VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH UNIVERSITY OF TECHNOLOGY

Probability and Statistics
assignment

Instructor:
Nguyễn Tiến Dũng

Author:

Name	Student ID	Class
Nguyễn Nam Kha	2052515	CC09
Nguyễn Ngọc Hòa	2052485	CC09
Dương Ngọc Quang Huy	2052489	CC09
Hứa Hoàng Nguyên	2052619	CC09
Trần Khoa	2052541	CC09

Ho Chi Minh city, 05/2022

Group contribution

Name	**Student ID**	**Task**	**Contribution percentage**	**Completion percentage**
Nguyễn Nam Kha	2052515	Part 1, 3.1, graphs, 4	20%	100%
Nguyễn Ngọc Hòa	2052485	Part 2.1, 2.5, 3.2.2, 3.8	20%	100%
Dương Ngọc Quang Huy	2052489	Part 2.6, 3.4, 3.9	20%	100%
Hứa Hoàng Nguyên	2052619	Part 2.3, 2.4, 3.6.4, 3.7.2	20%	100%
Trần Khoa	2052541	Part 2.2, 3.3.1, 3.3.2	20%	100%

Contents

1 Introduction

In this day and age, the world has become more globalized and along with it is the rapid growth of the Internet. The more people getting connected to the Internet, the more websites appear. Therefore, a webmaster needs to improve their websites quality so that they can attract more visitors as well as compete with other websites in the same field. To do this, the webmaster often uses Google Analytics (GA) – a web analytics service offered by Google that tracks and reports website traffic. GA can statistics some important data of a website such as number of Users, Sessions, Pageviews,… Among those data that GA can statistics of, there are three important data that are often used to evaluate the quality of a website, which are Exit Rate, Bounce Rate and Page Value.

Page Value is the average value that the page contributes to our website's revenue. An Exit occurs when a user exits your website after landing on a page. A Bounce, on the other hand, occurs whenever a user enters the page and subsequently exits without visiting any other pages on that website or interacting with any of the elements on the page (e.g. commenting). Therefore, Exit Rate is determined by calculating total number of Exits over total number of views the page received, and Bounce Rate is determined by calculating total number of Bounces over total number of times that the page is visited first.

In short, Bounce Rate is a type of Exit Rate, so Bounce Rate can affect Exit Rate, but not vice versa. The statistics of Bounce Rate and Exit Rate as well as Page Value can help webmasters know which pages are having problems that make users tend to leave at that page, therefore they could upgrade and improve the website promptly. In this essay, we will analyze the relationship between those three variables and see how they can reflect the quality of a website.

About the data

We use the dataset compiled by Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). The dataset predicts the correlation between visitor's shopping intention and website abandonment likelihood, using variables like ExitRates, BounceRates and PageValues. It has 12330 observations (rows) and 18 variables (columns). All the variables are stated in the table below:

No.	Name	Type of data
1	Administrative	Number
2	Administrative_Duration	Number
3	Informational	Number
4	Informational_Duration	Number
5	ProductRelated	Number
6	ProductRelated_Duration	Number
7	BounceRates	Number
8	ExitRates	Number
9	PageValues	Number
10	SpecialDay	Number
11	Month	Text
12	OperatingSystems	Number
13	Browser	Number
14	Region	Number
15	TrafficType	Number
16	VisitorType	Text
17	Weekend	Text
18	Revenue	Text

This essay will focus on BounceRates, ExitRates and PageValues, using statistics to analyze the relationship between those three variables. The progress will be conducted by the following steps:

Step 1: *Sample the data*. This step includes importing data, cleaning data and sampling data.

Step 2: *Compare the means of the data*, using Boxplot and t-test

Step 3: *Build the linear model*, using R built-in command.

Step 4: *Test the linearity of the data*, using Residuals vs Fitted plot

Step 5: *Test the normality of the data*, using Histogram, Density curve, Q-Q plot and Shapiro-Wilk test

Step 6: *Test the homoscedasticity of the data*, using Scale-location plot and Breusch-Pagan test

Step 7: *Test the independence of the data*, using Durbin - Watson test

Step 8: *Analyze the linear model*, using R built-in command.

For all the tests, we will use the confidence interval of 95% which is assumed by RStudio program by default.

2 Theory

2.1 t-test

The Paired Samples t Test compares the means of two measurements taken from the same individual, object, or related units. The purpose of the test is to determine whether there is statistical evidence that the mean difference between paired observations is significantly different from zero.

The variable used in this test is known as: dependent variable, or test variable (continuous), measured at two different times or for two related conditions or units. The data must meet the following requirements:

- Dependent variable that is continuous
- Related samples/groups (i.e., dependent observations)
- Random sample of data from the population
- Normal distribution (approximately) of the difference between the paired values
- No outliers in the difference between the two related groups

Hypotheses

- H_0: $\mu_1 - \mu_2 = 0$ (the difference between the paired population means is 0)
- H_1: $\mu_1 - \mu_2 \neq 0$ (the difference between the paired population means is not 0)

Performing t-test

The test statistic for the Paired Samples $t-test$, denoted t, follows the same formula as the one sample $t-test$.

$$t = \frac{\overline{x}_{diff} - 0}{s_{\overline{x}}}$$

where

$$s_{\overline{x}} = \frac{s_{diff}}{\sqrt{n}}$$

where

$\overline{x}_{diff}$: sample mean of the differences.

n : sample size.

s_{diff}: sample standard deviation of the differences.

$s_{\overline{x}}$: estimated standard error of the mean ($s/\sqrt{n}$).

Rejection region: $t > t_{\alpha,n-1}$

2.2 Shapiro-Wilk test

The Shapiro–Wilk test is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk.

This test is a way to tell if a random sample comes from a normal distribution through testing the null hypothesis. The test statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

where:

x_i: the ordered random sample values

a_i: constants generated from the covariances, variances and means of the sample from a normally distributed sample.

The null hypothesis of this test is that the population is normally distributed. Thus, when the p-value of the test is smaller than the chosen alpha level, then the null hypothesis is rejected and there will be enough evidence to prove the tested data is not normally distributed. On the other hand, if p-value is greater than the chosen alpha level, then we can not reject the null hypothesis.

2.3 Breusch-Pagan Test

The Breusch-Pagan test developed in 1979 by Trevor Breusch and Adrian Pagan, is used to test the heteroskedasticity n a linear regression model. The test is used to determine whether or not heteroscedasticity is present in a regression model.

The test uses the following null and alternative hypotheses:

- Null Hypothesis (H_0): Homoscedasticity is present (the residuals are distributed with equal variance)
- Alternative Hypothesis (H_1): Heteroscedasticity is present (the residuals are not distributed with equal variance)

There are 3 steps for this test statistics:

- Step 1: Apply OLS in the model:

$$y_i = X_i\beta + \varepsilon_i, \qquad i = 1, \dots, n$$

- Step 2.1: Compute the regression residuals, $\widehat{\varepsilon_i}$, square them, and divide by the Maximum Likelihood estimate the error variance from the step 1 regression, to obtain what the test call g_i.

$$g_i = \frac{\widehat{\varepsilon_i}}{\hat{\sigma}^2}, \qquad \hat{\sigma}^2 = \sum \widehat{\varepsilon_i}^2/n$$

- Step 2.2: Estimate the auxiliary regression

$$g_i = \gamma_1 + \gamma_2 z_{2i} + \cdots + \gamma_p z_{pi} + n_i$$

  Where the z terms will typically but not necessarily be the same as the original covariates x.

- Step 3: The LM test statistic is then half of the explained sum of squares from the auxiliary regression in Step 2:

$$LM = \frac{1}{2}(TSS - SSR)$$

Where TSS is sum of squared deviations of the g_i from the mean of 1, and SSR is the sum of squared residuals from the auxiliary regression.

2.4 Durbin-Watson test

The Durbin-Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The DW statistic ranges from zero to four, with a value of 2.0 indicating zero autocorrelation. Values below 2.0 mean there is positive autocorrelation and above 2.0 indicates negative autocorrelation.

Hypotheses

- H_0: First-order autocorrelation does not exist.
- H_1: First-order autocorrelation exists.

Assumptions

- Errors are normally distributed with a mean value of 0
- All errors are stationary

Performing DW test

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} {e_t}^2}$$

where

- e_t is the residual figure.
- T is the number of observations of the experiment.

2.5 Linear regression

Linear regression is fundamental and commonly used in predictive analysis to observe and model the relationship between variables. By building a linear regression model, we can understand how strong the relationship is between variables, and find out how much the dependent variable would change by changing the independent variables, thus we can estimate the dependent value from the collected independent values.

Simple linear regression

As its name implies, this model is used to evaluate the relationship between the dependent variable or response variable Y and only one predictor X.

The expected or estimated value of y:

$$E(y) = \beta_0 + \beta_1 x$$

- y : Dependent variable
- x : Independent variable
- β_0, β_1: Intercepts

The value of y can also be present by:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- ε : The error or residual of the estimation. $\varepsilon = y - E(y)$

In order to apply the linear regression model, the collected dataset must satisfy some assumptions:

1. **Homogeneity (homoscedasticity)** : the range value of error or residual remains stable and does not have significant change across the dataset.
2. **Independence:** The observation in the dataset must have no relationship between each other.
3. **Normality:** The dataset collected must be normally distributed.
4. **Linearity:** The relationship between variables must be linear across the dataset.

Least-squares method

This is the most common method to create the line of best fit for the dataset. This method calculates the best fit line for the data by minimizing the sum of squares of the residual from each point in the dataset to the line, since the residual or the deviation of the point to the line can be negative, we must take the squares of them to use for calculation.

The least-squares estimates of the intercept and slope in the simple linear regression model are:

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$b_0 = \overline{y} - \beta_1 x$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} {x_i}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} {y_i}^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2$$

Sum of squares of the residuals:

$$SSE = S_{yy} - b_1 S_{xy}$$

Sum of squares of the response variable:

$$SST = S_{yy}$$

Sum of squares for regression:

$$SSR = b_1 S_{xy}$$

It can be shown that:

$$E(SSE) = (n-2)\sigma^2$$

An unbiased estimator of σ^2

$$S^2 = \frac{SSE}{n-2}$$

Coefficient of determination

$$r^2 = 1 - \frac{SSE}{SST}$$

Properties of Least Squares Estimator

- Slope properties

$$E(b_1) = \beta_1, V(b_1) = \frac{\sigma^2}{S_{xx}}$$

Estimates standard error of the slope

$$S_{b_1} = \sqrt{\frac{S^2}{S_{xx}}}$$

Moreover

$$T = \frac{b_1 - \beta_1}{S_{b_1}} \sim t(n-2)$$

CI

$$b_1 \pm t_{v/2,n-2} \times S_{b_1}$$

- Intercept properties

$$E(b_0) = \beta_0, V(b_0) = \frac{\sigma^2 \mu_{xx}}{S_{xx}} \quad (with\ \mu_{xx} = \frac{1}{n}\sum {x_i}^2)$$

Estimates standard error of the slope

$$S_{b_0} = \sqrt{\frac{S^2 \mu_{xx}}{S_{xx}}} = \sqrt{{S_b}^2 \mu_{xx}}$$

Moreover

$$T = \frac{b_0 - \beta_0}{S_{b0}} \sim t(n-2)$$

CI

$$b_0 \pm t_{v/2,n-2} \times S_{b_0}$$

3 Implementation in R

3.1 Sampling data

First we have to import the required packages. `install()` command is used to install the packages, `library()` command is used to load the packages into our R file.

```
#Load packages
install.packages("ggplot2")
install.packages("dplyr")
install.packages("broom")
install.packages("ggpubr")
install.packages("moments")
install.packages("hms")
install.packages("car")
install.packages("lmtest")
library("ggplot2")
library("dplyr")
library("broom")
library("ggpubr")
library("moments")
library("hms")
library("car")
library("lmtest")
```

Then we import the dataset, using command `read.csv()`.

```
#Load file
source <- read.csv("D:/dataset.csv")
```

ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems
1	0.00000	0.00000000	0.00000000	NA	0.0	May	1
1	0.00000	0.00000000	0.00000000	NA	0.0	May	2
3	311.00000	0.11428731	0.03309439	NA	0.6	May	1
160	2418.52904	0.41710348	0.31412998	0.5789763	0.0	May	2
39	905.60000	0.34106330	0.29529537	NA	0.0	May	3
8	100.00000	0.15411000	0.08066976	NA	1.0	May	2
30	986.50000	0.39394726	0.28624381	NA	0.0	May	2
32	3121.55556	0.36818528	0.23946547	0.3292828	0.0	May	2
34	1378.96667	0.38087888	0.31519596	0.3602072	0.0	May	3
49	3003.41905	0.35769737	0.32321338	0.4649747	0.4	May	1
43	2053.93333	0.34067310	0.23294447	NA	0.0	May	2
4	212.50000	NA	0.23018571	NA	0.0	May	2
29	785.47436	NA	0.49559067	0.3041587	0.0	May	2
23	316.33333	0.37661551	0.25816048	NA	1.0	May	2
28	769.50000	0.28249839	0.22360389	0.3562230	0.0	May	2
15	298.16667	0.33767711	0.21832869	NA	0.4	May	2
32	1808.33333	0.20467877	0.15374294	0.3912999	0.4	May	3

The image above is a fraction of the dataset. As we can see, there are cells in the dataframe whose value is NA (non-available). These values are not useful in the process of statistical analysis so we have to drop out all the rows containing these cells, using command `na.omit()`.

```
#Clean data
source <- source %>%
  na.omit()
```

After data cleaning, there are 1538 observations left in the dataset. However this size is still too large for statistical purposes so we have to do the sampling. The method we use is using command `filter()` to take the rows whose number is divisible by 13, the new dataset will be saved in the dataframe called `mini`.

```
#Sampling
mini <- source %>%
  filter(row_number() %% 13 == 0)
```

After sampling, there are 118 observations left in the dataset. This is an ideal size for conducting the statistical analysis.

We can use command `summary()` to get some important numerical summaries of the data.

```
#Summary statistic value of the dataframe
summary(mini)
```

```
> summary(mini)
 Administrative   Administrative_Duration Informational  Informational_Duration
 Min.   : 0.000   Min.   :   0.00         Min.   :0.00   Min.   :   0.0
 1st Qu.: 2.000   1st Qu.:  47.75         1st Qu.:0.00   1st Qu.:   0.0
 Median : 5.000   Median : 115.10         Median :0.00   Median :   0.0
 Mean   : 5.525   Mean   : 198.88         Mean   :1.28   Mean   : 122.8
 3rd Qu.: 8.000   3rd Qu.: 251.55         3rd Qu.:2.00   3rd Qu.: 101.1
 Max.   :18.000   Max.   :1443.10         Max.   :9.00   Max.   :1665.1
 ProductRelated   ProductRelated_Duration  BounceRates       ExitRates
 Min.   : 10.00   Min.   : 276.7          Min.   :0.2120   Min.   :0.1326
 1st Qu.: 26.25   1st Qu.: 902.7          1st Qu.:0.3350   1st Qu.:0.2628
 Median : 47.00   Median :1721.0          Median :0.3860   Median :0.2900
 Mean   : 66.46   Mean   :2422.7          Mean   :0.3848   Mean   :0.2955
 3rd Qu.: 81.00   3rd Qu.:3241.9          3rd Qu.:0.4362   3rd Qu.:0.3351
 Max.   :292.00   Max.   :9595.6          Max.   :0.5938   Max.   :0.4185
   PageValues       SpecialDay        Month           OperatingSystems    Browser
 Min.   :0.1168   Min.   :0.00000   Length:118         Min.   :1.000    Min.   :1.000
 1st Qu.:0.3304   1st Qu.:0.00000   Class :character   1st Qu.:2.000    1st Qu.:2.000
 Median :0.3898   Median :0.00000   Mode  :character   Median :2.000    Median :2.000
 Mean   :0.3884   Mean   :0.05254                      Mean   :2.144    Mean   :2.178
 3rd Qu.:0.4357   3rd Qu.:0.00000                      3rd Qu.:3.000    3rd Qu.:2.000
 Max.   :0.6411   Max.   :0.80000                      Max.   :4.000    Max.   :8.000
    Region       TrafficType     VisitorType         Weekend          Revenue
 Min.   :1.00   Min.   : 1.000   Length:118         Mode :logical   Mode :logical
 1st Qu.:1.00   1st Qu.: 1.000   Class :character   FALSE:96        FALSE:58
 Median :2.00   Median : 2.000   Mode  :character   TRUE :22        TRUE :60
 Mean   :2.78   Mean   : 3.678
 3rd Qu.:4.00   3rd Qu.: 4.000
 Max.   :9.00   Max.   :20.000
```

The table shows us the minimum, 1st quartile, median, mean, 3rd quartile and maximum value of the variables.

3.2 Comparisons of means

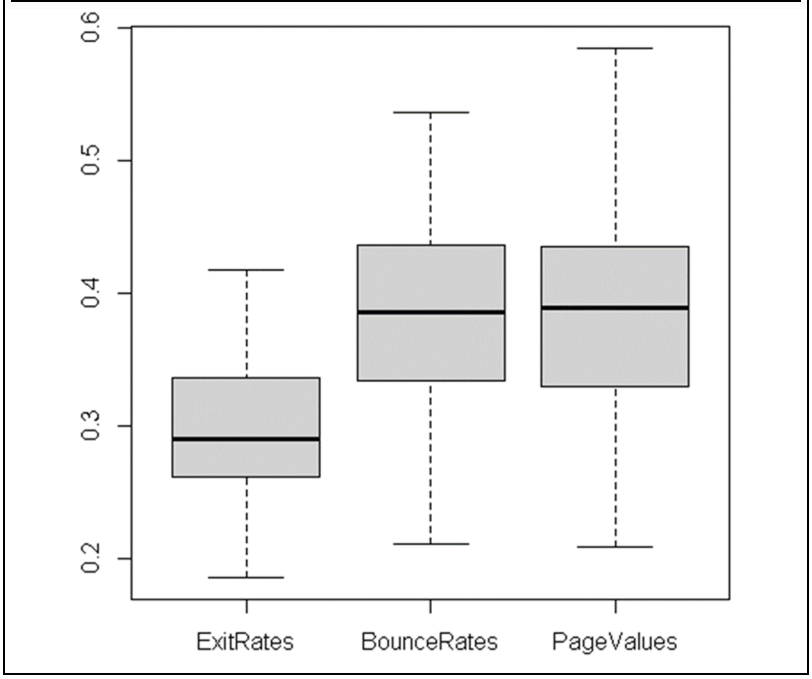
In this section, we compare the means between BounceRates, ExitRates and PageValues. There are many ways to compare the means of the dataframe. In this essay we will use 2 ways:

- Drawing boxplot
- t-test

3.2.1 Drawing boxplot

Boxplot can be used to compare the means between two data. We draw the boxplot by using command `boxplot()`.

```
#Drawing boxplot
boxplot(mini$ExitRates, mini$BounceRates,
mini$PageValues, outline = FALSE, names =
c("ExitRates", "BounceRates", "PageValues"))
```



The blacklines in the center of three boxes are the medians of the data. If we project those blacklines onto the y-axis, we can see that BounceRates stays near PageValues, while ExitRates stays quite far from the other two. So we can conclude that there is no difference between the medians of the BounceRates and PageValues, while there is a significant difference between ExitRates and the other two variables.

The dataset follows a normal distribution (which will be proved in section 3.6), so the value of median is approximately equal to the value of mean. Therefore, we can assume that there is no difference between the means of the BounceRates and PageValues, while there is a significant difference between the mean of ExitRates and the other two variables. However, to be more accurate, we will conduct the t-test in the following section.

3.2.2 t-test

t-test is used to compare the means between 2 related groups of samples. First we will test if there is a significant difference between the means of ExitRates and BounceRates, using command `t.test()`.

```
#t-test ExitRates and BounceRates
t.test(mini$ExitRates, mini$ BounceRates, mu = 0, alt
= "two.sided", cont = 0.95, var.eq = T, paired = F)
```

```
> t.test(mini$ExitRates, mini$BounceRates, mu = 0, alt = "two.sided",
 cont = 0.95, var.eq = T, paired = F)

        Two Sample t-test

data:  mini$ExitRates and mini$BounceRates
t = -10.43, df = 234, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.10615641 -0.07242519
sample estimates:
mean of x mean of y
0.2954675 0.3847583
```

Hypothesis:

- H_0: $\mu_1 = \mu_2$, ExitRates and BounceRates means are equal.
- H_1: $\mu_1 \neq \mu_2$, ExitRates and BounceRates means are not equal.

Results:

- t is the t-statistic value (t = -10.43).
- df is the degrees of freedom (df = 234).
- p-value is the significance level of the t-test.
- The mean value is 0.2954674 for ExitRates and 0. 3847583 for BounceRates.

The resulting p-value is much less than 0.05. Therefore, we can reject the null hypothesis H_0 and conclude that there is difference in ExitRates and BounceRates means.

Then we will test if there is a significant difference between the means of ExitRates and PageValues.

```
#t-test ExitRates and PageValues
t.test(mini$ExitRates, mini$PageValues, mu = 0, alt =
"two.sided", cont = 0.95, var.eq = T, paired = F)
```

```
> t.test(mini$ExitRates, mini$PageValues, mu = 0, alt = "two.sided",
 cont = 0.95, var.eq = T, paired = F)

        Two Sample t-test

data:  mini$ExitRates and mini$PageValues
t = -9.95, df = 234, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.11137293 -0.07455754
sample estimates:
mean of x mean of y
0.2954675 0.3884327
```

Hypothesis:

- H_0: $\mu_1 = \mu_2$, ExitRates and PageValues means are equal.
- H_1: $\mu_1 \neq \mu_2$, ExitRates and PageValues means are not equal.

Results:

- t is the t-statistic value (t = -9.95).
- df is the degrees of freedom (df = 234).
- p-value is the significance level of the t-test.
- The mean value is 0.2954675 for ExitRates and 0.3884327 for PageValues.

The resulting p-value is much less than 0.05. Therefore, we can reject the null hypothesis H_0 and conclude that there is difference in ExitRates and PageValues means.

Finally we will test if there is a significant difference between the means of BounceRates and PageValues.

```
#t-test BounceRates and PageValues
t.test(mini$BounceRates, mini$PageValues, mu = 0, alt
= "two.sided", cont = 0.95, var.eq = T, paired = F)
```

```
> t.test(mini$BounceRates, mini$PageValues, mu = 0, alt = "two.side
d", cont = 0.95, var.eq = T, paired = F)

        Two Sample t-test

data:  mini$BounceRates and mini$PageValues
t = -0.35079, df = 234, p-value = 0.7261
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02431106  0.01696219
sample estimates:
mean of x mean of y
0.3847583 0.3884327
```

Hypothesis:

- H_0: $\mu_1 = \mu_2$, BounceRates and PageValues means are equal.
- H_1: $\mu_1 \neq \mu_2$, BounceRates and PageValues means are not equal.

Results:

- t is the t-statistic value (t = -0.35079).
- df is the degrees of freedom (df = 234).
- p-value is the significance level of the t-test (p-value = 0.7261).
- The mean value is 0. 3847583 for BounceRates and 0. 3884327 for PageValues.

The resulting p-value is higher than 0.05 (p-value = 0.7261 > 0.05). Therefore, we cannot reject the null hypothesis H_0 and conclude that there is no difference in BounceRates and PageValues means.

3.3 Building the linear regression model

In the previous section, we have proven that there is a significant effect of BounceRates on ExitRates, while PageValues shows no effect on ExitRates. Therefore in this section, we will build a linear regression model which takes ExitRates as dependent variable (Y) and BounceRates as independent variable (X). The linear model is built using command `lm()`.

```
#Build linear regression model
mini.model <- lm(mini$ExitRates ~ mini$BounceRates,
data=mini)
```

There are four assumptions associated with a linear regression model:

- *Linearity*: The relationship between X and the mean of Y is linear.
- *Normality*: For any fixed value of X, Y is normally distributed.
- *Homoscedasticity*: The variance of residual is the same for any value of X.
- *Independence*: Observations are independent of each other.

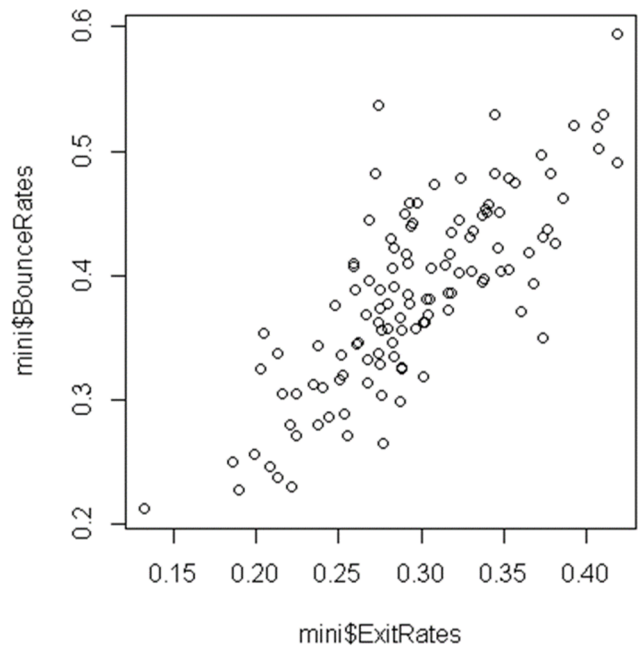
Now we will test if these four assumptions are met.

3.4 Testing for linearity

3.4.1 Drawing Scatter plot

Scatter plot can be used to test the linearity of the data, using this command

```
#Drawing Scatter plot
plot(mini$ExitRates,mini$BounceRates)
```

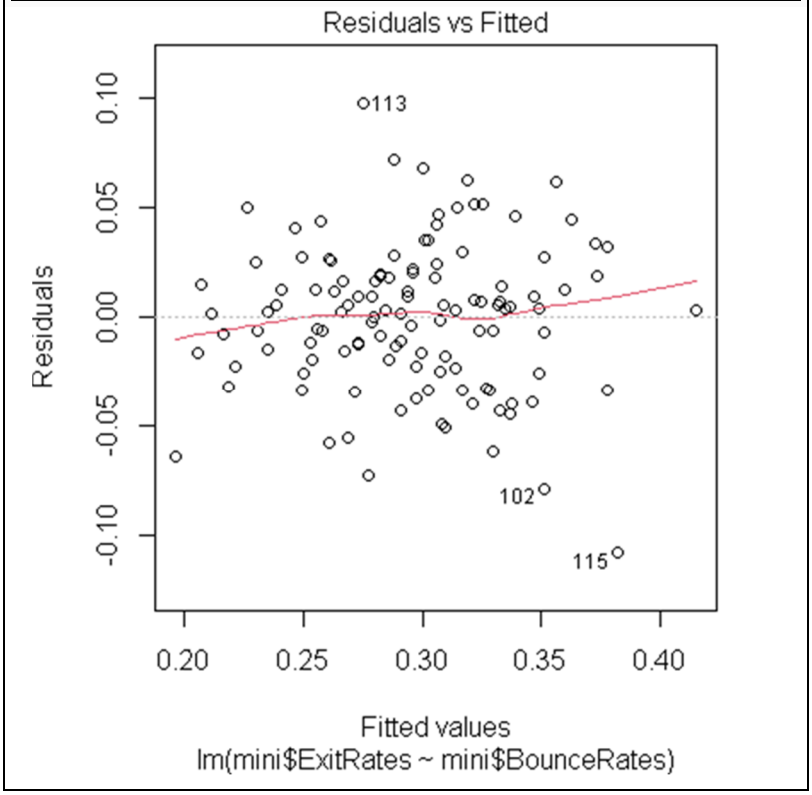


As we can see, the data follows a quite straight pattern, so we can assume the linearity of the dataset. However, to make sure, we will have a look at the Residuals vs Fitted plot.

3.4.2 Drawing Residuals vs Fitted plot

The linearity assumption can also be checked by inspecting the Residuals vs Fitted plot, using command `plot()` with parameter `1`.

```
#Drawing Residuals vs Fitted plot
plot(mini.model, 1)
```



Ideally, the residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. The residual plot above only has a few patterns. This suggests that we can assume a linear relationship between the predictors and the outcome variables.

3.5 Testing for normality

The second assumption of linear regression is normality, which will be tested in this section. First we have to state the null hypothesis and alternative hypothesis. These hypothesis will be used throughout section 3.5:

- H_0: The data follows a normal distribution.
- H_1: The data does not follow a normal distribution.

There are many ways to test the normality of the dataset. In this essay we will use 4 ways:

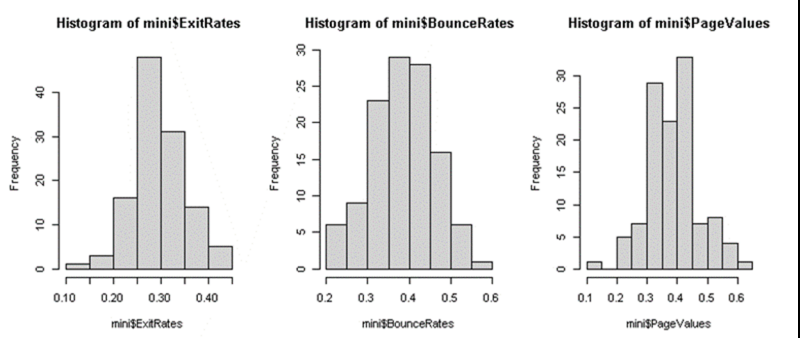
- Drawing histogram

- Drawing density curve
- Drawing Q-Q plot
- Shapiro-Wilk test

3.5.1 Drawing histogram

Histogram is a great way to check the distribution of the data visually, using command `hist()`.

```
#Drawing histogram
hist(mini$ExitRates)
hist(mini$BounceRates)
hist(mini$PageValues)
```

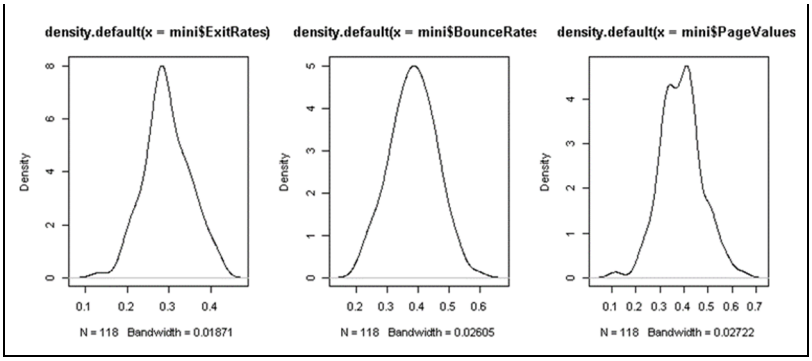


As we can see, the distribution of three variables all follow the same shape, which is the bell-shaped. Therefore, we cannot reject the null hypothesis H_0 and we can conclude that the data follow normal distribution.

3.5.2 Drawing density curve

Density curve is the same as histogram, we can easily have a glance at how the data is distributed by drawing this type of plot. This command `plot(density())` is used to draw a density curve.

```
#Drawing density curve
plot(density(mini$ExitRates))
plot(density(mini$BounceRates))
plot(density(mini$PageValues))
```

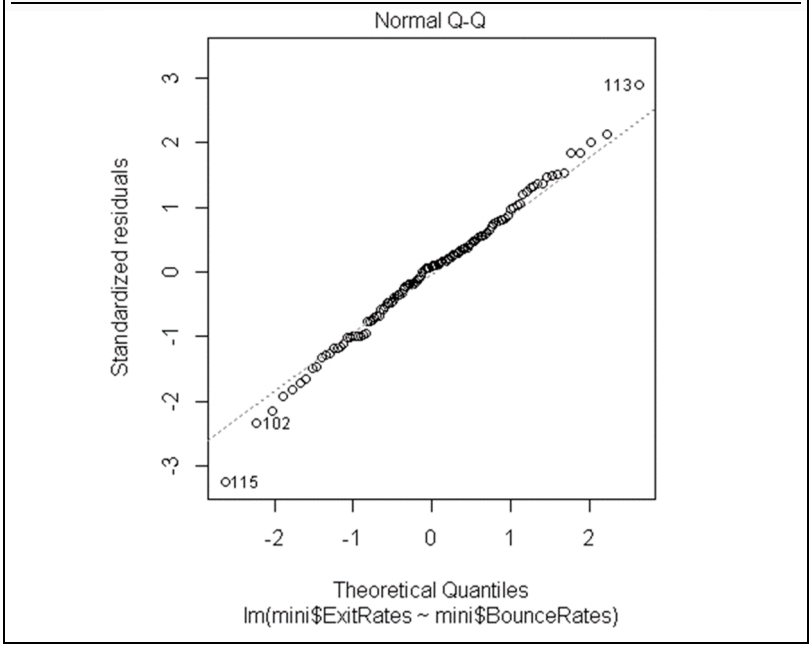


Same as histogram, the bell-shaped curves above indicate that the data follow a normal distribution.

3.5.3 Drawing Q-Q plot

Q-Q plot is another way to test the normality of data, using command `plot()` with parameter `2`.

```
#Drawing Q-Q plot
plot(mini.model, 2)
```



If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line. On the plot we draw above, all the points fall roughly on the reference line. Therefore, we can conclude that the data follow normal distribution.

3.5.4 Shapiro-Wilk test

Shapiro-Wilk test is the final solution that we use to test the normality of data, using command `shapiro.test()`.

```
#Shapiro-Wilk test
shapiro.test(mini$ExitRates)
shapiro.test(mini$BounceRates)
shapiro.test(mini$PageValues)
```

```
> shapiro.test(mini$ExitRates)

        Shapiro-Wilk normality test

data:  mini$ExitRates
W = 0.9898, p-value = 0.5292

> shapiro.test(mini$BounceRates)

        Shapiro-Wilk normality test

data:  mini$BounceRates
W = 0.99573, p-value = 0.9784

> shapiro.test(mini$PageValues)

        Shapiro-Wilk normality test

data:  mini$PageValues
W = 0.9896, p-value = 0.5116
```

For all datasets, the p-values are greater than alpha, which is 0.05, respectively. Thus we cannot reject the null hypotheses and come to the conclusion that all the data follow a normal distribution.

3.6 Testing for homoscedasticity

In this section, we test the homoscedasticity (i.e. homogeneity of variance) of the dataset. First we have to state the null hypothesis and alternative hypothesis. These hypotheses will be used throughout section 3.5:

- H_0: Residuals are distributed with equal variance (i.e. homoscedasticity)
- H_1: Residuals are distributed with unequal variance (i.e. heteroscedasticity)

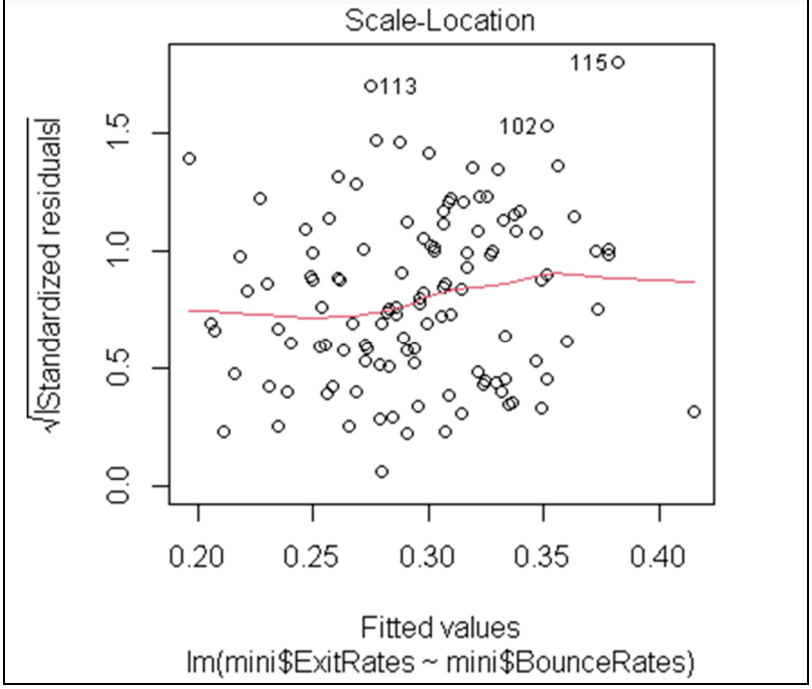
There are many ways to test the homoscedasticity of the dataset. In this essay we will use 2 ways:

- Drawing Scale-location plot
- Using Breusch-Pagan test

3.6.1 Drawing Scale-Location plot

Scale-Location plot, also known as the spread-location plot, can be drawn by using command `plot()` with parameter `3`.

```
#Drawing Scale-Location plot
plot(mini.model, 3)
```



The Scale-Location plot is quite similar to the Residual vs Fitted plot that we used in section 3.4: the flat red line assumes that the data is homoscedasticity. Although the line on the plot above is not flat, the variability among the $\sqrt{Standardized\ residuals}$ seems stable. The variability does neither increase nor decrease with the fitted values. Therefore, we can assume that this regression model does not violate the homoscedasticity assumption.

However, to be completely sure, now we will perform the Breusch-Pagan test.

3.6.2 Breusch-Pagan test

Breusch-Pagan test can be implemented using command `bptest()`.

```
#Breusch-Pagan Test
bptest(mini.model)
```

```
> bptest(mini.model)

        studentized Breusch-Pagan test

data:  mini.model
BP = 3.2942, df = 1, p-value = 0.06952
```

For the mini.model linear regression model, the p-value is greater than 0.05, which leads to failing to reject the null hypothesis, so we can assume the homoscedasticity of data.

3.7 Testing for independence

In this section, we test the independence of the dataset. First we have to state the null hypothesis and alternative hypothesis:

- H_0: Residuals are not autocorrelated (i.e., independence)
- H_1: Residuals are auto correlated (i.e. dependence)

Durbin-Watson test can be used to test for independence of dataset, using command `durbinWatsonTest()`.

```
#Durbin-Watson test
durbinWatsonTest(mini.model)
```

```
> durbinWatsonTest(mini.model)
 lag Autocorrelation D-W Statistic p-value
   1       0.04688997       1.900991   0.544
 Alternative hypothesis: rho != 0
```

As we can see, the p-value we get from the Durbin-Watson test is greater than 0.05 (p-value = 0.544 > 0.05) so we cannot reject the null hypothesis H_0. Therefore, we can conclude that residuals are not autocorrelated, which assumes the independence of data.

3.8 Analyzing the linear regression model

We use the following command to have descriptive statistics about the linear regression model that we built in section 3.3.

```
#Summarize the detail of the linear regression model
summary(mini.model)
```

```
> summary(mini.model)

Call:
lm(formula = mini$ExitRates ~ mini$BounceRates, data = mini)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10801 -0.02209  0.00308  0.01917  0.09790

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.07495    0.01640   4.571 1.22e-05 ***
mini$BounceRates 0.57313    0.04183  13.701  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.034 on 116 degrees of freedom
Multiple R-squared:  0.6181,    Adjusted R-squared:  0.6148
F-statistic: 187.7 on 1 and 116 DF,  p-value: < 2.2e-16
```

The above table shows us some value from the calculation of a best-line for the dataset, these values help us to build a simple linear regression model that minimizes the total error.

From the table, we can observe:

- The range of residual is from -0.10801 to 0.09790 with the median of 0.00308, which is relatively low, proving the accuracy of the model.
- The estimate value of intercepts b_0 and $b_1$are 0.07495 and 0.57313 respectively.
- The value of p-value is very small, showing that there is a significant relationship between two variables in the model.

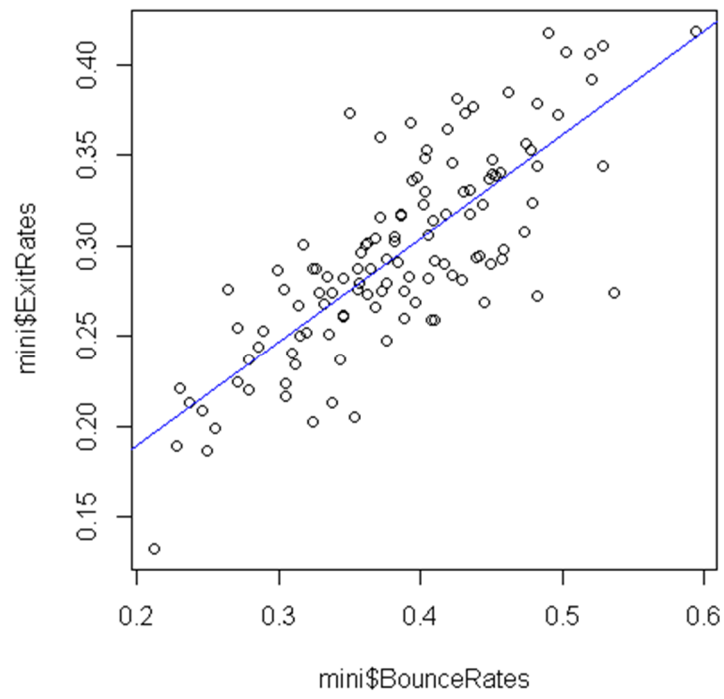
From the result above, we can achieve the estimate value of Exit rate by the formula

$$ExitRates = 0.07495 + 0.57313 \times BounceRates$$

As the formula implies, the relationship between ExitRates and BounceRates is significant positive. A change in 1 unit of BounceRates would result in 0.57313 unit change in ExitRates.

Finally, we use the following command to draw the linear regression model

```r
#Draw the plot of the linear regression model
plot(mini$BounceRates, mini$ExitRates)
abline(mini.model, col = "blue")
```



To draw the plot above, first we draw the Scatter plot of ExitRates and BounceRates (same as the plot we drawn in section 3.5.1). Then we use command `abline()` to draw the blue line, which denotes the linear function that models the relationship between dependent variable (ExitRates) and independent variable (BounceRates). From the plot, we can see that ExitRates depends significantly on BounceRates.

4 Conclusion

In this essay, we conducted some tasks such as collecting and processing data; performing t-test; testing four assumptions of linear regression using some different ways from drawing plots to taking hypothesis tests. Finally, we built and analyzed the linear regression model between two variables from our dataset.

Through the essay, we come to the conclusion that Bounce Rate has an significant impact on Exit Rate. High Bounce Rates indicates that the quality of that website is very bad which can severely affect the website's revenue. There are many ways to improve the quality of the website like designing better user experience, engaging the visitors with full of fresh content, planning better content marketing,... A good website can attract many customers and get many opportunities for future growth.

References

[1] Aryansh Gupta (2020), *Linear Regression Assumptions and Diagnostics in R*, https://rpubs.com/aryn999/LinearRegressionAssumptionsAndDiagnosticsInR

[2] Coding Prof (2022), *3 Easy Ways to Test for Heteroscedasticity in R [Examples]*, https://www.codingprof.com/3-easy-ways-to-test-for-heteroscedasticity-in-r-examples/

[3] Rebecca Bevans (2020), *Simple Linear Regression | An Easy Introduction & Examples*, https://www.scribbr.com/statistics/simple-linear-regression/

[4] Rebecca Bevans (2020), *Linear Regression in R | An Easy Step-by-Step Guide*, https://www.scribbr.com/statistics/linear-regression-in-r/

[5] Will Kenton (2021), *Durbin Watson Statistic Definition*, https://www.investopedia.com/terms/d/durbin-watson-statistic.asp

[6] *SPSS Tutorials: Paired samples t test*, https://libguides.library.kent.edu/spss/pairedsamplesttest