Example of Analysis of Data from an Experiment

The necessary libraries

```
library(tidyverse)
library(knitr)
library(kableExtra)
library(cowplot)
library(stargazer)
```

1. Input the data from the file `fakeMayo.rds` file.

```
fakeMayo <- readRDS("fakeMayo.rds")
dim(fakeMayo)
```

```
## [1] 217  58
```

2. Input the data from the file `fakeMayoMDScores.rds` file.

```
fakeMayoMDScores <- readRDS("fakeMayoMDScores.rds")
dim(fakeMayoMDScores)
```

```
## [1] 217  11
```

3. Do EDA on the data to look for anomalies, suspicious values, etc. No need to do extensive examinations of whether different variables covary, a multitude or graphs, or other such things. Just describe the data and look for problems.

We start with some numerical summaries:

```
summary(fakeMayoMDScores[,-11]) %>% kable() %>%
kable_styling(latex_options = c("striped", "scale_down"))
```

	stmtMD1	stmtMD2	stmtMD3	stmtMD4	stmtMD5	stmtMD6	stmtMD7	stmtMD8	stmtMD9	stmtMD10
	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00	Min. :-3.00
	1st Qu.:-1.00	1st Qu.:-2.00	1st Qu.:-1.00	1st Qu.:-1.00	1st Qu.:-2.00	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.:-2.00	1st Qu.:-2.00	1st Qu.: 0.00
	Median : 1.00	Median :-1.00	Median :-1.00	Median : 0.00	Median :-1.00	Median : 2.00	Median : 1.00	Median :-1.00	Median :-1.00	Median : 1.00
	Mean : 0.57	Mean :-0.79	Mean :-0.25	Mean :-0.04	Mean :-1.05	Mean : 1.17	Mean : 1.11	Mean :-0.77	Mean :-0.89	Mean : 1.16
	3rd Qu.: 2.00	3rd Qu.: 0.00	3rd Qu.: 1.00	3rd Qu.: 1.00	3rd Qu.: 0.00	3rd Qu.: 3.00	3rd Qu.: 2.00	3rd Qu.: 0.00	3rd Qu.: 0.00	3rd Qu.: 2.00
	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00	Max. : 3.00

The above table draws the usual univariate summary information. At this stage, we are looking for anything unusual or unexpected, perhaps indicating a data-entry error. For this purpose, let's take a look at the minimum and maximum values of each variable. We see that all variables have minimum values of -3 and maximum values of 3. The dataset does not contain a missing values.

Figure below shows a correlation matrix of the data set. Each pairwise correlation is computed from the `fakeMayoMDScores` and colored according to its magnitude. This visualization is symmetric: the top and bottom diagonals show identical information. Dark blue colors indicate strong positive correlations, dark red is used for strong negative correlations, and white implies no empirical relationship between the variables.

```
# Correlation Matrices from the corrplot Package
library(corrplot)
M <- cor(fakeMayoMDScores[,-11])
ord <- corrMatOrder(M, order = "AOE")
M2 <- M[ord,ord]
```

```
corrplot.mixed(M2, tl.pos = "lt", diag = "l",upper ="shade")
```

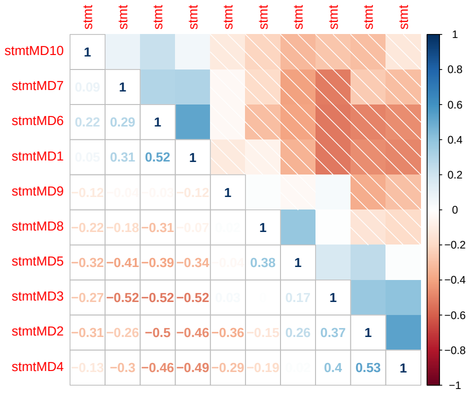

Figure 1: Details correlation coef. between MaxDiff scores

Based on the above figure, the pairwise correlation of all the variables is less than 0.5.

4. Describe how the MaxDiff scores in the `fakeMayoMDScores.rds` file are distributed across the consumers. These are named stmtMD1 trough stmtMD10.

First, we need to transform the data from wide format to long format. In data set `fakeMayoMDScores.long`, the key column contains only keys. Conveniently, the value column contains the values associated with those keys.

```
fakeMayoMDScores.long <-
  fakeMayoMDScores[,-11]%>%
  gather(key, value) #wide format
```

The Table 1 shows Frequency Table for stmtMD1 through stmtMD10 of fakeMayoMDScores Data:

```
addmargins(table(fakeMayoMDScores.long)) %>%
  kable(caption = "Frequency Table for the fakeMayoMDScores data") %>%
kable_styling(latex_options = c("striped", "scale_down"))
```

The Figure below draw a bar graph of the data:

```
fakeMayoMDScores[,-11] %>%
  gather(key, value) %>%
  ggplot()+
  geom_bar(aes(value), stat = "count", fill="blue")+
  facet_wrap(key~.)
```

Table 1: Frequency Table for the fakeMayoMDScores data

	-3	-2	-1	0	1	2	3	Sum
stmtMD1	11	22	35	27	35	60	27	217
stmtMD10	8	9	8	34	51	71	36	217
stmtMD2	41	45	46	33	23	22	7	217
stmtMD3	9	37	70	40	21	25	15	217
stmtMD4	21	22	42	48	42	27	15	217
stmtMD5	24	70	59	34	18	11	1	217
stmtMD6	13	14	20	24	24	49	73	217
stmtMD7	1	6	23	42	41	74	30	217
stmtMD8	25	57	43	50	24	13	5	217
stmtMD9	49	44	44	28	27	16	9	217
Sum	202	326	390	360	306	368	218	2170

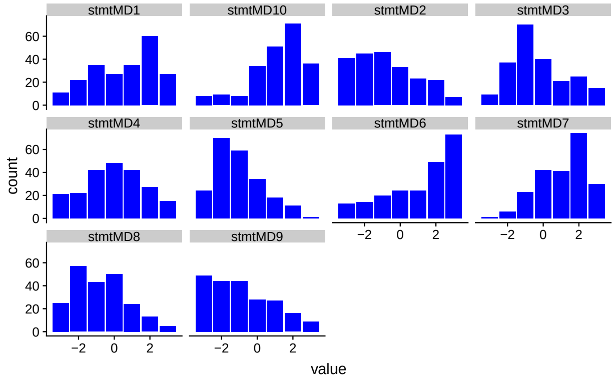

Figure 2: Bar Graph for fake Mayo MD Scores (stmtMD1 through stmtMD10)

Bar graphs or charts consist of the frequencies on one axis and the categories on the other axis for each individual maxdiff score, which are useful for comparing sizes of categories.

Reviewing the graph (Figure 2) we can see that most of the customers score around 2 or 3 the maxdiff STMTMD6 (Real mayo taste with 10 calories and zero fat).

Notice from this graph, we can see that STMTMD6 (Real mayo taste with 10 calories and zero fat), STMTMD7 (Tastes as good as regular mayo but with half the fat and calories), STMTMD10 (Tastes more like regular mayo), and , STMTMD1 (Half the fat and calories of mayonnaise) are highly scored compared to the remaining statements. However, The statements STMTMD3 (New great taste) and STMTMD5 (Preferred over the leading brand of Light mayonnaise) seem to be the type of statement that we can tell was the least scored.

Table 2: correlation matrix

	stmtMD10	stmtMD7	stmtMD6	stmtMD1	stmtMD9	stmtMD8	stmtMD5	stmtMD3	stmtMD2	stmtMD4
stmtMD10	1.00	0.09	0.22	0.05	-0.12	-0.22	-0.32	-0.27	-0.31	-0.13
stmtMD7	0.09	1.00	0.29	0.31	-0.04	-0.18	-0.41	-0.52	-0.26	-0.30
stmtMD6	0.22	0.29	1.00	0.52	-0.03	-0.31	-0.39	-0.52	-0.50	-0.46
stmtMD1	0.05	0.31	0.52	1.00	-0.12	-0.07	-0.34	-0.52	-0.46	-0.49
stmtMD9	-0.12	-0.04	-0.03	-0.12	1.00	0.02	-0.04	0.03	-0.36	-0.29
stmtMD8	-0.22	-0.18	-0.31	-0.07	0.02	1.00	0.38	0.00	-0.15	-0.19
stmtMD5	-0.32	-0.41	-0.39	-0.34	-0.04	0.38	1.00	0.17	0.26	0.02
stmtMD3	-0.27	-0.52	-0.52	-0.52	0.03	0.00	0.17	1.00	0.37	0.40
stmtMD2	-0.31	-0.26	-0.50	-0.46	-0.36	-0.15	0.26	0.37	1.00	0.53
stmtMD4	-0.13	-0.30	-0.46	-0.49	-0.29	-0.19	0.02	0.40	0.53	1.00

5. Describe how the 10 statements' MaxDiff scores are associated with one another.

We can use correlatons to investgate whether the 10 statements' MaxDiff scores are associated with one another or not. The table 2 shows the correlatons between all of the variables in the dataset:

```r
round(M2,2) %>% kable(caption="correlation matrix") %>%
kable_styling(latex_options = c("striped", "scale_down"))
```

Based on the correlation matrix (Table 2), Correlatons range in value from -0.52 to 0.53.

A value of negative one means that two variables are negatively correlated. That is, a high value in one is associated with a low value in the other, and vice versa. On the other hand, a value of positive one means that two variables are are positvely correlated. As such, high values in one are associated with high values in the other, and vice versa. For instance, the correlation of stmtMD2 and stmtMD4 is of 0.53, which indicates that the two variable are positively correlated.

We can form two sets of variables, mainly defined as follows:

- set one: {(stmtMD8,stmtMD5) , (stmtMD3,stmtMD2, stmtMD4)}: Each pairwise correlation is positive
- set two: {(stmtMD10,stmtMD7,stmtMD6 and stmtMD1)}: Each pairwise correlation is positive
- set one and set two are negatively correlated. (Each variable of the first group is negatively related to the variable of the second set)

Further, stmtMD9have zero correlation the the remaining variables respectively, except the correlation with stmtMD4 and stmtMD2 is negative. A correlation of zero indicates that two variables are perfectly uncorrelated. This means that their values do not associate with one another.

6. Use *linear regression models* to quantify how each statement's MaxDiff score is related to its purchase likelihood ratings. You can supplement these models graphically. Interpret the results in terms of differences in the strengths of the relationships between the scores and their ratings.

For these regression models, the purchase likelihood rating is used as the dependent variable and the MaxDiff score is used as the independent (or predictor) variable.

The simple linear regression is conducted in order to predict the purchase likelihood rating based on the MaxDiff score:

$$Y_i = \alpha + \beta * x_i + \epsilon_i$$

Let's get the data we want to use for regression:

```r
data.model <- fakeMayo %>% select(CASEID,PIS1:PIS10) %>%
  inner_join(fakeMayoMDScores, by = "CASEID")
```

```r
#predict the purchase likelihood rating based using the MaxDiff score
lm1 <- lm(PIS1 ~ stmtMD1, data=data.model)
lm2 <- lm(PIS2 ~ stmtMD2, data=data.model)
lm3 <- lm(PIS3 ~ stmtMD3, data=data.model)
lm4 <- lm(PIS4 ~ stmtMD4, data=data.model)
lm5 <- lm(PIS5 ~ stmtMD5, data=data.model)
lm6 <- lm(PIS6 ~ stmtMD6, data=data.model)
lm7 <- lm(PIS7 ~ stmtMD7, data=data.model)
lm8 <- lm(PIS8 ~ stmtMD8, data=data.model)
lm9 <- lm(PIS9 ~ stmtMD9, data=data.model)
lm10 <- lm(PIS10 ~ stmtMD10, data=data.model)
```

The summary of the models are as follows:

```r
stargazer(lm1,lm2,lm3,lm4,lm5,lm6,lm7,lm8,lm9,lm10, type = "text",
          title="Regression Results", single.row=F,font.size ="tiny",
           align = TRUE,
          omit.stat=c("f", "ser"),
          column.sep.width = "-15pt",nobs = FALSE,
          omit="Observations") # Well... you can tweak this)
```

```
##
## Regression Results
## ====================================================================================================
##                                                  Dependent variable:
##              --------------------------------------------------------------------------------------
##                PIS1     PIS2     PIS3     PIS4     PIS5     PIS6     PIS7     PIS8     PIS9    PIS10
##                (1)      (2)      (3)      (4)      (5)      (6)      (7)      (8)      (9)      (10)
## ----------------------------------------------------------------------------------------------------
## stmtMD1      0.270***
##              (0.043)
##
## stmtMD2                0.050
##                       (0.045)
##
## stmtMD3                        -0.100**
##                                 (0.047)
##
## stmtMD4                                   0.014
##                                          (0.047)
##
## stmtMD5                                           -0.026
##                                                  (0.059)
##
## stmtMD6                                                   0.260***
##                                                            (0.041)
##
## stmtMD7                                                            0.280***
##                                                                     (0.058)
##
## stmtMD8                                                                      0.027
##                                                                             (0.054)
##
## stmtMD9                                                                               0.037
##                                                                                      (0.046)
##
## stmtMD10                                                                                       0.005
##                                                                                               (0.054
##
## Constant     3.200*** 3.200*** 3.300*** 3.300*** 3.100*** 3.200*** 3.100*** 3.200*** 3.100*** 3.400**
##              (0.080)  (0.085)  (0.075)  (0.078)  (0.100)  (0.091)  (0.100)  (0.090)  (0.089)  (0.100
##
## ----------------------------------------------------------------------------------------------------
## Observations   217      217      217      217      217      217      217      217      216      217
## R2            0.150    0.006    0.021    0.0004   0.001    0.150    0.100    0.001    0.003   0.0000
## Adjusted R2   0.150    0.001    0.016    -0.004   -0.004   0.150    0.096    -0.003   -0.002   -0.005
## ====================================================================================================
## Note:                                                                      *p<0.1; **p<0.05; ***p<0.0
```

The regression equations are as the following:

$$\begin{gathered}
PIS1 = 3.2 + 0.270 * stmtMD1; \quad t = 6.23, p < 0.001 \\
PIS2 = 3.2 + 0.050 * stmtMD2; \quad t = 1.11, p = 0.27 > 0.05 \\
PIS3 = 3.3 - 0.100 * stmtMD3; \quad t = -2.13, p = 0.034 < 0.05 \\
PIS4 = 3.1 + 0.014 * stmtMD4; \quad t = 0.3, p = 0.76 > .05 \\
PIS5 = 3.2 - 0.026 * stmtMD5; \quad t = -0.44, p = 0.66 > .05 \\
PIS6 = 3.2 + 0.260 * stmtMD6; \quad t = 6.29, p < 0.001 \\
PIS7 = 3.1 + 0.280 * stmtMD7; \quad t = 4.89, p < 0.001 \\
PIS8 = 3.2 + 0.027 * stmtMD8; \quad t = .51, p = 0.61 > .05 \\
PIS9 = 3.1 + 0.037 * stmtMD9; \quad t = 0.82, p = 0.42 > .05 \\
PIS10 = 3.4 + 0.005 * stmtMD10; \quad t = 0.1, p = 0.92 > .05
\end{gathered} \tag{1}$$

The purchase likelihood rating for statements 1, 6, and 7(PIS1, PIS6 and PIS7) were significantly explained by the MaxDiff score 1, 6 and 7 (stmtMD1, stmtMD6, stmtMD7), respectively. Since the p-value for the corresponding t-test is lower than 5% level of significance. For instance, the is a positive association between purchase likelihood rating for statement PIS1 and the MaxDiff score stmtMD1, but it not not very strong. For each additional maxdiff score stmtMD1 by one, the purchase likelihood rating PIS1 increases by about 0.05. Knowing the MaxDiff score stmtMD1 for only about 15% of the variation in purchase likelihood rating PIS1.

In contrast, all remaining relationships are insignificant (PIS2 and stmtMD2, PIS3 and stmtMD3, PIS4 and stmtMD4, PIS5 and stmtMD5, PIS8 and stmtMD8, PIS9 and stmtMD9, PIS10 and stmtMD10).

7. Do a *cluster analysis* on the stmtMD1 through stmtMD10 MaxDiff scores. Can you identify two or more "types" of customers? How do these types differ? Be sure to get the best solution you can.

A hierarchical clustering on the stmtMD1 through stmtMD10 MaxDiff scores is performed. We want to understand systematic patterns and relationships between the Maxdiff scores in order to identity the customer types.

```r
# Visualize kmeans clustering
library(factoextra)

row.names(fakeMayoMDScores) <- fakeMayoMDScores$CASEID
fakeMayoMDScores$CASEID <- NULL #remove ID
# # PAM clustering
# # ++++++++++++++++++++
 library(cluster)
 pam.res <- pam(fakeMayoMDScores, 3)
#  # Visualize pam clustering
 fviz_cluster(pam.res, geom = "point", ellipse.type = "norm")
```

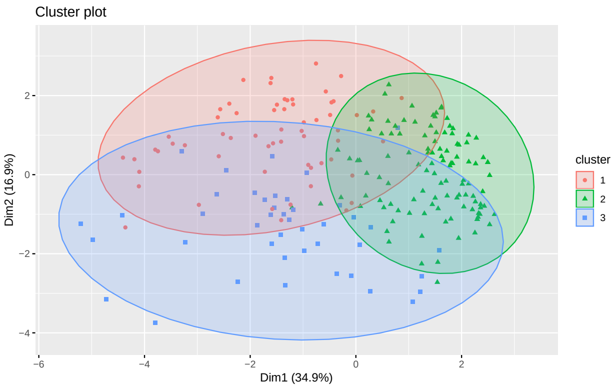

Figure 3: Cluster plot

```
#
# # Hierarchical clustering
# # ++++++++++++++++++++++++
# # Use hcut() which compute hclust and cut the tree
 hc.cut <- hcut(t(fakeMayoMDScores), k = 3, hc_method = "complete")
# # Visualize dendrogram
 fviz_dend(hc.cut, show_labels = TRUE, rect = TRUE)
```

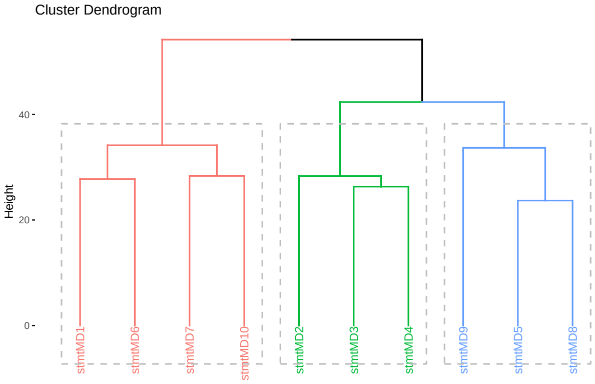

Figure 4: Displays all possible clusters of customers

Based on the cluster dendrogram (Figure 4), which displays all possible clusters of customers. “The height indicates the distance between the objects, i.e. it is a measure of closeness”. Three rectangles have been drawn, highlighting a cluster group of Maxdiff scores - (STMTMD1, STMTMD6, STMTMD7, and STMTMD10), (STMTMD2, STMTMD3, STMTMD4), and (STMTMD9, STMTMD5 STMTMD8).

We highlight three biggest clusters in red, green and blue rectangles, which show three groups of customers. Finally, We want to present a line chart, see figure 5, which displays means of Maxdiff scores for three abovementioned clusters:

```
fakeMayoMDScores$cluster <- factor(pam.res$cluster)
fakeMayoMDScores%>%
  gather(key, value, -cluster) %>% #wide format
  group_by(cluster,key) %>%
  summarise_at(vars(value),funs(mean)) %>%
  ggplot(aes(x=key, y=value))+
  geom_line(aes(colour = cluster, group = cluster)) +
  geom_point(aes(colour = cluster, group = cluster))+
  labs(x= "Maxdiff scores", y="Mean of each variable in the 3 clusters")+
  theme_bw()+
  theme( legend.position="top")
```

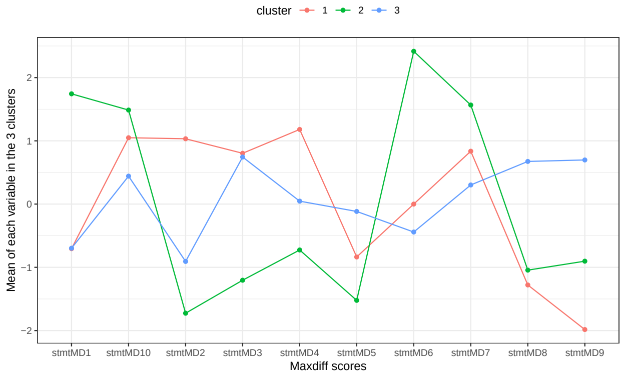

Figure 5: Means plot of clusters

Figure 5 presents a mean of each individual indicator for the three cluster (red green and blue points), i.e. types of customers:

- Cluster one (green points): Highest (STMTMD1, STMTMD6, STMTMD7, and STMTMD10) scores and lowest (STMTMD2, STMTMD3, STMTMD4, and STMTMD5) scores,
- Cluster two (red points): Highest (STMTMD2, STMTMD3, STMTMD4) scores, and lowest (STMTMD8, and STMTMD9) scores
- Cluster three (blue points): Highest (STMTMD9, STMTMD5 STMTMD8) and zero mean scale for the remaining Maxdiff scores.

Finally, we can conclude that we found in the Maxdiff score dataset three patterns of customer behaviour. Further cluster analysis revealed two major differences and a minor difference. (Figure 5) Cluster 1 (green dots) was clearly opposed to cluster 2 (red dots). While group three (blue dots) is the only group that preferred the statements: STMTMD5 (Preferred over the leading brand of Light mayonnaise), STMTMD8 (Tastes as good as the leading brand of Light mayonnaise) and STMTMD9 (Tastes better than Ecclesâs Light mayonnaise).