Northeastern University

ASSIGNMENT FRONT SHEET

Course Name: **ALY6040 Data Mining Applications**

Professor Name: **Nagadeepa Shanmuganathan**

Student Name: **Dong Quoc Tuong (Lukas)**

Student Class: **Fall 2019 CPS** **Term: Winter 2021**

Module 2: Description Analysis

Completion Date: January 31st **Due Time:12:00am**

Statement of Authorship

I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University's regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.

☒ **Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

1. Loading data and packages

Exploratory data analysis (EDA) and Linear Regressions are the most fundamental pillars for anyone who are entering Machine Learning (Prasad Patil, 2018). EDA can be defined as the critical process of performing initial investigations on data to seek patters, anomalies and the validity of the hypothesis for graphical representations. In another words, familiarizing yourself with data before getting dirty.

As technology advanced throughout the years, marketers are faced with the difficult task to see which channel of marketing is the most effective. Thus, in this assignment, we will use the preloaded dataset in R named "marketing" in the "datarium" R package, dedicated Data Bank for Statistical Analysis and Visualization. The dataset studies the impact of three advertising medium (Youtube, Facebook, Newspaper) on Sales. The first three columns are the advertising budget in thousands of dollars along with the fourth column as sales. The experiment has been repeated 200 times which explains why we have 200 inputs at the end. Then we load the packages that we need to use, including : caTools, ggplot2, GGally

2. Exploratory data analysis (EDA)

Once it is loaded it makes sense to look at a few properties of the data. The very first thing people do is to look at a few samples of the dataset. A command called **head** is used.

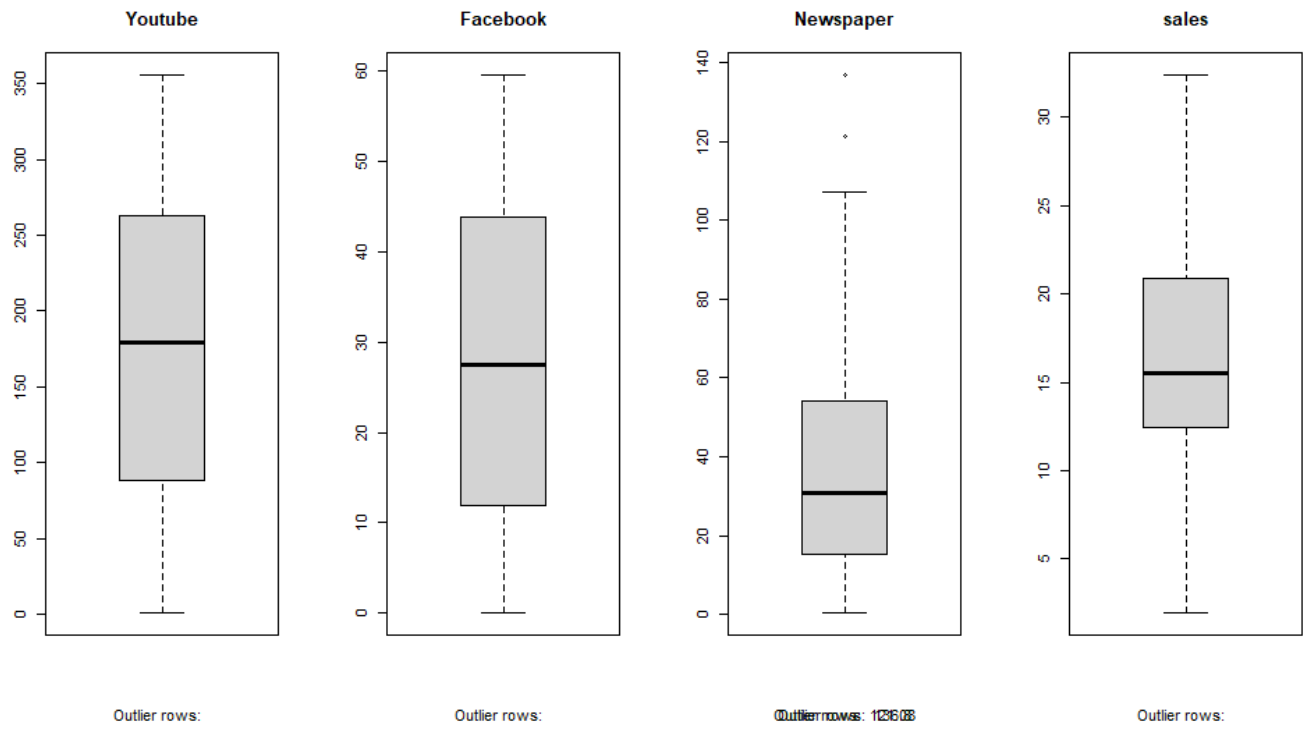
```
> head(marketing)
  youtube facebook newspaper sales
1  276.12    45.36     83.04 26.52
2   53.40    47.16     54.12 12.48
3   20.64    55.08     83.16 11.16
4  181.80    49.56     70.20 22.20
5  216.96    12.96     70.08 15.48
6   10.44    58.68     90.00  8.64
```

A command called **summary** gives you the basic statistics of your dataset like mean, median, 1st quartile, 2nd quartile etc. As we can see, Youtube, Facebook, Newspaper all have the same minimum starting value at 0 but then Yotube marketing expenditure rises dramatically to achieve the maximum at 355, nearly 7 times that of Facebook. Newspaper comes in second in term of maximum expenditure in experiment, just only half of that of YouTube's.

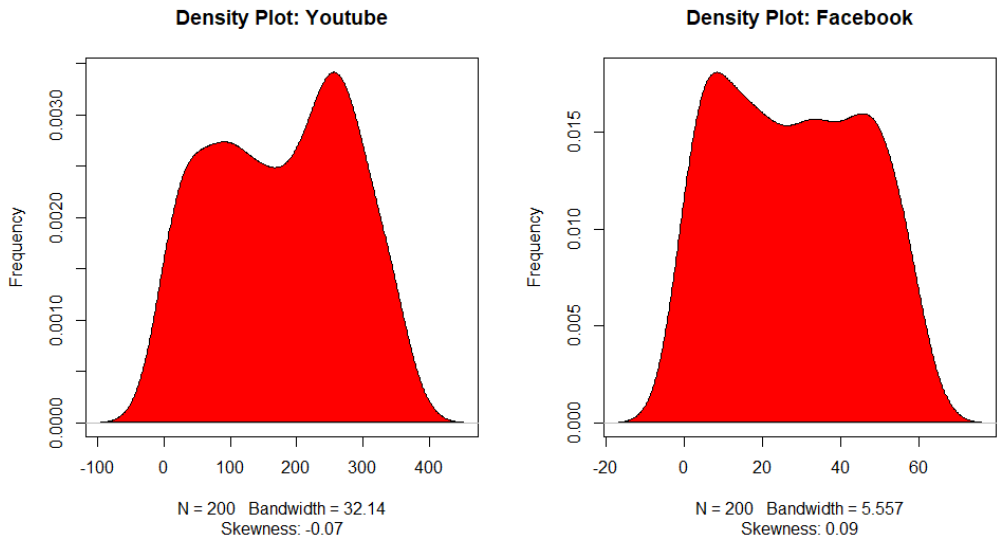
```
> summary(marketing)
    youtube          facebook       newspaper          sales      
 Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92  
 1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45  
 Median :179.70   Median :27.48   Median : 30.90   Median :15.48  
 Mean   :176.45   Mean   :27.92   Mean   : 36.66   Mean   :16.83  
 3rd Qu.:262.59   3rd Qu.:43.83   3rd Qu.: 54.12   3rd Qu.:20.88  
 Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40  
```

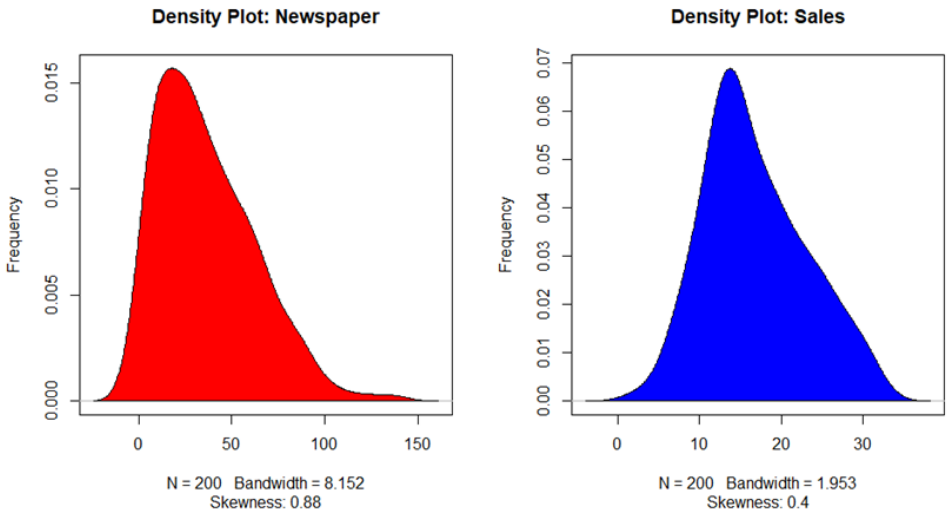
Next, we created a boxplot to examine the distribution of the dataset for each com components. It is clearly shown that only Youtube, Facebook, Sales are equally distributed (Sales are slightly right skewed). Newspaper boxplot, on the other hand, heavily skews towards to the right, indicating that the majority of marketing expenditure for Newspaper is relatively small amount, and only a few expenditures are significant. The two outliners only appear at the right end of the Newspaper boxplot, signifying that there could be some big spending for Newspaper marketing

throughout the years, perhaps to advertise annual discounts or product launches. Normally, we need to eliminate the extreme outliners in our project to ensure the most accurate prediction at later on. Nevertheless, there are only 2 outliners in the Newspaper, it is not worth worrying about

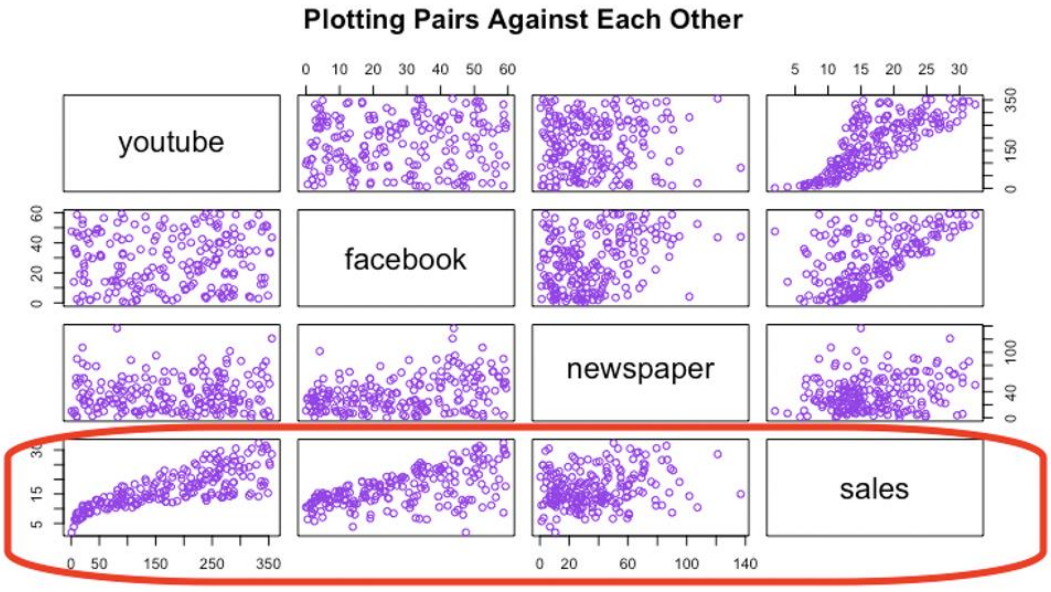


When we look at the density plot s, we also see that they resonate with the information we extracted above in term of distribution and skewness. Youtube and Facebook plots are closed to zero skewness, follow by Sales plot with 0.4, Newspaper is the most extreme with 0.8 for skewness ratio

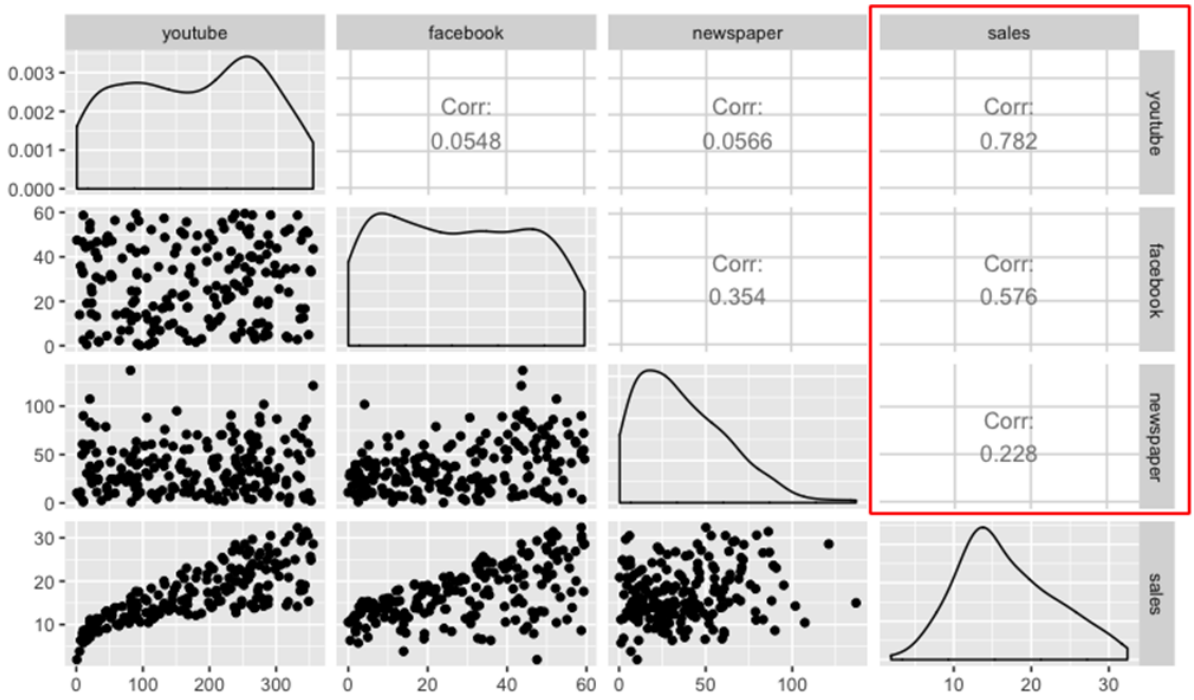




Given to the fact that this is a linear regression assignment to examine how sales vary with the advertising budget, we should look to plot the data pairwise. The last (highlighted) of [plots is the most useful as it illustrates how various advertising budgets affect sales. From a respective point of view, Youtube and Facebook budgets (1st two plots in the highlighted row) show strong correlations with the growth of the advertising budget. We cannot, unfortunately, see any particular trend in the third plot (newspaper)



Another way to see the correlations with different manner is using “ggplot” library. Correlation between each media and corresponding sales numbers. The diagonal consists of the densities of the three variables and the upper panels consist of the correlation coefficients between the variables. Looking to the highlighted part on the right, we can see the Correlation ratio between sales and respective columns, Youtube at 0.78, Facebook at 0.58 and Newspaper at the lowest with around 0.23



3. Preparing the data

To prepare the data, we split them into training set and test set with the normal ratio of 75% (150 rows will go to the train set and 50 rows will go to the test set. We also make sure to assign a seed

(101) to replicate so that we can generate the same sample with caTools in the future. The final product I the RStudio section looks like this:

test_size	int [1:2] 50 4
train_size	int [1:2] 150 4

4. Creating the model

This is a case of multiple linear regressions. We have **'n'** variables that combine linearly to provide us with our output. Its equation looks like the following. Beta **(ß)** are the coefficients that control the effect of any variable **(x)** has on the output **(y)** (Rebecca Bevans, 2020)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n$$

If n= 1 then it reduced to a simple linear regression: **y=mx+c. For** the **marketing** dataset however, we have three different variables: Youtube, Facebook and Newspaper. Hence the final equation would look something like this:

$$sales = \beta_0 + \beta_1 * youtube + \beta_2 * facebook + \beta_3 * newspaper$$

As a result, our goal now is for us to find the betas **(ß)** and examine how accurately they sales number predictions are from those betas. With the summary function in R return the model properties like this:

```
> summary(Model)

Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Firstly, let’s have a look at the F-statistic and associated P-value at the bottom of the summary. In our example, it can be seen that P-value of the F-statistic is < 2.2e-16. This small number means that, at least, one of the predictor variables (Youtube, FB or newspaper) is significantly related to the outcome variable (Sales).

Secondly, the estimates column suggests that changes in the Youtube and Facebook advertising budget are significantly linked to changes in Sales, on the contrary tp the newspaper budget. Such phenomenon has been pointed out above. As an example:

- Spending an additional 10.000 dollars on Facebook advertising can result in a growth of 0.1885*10.000=1890 sales units, on average.

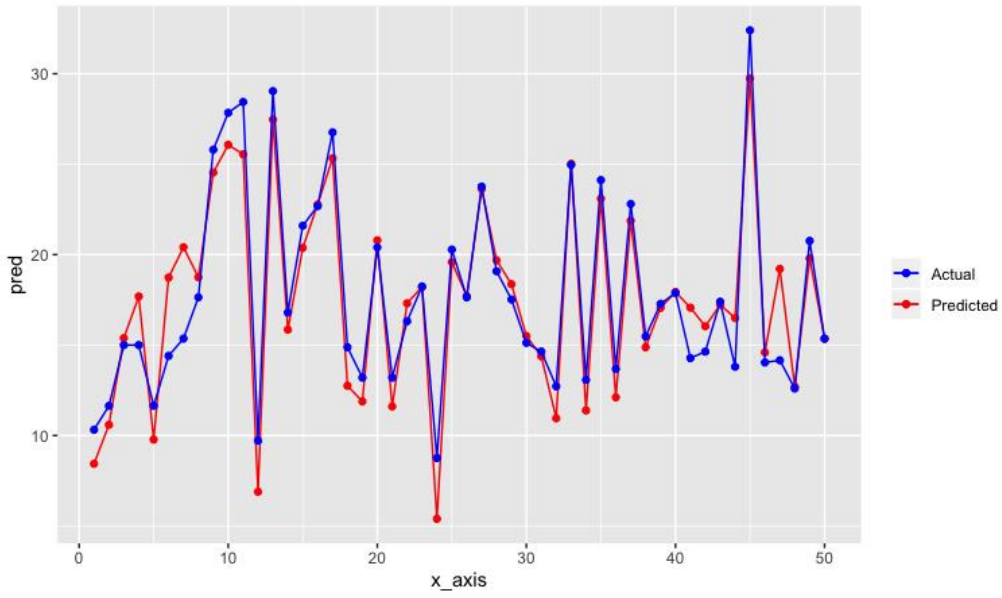
- Spending an additional 10.000 dollars on Youtube advertising can result in a growth of 0.045*10.000=450 sales units, on average.

The same value for newspaper is small and not significant enough to count in the model. Hence, it would be much better to eliminate it completely from our analysis and have the final model be:

Sales = 3.5 + 0.045* Youtube + 0.187* Facebook

5. Model accuracy analysis

One way to know how good a model is to use our test set that we had separated earlier and compute the prediction values from that. After that, compare them to the actual qualitative values to see our model performance. The analysis can be assessed below with Red points representing predictions and Blue illustrating the actual results



It seems that our predictions are very close to the real quantitative results. But it is much better to look at actual regression model valuation metrics like Mean absolute error, Mean Squared Error,

Root Mean Squared Error, Coefficient of determination (R^2) that are mainly used for to evaluate the prediction error rates and model performance in regression analysis. (Vimarsh Karbhari, 2018)

- **Mean absolute error:** the difference between the original and predicted values. Obtain by averaging the absolute difference over the data set.
- **Mean Squared Error:** the difference between the original and predicted values. Obtain by averaging the squared difference over the data set.
- **Root Mean Squared Error:** the square root of the arithmetic mean of the squares of difference over the set.
- **Coefficient of determination:** the coefficient of how well the values fit compared to the original values. The value from 0 - 1 interpreted as percentages. The higher R^2, the better

```
MAE: 1.53983
MSE: 4.187806
RMSE: 2.046413
R-squared: 0.9001034
```

For our model R-Squared error comes out to be 0.90 which is a pretty significant and as a result we can trust in our result.

References

Prasad Patil. (2018). What is Exploratory Data Analysis? Retrieved January 31, 2021, from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Rebecca Bevans. (2020). An introduction to multiple linear regression. Retrieved January 31, 2021, from https://www.scribbr.com/statistics/multiple-linear-regression/

Vimarsh Karbhari. (2018, December 18). How to evaluate regression models? . Retrieved January 31, 2021, from https://medium.com/acing-ai/how-to-evaluate-regression-models-d183b4f5853d