STQD6134: Business Analytics

Data Exploration

Exploratory data analysis

attributed to the statistician John Tukey in a book of the same name (Tukey, 1977).

Exploratory data analysis means examining a dataset to discover its underlying characteristics with an emphasis on visualization. It helps you during analysis design to determine if you should gather more data, suggest hypotheses to test, and identify models to develop. In this chapter, we will cover the following four topics related to exploratory data analysis:

- Understanding exploratory data analysis
- Analyzing a single data variable
- Analyzing two variables together
- Exploring multiple variables simultaneously

Question → Models → Answer

Types of measurement scales

Scale	Basic empirical operations	Permissible statistics
Nominal	Determining equality or membership	Number of cases Mode Contingency correlation
Ordinal	Determining of greater than or less than	Median Percentiles
Interval	Determining equality of interval or difference	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determining equality of ratios	Coefficient of variation

Gender	Level	Age	GPA
Male	Sophomore	15	3.6
Female	Freshman	14	3.2
Female	Sophomore	14	3.3
Male	Junior	16	3.7
Female	Senior	18	3.1
Male	Senior	17	2.8

- Gender is **nominal**. A student can be either a male or female. You can calculate the *count* and *mode* of nominal scale data.
- Level is **ordinal**. Seniors are above juniors, juniors are above sophomores, and sophomores are above freshmen. You can calculate the *median* and *percentiles* of ordinal scale data when represented numerically.
- Age is **interval**. Values are continuous with an understood increment between them. In this example, an integer represents a student's age and they differ by increments of one year. You can calculate the *mean* and *standard deviation* of interval scale data.
- GPA is a **ratio**. It is a combination of two other numbers expressed as a rational number.

Analyzing a single data variable

```
marketing <- read.csv("./data/Ch3_marketing.csv", stringsAsFactors = TRUE)
str(marketing)
```

```
marketing$pop_density <- factor(marketing$pop_density,
                                ordered = TRUE,
                                levels = c("Low", "Medium", "High"))
```

Focus on two variables, `google_adwords` (interval, numeric) and `pop_density` (ordinal, factor). You can learn their distributions using a tabular or graphical approach.

Tabular exploration

```
summary(marketing$google_adwords)
```

- **Minimum**: This is the smallest observation in the dataset
- **First Quartile**: 25% of the data lies below this value
- **Median**: This is the middle observation
- **Mean**: The average of the observations
- **Third Quartile**: 75% of the data lies below this value
- **Maximum**: This is the argest observation in the dataset

```
sd(marketing$google_adwords)
```

```
var(marketing$google_adwords)
```

- The values range from a minimum of 23.65 to a maximum of 321
- 25% of the observations are below 97.25
- 50% of the observations are below 169.50 (the median)
- 75% of the observation are below 243.10

```
summary(marketing$pop_density)
```

The output is as follows:

```
   Low Medium   High
    68     52     52
```

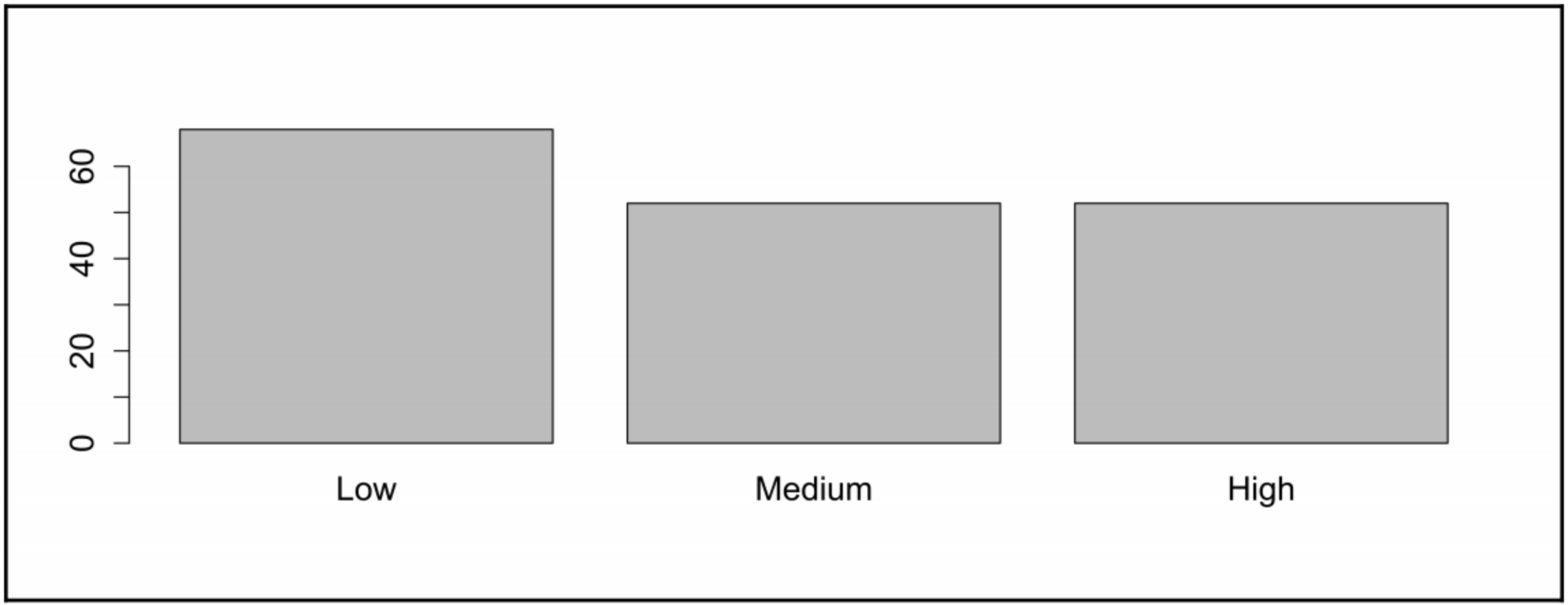
Graphical exploration

"The greatest value of a picture is when it forces us to notice what we never expected to see."

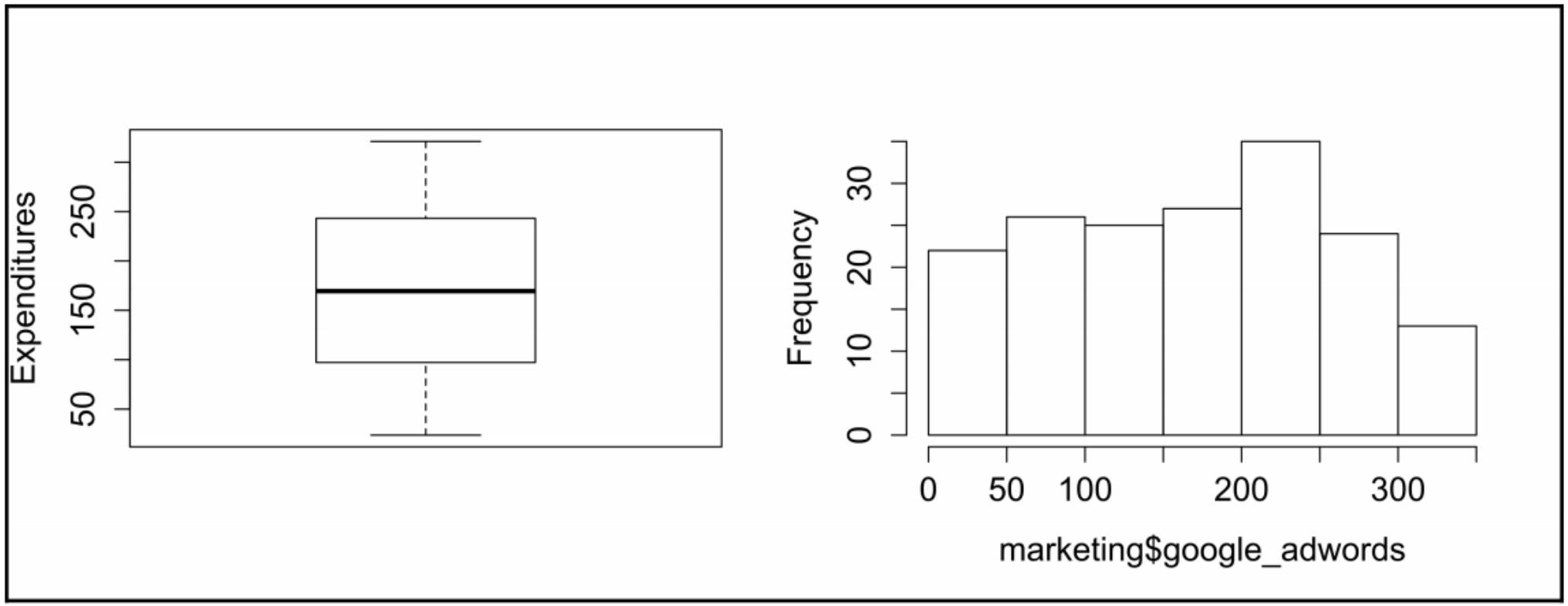
– John Tukey

```
plot(marketing$pop_density)
```

It will return the following output:



```
boxplot(marketing$google_adwords, ylab = "Expenditures")
hist(marketing$google_adwords, main = NULL)
```



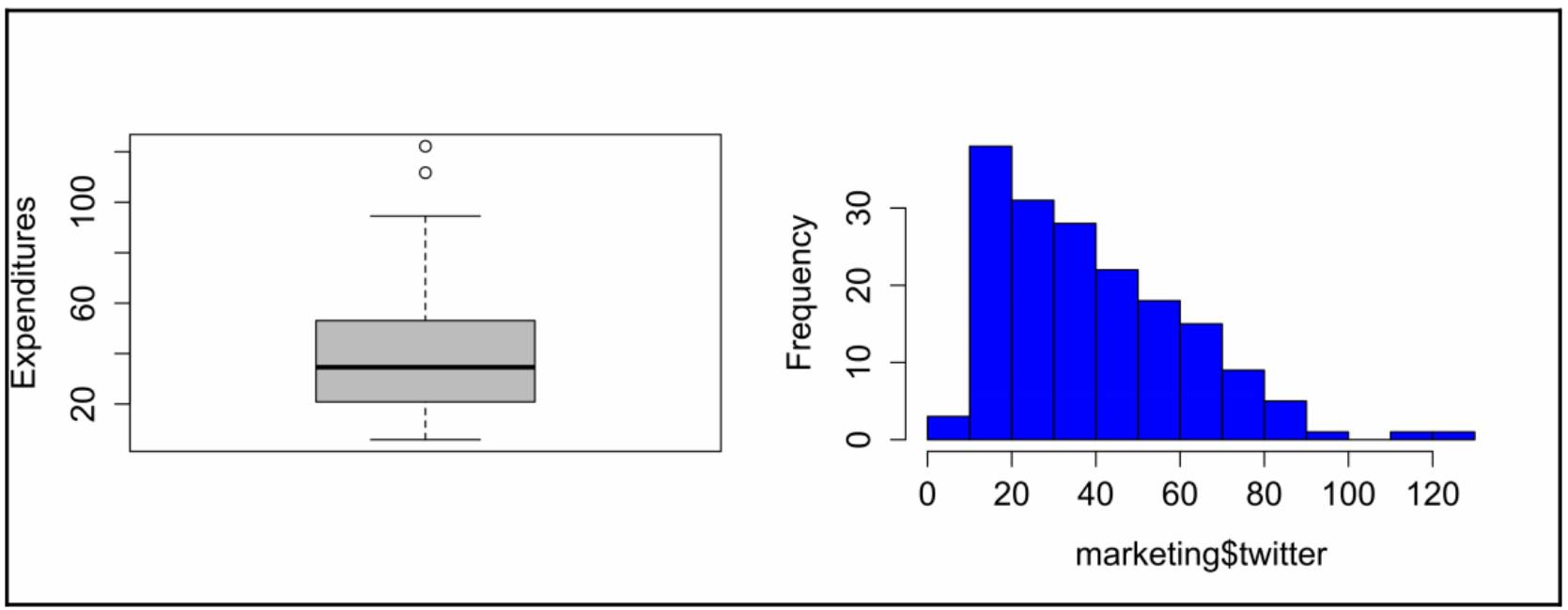
tabular output using `summary(marketing$twitter)`:

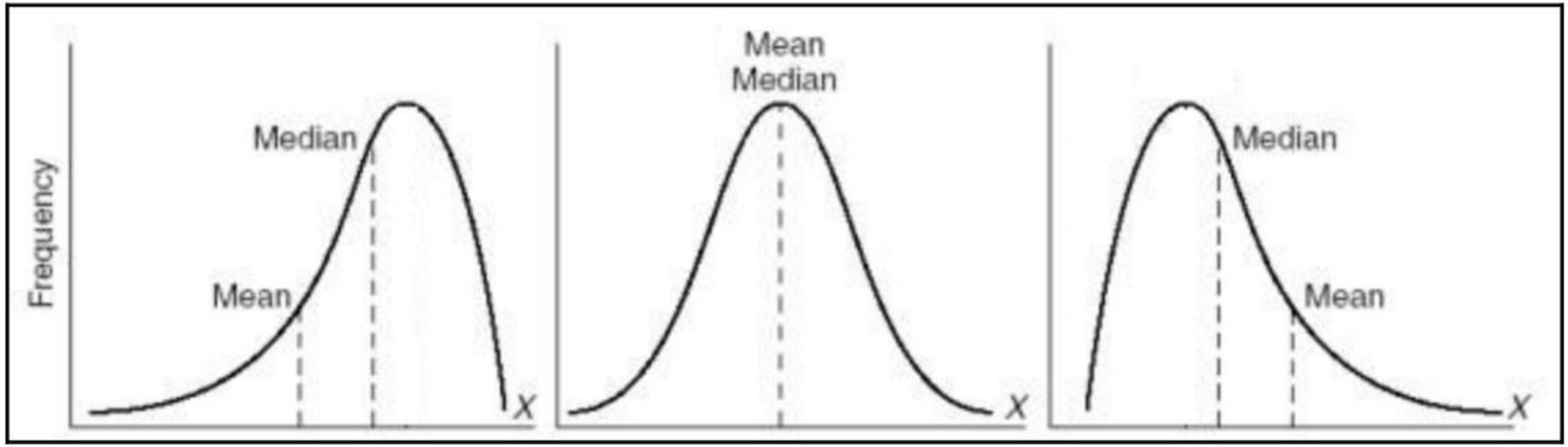
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.89   20.94   34.60   38.98   52.94  122.20
```

Notice the difference between the median and the mean. This difference indicates skewed data. Plot the boxplot and histogram to see the skew using graphical exploration:

```
boxplot(marketing$twitter, ylab = "Expenditures", col = "gray")
hist(marketing$twitter, main = NULL, col = "blue")
```

We will get the following output:




Frequency
Median
Mean
X
Mean
Median
X
Median
Mean
X

Analyzing two variables together

- What does the data *look* like?
- Is there any *relationship* between two variables?
- Is there any *correlation* between the two?
- Is the correlation *significant*?

```
summary(marketing)
```

Adding and removing variables: You may have noticed an extra column in this data. We added the `emp_factor` variable to show you plots between two variables that are factors. Here is how you can add variables to a data frame:

```
marketing$emp_factor <- cut(marketing$employees, 2)
```

There are two elements in this line of code. Look at what is going on here:

- The `cut()` function: This converts a number to a factor by dividing the values into a number of intervals you choose. In this case, you will pass in `marketing$employees` and tell R to cut it into two (`2`) factors. It splits the data at the midpoint between the minimum value (`3`) and the maximum value (`12`) resulting in a factor for 3 to 7 and 8 to 12 employees.
- Create a new variable: You can add new columns (variables) to a data frame by assigning the result of some operation to a new variable name. In this case, you used the `cut()` function and assigned the result to a new `emp_factor` variable using this new name as a variable.

The new variable is only there temporarily to show you some functionality in this section. You will remove it at the end of this section using the following code:

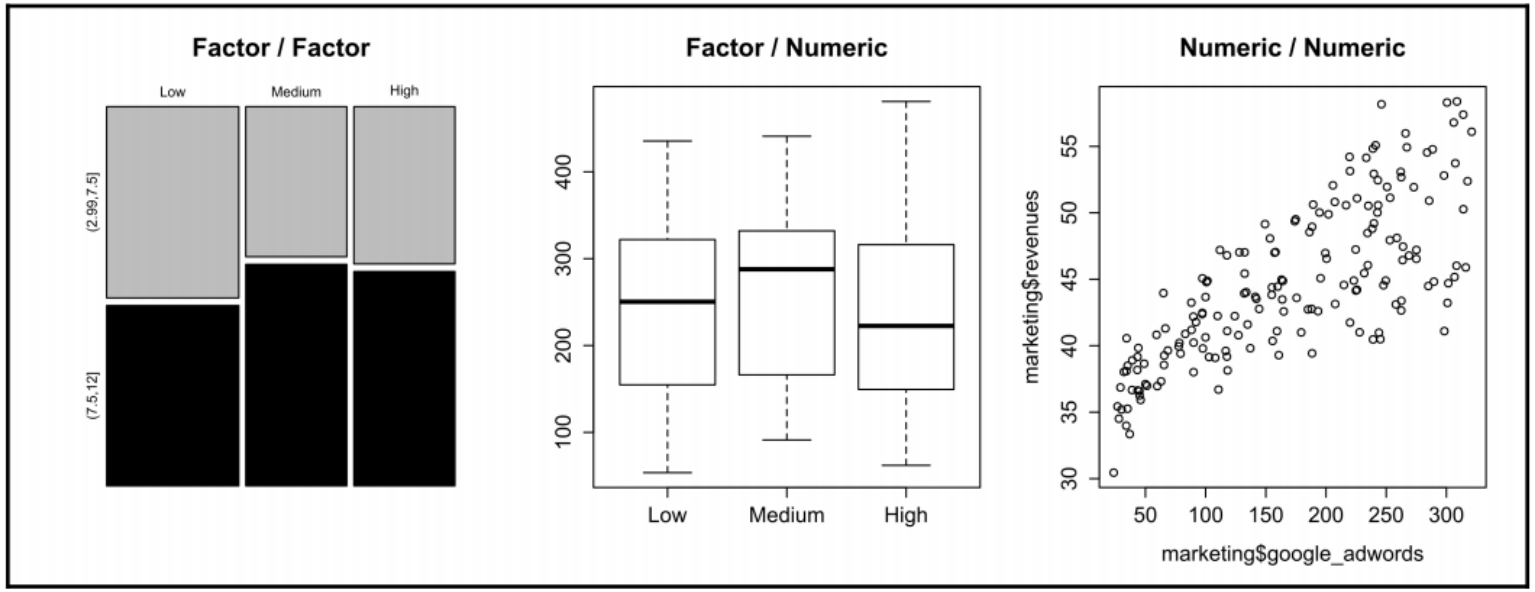
```
marketing$emp_factor <- NULL
```

Relationship between two variables

```
table(marketing$emp_factor, marketing$pop_density)
```

```
mosaicplot(table(marketing$pop_density, marketing$emp_factor),
           col = c("gray","black"), main = "Factor / Factor")
boxplot(marketing$marketing_total ~ marketing$pop_density,
        main = "Factor / Numeric")
plot(marketing$google_adwords, marketing$revenues,
     main = "Numeric / Numeric")
```

The three combinations of graphics are shown here:



```
cor(marketing$google_adwords, marketing$revenues)

cor(marketing$google_adwords, marketing$facebook)
```

- **Sign** will either be positive (+) or negative (-). Positive correlation means that as the first variable increases, the second variable increases as well. Negative correlation means that when the first variable increases, the second variable decreases. Both the previous examples are positively correlated.
- **Value** of the correlation result ranges from zero (0) to one (1). The value increases as the correlation strength increases. The first example has a correlation of `0.766`, which is much stronger than the second example (`0.076`). A value of zero indicates no correlation between two variables.

```
cor.test(marketing$google_adwords, marketing$revenues)
```

- http://stattrek.com/hypothesis-test/hypothesis-testing.aspx
- https://www.khanacademy.org/math/probability/statistics-inferential/hypothesis-testing/v/hypothesis-testing-and-p-values

A **null hypothesis** typically represents the status quo-meaning that there is nothing going on. In our example, the null hypothesis is that the true correlation between `google_adwords` and `revenues` is equal to zero. In other words, the variables have no correlation. The **alternative hypothesis** is that the correlation between the two is not equal to zero and that the correlation is significant.

A **t-test** gets at the question: *how surprising is it to see this degree of correlation if the two variables truly are not correlated?* Going into the details of t-test is beyond the scope of this book, but the result of the t-test is `15.548`. How surprising is this result? The **p-value** associated with this value of `t` is `2.2e-16`. This essentially says, *if the null hypothesis were indeed true, the probability of getting t = 15.548 would be 0.00000000000000022*.

In other words, the probability of getting this result, if the null hypothesis were true, would essentially be zero. Therefore, you will reject the null hypothesis, stating that the correlation is zero. A typical value used for rejecting a null hypothesis is a p-value less than 0.05.

BI tip: Keep all this statistics stuff in perspective. Your goal is not to become a statistician. Rather, your goal is to become more aware of what your analysis means and what are its limitations. Business decisions are not the same as clinical drug trials, but you want to know when you can have faith in the analysis you provide to business leaders that informs their decisions.

Putting your business hat back on, you decide to determine whether the other two marketing channels (Twitter and Facebook) are correlated and significant. You run `cor.test()` on these variables against `revenues`. The following is a summary of the test results:

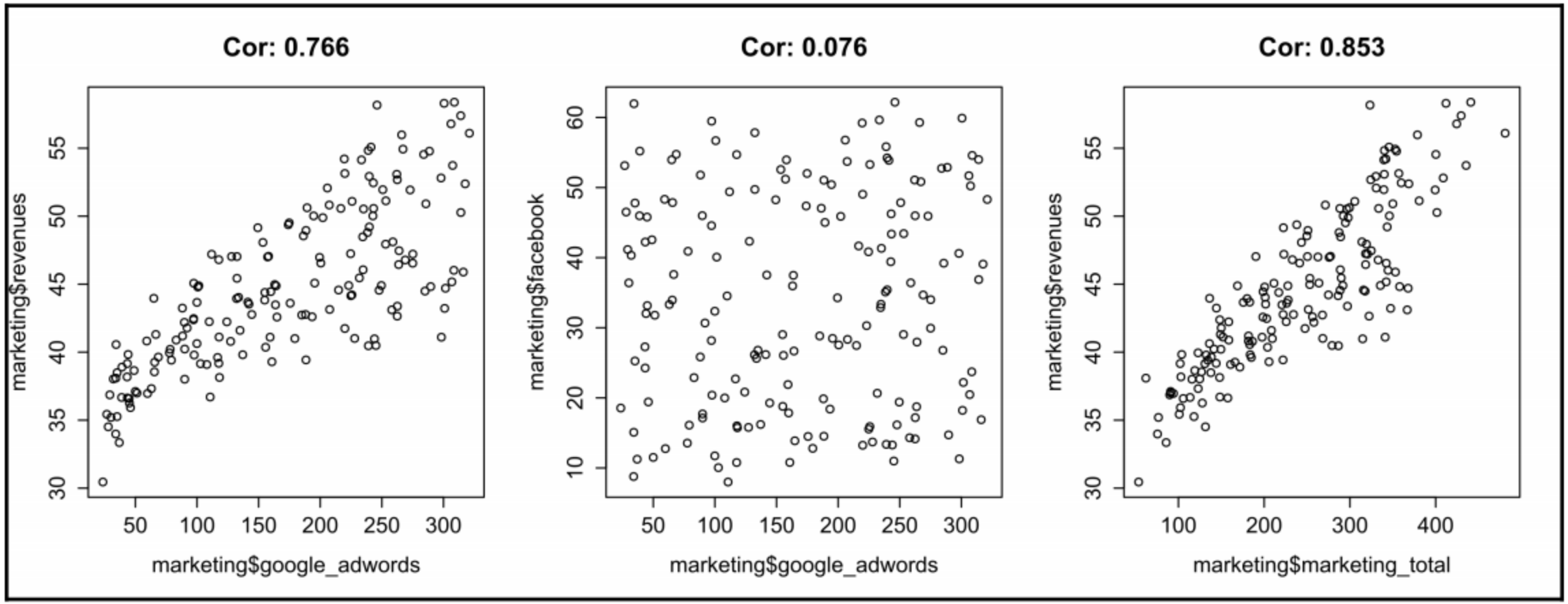
Relationships	`twitter` **and** `revenues`	`facebook` **and** `revenues`
t	`3.6516`	`9.2308`
p-value	`0.0003467`	`2.2e-16`
correlation	`0.2696854`	`0.5778213`
Significant?	Yes	Yes

```
cor.test(marketing$google_adwords, marketing$facebook)
```

```
cor.test(marketing$revenues, marketing$marketing_total)
```

```
plot(marketing$google_adwords, marketing$revenues)
plot(marketing$google_adwords, marketing$facebook)
plot(marketing$marketing_total, marketing$revenues)
```

The output is shown in the following figure:

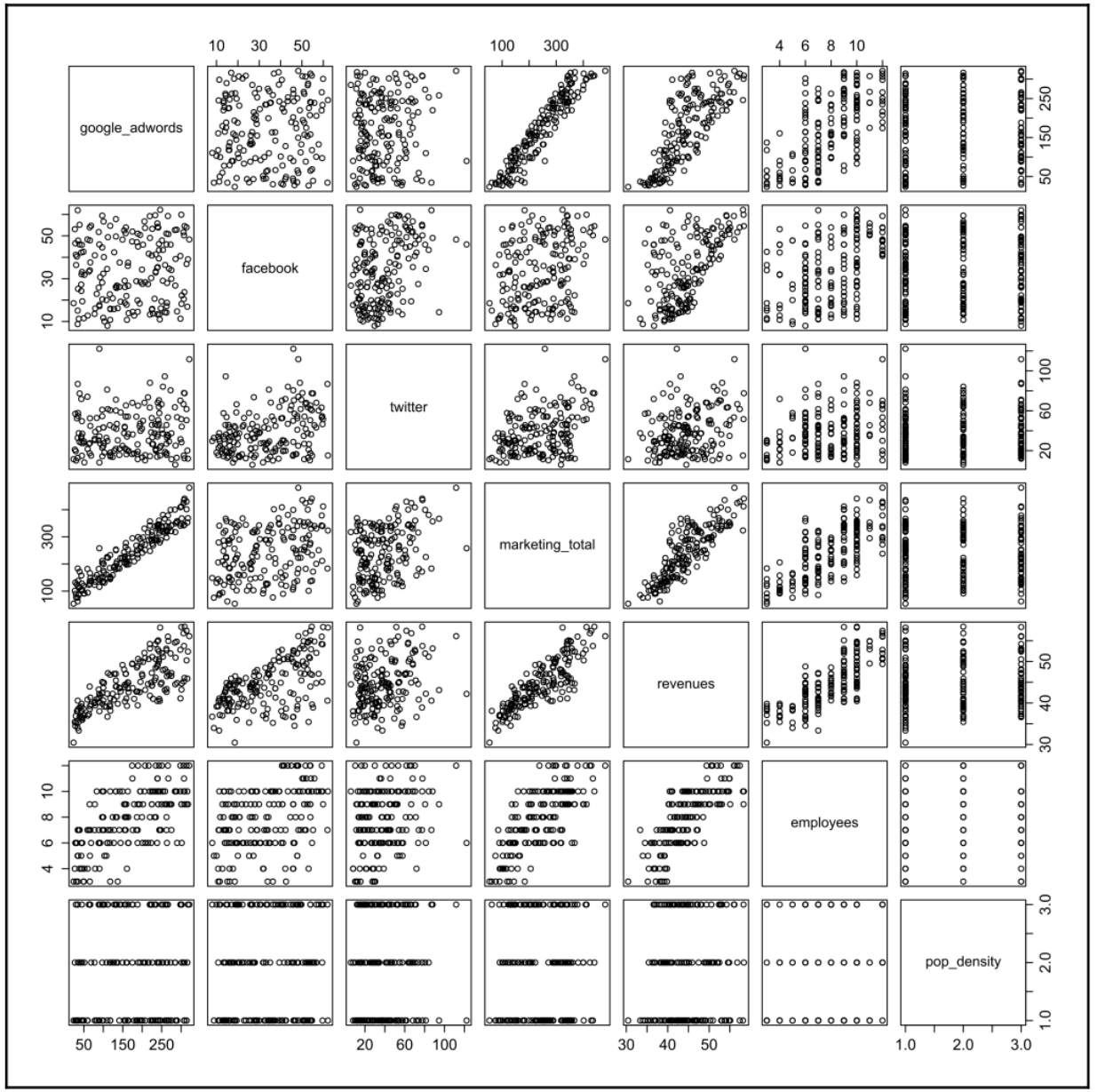


Exploring multiple variables simultaneously

```
pairs(marketing)
```

```
cor(marketing[ ,1:6])
```

```
library(psych)
corr.test (marketing[ ,1:6])
```

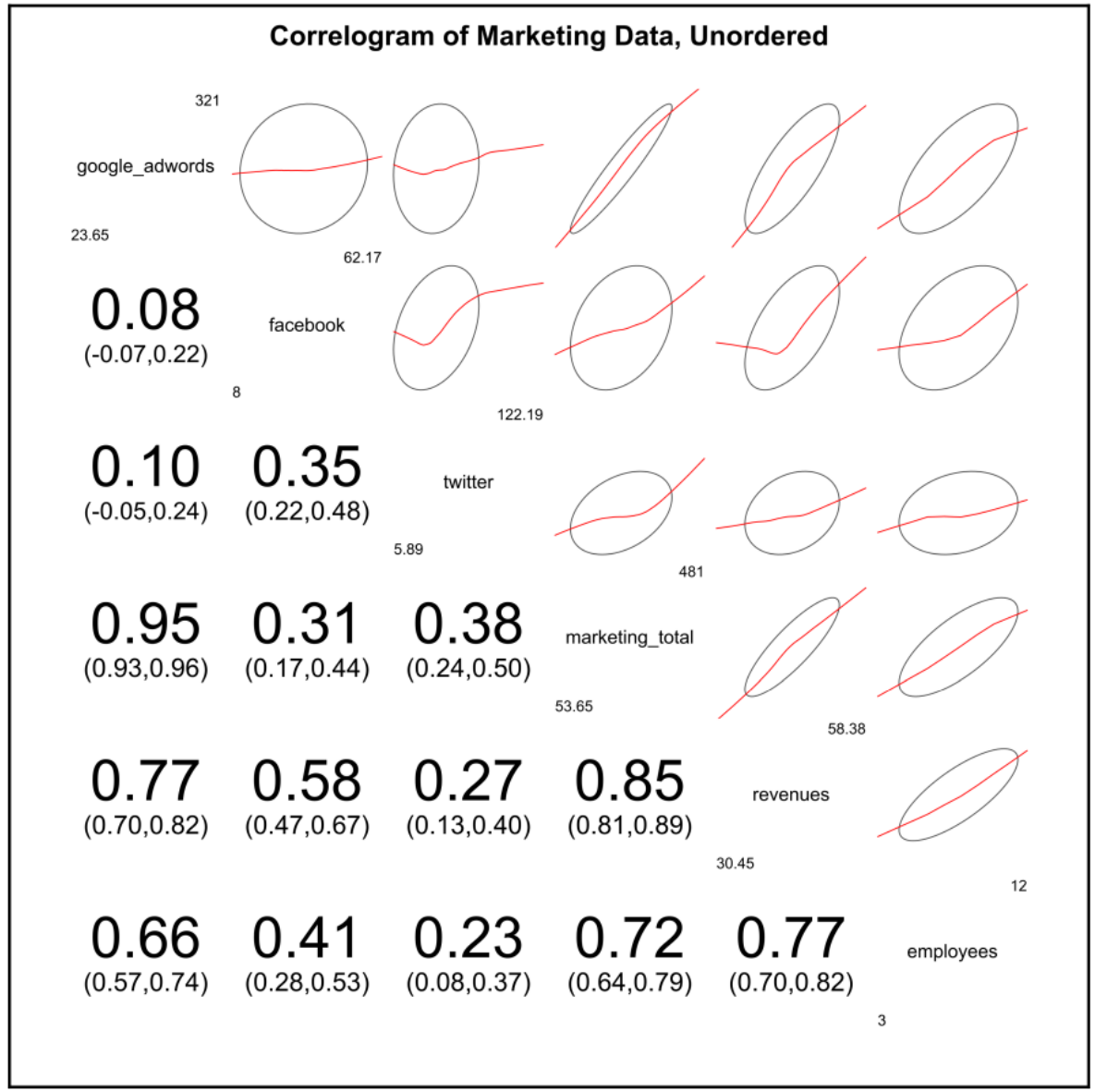


Correlogram

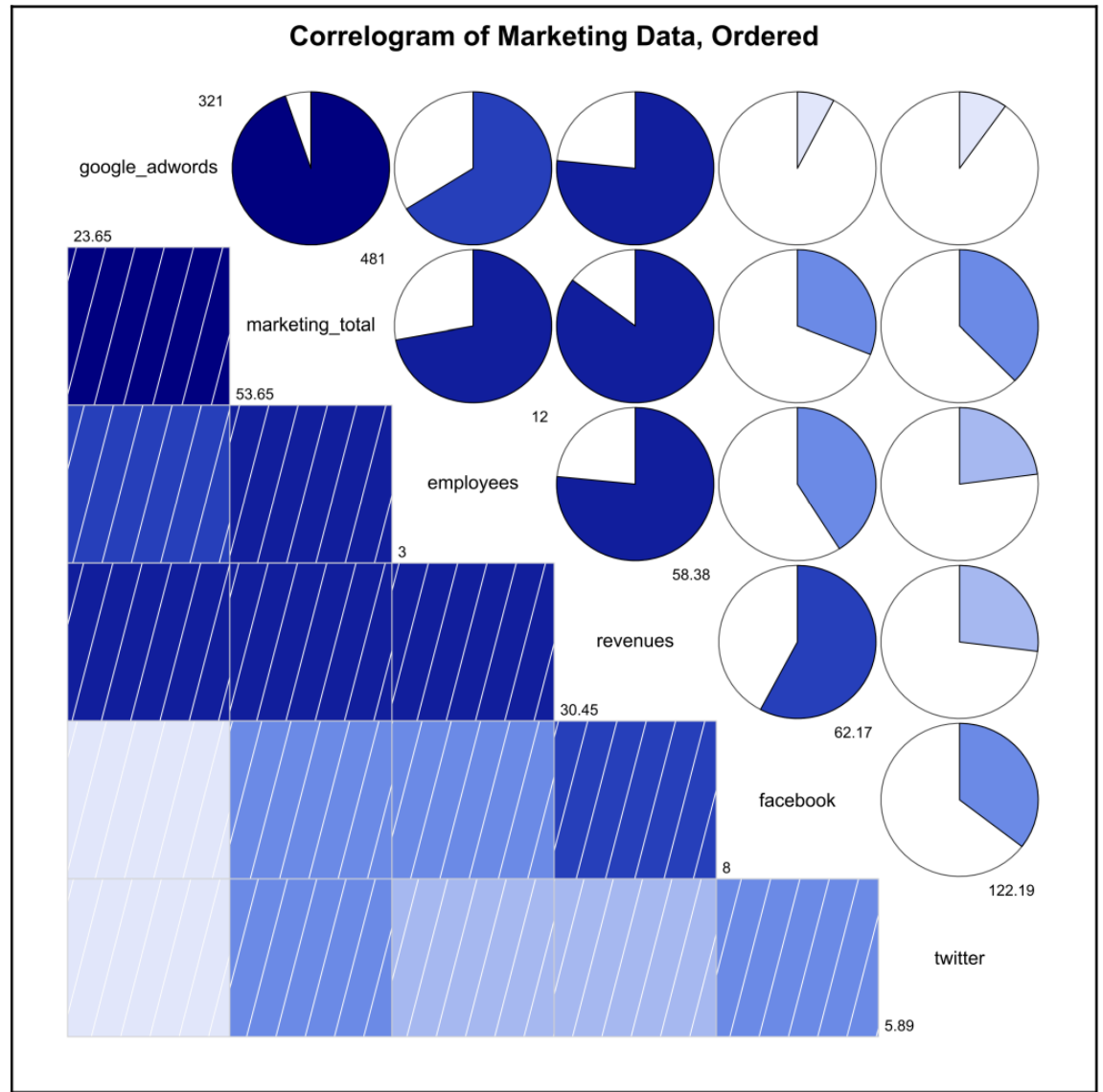
```
library(corrgram)
corrgram(marketing[ ,1:6], order = FALSE,
         main = "Correlogram of Marketing Data, Unordered",
         lower.panel = panel.conf, upper.panel = panel.ellipse,
         diag.panel = panel.minmax, text.panel = panel.txt)
```

Four parameters appear in the code. They control the content of the regions of the correlogram. These parameters in this example are explained as follows:

- **Lower Panel**: The bottom-left half of the graphic was set to display correlation coefficients and confidence intervals using `panel.conf`
- **Upper Panel**: The upper-right half of the graphic was set to display ellipses and smooth lines using `panel.ellipse`
- **Diagonal and Text**: The diagonal contains the name of the variable and its minimum and maximum values using `panel.minmax` and `panel.txt`, respectively


Correlogram of Marketing Data, Unordered
google_adwords
facebook
twitter
marketing_total
revenues
employees
0.08
(-0.07,0.22)
0.10
(-0.05,0.24)
0.35
(0.22,0.48)
0.95
(0.93,0.96)
0.31
(0.17,0.44)
0.38
(0.24,0.50)
0.77
(0.70,0.82)
0.58
(0.47,0.67)
0.27
(0.13,0.40)
0.85
(0.81,0.89)
0.66
(0.57,0.74)
0.41
(0.28,0.53)
0.23
(0.08,0.37)
0.72
(0.64,0.79)
0.77
(0.70,0.82)
321
23.65
62.17
8
122.19
5.89
481
53.65
58.38
30.45
12
3

```
corrgram(marketing[ ,1:6], order = TRUE,
         main = "Correlogram of Marketing Data, Ordered",
         lower.panel = panel.shade, upper.panel = panel.pie,
         diag.panel = panel.minmax, text.panel = panel.txt)
```



Task 1 (due on 6th Nov 2020, 5pm) – email to zamira@ukm.edu.my

- Using the bike data last week (the cleaned version), answer the following questions:

  - Were you able to load the data and reconvert the data types? How many variables are in the dataset? How many observations?
  - How many observations are there for each season?
  - What is the mean and standard deviation of the `temp` variable?
  - Which variable is distributed nearly normally?
  - Which variable is most left skewed and most right skewed?
  - Which pair of variables has the greatest positive correlation and greatest negative correlation?
  - Which variable(s) shows significant correlation?