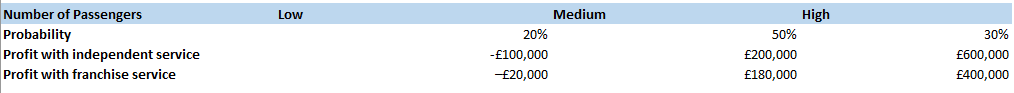
**Fig 2.3**



**Fig 2.4**

**Assumptions:**

I referred to the initial probability and profit table in order to get the profit based on probabilities discovered by the transport consultant. The Fig 2.5 refers to the base values given in the problem. The Fig 2.6 refers to the profit calculation for new probabilities.



**Fig 2.5**

A screenshot of a computer

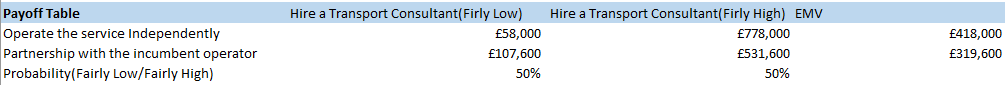
Description automatically generated with medium confidence

**Fig 2.6**

**Note:** Kindly zoom in for clear visibility.



**Fig 2.7**



**Fig 2.8**

**The company’s optimal course of action using a decision tree.**

London Coaches can maximize the expected gain by deciding to hire a transport consultant and to operate the service independently. The optimal expected payoff associated with this decision is £418,000. Further, if consultant’s prediction is true for ‘Fairly High’, company will have more profit.

## Sensitivity analysis

* I have highlighted four input variables for sensitivity analysis. The pairs of Low, medium and high probability when result is ‘Fairly Low’ and ‘Fairly High’ while working independently.

**Fig 2.8 Fig 2.9**

**Fig 2.10 Fig 2.11**

* After analysing the sensitivity of given optimal decision (EMV) by change in the initial profits. I got to know that **if the ‘Fairly High’ result will be true** and further the probability of having High number of passengers increases it will generate more income.(**Fig 2.11**)
* Although, the overall profit (EMV) will decrease by 21.53% and the probability decreases by 25% still it will not have any impact on the optimal decision. (**Fig 2.11**)
* **If the ‘Fairly High’ result will be true** and the possibility of having medium number of customers either increase or decrease by 25%, will not make much difference in the optimal decision. As it will lead to a change of just 1.91% in the EVM. **(Fig 2.10)**
* **However, if ‘Fairly Low’ result is true**, a change of -25% and 25% in the probability of low number of passengers will just make a difference of 1000 pounds in the EMV. **(Fig 2.8)**
* A change of -25% and 25% in the probability of medium number of passengers will just make a difference of 18,000 pounds in the EMV. **(Fig 2.9)**

**Conclusion:**

As the charts displayed above and available in the excel workbook reveal, plus or minus 25% changes in the values of each of the model inputs lead to no changes in London Coaches’ optimal decision to run the service independently and hire a consultant.

**Fig 2.12**



**Fig. 2.13**

**Conclusion:**

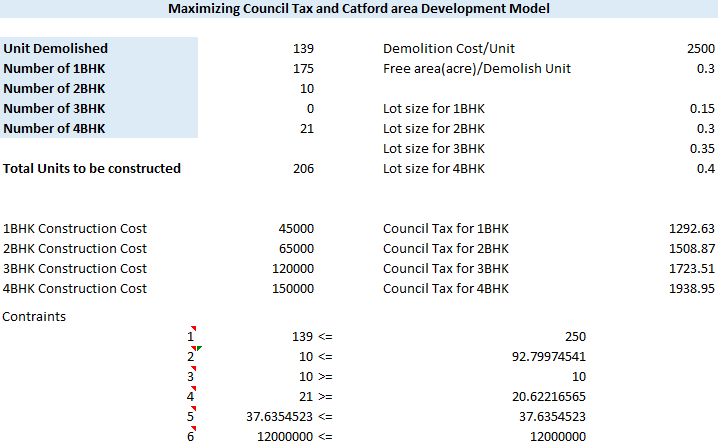
This tornado diagram indicates that the probability of high number of users (if they operate the service independently and ‘Fairly High’ result is true) is the most influential model inputs. To a lesser extent, Medium and Low number of users (if ‘Fairly Low’ is true and service is independent) also has some impact on the optimal expected gain.

## Findings

* As per overall analysis, I recommend London Coaches to run the service independently. If the given probabilities about the number of passengers are true in real, then the expected profit will be £260,000. However, if they will still choose to go for partnership with the incumbent operator, they may have loss of £54,000.
* Further, I will advise to hire a transport consultant for having more accurate assumption of service users. If his predicted probabilities will be true, they can expect to have $418,000 profit.
* The maximum profit they can earn is 508,000 pounds after achieving the target of having 81.53% high number of customers (based on -25% to 25% change).
* However, if consultant’s predictions get wrong or reduced to 25%. Still the optimal solution would not change.
* Getting into a partnership will not help them to generate the maximum profit.
* As the option of partnership is eliminated from the plan. So, London Coaches should be prepare to handle the sales and marketing on their own.

# Problem 3

## Linear Programming Model



**Fig. 3.1**

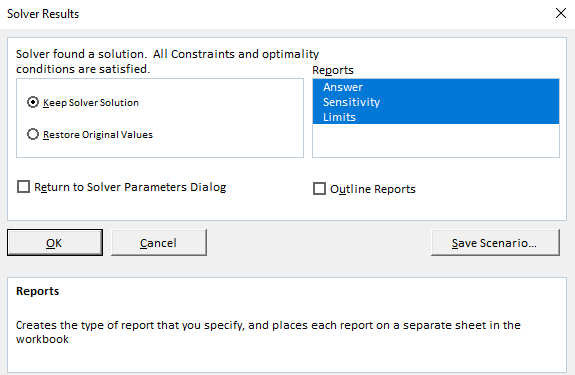
These five are the core variables. The Value of the whole model depends on these variables. It reveals that data about number of 1BHK, 2BHK, 3BHK and 4BHK can be constructed and total unit demolish.

**Constraints:**

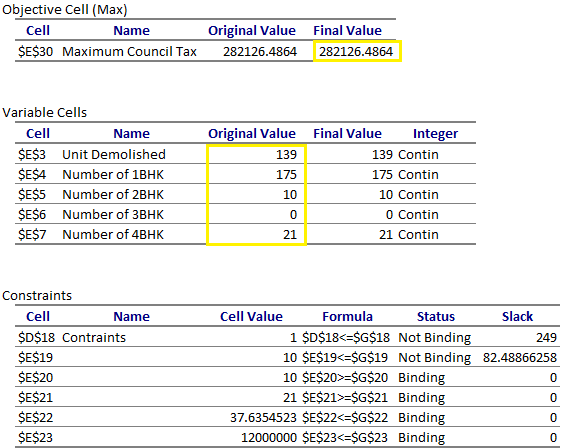
1. Unit demolished <= 250
2. 2BHK+3BHK < 45%
3. 2BHK > 5%
4. 4BHK >10%
5. Construction size <= (Unit demolished\* Free area(acre)/demolish unit)\*0.9
6. Construction cost <= 12,000,000

**Goal Maximum: (Number of 1BHK\*Council tax for 1 BHK) + (Number of 2BHK\*Council tax for 2BHK) + (Number of 3BHK\*Council tax for 3BHK) + (Number of 4BHK\*Council tax for 4BHK)**

## Solve the LP model



**Fig. 3.2**



**Fig. 3.3**

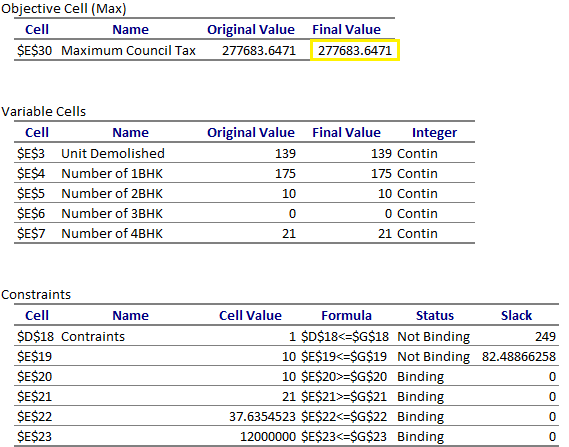
## Sensitivity report

1. There can be two reasons that why solver has not considered 3BHK in an optimal plan. First, 3BHK covers more than a double of area covered by 1BHK. It means two 1BHK can be constructed in the same space instead of one 3BHK. Moreover, there is not much difference in the council tax that can be generated by 3BHK and 1BHK and 2BHK. Also the cost of 3BHK construction is extremely high as compared to 1BHK. So, after considering these points it is obvious that 3BHK is not a good option to be included in the plan.

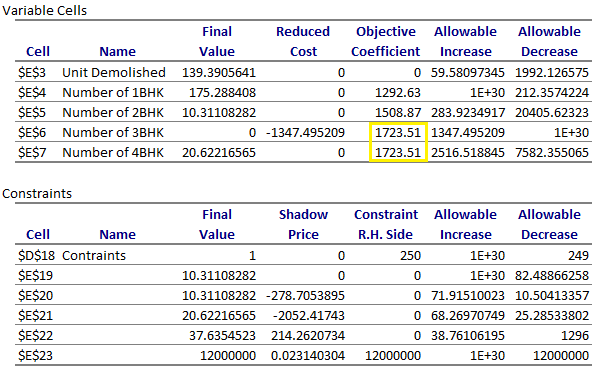
**Advice to the Council:**

* One solution can be increasing the council tax of 3BHK.
* Following the allowable increase, the number of 3BHK can be increase to check if there is any improvement in the council tax.

ii. Council Tax 3BHK= 4BHK



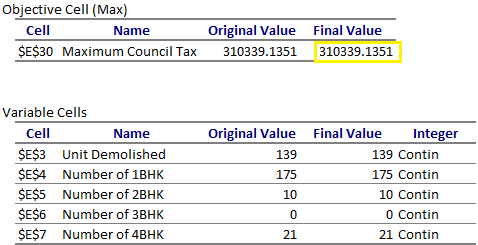
**Fig. 3.4**



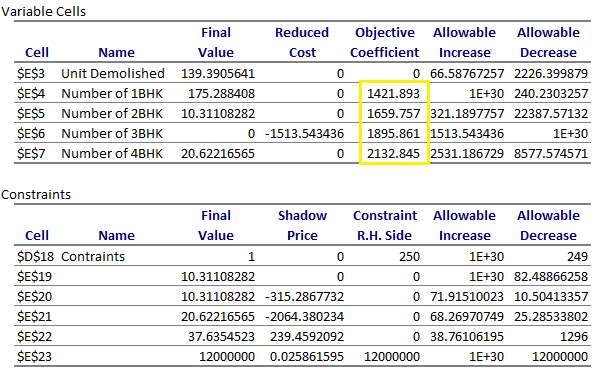
**Fig. 3.5**

There is no effect on the optimal plan. Even the maximum council tax is reduced by $4,442.8393.

### iii. Increasing council taxes by 10%



**Fig. 3.6**



**Fig. 3.7**

It can be noticed that the maximum council tax is increased to $310,339.1351 by increasing the council tax by 10%.

But, there is no change in the optimal plan in regards to adding 3BHK.

## Findings:

The Lewisham City Council can work on this proposal to maximize the tax and to develop the area. But, the optimal plan suggest to not add any 3BHK in the plan. As, it covers more space and costs more as compared to other housing type. Further ROI (Return on Investment) is low as well. So, the council can consider to build just the 1BHK, 2BHK and 4BHK in that area. Further to check for any increase the council tax, allowable increase and decrease matrices can be considered to have a trial if there is any improvement in the tax. To add the option of 3BHK into plan and maximize the overall council tax, tax for 3BHK can be increased.

# Problem 4

## Explore Time Series

### Graphs

### Trend/Seasonality

A picture containing text, screenshot, font, line

Description automatically generated

**Fig. 4.1**

### Seasonal indexes for each region

A picture containing text, screenshot, number, font

Description automatically generatedA picture containing text, screenshot, number, font

Description automatically generatedA picture containing text, screenshot, number, font

Description automatically generated

**Fig. 4.2**

### Appropriate forecasting methods

A picture containing text, screenshot, font, document

Description automatically generated

## Two appropriate forecasting methods

North

A screenshot of a report

Description automatically generated with medium confidence

**South**

A screenshot of a report

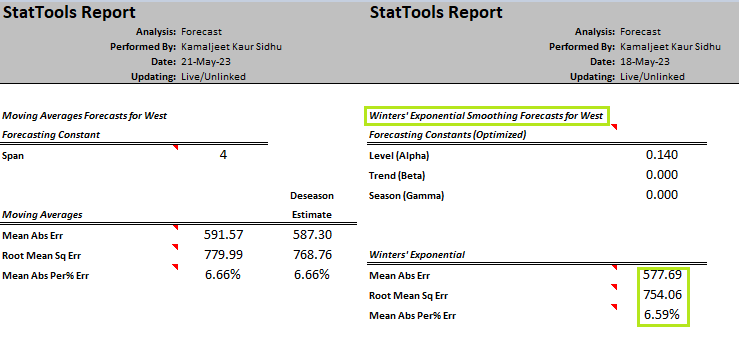
Description automatically generated with medium confidence

**East**

A screenshot of a report

Description automatically generated with medium confidence

**West**



Having less errors means the model is able to understand and interpret the past value. But, it is not guaranteed that the model is better at forecasting as well solely based on less errors. It can be analysed from abovementioned models, winter’s Exponential Smoothing model is good at understanding the past values where there is a seasonality in the dataset. On other hand, west region does not show any seasonality. Simple exponential smoothing model did good job for west region dataset.

### Forecasting Errors

**Autocorrelation:** It refers to the measurement of relationship between present and past values of a variable. Autocorrelation of +1 refers to a perfect positive correlation, however, -1 refers to a perfect negative correlation. In this technical report I have used it to measure the influence of past values on new values.

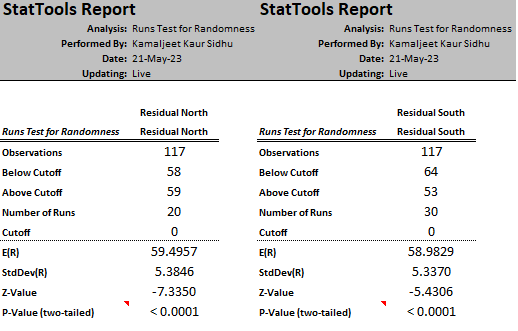
It can be analysed from the graph, in case of positive correlation an incline in one time series leads to a proportionate incline in other time series. Whereas, in case of negative correlation and increase noticed in one series results in a proportionate decrease in other time series.

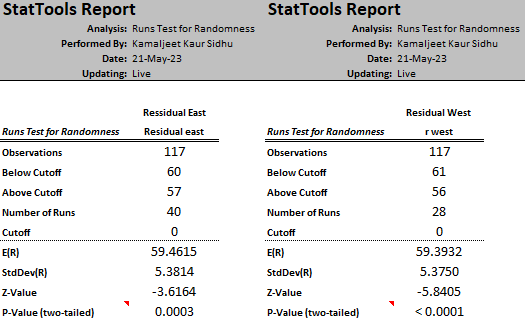
**Note:** Move cursor over bars to see correlation values.

**Randomness:**

Runs-test evolves around the measures described below. Null hypothesis of randomness can be rejected if the value of p is < 0.0001. It can be concluded that the series does not alternate too much (if number of runs are low) or vice versa.

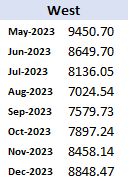
Now, it is clear that the data is not random for any of the region.





A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run. (Commerce, n.d.)

### Monthly Forecast (May 2023- December 2023)

**Visual Representation of Forecasted values(North,South,East and West)**

## Findings:

* From the historical data of 9 continuous year, it can be noticed that in South region the demand of AC (Air Conditioners) was higher as compared to other regions.
* The time period from January to June is noticed as a peak season for sales of AC in North.
* In South, it approximately always remain in demand. And, demand is increasing as time passes.
* In West and East, there is no specific time frame to be considered important for sales. But, a difference has been observed in the sales during first quarter and rest of the year.
* After analysing autocorrelation and randomness for forecasting errors. The insights I got about the data set is that in North, South and West region the increasing and decreasing trend was almost same during same time period for 9 years.
* For East region, the number of runs in runs-test for randomness are the highest in comparison with other regions. It means series alternate too much.

**Forecast:**

* As usual, the company will have maximum sales in South region.
* Although, a minor increase can be expected in the sales in West region still sales will be the lowest among other regions.
* The second higher sales can be expected in east followed by North.

# References

Commerce, U. D. (n.d.). *National Institute of Standards and Technology*. Retrieved from National Institute of Standards and Technology: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35d.htm#:~:text=The%20runs%20test%20(Bradley%2C%201968,the%20length%20of%20the%20run.