Yenneti, Uday Bhaskar Santosh Kumar

K2140746@kingston.ac.uk

CI7330 – Data Analytics & Visualisation

Summative Assessment 2021/22 - K2140746

I’ve used R programming language for this assignment.

**Imports:**

Text

Description automatically generated

Added ‘psych’ library for using describe function to get descriptive statistics.

Added ‘rstatix’ library for using statistical t\_test function.

Added ‘ggpubr’ library for using ggboxplot.

**Loading dataset from GIT repository:**

Reading the k2140746.csv file from git repository and loading the dataset to the sales\_dataset.



**Checking for null values in the columns:**



Using is.na() method check for null values in the columns in the dataset and found no null values in the data.

**1: Statistical summary of all the variables**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Count | Mean | SD | Min. | 1st Qu. | Median | 3rd Qu. | Max. | Range | SE |
| deprivation | 16000 | 32.798 | 10.05 | 5.509 | 25.477 | 33.315 | 40.011 | 60.518 | 55.01 | 0.08 |
| covid | 16000 | 0.5 | 0.5 | 0 | 0 | 0.5 | 1 | 1 | 1 | 0 |
| logspend | 16000 | 2.434 | 0.3 | 1.47 | 2.224 | 2.397 | 2.602 | 3.778 | 2.31 | 0 |
| spend | 16000 | 11.97 | 4.08 | 4.33 | 9.25 | 10.99 | 13.49 | 43.71 | 39.38 | 0.03 |

A screenshot of a computer

Description automatically generated with medium confidence

By using both ‘**summary**’ and **‘describe’** (psych library) method for getting above descriptive statistics. The dataset contains 4 columns, of which ‘deprivation’, ‘logspend’ and ‘spend’ are numeric continuous variables and ‘covid’ is Boolean. The dataset contains 16000 records, of which 8000 records are pre-covid and remaining are post-covid.

**‘Name’** represents the variable names. **‘Count’** gives the number of columns. **‘Mean’** is the average. Standard deviation(**'SD'**) is a measure of the spread of scores within a set of data. **‘Min’, ‘1st Qu.’,** **‘3rd Qu.’** and **‘Max’** represents minimum, first quartile (25%), third quartile (75%) and maximum values of the variable in the dataset. **‘Median’** is the middle value. **'Range'** indicates the difference between the lowest and highest values. Standard error(**'SE'**) indicates how different the population mean is likely to be from a sample mean.

**2: Are they correlated?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | deprivation | covid | Logspend | Spend |
| deprivation | 1.0000000 | 0.00000000 | -0.66274483 | -0.66816249 |
| covid | 0.0000000 | 1.00000000 | -0.06834616 | -0.08098647 |
| logspend | -0.6627448 | -0.06834616 | 1.00000000 | 0.97510269 |
| spend | -0.6681625 | -0.08098647 | 0.97510269 | 1.00000000 |

Text

Description automatically generated

The correlation coefficient of **1.0** indicates a perfect positive linear relationship between the two variables. The correlation coefficient of **-1.0** indicates a perfect negative linear relationship between the two variables. **0** indicates no linear correlation between the variables. Above table shows the correlation between the variables in the dataset. As logspend is the logarithmic value of spend it shows high positive correlation (about **0.97510269**) between them. Deprivation and logspend also displays strong negative correlation (about **-0.6627448**). Similarly, deprivation and spend also displays strong negative correlation (about **-0.6681625**). Attached the correlation plot below.



Diagram

Description automatically generated

**3: Did the spend change between pre-COVID and lockdown? Can you please do one of those tests that check if it is significant? I want to know if the difference is real.**

Text

Description automatically generated

Used Welch Two sample t-test statistical analysis in this case. From the above output, we can see that the difference in means for our sample data is 0.66149 (12.30327 – 11.64178), and the confidence interval shows that the true difference in means is between 0.5353281 and 0.7876544, the difference is not significant. t = 10.277 is the calculated t-statistics. The degrees of freedom are df = 15197. So, 95% of the time, the true difference in means will be different from 0. Our p-value of 2.2e-16 is much smaller than 0.05, so we can reject the null hypothesis of no difference and say with a high degree of confidence that the true difference in means is not equal to zero.

**4: can we use covid status and deprivation to predict average spend? can you give me a formula that would do that, and tell me how uncertain it is?**

A screenshot of a computer

Description automatically generated with medium confidence

I have used R inbuild function lm() to calculate the b0, b1 and b2 values. Also used excel to cross verify the values as attached in the below image.

A screenshot of a computer

Description automatically generated with low confidence



Ŷ = b0 + b1 \* (covid status) + b2 \* (deprivation value)

b0(intercept estimate) = 21.20921429

b1(covid estimate) = -0.661491

b2(deprivation estimate) = -0.27154063

Ŷ (average spend prediction) = 21.20921429 - 0.661491 \* (covid status) - 0.27154063 \* (deprivation value)

In the covid status insert 1 for non-covid and 0 for lockdown. In the deprivation value insert the relevant value of the neighbourhood.

**5: Also, can you give me two graphs, one for question 3, and one for question 4. I want to show them at the next board meeting, so please make them look professional.**

**Plot for Q3:**

Text

Description automatically generated with medium confidence

Text, letter

Description automatically generated

A picture containing chart

Description automatically generated

The above report can be interpreted as follows. The mean deprivation during covid lockdown group was 11.64178 (SD = 3.57), whereas the mean in non-covid group (**“0”**) was 12.30327 (SD = 4.51). A Welch two-samples t-test showed that the difference was statistically significant, t(15197.18) = 10.28, p < 0.0001, d = 6.57; where, t(26.9) is shorthand notation for a Welch t-statistic that has 26.9 degrees of freedom.

**Plot for Q4:**

Text

Description automatically generated

Graphical user interface, chart

Description automatically generated with medium confidence

When using the **plot()** function, the first plot is the **Residuals vs Fitted plot**, second plot is the **Q-Q Plot**, third plot is the **Scale-Location plot**, fifth plot is the **Residuals vs Leverage plot**.

**Residuals VS Fitted:**

For a correct linear regression, the data needs to be linear so this will test if that condition is met. With the sales data, we can see that it does not have any obvious pattern. It is curved, but the residuals are evenly distributed around the horizontal line without a distinct pattern. This indicates that it is not a nonlinear relationship.

**Normal Q-Q:**

This show how residuals are distributed. The plot indicates that residuals are normally distributed if they follow a straight line. However, in our case the distribution moves off the line considerably, which indicates a problem.

**Scale-Location:**

For our sales dataset the residuals are reasonably well spread above and below a slightly curved line however the beginning and ending of the line does have fewer points so slightly less variance there.

**Residuals vs Leverage:**

Cooks distance is not available in the plot to identify if there are any records available out of it. So, there is no influential data available.



Attached the R file.

# References

Bevans, R., 2020. *Choosing the Right Statistical Test | Types and Examples.* [Online]   
Available at: https://www.scribbr.com/statistics/statistical-tests/  
[Accessed 01 2022].

Brady, I., 2018. *Vignette - Residual analysis.* [Online]   
Available at: https://rpubs.com/iabrady/residual-analysis  
[Accessed 15 01 2022].

Data Flair, 2022. *Introduction to Hypothesis Testing in R – Learn every concept from Scratch!.* [Online]   
Available at: https://data-flair.training/blogs/hypothesis-testing-in-r/  
[Accessed 01 2022].

Kabacoff, R. I., 2017. *Descriptive Statistics.* [Online]   
Available at: https://www.statmethods.net/stats/descriptives.html  
[Accessed 01 2022].

Kabacoff, R. I., 2017. *Multiple (Linear) Regression.* [Online]   
Available at: https://www.statmethods.net/stats/regression.html  
[Accessed 01 2022].

Kabacoff, R. I., 2017. *t-tests.* [Online]   
Available at: https://www.statmethods.net/stats/ttest.html  
[Accessed 01 2022].

Prabhakaran, S., 2017. *How to implement common statistical significance tests and find the p value?.* [Online]   
Available at: https://www.machinelearningplus.com/statistics/statistical-significance-tests-r/  
[Accessed 01 2022].