|  |  |
| --- | --- |
| Lab 1 – R Language | Name: SHUYUE JIAID: 56846018 |

1. Download the “yearly\_sales.csv” file from CANVAS
2. Open the csv file using a spreadsheet program. What is the csv file?

> data <- read.csv('yearly\_sales.csv')

> data

Answer:

This CSV file consists of customer IDs, total sales, number of orders, and gender of the customers. It might be a **sales summary spreadsheet** in a company, which can be utilized to further analyses the sales data of the company (or department) and help managers make decisions to profit in the future.

1. Open the program “RStudio”
2. Load the previously downloaded csv file into R with the command:

sales <- read.csv('yearly\_sales.csv')

1. Type “head(sales)”. What can you observe?

> head(data)

cust\_id sales\_total num\_of\_orders gender

1 100001 800.64 3 F

2 100002 217.53 3 F

3 100003 74.58 2 M

4 100004 498.60 3 M

5 100005 723.11 4 F

6 100006 69.43 2 F

Answer:

A (small) part of the whole sheet, the first six rows in default.

Besides, the titles of each column and the indexes of the row data are also shown.

1. Type “summary(sales)”. What can you observe?

> summary(data)

cust\_id sales\_total num\_of\_orders gender

Min. :100001 Min. : 30.02 Min. : 1.000 Length:10000

1st Qu.:102501 1st Qu.: 80.29 1st Qu.: 2.000 Class :character

Median :105000 Median : 151.65 Median : 2.000 Mode :character

Mean :105000 Mean : 249.46 Mean : 2.428

3rd Qu.:107500 3rd Qu.: 295.50 3rd Qu.: 3.000

Max. :110000 Max. :7606.09 Max. :22.000

Answer:

This function produces result summaries of the sheet data.

In particular, with regard to **each column**, for the numerical values, the minimum value, the first quartile value, the median value, the mean value, and the third quartile value, and the maximum value are summarized. Meanwhile, **as for the string values**, the length of the rows, the class of the column, and the mode of the column are shown.

1. Type the following command. What can you observe?

plot(sales$num\_of\_orders,sales$sales\_total)

Chart, scatter chart

Description automatically generated

Answer:

The picture has shown **the total amount of sales with regard to the number of sales orders**.

As for the different orders, the corresponding sales amount was plotted in the form of dots.

We can see from the figure that as the number of orders increases, the amount of sales also increases, and when the number is larger, e.g., after 15 orders, the sales amount surges.

1. Type the following commands. What have you done to the “sales” data?

sales$per\_order <- sales$sales\_total/sales$num\_of\_orders

head(sales)

> data$per\_order <- data$sales\_total / data$num\_of\_orders

> head(data)

cust\_id sales\_total num\_of\_orders gender per\_order

1 100001 800.64 3 F 266.8800

2 100002 217.53 3 F 72.5100

3 100003 74.58 2 M 37.2900

4 100004 498.60 3 M 166.2000

5 100005 723.11 4 F 180.7775

6 100006 69.43 2 F 34.7150

Answer:

Using this script, a new column called “per\_order” can be added in the right of the sheet. Specially, it shows the **amount of sales per order** for each customer (each row).

1. Type the following command. What have you done?

write.table(sales,"sales\_modified.txt", sep="\t", row.names=FALSE)

Answer:

Based on this command, the “sales” data can be saved into a TXT file called “sales\_modified.txt” with TAB as a separation and the row indexes are discarded.

1. Type the following commands. What have you done?

jpeg(file="sales\_hist.jpeg")

hist(sales$num\_of\_orders)

dev.off()

Chart, histogram

Description automatically generated

Answer:

First of all, a JPEG file called “sales\_hist.jpeg” was created.

Then, a histogram of number of orders with regard to the frequency was drawn.

Finally, the figure was saved.

1. Type the following commands. What have you done?

x <- sales$sales\_total

y <- sales$num\_of\_orders

Answer:

The sales\_total column of the sales sheet is assigned to a variable x.

And the num\_of\_orders of the sales is assigned to variable y.

1. Type the commands in the leftmost column and fill in the following table:

|  |  |  |
| --- | --- | --- |
| R Command | Return Value | Semantic Meaning (hint: use the help panel in RStudio) |
| cor(x,y) | 0.7508015 | The correlation between the total amount of sales and the number of orders. |
| cov(x,y) | 345.2111 | The covariance between the total amount of sales and the number of orders. |
| IQR(x) | 215.21 | Computes interquartile range of the total amount of sales. |
| mean(x) | 249.4557 | The mean (average) value of the total amount of sales. |
| median(x) | 151.65 | The median value of the total amount of sales. |
| range(x) | 30.02 7606.09 | A vector containing the minimum and maximum of the total amount of sales. |
| sd(x) | 319.0508 | Computes the standard deviation of the total amount of sales. |
| var(x) | 101793.4 | Computes the variance of the total amount of sales. |

1. Apply the R knowledge you have learned in the previous steps to the dataset “zipIncome.csv” in CANVAS and state your data analytics insights in less than 50 words.

Answer:

The insight that I have found as the mean education level promotes, the mean householding incomes are growing. However, when the edition level is between 10 and 15, the income is also spanning a larger range compared with the lower education level.

1. **People are encouraged to obtain a higher degree to get a relatively higher income.**
2. **Even though some people keep a higher degree, a higher income is not guaranteed.**
3. This is the end; please upload this sheet with your answers to the submission system..