1. ANOVA – Bank Marketing Campaign
2. LINEAR REGRESSION - Household expenditure
3. PRINCIPAL COMPONENT ANALYSIS – IPL Player Ranking

Mini Project – 3 /Advanced Statistics

Shyam Krishnan k

BACP DEC-2017 Batch-2

Date: 01-02-2018

**ANOVA – Bank Marketing Campaign**

Table of Contents

1. [Overview 2](#_Toc505246858)

I.a) Introduction

I.b) Problem Summary

1. [Methodology 3](#_Toc505246859)
2. [Data Analysis and Reporting 4](#_Toc505246860)

III.a) Hypothesis rationale formulation

III.b) P – Value Calculation

III.c) Descriptive Statistical Summaries

III.d) Confidence Interval

1. [Conclusion 9](#_Toc505246861)
2. [Research Appendix 10](#_Toc505246862)

# Overview

a) Introduction

A bank decided to offer a new product to the customers. Part of the product design and market introduction bank decided to conduct a marketing campaign and collected some relevant and non-biased data from the customer for the purpose of analysis and decision making. So they targeted 20000 (twenty thousand) customers for the campaign and collected corresponding data from the customers. Bank want to conduct - analysis of variance on the collected data to understand about dependence among some features in the dataset.

b) Project objective

Part of new product introduction a bank conducted a market campaign among its customers. Bank targeted 20000 (twenty thousand) customers and collect data from them. The collected data is contained in the data set PL\_X\_SELL in CSV format. We need to conduct the analysis of variance on the variables occupation of the customers and gender of the customer against the average quarterly balance of the customer.

1. Conduct a one-way ANOVA analysis to study whether the occupation of the account holder affects quarterly average balance in the account
2. Conduct two-way ANOVA analysis on gender and occupation on quarterly average balance.

# Assumptions

We assume that the nature of the data provided in the dataset PL\_X\_SELL is non-biased and error free. Which means – data is collected from an experimental method or quasi-experimental like conditions, implies that data is collected directly from sources and not in a laboratory setting. It is assumed that the data is related to the customers, who had participated in the campaign during the campaign period. We also assume that the dataset will be reflective of the reality and the features in the dataset are linked to the characteristics of the customers.

ANOVA is an extension of independent two-sample t-test for comparing means in a situation where there are more than two groups. In one-way ANOVA, the data is organized into several groups base on one single grouping variable (also called factor variable). But the results of an ANOVA can be considered reliable as long as the following assumptions are met:

1. The observations are obtained independently and randomly from the population defined by the factor levels
2. The data of each factor level are normally distributed.
3. These normal populations have a common variance.

Also, the following data dictionary is considered for the 10 features in the dataset:

# Methodology

Solutions to the above-stated project objective can be accomplished through the process of statistical analysis on the given data set – PL\_X\_SELL. The process of data analysis can start with exploratory data analysis, which will give us the outline of the data set. It includes setting up the working directory, fetching information on the number of rows and column, listing features of the data set. In the next step, we will perform the factor variable identification. It is the process of setting the independent variables and the dependent variable – based on project objective, which will later use for hypothesis formation and ANOVA analysis. Then we will perform the descriptive analysis. Descriptive analysis will help us to understand, what these sample data say. Using descriptive analysis we can get overview or summary statistic of the data, which includes- measurement of center tendency, averages, mean, standard deviation, histogram, boxplots etc.

To test the dependencies we will perform ANOVA analysis. Using ANOVA analysis we will find out, whether the features ‘occupation' and ‘gender' have any specific and reliable effect on the dependent variable ‘average quarterly balance'. We are performing one-way ANOVA for understanding the effect of occupation on the average balance of the of an account holder, and in the same manner, we will measure the effect of gender and occupation on the average balance using two- way ANOVA.

Steps to conduct ANOVA

1. Identify the independent variable and dependent variable. Factorize the independent variable into corresponding groups.
2. Formulate the null hypothesis and alternative hypothesis. Commonly null hypothesis Ho: the mean of the different groups are same and alternative hypothesis Ha: at least one sample or group mean is not equal to others.
3. Test of assumptions. The normality can be tested using Shapiro test. Same way the homogeneity in the variance across the categories in the factor variable is tested using Levene's test. Shapiro test rejects the null hypothesis Ho: each factor are normally distributed when P-value is <=0.05. Failing the normality test allows us to state with 95% confidence that – the data do not have a normal distribution. Levene's test is using to check if the variance is equal for all the sample when our data come from a non-normal distribution. If the p-value is <= 0.05, then we reject the Ho: σ12= σ22= σ32. We can also use Bartlett's test for checking the homogeneity of variance.
4. If assumptions hold, then we can move to ANOVA test. The ANOVA test using R language will give us between group of variance and within group of variance (I.e.: SSB & SSE) as an output in tabular format. In two way ANOVA apart from direct effects we will also get interaction effect. If assumptions are violated we need to conduct Robust ANOVA tests like Huber – M – estimates or we need to scale each and every observation by its respective group variance – which can also possible to perform using a specific test in R language.
5. The 5th step is POST – HOC – TEST, which is also called as Tukey test. This test will show where the difference between each factor is. The Post – Hoc – Test will conduct when the null hypothesis is rejected. Otherwise, we don't need to conduct a Post – Hoc – Test.

# Data Analysis and Reporting

The entire process of data analysis can be divided into following steps. We can follow step by step approach to arrive at the conclusion.

1. Exploratory Data Analysis
2. Factor Variable Identification
3. Descriptive Statistics
4. ANOVA analysis

a) Exploratory Data Analysis

Dataset is a data frame with 20000 observations of 10 variables, in other words - the dataset PL\_X\_SELL is consist of 20000 (twenty thousand) rows and 10 (ten) columns. The dataset contains 10 features as follows in the exact order.

|  |  |  |
| --- | --- | --- |
| Feature Code | Type | Continuous/ Categorical |
| Cust\_ID | Factor | Categorical |
| Target | int | Continuous |
| Age | int | Continuous |
| Gender | Factor | Categorical |
| Balance | num | Continuous |
| Occupation | Factor | Categorical |
| No\_OF\_CR\_TXNS | int | Continuous |
| AGE\_BKT | Factor | Categorical |
| SCR | int | Continuous |
| Holding\_Period | int | Continuous |

Using R code we can check for missing values in the data set - output indicating that there are no missing values or placeholders in the data. Customer ID (Cust\_ID) is the unique data\_id given to the customers. It is a factor variable which is not important in the following analysis, and we can discard it during the coming steps. Target is an integer - binary variable consists of two values 0 & 1, which helps to differentiate the customer who responds to the campaign and who does not respond to the campaign. Zero representing non-respondent segment and 0ne represents responder segment. Age is a continuous variable. The column gender is a factor or categorical variable with 3 levels: F/ M/ O. O is an indication of firms participated in the campaign. The balance is a numerical variable which is continuous in nature. Occupation is a factor variable with 4 levels which are – PROF (professional), SAL (salaried), SELF-ENP (self-employed), SENP (self-employed nonprofessional). Age is again using as a categorical variable after transforming it into levels of slabs (7 levels). SCR is continuous variable representing the marketing scores of each individual, which is assigned and collected from adscititious resources.

b) Factor Variable Identification

The project objective is to conduct a one-way ANOVA analysis to study whether the occupation of the account holder affects quarterly average balance in the account and we need to conduct two-way ANOVA analysis on gender and occupation on quarterly average balance. Based on the objective it is clear that In one way ANOVA: Occupation is the treatment variable or independent variable, which is categorical in nature and the variable balance is the dependent variable. Here we will use ANOVA to determine whether the occupation has any effect on the dependent variable balance. In two way ANOVA: occupation and gender are the independent variables which are categorical in nature and the variable balance is the dependent variable. Here we will use ANOVA to determine whether the occupation has any effect on the dependent variable balance. In two way ANOVA –apart from the direct effect of the independent variable on the dependent variable, we need to take in consideration of the interaction effect.

The factor levels of each independent variable are as follows:

|  |  |
| --- | --- |
| occupation | PROF (professional) |
| SAL (salaried) |
| SELF-ENP (self-employed) |
| SENP |

|  |  |
| --- | --- |
| gender | F (female) |
| M (male) |
| O (others- firms) |

In the descriptive analysis, we will concentrate on the independent variable and dependent variable, which will help us to fetch out the valuable insights on dataset and ANOVA results.

c) Descriptive Statistics

Descriptive Statistics provides simple summaries about the sample and the measures. Together with simple graphics analysis. Using descriptive analysis we can analyze the measures of Central Tendency and measure of dispersion of continuous variables. Here in the descriptive analysis, we can discard the unique customer id – which had no use in the following analysis. The feature target which is numeric in nature but contains binary values can't be evaluated using the measure of dispersion or central tendency, need to be treated along with the categorical variables.

Descriptive Statistics – continuous Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| The measure of Central Tendency | Age | Balance | No\_OF\_CR\_TXNS | SCR | Holding\_Period |
| Mean | 38.4 | 146181 | 16.65 | 557.1 | 15.34 |
| Median | 38 | 79756 | 13.00 | 560.0 | 16.00 |
| Minimum | 21 | 0 | 0.00 | 100.0 | 1.00 |
| Maximum | 55 | 1246967 | 50.00 | 999.0 | 31.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Measure of Dispersion | Age | Balance | No\_OF\_CR\_TXNS | SCR | Holding\_Period |
| Range | 34 | 1246967 | 50 | 899 | 30 |
| 1st Quartile | 30.0 | 23737 | 7.00 | 333.0 | 8.00 |
| 3rd Quartile | 47.0 | 217311 | 22.00 | 784.0 | 23.00 |
| IQR | 17 | 193574 | 15 | 451 | 15 |
| Variance | 92.16343 | 28836294478 | 168.4208 | 67871.43 | 80.1406 |
| Standard Deviation | 9.600179 | 169812.5 | 12.9777 | 260.5215 | 8.952128 |

Descriptive Statistics – Categorical Variables

|  |  |
| --- | --- |
| Target | Count |
| 0 | 18267 |
| 1 | 1733 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age slab | <25 | 26-30 | 31-35 | 36-40 | 41-45 | 46-50 | >50 |
| count | 1784 | 3404 | 3488 | 2756 | 3016 | 2532 | 3020 |

Summary statistics by groups - count, mean, sd

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | count | mean | sd |
| F | 5525 | 197405 | 191814 |
| M | 14279 | 126888 | 155568 |
| O | 196 | 107775 | 200075 |

|  |  |  |  |
| --- | --- | --- | --- |
| Occupation | count | mean | sd |
| PROF | 5463 | 146952 | 173427 |
| SAL | 5839 | 123800 | 160650 |
| SELF-EMP | 3366 | 128359 | 157177 |
| SENP | 5332 | 181152 | 177573 |

From the descriptive data analysis we have the following brief observations:

The dependent variable follows a non-normal distribution, specifically speaking it is a right-skewed distribution where more points are on the right of the median. The graphical representation directly showing the existence of a large number of outliers. We can observe that the out liars are distributed from just above the 3rd quartile region to a very large extent. The two variables i.e. independent variables occupation and gender they are unbalanced. Each level or groups in the variables are of different size. The factor level SELF\_EMP in occupation is about 3366 which is far less than other groups. The category other in the gender variable is not even close to the other categories. From the table, we can observe that only 1733 customers out of 20000 belong to the responder category.

Data Visualization

d) ANOVA analysis

ANOVA analysis is using to measure the effect of a factor or treatment on the dependent variable of interest. Here we need to conduct One-way ANOVA analysis of independent variable occupation and dependent variable balance. Same way Two-way ANOVA analysis is framed of independent (categorical) variables gender and occupation over continuous variable balance. Here we need to check whether the independent variables have any effect on the dependent variable balance.

One-way ANOVA

The one-way analysis of variance is using to determine whether there is any statistical difference between mean of three or more independent groups. Our objective is to conduct a one-way ANOVA analysis to study whether the occupation of the account holder affects quarterly average balance in the account. Here occupation is the independent variable which is continuous in nature. Balance is the dependent variable. To perform the ANOVA first we need to formulate a null and alternative hypothesis.

Ho: µ PROF = µ SAL = µ SELF-EMP = µ SENP

Ha: At least one of the mean is different from the rest.

Here we will decompose total variation into between-group variation and within group variation. The ratio of between-group variations to total variation is a measure of the strength of treatment effect (occupation) on the dependent variable avg balance. The significance of null hypothesis can be tested using F- statistic

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ANOVA/lessonimages/equation_image35.gif

= MSB / MSE

= [SSB/ (K-1)]/ [SSW/ (N-K)]

Where: K is the degree of freedom of occupations (no of job types) =4

N is the total degree of freedom=20000

If F Calculated > F critical, then we will reject the null hypothesis.

Here SSB is the between group of sum of square and SSW is the within group of sum of square. Using ANOVA test in R we can calculate each individual values including the P-value which can be used as a scale of measurement of significance

Test of Normality

Here normality assumptions can't be tested using Shapiro test because the number of dataset observations is > 5000. So we can check the normality using ks.test (Kolmogorov Simonov test) or ad.test (Anderson-Darling test) by computing the p-value. The test rejects the hypothesis of normality when p value is less than or equal to 0.05. Failing the normality test allows us to state with 95% confidence the data does not fit the normal distribution.

Ho: Each category in the factor variable occupation is normally distributed.

I.e.: µ PROF= µ SAL = µ SELF-EMP = µ SENP

Ha: Distribution is not normal in all categories

Computing P-value using R functions indicating that: P- value is < 2.2e-16 for each group in the variable occupation, which is far less than 0.05. So we reject the null hypothesis. So normality assumption violated.

Note: refer appendix for code and output

Test of Homogeneity of Variance

Homogeneity of variance across different categories in the factor variable occupation against variable Balance can be tested using Levene’s test or Bartlett test. Levene's test uses when data comes from a non-normal distribution.

Ho: σ2 PROF= σ2 SAL = σ2 SELF-EMP = σ2 SENP  (variance across the occupation are same)

Ha: Variance across the variable occupation are not equal

Let us calculate the p-value and F-statistic using ‘leveneTest’ and ‘bartlett test’.

leveneTest(Balance~Occupation)

bartlett.test(Balance~Occupation)

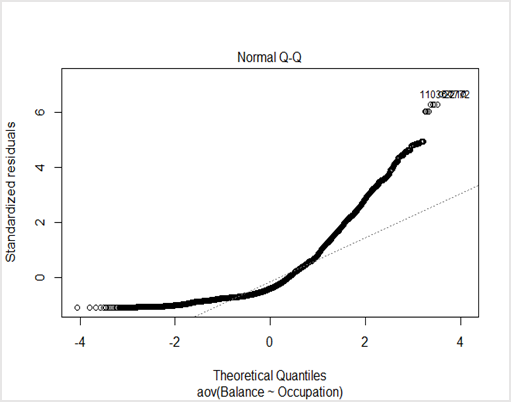
From the analysis, it is very clear that P-values is <0.05 and F-calculated > F-critical. We reject the null hypothesis and assumption of homogeneity of variance is violated. So we can conclude that variance is not homogeneous across the variable occupation.

Test of assumptions directly indicating that both assumptions are violated so in the practical sense we need to use the robust ANOVA methods. When the number of observations in the group is unbalanced and the variance is different – then we can't reach a conclusion on the hypothesis from ANOVA test output. Even though here we are running ANOVA analysis on the purpose of reference and comparison.

ANOVA test

Function to run ANOVA analysis is aov(dependent variable ~ independent variable). The function summary() is used to summarize the analysis of variance model.

In the output between group of sum of square is ‘sum sq' of occupation which is equal to 1.052e+13. Within group of sum of square is given by the value of ‘sum sq' of residuals. Here F calculated > F critical and P-value< 0.05 so we need to reject the null hypothesis (Ho: µ PROF = µ SAL = µ SELF-EMP = µ SENP). Here we concluded that treatment effect is significant, in other words, occupational categories have different variance or effect on the avg quarterly balance.



In the plot, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted. The normal probability plot of residuals is used to check the assumption that the residuals are normally distributed. It should approximately follow a straight line. But here most of the points not following this reference line, we can ensure that normality assumption is violated.

Robust ANOVA test

In the robust method, we will adjust for the difference the variance for the variable occupation. So we will scale each and every observations by its respective group variance i.e. Xi / σi or we can use the Huber-M-estimates. These methods are using to

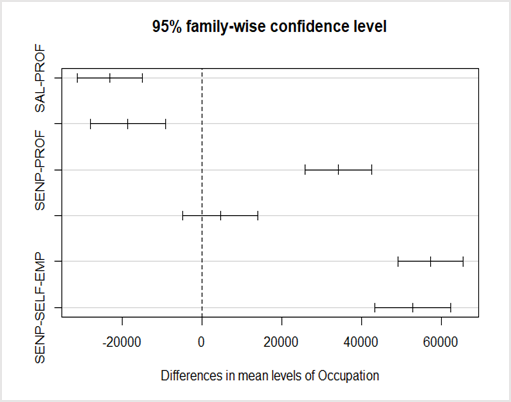
Let us calculate the p-value Huber-M-estimate method

Else we can also use the following code, which will give us the adjusted value result. In the code white.adust=TRUE will perform Xi / σi scaling in the background. Still P-value< 0.05, so we will reject the null hypothesis and go forward with the previous conclusion. Occupational categories have different variance or effect on the avg quarterly balance.

Post – Hoc Test (Tukey)

Tukey test is using to find out, where is the difference between each occupation. The overall significance level of the test is at 0.05.

Here from the output, we can see that each of the occupations are different from other except for SELF-EMP and SAL. For each set of groups, P-adj values are equal to zero except for above-mentioned groups - SELF-EMP and SAL with an adjusted p-value of 0.5936835 which is > 0.05. In the output columns, lwr and upr contain no zero value which means the confidence intervals are either to the left or to the right.



Two-way ANOVA

The primary purpose of two way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable. The two way ANOVA compare the mean differences groups that have been split into two independent variables. Here we have two independent variables ‘gender' and occupation. So we need to study about the effect of gender and occupation on balance.

Here we have 4 categories in occupation and 3 groups in gender. In two way ANOVA we need to consider the

1. Direct effect from occupation
2. Direct effect from gender
3. The interaction effect between occupation and gender.

Direct effects are same as in one way ANOVA. Interaction effect defines how 2 independent factors, altogether affecting the dependent variable.

Interaction plot gives an overview of the interaction between the factors. In the graph, we can rapid or non-normal changes, which indicated an interaction effect.

Normality Test

Here normality assumption can be tested using ks.test (Kolmogorov Simonov test) for both independent variables. Which gives output P-value<2.2e-16 which is less than 0.05. So we reject the null hypothesis which means normality assumption is violated.

Test of Homogeneity of Variance

Homogeneity of variance across different categories in the factor variable occupation and gender against variable Balance can be tested using Levene’s test or Bartlett test.

Let us calculate the p-value and F-statistic using ‘levene’s Test’.

The output indicating that P-value <α. So variance assumption is violated.

Both test of assumptions indicating that

1. Normality assumption is violated
2. Homogeneous variance across the group is also violated.

Test of assumptions directly indicating that both assumptions are violated so in the practical sense we need to use the robust ANOVA methods. Even though here we are running ANOVA analysis on the purpose of reference and comparison.

ANOVA test

Function to run ANOVA analysis is: aov(dependent variable ~ factor1+ factor2+ factor1: factor2,data=mydata). The function summary () is used to summarize the analysis of variance model.

The output contains F and P values of gender occupation and occupation and gender together. The output indicating that occupation wise differences i.e. balance across occupation differ based on the type of occupation. (P-value < 0.05). So the effect is significant. Also, avg balance is differ based on the gender of the customer.

Occupation: Gender is indicating the interaction effect. P-value=0.00225 which is <0.05 implies that interaction effect is also significant.

Robust ANOVA test

Here both assumptions: normality assumption and homogeneous variance across the variables are also violated. Then we can we can use the following code, which will give us the adjusted P-value result. In the code white.adust=TRUE will perform Xi / σi scaling in the background. Here P-value of each variable including the interaction effect is> 0.05, so we will reject the null hypothesis.

The output indicating that each independent variable occupation and gender and there corresponding factors along with the interaction effect does have the significant effect across the dependent variable balance.

Post – Hoc Test (Tukey)

Tukey test is using to find out, where is the difference between each occupation. The overall significance level of the test is at 0.05.

We can use the following code: TukeyHSD(aov2) for running the test

The output will give a detailed view of each variable – where the difference present is or not. From the output we can observe that there is no difference between SELF-EMP and SAL in the variable occupation also in the variable gender there is no difference in effects of O-M.

Note: please refer appendix for output

# Conclusion

Descriptive analysis on dataset indicating that the dependent variable follows a non-normal distribution, in right words – it follows a right-skewed distribution where more points are on the right of the median. Normal Q-Q plot from ANOVA analysis is a solid proof of above observation. The box plot clearly picturing the existence of a large number of outliers in the variable balance, which playing as a major cause of violations of normality and homogeneity of variance.

One way ANOVA analysis showing that – Normality assumption is violated and the variance of the variable balance is not homogeneous across the variable balance. The output ANOVA test and Robust methods indicating that treatment effect is significant or occupational categories have different variance or effect over the average quarterly balance. From the post-hoc test, we can observe that – each of the occupation categories have a different effect on the variable balance except for the group SELF-EMP and SAL (i.e. no difference in the treatment effect) with a P-value=0.5936 >0.05.

From the two way ANOVA analysis we can observe that normality assumption is violated with a P-value<2.2e-16. Leneve’s test indicating that homogeneous variance across the group is also violated. The ANOVA and Robust ANOVA methods indicating the direct effect of occupation, direct effect of gender and the interaction effect of occupation and gender are significant over the variable balance.

# Research Appendix

################# Source Code – R programming ######################

setwd("F:/BACP/w4-Mini project") #### setting the working directory getwd()

getwd()

mydata<-read.csv("PL\_X\_SELL.csv",header = TRUE)

########## Descriptive analysis ################

dim(mydata)

names(mydata)

str(mydata)

head(mydata)

tail(mydata)

colSums(is.na(mydata))

summary(mydata)

attach(mydata)

range(Age)

sd(Age)

var(Age)

table(Gender)

table(Gender,Occupation)

################ plots #################

library(lattice)

plot(density(Balance),col='green')

hist(Balance,main = 'Avg Balance',xlab = 'Balance',ylab = 'Frequency',col = 'orange')

plot(Occupation,Balance,main = 'Occupation vs Balance',xlab = 'Occupation',ylab = 'balance',col = 'green')

plot(Gender,Balance,main = 'Gender vs Balance',xlab = 'gender',ylab = 'balance',col = 'pink')

par(mfrow=c(1,2))

boxplot(Balance,main = 'Avg Balance',xlab = 'Balance',ylab = 'Frequency',col = 'yellow')

plot(Occupation,main = 'Occupation',xlab = 'occupation',ylab = 'Frequency',col = 'green')

plot(Gender,main = 'gender',xlab = 'gender',ylab = 'Frequency',col = 'pink')

plot(Occupation,Gender,main = 'Occupation Vs Gender',xlab = 'Occupation',ylab = 'Gender')

dev.off()

install.packages('dplyr')

library(dplyr)

levels(Occupation)

group\_by(mydata, Occupation) %>%

summarise(

count = n(),

mean = mean(Balance, na.rm = TRUE),

sd = sd(Balance, na.rm = TRUE)

)

group\_by(mydata, Gender) %>%

summarise(

count = n(),

mean = mean(Balance, na.rm = TRUE),

sd = sd(Balance, na.rm = TRUE)

)

############# ANOVA & Test of Assumptions ##############

library(psych)

install.packages("car")

library(car)

library(foreign)

install.packages("MASS")

library(MASS)

for (i in unique(factor(mydata$Occupation))){ cat(shapiro.test(mydata[mydata$Occupation==i, ]$Balance)$p.value," ") } #### not working

ks.test(mydata$Occupation,mydata$Balance)

levels(Occupation)

newdata1<- subset(mydata, Occupation == 'PROF', select=c(Balance))

newdata2<- subset(mydata, Occupation == 'SAL', select=c(Balance))

View(newdata1)

View(newdata2)

######################################

library(nortest)

ad.test(newdata1$Balance)$p.value

ad.test(newdata2$Balance)$p.value

ad.test(newdata3$Balance)$p.value

ad.test(newdata4$Balance)$p.value

m <- mean(newdata1$Balance)

s <- sd(newdata1$Balance)

ks.test(newdata1$Balance, "pnorm", m, s)

############## Test of variance ##############

leveneTest(Balance~Occupation)

bartlett.test(Balance~Occupation)

############# ANOVA ##############

aov1<-aov(Balance~Occupation)

summary(aov1)

######## Robust method ###############

oneway.test(Balance~Occupation,var.equal = FALSE)

model1<-lm(Balance~Occupation)

Anova(model1,type = "II",white.adjust = TRUE)

######## Post hoc ######################

TukeyHSD(aov1)

plot(TukeyHSD(aov1))

plot(aov1,1)

plot(aov1,2)

######################## 2 way ANOVA ######################

library(robustHD)

library(rcompanion)

library(WRS2)

library(psych)

library(car)

library(foreign)

library(MASS)

Occupation<-factor(Occupation,labels = c("PROF","SAL","SELF-EMP","SENP"))

Gender<-factor(Gender,labels = c("F","M","O"))

tapply(Balance,list(Occupation,Gender),mean)

tapply(Balance,list(Occupation,Gender),sd)

interaction.plot(Gender,Occupation,Balance)

ks.test(mydata$Occupation,mydata$Balance)

ks.test(mydata$Gender,mydata$Balance)

leveneTest(Balance~Occupation\*Gender)

aov2<- aov(Balance~Occupation+Gender+Occupation:Gender, data=mydata)

summary(aov2)

model2<-lm(Balance~Occupation + Gender + Occupation:Gender)

Anova(model2,type = "II",white.adjust = TRUE)

TukeyHSD(aov2)

############# Other References ##################

1. <http://www.kean.edu/~fosborne/bstat/06b2means.html>
2. <https://www.r-bloggers.com/two-sample-students-t-test-1/>
3. <https://onlinecourses.science.psu.edu/stat100/node/58>

##################################### END ########################################

**ANOVA – Bank Marketing Campaign**

Table of Contents

1. [Overview 2](#_Toc505246858)

I.a) Introduction

I.b) Problem Summary

1. [Methodology 3](#_Toc505246859)
2. [Data Analysis and Reporting 4](#_Toc505246860)

III.a) Hypothesis rationale formulation

III.b) P – Value Calculation

III.c) Descriptive Statistical Summaries

III.d) Confidence Interval

1. [Conclusion 9](#_Toc505246861)
2. [Research Appendix 10](#_Toc505246862)