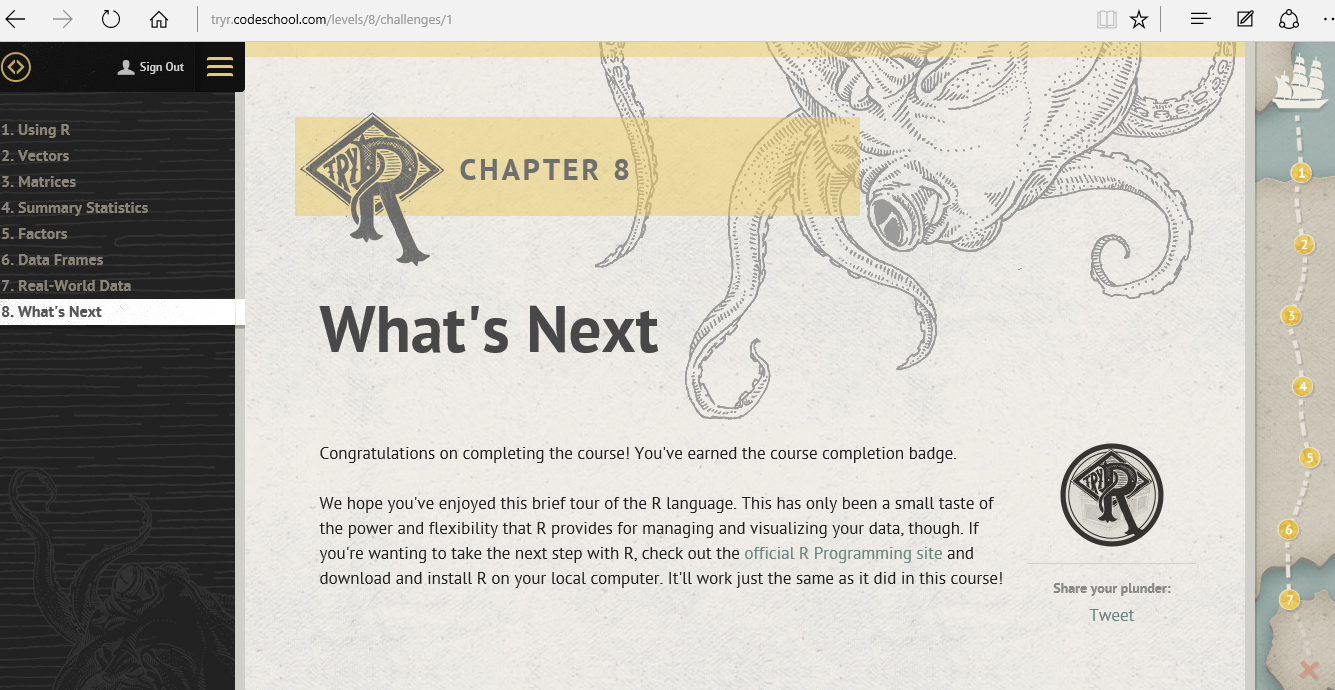
**Data Analysis & Visualisation Using R**

**Module: Tools for Data Analytics**

**Student: Galina Lopez 10333429**

Try R course from Code School completion badge.



**Report:**

* how you achieved the graphics based on your dataset

We downloaded R and the R Studio.

We opened cookbook and selected Graphs.

<http://www.cookbook-r.com/Graphs/>

From the Graphs menu we then selected Bar and Line graphs from ggplot2.

http://www.cookbook-r.com/Graphs/Bar\_and\_line\_graphs\_(ggplot2)

We decided to become familiar with them inside R using R Studio.

We selected sample data derived from the tips dataset in the reshape2 package.

For this we downloaded the reshape2 package in R Studio:

install.packages('reshape2')

and loaded the reshape2 package:

library(reshape2)

We created a data frame with our sample data to explore it using ggplot2 graphs.

dat <- data.frame(

  time = factor(c("Lunch","Dinner"), levels=c("Lunch","Dinner")),

  total\_bill = c(14.89, 17.23)

)

We downloaded the gglot2 package in R Studio:

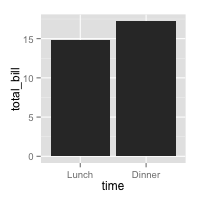
install.packages('ggplot2', dep = TRUE)

and loaded the ggplot2 package:

library(ggplot2)

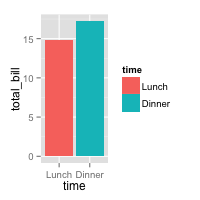
We then plotted a very basic bar graph of values using geom\_bar of ggplot2.

ggplot(data=dat, aes(x=time, y=total\_bill)) + geom\_bar(stat="identity")



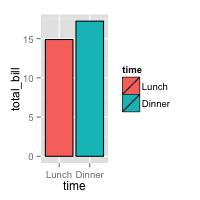
To map the time of day to different fill colours

ggplot(data=dat, aes(x=time, y=total\_bill, fill=time)) + geom\_bar(stat="identity")



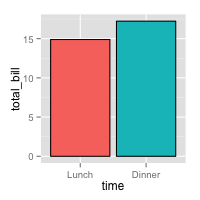
To add a black outline.

ggplot(data=dat, aes(x=time, y=total\_bill, fill=time)) + geom\_bar(colour=”black”, stat="identity")



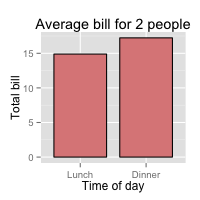
To remove the legend, since the information is redundant.

ggplot(data=dat, aes(x=time, y=total\_bill, fill=time)) + geom\_bar(colour=”black”, stat="identity") + guides(fill=FALSE)



To add title, narrower bars, fill colour and change axis labels.

ggplot(data=dat, aes(x=time, y=total\_bill, fill=time)) + geom\_bar(colour=”black”, fill=”#DD8888”, width=.8, stat="identity") + guides(fill=FALSE) + xlab(“Time of day”) + ylab(“Average bill for 2 people”)

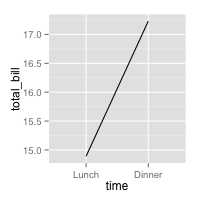


Second we plotted Line graphs.

For line graphs, the data points must be grouped so that it knows which points to connect. In this case, it is simple – all points should be connected, so group=1.

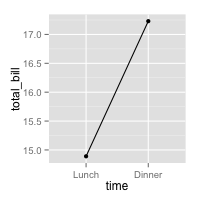
We then plotted a very basic line graph using geom\_line of ggplot2.

ggplot(data=dat, aes(x=time, y=total\_bill, group = 1)) + geom\_line()



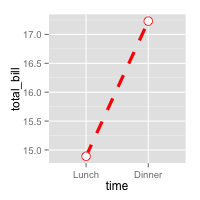
To add a point.

ggplot(data=dat, aes(x=time, y=total\_bill, group = 1)) + geom\_line() + geom\_point()



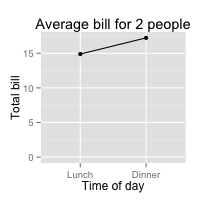
To change colour of both line and points, change line type and point type, use thicker line, larger points, change points to circles with white fill.

ggplot(data=dat, aes(x=time, y=total\_bill, group=1)) + geom\_line(colour=”red”, linetype=”dashed”, size=1.5) + geom\_point(colour=”red”, size=4, shape=21, fill=”white”)



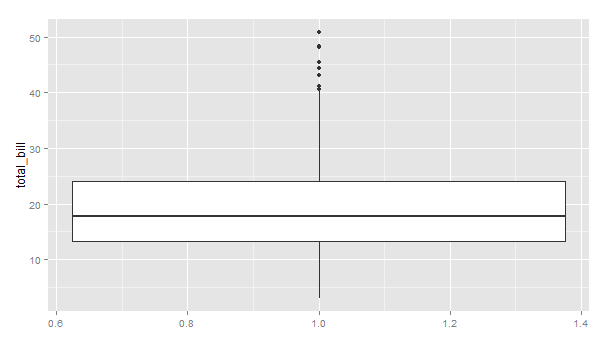
To change the y range to go from 0 to the maximum value in the total\_bill column and change axis labels.

ggplot(data=dat, aes(x=time, y=total\_bill, group=1)) + geom\_line() + geom\_point() + expand\_limit(y=0) + xlab(“Time of day”) + ylab(“Total bill”) + ggtitle(“Average bill for 2 people”)

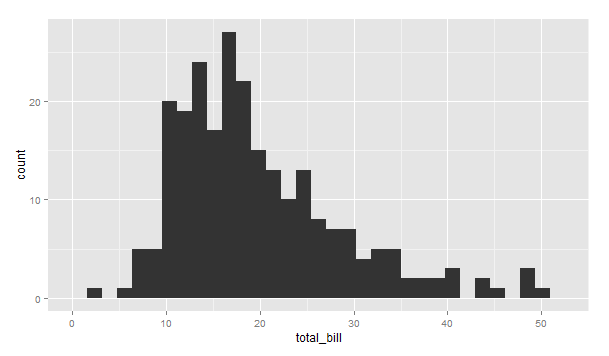


Third we decided to become familiar with the qplot function in ggplot2 to plot quickplots of our data set. So we plotted a boxplot and a histogram of the variable total\_bill in the data set tips and got more information about our data.

qplot(x = 1, y = total\_bill, data = tips, xlab = "", geom = 'boxplot')



qplot(x = total\_bill, data = tips, geom = 'histogram')



* what information could be gleamed from the dataset

We selected the tips data from the reshape2 package. Where one waiter recorded information about each tip he received over a period of a few months working in one restaurant. He collected several variables: tip in dollars, bill in dollars, sex of the bill payer, whether there were smokers in the party, day of the week, time of day, size of the party.

library(reshape2)

# Look at fist several rows

head(tips)

#> total\_bill tip sex smoker day time size

#> 1 16.99 1.01 Female No Sun Dinner 2

#> 2 10.34 1.66 Male No Sun Dinner 3

#> 3 21.01 3.50 Male No Sun Dinner 3

#> 4 23.68 3.31 Male No Sun Dinner 2

#> 5 24.59 3.61 Female No Sun Dinner 4

#> 6 25.29 4.71 Male No Sun Dinner 4

|  |
| --- |
| > str(tips)  'data.frame': 244 obs. of 7 variables:  $ total\_bill: num 17 10.3 21 23.7 24.6 ...  $ tip : num 1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...  $ sex : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...  $ smoker : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  $ day : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...  $ time : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...  $ size : int 2 3 3 2 4 4 2 4 2 2 ...  Using the str function in R we found out the structure of the tips data. It is a data frame of 244  objects and 7 variables: total\_bill, tip, sex, smoker, day, time and size. Total\_bill and tip are  numeric, size is integer and the rest are factors: sex, smoker, and time with 2 levels and day with  4 levels. |
|  |
| |  | | --- | |  | |

> summary(tips)

total\_bill tip sex smoker day time

Min. : 3.07 Min. : 1.000 Female: 87 No :151 Fri :19 Dinner:176

1st Qu.:13.35 1st Qu: 2.000 Male :157 Yes: 93 Sat :87 Lunch : 68

Median :17.80 Median : 2.900 Sun :76

Mean :19.79 Mean : 2.998 Thur:62

3rd Qu.:24.13 3rd Qu.: 3.562

Max. :50.81 Max. :10.000

size

Min. :1.00

1st Qu.:2.00

Median :2.00

Mean :2.57

3rd Qu.:3.00

Max. :6.00

From the above data we derived our sample data to put into a new data frame: dat

We performed our visual analysis on it using the R Graphics ggplot2 package.

dat <- data.frame(

time = factor(c("Lunch","Dinner"), levels=c("Lunch","Dinner")),

total\_bill = c(14.89, 17.23)

)

dat

#> time total\_bill

#> 1 Lunch 14.89

#> 2 Dinner 17.23

# Load the ggplot2 package

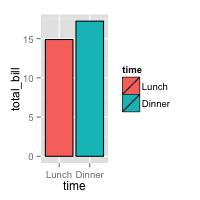
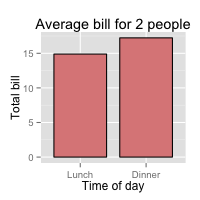
library(ggplot2)

|  |
| --- |
| > str(dat)  'data.frame': 2 obs. of 2 variables:  $ time : Factor w/ 2 levels "Lunch","Dinner": 1 2  $ total\_bill: num 14.9 17.2 |
|  |
| |  | | --- | |  | |

The dat data frame consists of 2 objects with 2 variables: time and total\_bill. Time is factor with 2 levels: lunch and dinner and total\_bill is numeric for the average spend of 2 people at lunch and dinner.

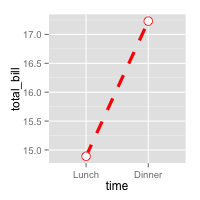
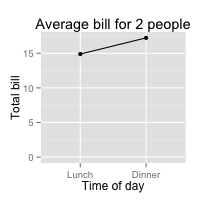
By plotting bar charts of values and line graphs of this data set we could visualize the relationship between these 2 variables clearly and get good insight into our data set.

First we did bar charts of value of the data frame: dat. Bar charts are a great visualization tool as the height of the bar will represent the value in a column of the data frame, in this case the total\_bill. We plotted time in the x axis against total\_bill in the y axis and used colours, black outline, legend, labels and title to aid our visualization.

From it we gleamed that there is a difference in the average bill of 2 people of 1.3 dollars from lunch to dinner as there is a difference in the height of the bars corresponding to their values. Lunch average spend or total bill of 14.9 dollars is less than dinner average spend or total bill 0f 17.2 dollars. So people spend less money on average eating out at lunch than dinner. The difference between the total\_bills of lunch and dinner of 1.3 dollars is the difference between the height of the bars. This is not surprising as restaurants tend to offer better value or deals at lunch time which results in lower total bills rather than at dinner time when usually you pay a la carte, thus more expensive menu, therefore higher total bills are incurred at dinner time.

Second we did line charts of this data set. Line chars are great to visualize changes, differences and trends between the variables. For line graphs, the data points must be grouped so that it knows which points to connect. In this case, all points should be connected, so group=1. Once again we plotted time in the x axis against total\_bill in the y axis and used colours, points, labels and title to aid our visualization.

From these line charts we clearly can visualize the difference between our variables as the single line with points at its ends indicates with its ascending direction that average bill for 2 people increases from lunch to dinner. People on average spend less at lunch around 14.9 dollars and more at dinner around 17.2 dollars. These values are clearly denoted by the points at the end of our line bar. So with this line chart we can clearly visualize a trend of average spend increasing from lunch to dinner. This is again not surprising to us as explained before but this is another great way to visualize more clearly this fact.

From our qplots quickplots: boxplot and histogram of our total\_bill variable in our tips data set we gleamed some useful information about its shape, span and presence of outliers. Our boxplot showed that there are outliers in our data set that need further investigation or consideration when performing further analysis so as not to affect the findings. The presence of outliers tell us to use the median rather than the mean as the measure of the average of our data set as it is a truer representation of the data. The shape of the graph in our histogram also points out to this fact.

Histograms can tell us about the shape, span, and presence of outliers ie extreme low or high values that may need further investigating. The shape of our histogram graph is not symmetric, it is right skewed so most values are to the left side of the graph, so the mean is greater than the median. So as before we should use the median rather than mean as the measure of central tendency. The histogram plot also showed us the range of our data set from 3.07 to 50.81 dollars and that most of our total\_bills are between the range of 10 to 25 dollars ie around the average, which accounts for the bell shape of the graph and then again mainly placed between 10 and 19 dollars ie below the average, hence the right skew shape of the graph. And we see 7 extreme high total\_bill values, above the 41 dollars total\_bill value and 1 extreme low value of 3.07 dollars. Checking against the boxplot we see that the low value is not an outlier but the 7 extreme high values are outliers. Therefore we have 7 outliers in our data set, ie total\_bills above 40 dollars.

summary(tips$total\_bill)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.07 13.35 17.80 19.79 24.13 50.81

The boxplot clearly visualised the above summary information, the fact that the total\_bill range is between 3.07 and 50.81 dollars with an average of 19.79 dollars and a median of 17.8 dollars. It also pointed out to a right skew shape like the histogram. There are 7 outliers above the 40 dollar total\_bill and that’s clearly shown in both the boxplot and histogram graphs. From these two graphs we gleamed that customers in general spent as little as 3.07 dollars and as much as 50.81 dollars but mainly between 10 and 25 dollars which is around average and then again more people between 10 and 19 dollars which comes below average with 7 outlier bills ie way above average bills, with total\_bills above 41 dollars.

Therefore more people paid total\_bills below average but overall most people paid around average bills with just 7 cases paying way above average in our tips data set.

* what other ideas/concepts could be represented via R Graphics if you had more time

In our example of using R Graphics we represented one variable, the time variable, against the total\_bill variable, in our Bar graphs of value and Line graphs. This was our initial primary research of the dataset: tips. This allowed to us to visualize the relationship between time of the meals: lunch and dinner and their corresponding total\_bill.

dat <- data.frame(

time = factor(c("Lunch","Dinner"), levels=c("Lunch","Dinner")),

total\_bill = c(14.89, 17.23)

)

dat

#> time total\_bill

#> 1 Lunch 14.89

#> 2 Dinner 17.23

If we wanted to conduct further analysis, we could add another variable to our data frame and conduct secondary research. We could do Bar graphs and Line Graphs with more variables and get further insights into our data. For example we could add the sex variable to the time variable and total\_bill variable and create a new data frame: dat1. We would be able to visualize the same information as before for each sex now, female and male, thus conducting secondary analysis of the tips data set.

dat1 <- data.frame(

sex = factor(c("Female","Female","Male","Male")),

time = factor(c("Lunch","Dinner","Lunch","Dinner"), levels=c("Lunch","Dinner")),

total\_bill = c(13.53, 16.81, 16.24, 17.42)

)

dat1

#> sex time total\_bill

#> 1 Female Lunch 13.53

#> 2 Female Dinner 16.81

#> 3 Male Lunch 16.24

#> 4 Male Dinner 17.42

We could plot new Bar graphs of values of this new data frame dat1.

These are the variable mappings we could use:

* time: x-axis
* sex: color fill
* total\_bill: y-axis.

This will give us a bar visualization of the total\_bill of females and males for each time of the meals, lunch and dinner which will allows to quickly make comparisons of their values based on the height of the bars.

We can change which variable is mapped to the x-axis and which is mapped to the fill.

* sex: x-axis
* time: colour fill
* total\_bill: y-axis

This will give us bars of the total\_bill of lunch and dinner for each sex: female and male to make further comparisons based on their height.

We could also plot new Line graphs of this new data frame dat1.

These are the variable mappings we could use:

* time: x-axis
* sex: line colour
* total\_bill: y-axis.

To draw multiple lines, the points must be grouped by a variable; otherwise all points will be connected by a single line. In this case, we could group them by sex.

* sex: group

This kind of visualization by using line graphs is more insightful to track trends or differences of payments from lunch to dinner by females and males. The nature of the graph aids by utilising a line to join the total\_bill values from lunch to dinner and the line colour to differentiate male from female.

We can change which variable is mapped the x-axis and which is mapped to the colour or shape as well.

* sex: x-axis
* time: line colour
* total\_bill: y-axis
* time: group
* time: shape

This will give us the trend of the total\_bill change from females to males for each time: lunch and dinner.

We could further explore colours, shapes and line types etc provided by R Graphics and in particular ggplot2 to further increase the impact of our visualizations.

##### We could also plot Bar graphs of counts instead of values like we did before. We could choose the variable day from the tips dataset in the reshape2 library and plot them against their count number. So day will be in the x-axis and y will be their count. In this case the height of the bar will represent the count of cases. This is done by using stat="bin" which is the default. To get a bar graph of counts, don’t map a variable to y, just use stat="bin" instead of stat="identity".

##### We could also choose to plot graphs where the variable on the x-axis is numeric. It is sometimes useful to treat that variable as continuous, and sometimes useful to treat it as categorical. A simple graph might put the variable tips from the tips data set on the x-axis as a numeric value. It is possible to make a line graph this way, but not a bar graph. If we wish to treat the tips variable as a categorical variable instead of a numeric one, it must be converted to a factor. This can be done by modifying the data frame, or by changing the specification of the graph. We could also choose a categorical variable from our tips data frame like size and plot against total\_bill and visualize the relationship between them. We could plot bar graphs when the variable is treated as categorical rather than numeric.

We could also plot our tips data set on other kinds of graphs in the ggplot2 package that could give you further visualizations of our data and more insights into it. We could use geom\_scatterplot to get scatterplots that further explore relationships between our variables and find correlations between them. We could use geom\_boxplot and geom\_histogram to represent our data in boxplots and histograms respectively. Boxplots will help identify outliers in our data that we need to take into account in our analysis ie investigate them even further or remove them so they don’t affect our findings. The histogram differs from a bar chart in that it is the area of the bar that denotes the value, not the height. Histograms help to visualise the shape or distribution of our data, if it is symmetric or not.

If the left side of a histogram resembles a mirror image of the right side, then the data are said to be symmetric. In this case, the mean (or average) is a good approximation for the centre of the data. And we can therefore safely utilize statistical tools that use the mean to analyse our data, such as t-tests.

If the data are not symmetric, then the data are either left-skewed or right-skewed. If the data are skewed, then the mean may not provide a good estimate for the centre of the data and represent where most of the data fall. In this case, we should consider using the median to evaluate the centre of the data, rather than the mean. If the data are left-skewed, then the mean is typically less than the median. If the data are right-skewed, then the mean is typically greater than the median. Histograms also give us information about the spread of our data and the presence of outliers.

All these graphs will greatly aid in our visualization of the data set and getting further useful insights into it.

**References:**

R Graphics Cookbook - <http://it-ebooks.info/book/1316/>

<http://www.cookbook-r.com/Graphs/>Bar\_and\_line\_graphs

Try R course from Code School - <http://tryr.codeschool.com/>