ggplot2 Lab Exercise - 1

# INSTRUCTIONS

Submit a word document containing the *question number, question text, R code, and the graphic produced (if applicable)* for each question, before moving on to the next question. It is important that you produce an easily readable Word document. These questions are based on the ggplot2 code presentation that we are studying in class. You should be able to find the answer by studying the PowerPoint (in PDF form) in most cases. Other alternative is to refer to the ggplot2 documentation available from [Official ggplot2 documentation](http://docs.ggplot2.org/current/)

# QUESTIONS

1. Read the satisfaction.csv file into R. What does head=TRUE mean while reading the CSV file? (R code and text answer)

1.1．R Code:

>setwd(C:\Users\lenovo\Desktop\NKU\2 Year\3rd semester\MBI664 Data Visualization\Assignment 4)

>getwd()

>w <- read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

1.2. Output:

>setwd("C:/Users/lenovo/Desktop/NKU/2 Year/3rd semester/MBI664 Data Visualization/Assignment 4")

>getwd()

[1] "C:/Users/lenovo/Desktop/NKU/2 Year/3rd semester/MBI664 Data Visualization/Assignment 4"

> w <- read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

1.3. Text Answer:

A logical value indicating whether the file contains the names of the variables as its first line. If missing, the value is determined from the file format: Header is set to True if and only if the first row contains one fewer filed than the number of columns.

1. Display the first few lines of the dataset. (R code + Output)

2.1. R code:

colnames(w)

2.2．Output:

[1] "ID" "Country" "Continent" "LE" "WellBeing"

[6] "Footprint" "HPI"

1. Display the column names of the dataset. (R code + Output)

3.1. R code:

w <- read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

head(w)

3.2. Console Tab:

ID Country Continent LE WellBeing Footprint HPI

1 109 Afghanistan Asia 48.7 4.8 0.5 36.8

2 18 Albania Europe 76.9 5.3 1.8 54.1

3 26 Algeria Africa 73.1 5.2 1.6 52.2

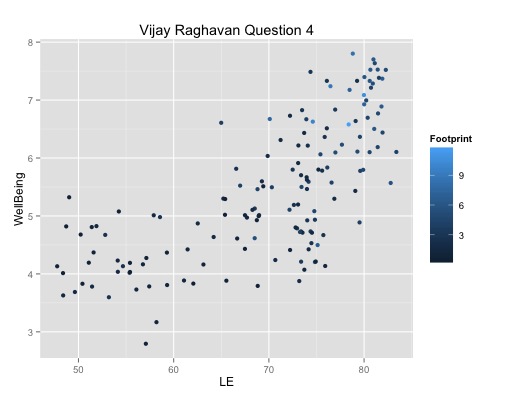
4 127 Angola Africa 51.1 4.2 0.9 33.2

5 17 Argentina South America 75.9 6.4 2.7 54.1

6 53 Armenia Europe 74.2 4.4 1.7 46.0

1. Write the R code necessary to **produce the following plot using the ggplot2 library.**

Note the Title in the graph for this question and all other question requiring a graph your title should include your name and the question that you are answering. (R code + graph) *All future graphs must include a title in similar format although it is NOT indicated in the graphs shown here.*



**4.1. R code:**

install.packages(“ggplot2”)

library (“ggplot2”)

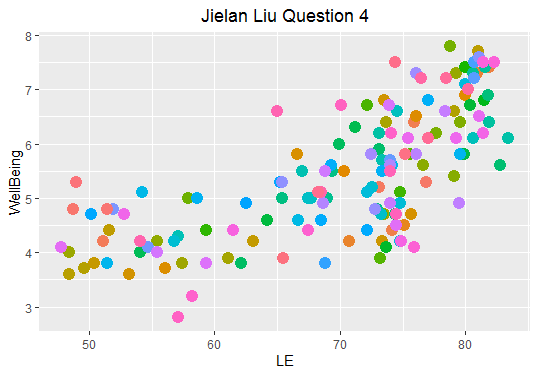
w<-read.csv(file= "Satisfaction.csv", head=TRUE,sep=",")

p<-ggplot(data=w,aes(x=LE,y=WellBeing,color=Country))

p + geom\_point(size=4,show.legend=FALSE)

print(p+geom\_point(size=4, show.legend=FALSE) + ggtitle("Jielan Liu Question 4") + theme(plot.title = element\_text(hjust=0.5)))

4.2. Plots Pane



1. Write a paragraph (at least four sentences) on your interpretation of the above plot. Use the correlation analysis discussion from Stephen Few, if you think it is appropriate .(Text)

**Text: Correlation analysis**

**WellBeing and LE are in positive direction. The correlation is weak as there are many scattered points around this trend. The shape is more closely to nonlinear monotonic, which looks like positive exponential correlation. There are two concentrations in this scatter plot. One is range from 70 to 75 in X axis, another is range from 80 to 85 in X axis.**

**Overall, the scatter plot correlation could be improved by change the scale of the axis to have stronger correlation, or group them by similarity based on the variable character.**

1. Jitter the above plot by adding a **geom layer** to your ggplot2 object. (R code + graph)

6.1. R Code:

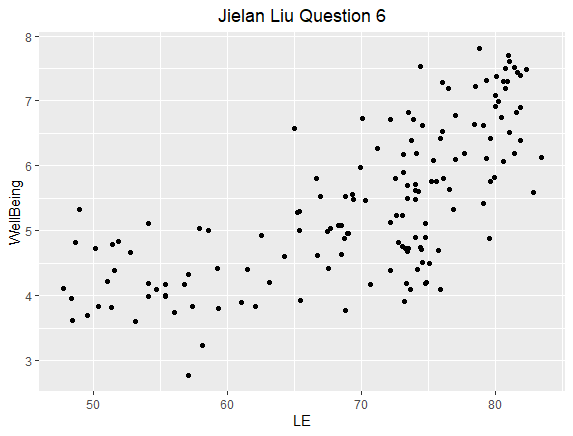
w <- read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

p<-ggplot(data=w,aes(x=LE,y=WellBeing,color=Country))

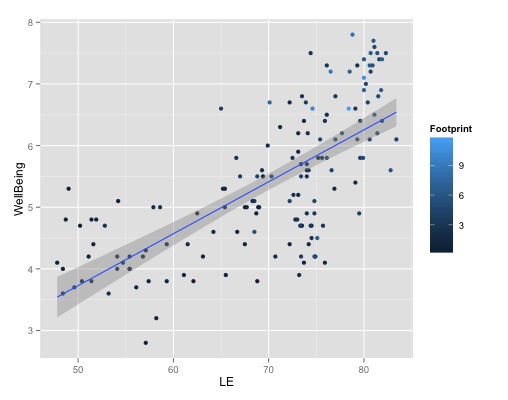
p+geom\_jitter()

print(p + geom\_jitter()+ggtitle("Jielan Liu Question 6")+theme(plot.title=element\_text(hjust=0.5)))

6.2. Graph:



1. Now write the code necessary using the ggplot2 library to produce the following plot. (R code + graph)



7.1. R code:

w<-read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

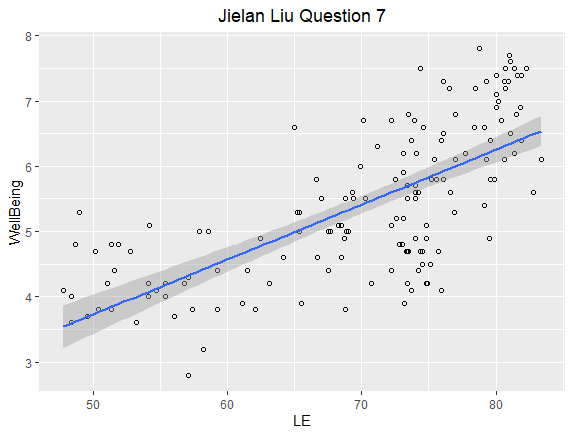
p<-ggplot(data=w,aes(x=LE,y=WellBeing))

p+geom\_point(shape=1) + geom\_smooth()

p+geom\_point(shape=1) +geom\_smooth(method="lm")

print(p+geom\_point(shape=1) +geom\_smooth(method="lm")+ggtitle("Jielan Liu Question 7")+theme(plot.title=element\_text(hjust=0.5)))

7.2. Graph:



1. What does the line represent in the above graph? What does a residual mean with reference to the above graph? What does the band around the line indicate? (Text)

The line represents a trend line which is a line that summarizes a pattern in a scatter plot in one continuous linear equation.

Residual value is a measure of how much a regression line vertically misses a data point. Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots. A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

The band around is the confidence band is the lines on a probability plot or fitted line plot that depict the upper and lower confidence bounds for all points on a fitted line within the range of data.

1. Explain the difference between aesthetic mapping and parameter setting – first using an example by writing R code and then write couple sentences to explain the difference. (code and two graphs)

**9.1. Aesthetic mapping:**

Data value determines visual characteristic. Use aes()

**-R Code:**

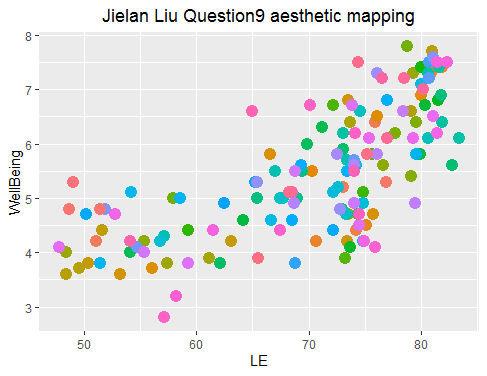
w<-read.csv(file= “Satisfaction.csv”, head=TRUE, sep= “,”)

P<-ggplot(data=w, aes(x=LE,y=WellBeing, color=Country))

p + geom\_point(size=4, show.legend=FALSE)

print(p + geom\_point(size=4, show.legend=FALSE) + ggtitle(“Jielan Liu Question9 Aesthetic Mapping) + theme(plot.title=element.text(hjust=0.5)))

**-Graph:**



**9.2. Parameter setting:**

Constant value determines visual characteristic. Use layer parameter

**-R code:**

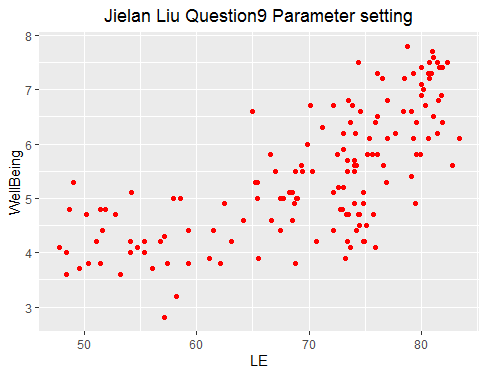
w<-read.csv(file="Satisfaction.csv", head=TRUE, sep= ",")

p<-ggplot(data=w,aes(x=LE,y=WellBeing)

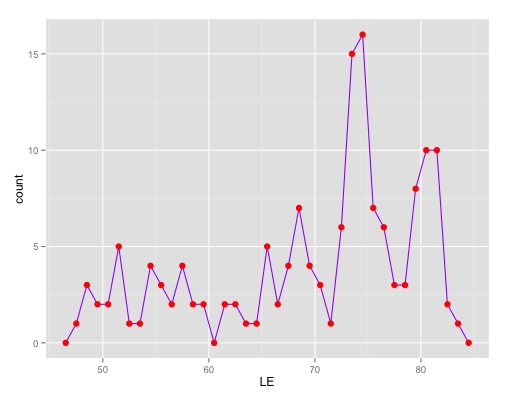
p + geom\_point(color="Red")

print(p+geom\_point(color="red")+ggtitle("Jielan Liu Question9 Parameter setting") + theme(plot.title=element\_text(hjust=0.5)))

**-Graph:**



1. Produce the following plot again using the ggplot2 library. (R code and graph)



10.1. R Code:

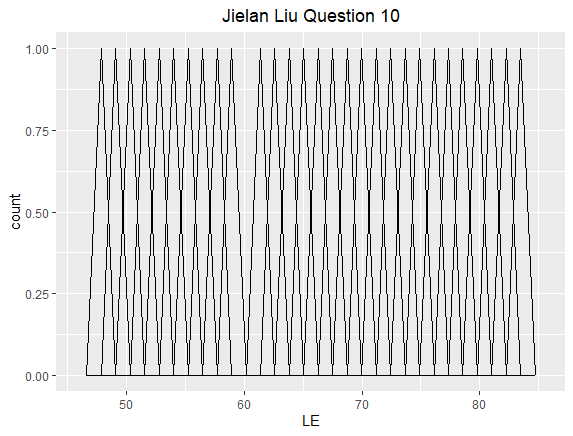
w<-read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

p<-ggplot(data=w,aes(x=LE))

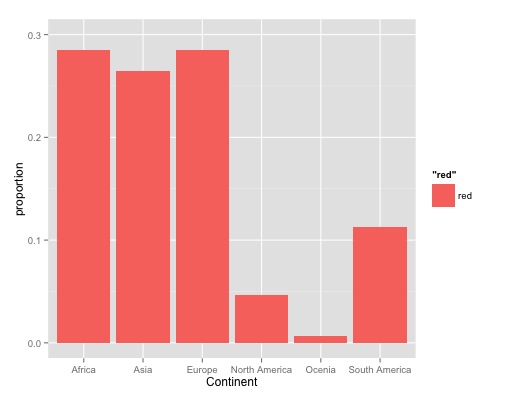
p+geom\_freqpoly(aes(color=Country), width=0.5)

print(p + geom\_freqpoly(aes(color=Country))+ggtitle("Jielan Liu Question 10")+theme(plot.title=element\_text(hjust=0.5)))

10.2. Graph:



1. **Produce the following graph** noting that the y-axis is now **the proportion of** the count of countries in each continent (from our dataset) and the y-axis **scale ranges** from 0.0 to 0.3, and the **fill color is now red**. (R code + graph)



**11.1. R code:**

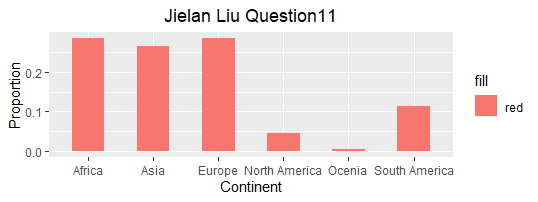
w<-read.csv(file= “Satisfaction.csv”, head=TRUE, sep= “,”)

p<-ggplot(data=w, aes(x = Continent, fill = "red"))

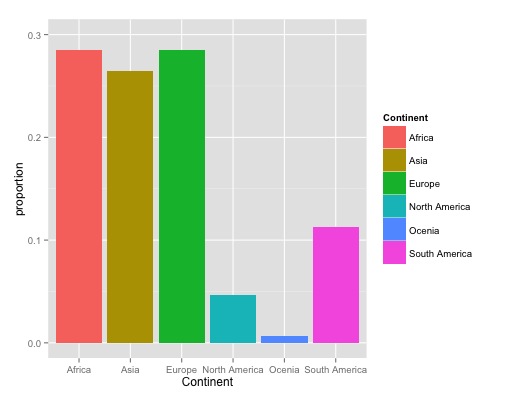
p+ geom\_bar(aes(y = stat(count/sum(count))), width=0.5) +labs(y = "Proportion")

print(p+geom\_bar(aes(y=stat(count/sum(count))),width=0.5)+labs(y="Proportion")+ggtitle("Jielan Liu Question11")+theme(plot.title=element\_text(hjust=0.5)))

**11.2. Graph:**



12) Use parameter mapping to produce the following graph where each continent is colored differently. (R code + graph)



12.1. **R code:**

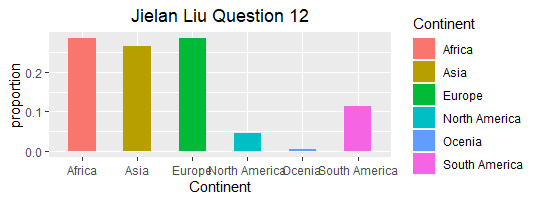
w<-read.csv(file="Satisfaction.csv", head=TRUE, sep=",")

p<-ggplot(data=w, aes(x = Continent, fill= Continent))

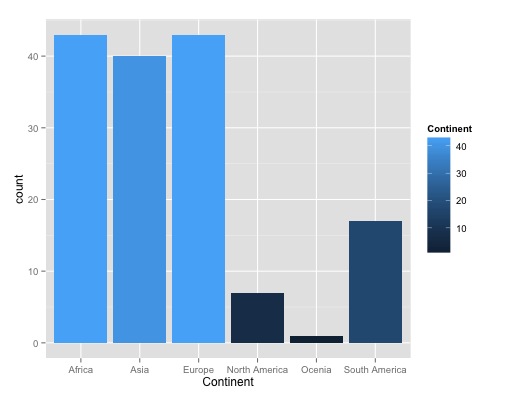
p + geom\_bar(aes(y=stat(count/sum(count))),width=0.5) + labs(y="proportion")

print(p + geom\_bar(aes(y=stat(count/sum(count))),width=0.5) + labs(y="proportion") + ggtitle("Jielan Liu Question 12") + theme(plot.title=element\_text(hjust=0.5)))

**12.2. Graph:**

****

1. Now, change the y-axis back to count and color intensity of the bars to reflect the count and produce the following graph. (R code + graph)



**13.1. R code:**

w<-read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

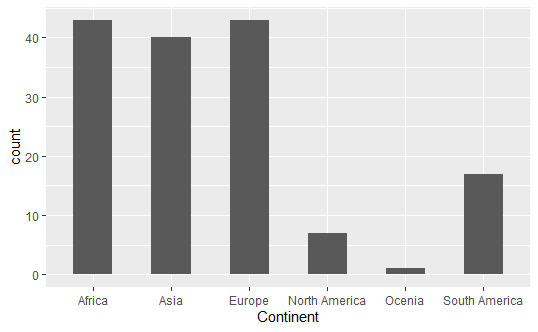
p<-ggplot(data=w, aes(x=Continent, fill=Continent))

p+geom\_bar(color= “blue”)

p+geom\_bar(width=.5)

print(p + geom\_bar(width=.5)+ggtitle(“Jielan Liu Question 13”) + theme(plot.title=element\_text(hjust=0.5))

13.2. R Graph:



1. Define a new variable for the **WellBeing index** for values greater than 5. Now, plot a stacked bar graph with Wellbeing >5 and <5 stacked, these bars next to each other, a filled graph that shows the proportions of these variables for each continent as shown below (R code and Graph):

|  |  |  |
| --- | --- | --- |
|  |  |  |

14.1. R code: (Position)

w<-read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

w$WellBeing2<-w$WellBeing>5

p<-ggplot(data=w,aes(x=Continent,fill=WellBeing2))

p+geom\_bar(color="black")+scale\_fill\_manual(values=c("red","blue"),labels=c("TRUE","FALSE"))

p+geom\_bar(position=”stack”)

print(p+geom\_bar(position="stack",color="black")+scale\_fill\_manual(values=c("red","blue"),labels=c("TRUE","FALSE"))+ggtitle("Jielan Liu Question 14")+theme(plot.title=element\_text(hjust=0.5)))

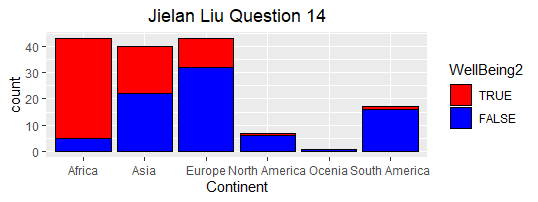
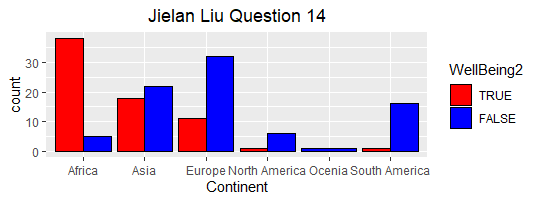
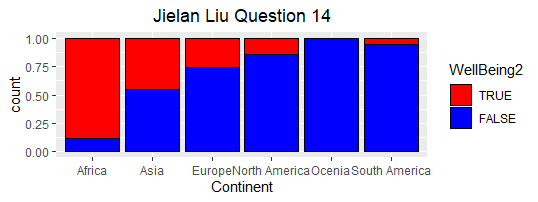
p+geom\_bar(position=”dodge”)

print(p+geom\_bar(position="dodge",color="black")+scale\_fill\_manual(values=c("red","blue"),labels=c("TRUE","FALSE"))+ggtitle("Jielan Liu Question 14")+theme(plot.title=element\_text(hjust=0.5)))

p+geom\_bar(position=”fill”)

print(p+geom\_bar(position="fill",color="black")+scale\_fill\_manual(values=c("red","blue"),labels=c("TRUE","FALSE"))+ggtitle("Jielan Liu Question 14")+theme(plot.title=element\_text(hjust=0.5)))

14.2. Graph

1. Produce the first graph shown above with NKU colors instead of as shown above (Black and Gold) R code and graphic object.
   1. R code

w<-read.csv(file="Satisfaction.csv",head=TRUE,sep=",")

w$WellBeing2<-w$WellBeing>5

p<-ggplot(data=w,aes(x=Continent,fill=WellBeing2))

p+geom\_bar(position="stack",color="black")+scale\_fill\_manual(values=c("black","gold"),labels=c("TRUE","FALSE"))

print(p+geom\_bar(position="stack",color="black")+scale\_fill\_manual(values=c("black","gold"),labels=c("TRUE","FALSE"))+ggtitle("Jielan Liu Question 15")+theme(plot.title=element\_text(hjust=0.5)))

* 1. Graph

