Question2

LI KONG

6/9/2020

## 1. Begin your markdown with an introduction explaining what is data visualization and its use.(3 marks)

### 1.1 What is data visualization?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

### 1.2 The advantages and benefits of good data visualization

Our culture is visual, including everything from art and advertisements to TV and movies.Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly.

Effective data visualization is a delicate balancing act between form and function. The plainest graph could be too boring to catch any notice or it make tell a powerful point; the most stunning visualization could utterly fail at conveying the right message or it could speak volumes. The data and the visuals need to work together, and there’s an art to combining great analysis with great storytelling.

### 1.3 Why data visualization is important?

It’s hard to think of a professional industry that doesn’t benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

According to the World Economic Forum, the world generates 25,000 megabytes of data every day, and 90% of the data is currently generated in the past two years. With so much data, management and understanding become more difficult. It is impossible for anyone to browse the data line by line to understand different patterns and draw observations. Data growth can be managed as part of the data science process, and data visualization is part of the process.

### 1.4 5 Types of Big Data Visualization Categories

**Temporal**: Data visualizations belong in the temporal category if they satisfy two conditions: that they are linear, and that they are one-dimensional. Temporal visualizations normally feature lines that either stand alone or overlap with each other, with a start and finish time.

* Examples of temporal data visualization include:
  + Scatter plots
  + Polar area diagrams
  + Time series sequences
  + Timelines
  + Line graphs

**Hierarchical**: Data visualizations that belong in the hierarchical category are those that order groups within larger groups. Hierarchical visualizations are best suited if you’re looking to display clusters of information, especially if they flow from a single origin point.

* Examples of hierarchical data visualizations include:
  + Tree diagrams
  + Ring charts
  + Sunburst diagrams

**Network**: Datasets connect deeply with other datasets. Network data visualizations show how they relate to one another within a network. In other words, demonstrating relationships between datasets without wordy explanations.

* Examples of network data visualizations include:
  + Matrix charts
  + Node-link diagrams
  + Word clouds
  + Alluvial diagrams

**Multidimensional**: Just like the name, multidimensional data visualizations have multiple dimensions. This means that there are always 2 or more variables in the mix to create a 3D data visualization. Because of the many concurrent layers and datasets, these types of visualizations tend to be the most vibrant or eye-catching visuals. Another plus? These visuals can break down a ton of data down to key takeaways.

* Examples of multidimensional data visualizations include:
  + Scatter plots
  + Pie charts
  + Venn diagrams
  + Stacked bar graphs
  + Histograms

**Geospatial**: Geospatial or spatial data visualizations relate to real life physical locations, overlaying familiar maps with different data points. These types of data visualizations are commonly used to display sales or acquisitions over time, and can be most recognizable for their use in political campaigns or to display market penetration in multinational corporations.

* Examples of geospatial data visualizations include:
  + Flow map
  + Density map
  + Cartogram
  + Heat map

## 2. Use the same dataset for all graphs. Provide some explanation regarding the dataset.(3 marks)

### 2.1 Diamonds

Analyze diamonds by their cut, color, clarity, price, and other attributes.

### 2.2 Context

This classic dataset contains the prices and other attributes of almost 54,000 diamonds. It’s a great dataset for beginners learning to work with data analysis and visualization.

### 2.3 Content

* carat: weight of the diamond (0.2–5.01)
* cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
* color: diamond colour, from J (worst) to D (best)
* clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
* depth: total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43–79)
* table: width of top of diamond relative to widest point (43–95)
* price: price in US dollars ($326–$18,823)
* x: length in mm (0–10.74)
* y: width in mm (0–58.9)
* z: depth in mm (0–31.8)

We can answer how many observations (53,940), how many variables (only 10), and how many of those variables are ordered factor (only 3).

library(ggplot2)  
data(diamonds)  
summary(diamonds)

## carat cut color clarity depth   
## Min. :0.2000 Fair : 1610 D: 6775 SI1 :13065 Min. :43.00   
## 1st Qu.:0.4000 Good : 4906 E: 9797 VS2 :12258 1st Qu.:61.00   
## Median :0.7000 Very Good:12082 F: 9542 SI2 : 9194 Median :61.80   
## Mean :0.7979 Premium :13791 G:11292 VS1 : 8171 Mean :61.75   
## 3rd Qu.:1.0400 Ideal :21551 H: 8304 VVS2 : 5066 3rd Qu.:62.50   
## Max. :5.0100 I: 5422 VVS1 : 3655 Max. :79.00   
## J: 2808 (Other): 2531   
## table price x y   
## Min. :43.00 Min. : 326 Min. : 0.000 Min. : 0.000   
## 1st Qu.:56.00 1st Qu.: 950 1st Qu.: 4.710 1st Qu.: 4.720   
## Median :57.00 Median : 2401 Median : 5.700 Median : 5.710   
## Mean :57.46 Mean : 3933 Mean : 5.731 Mean : 5.735   
## 3rd Qu.:59.00 3rd Qu.: 5324 3rd Qu.: 6.540 3rd Qu.: 6.540   
## Max. :95.00 Max. :18823 Max. :10.740 Max. :58.900   
##   
## z   
## Min. : 0.000   
## 1st Qu.: 2.910   
## Median : 3.530   
## Mean : 3.539   
## 3rd Qu.: 4.040   
## Max. :31.800   
##

dim(diamonds)

## [1] 53940 10

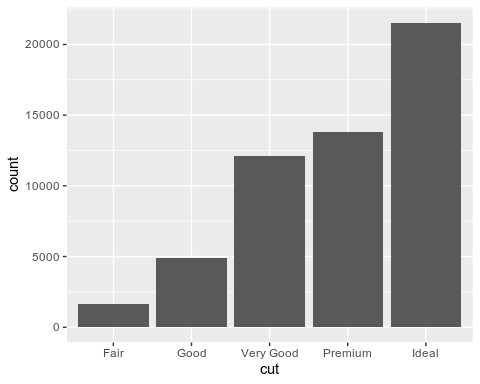
## 3. Write R codes to produce a scatterplot, a histogram, and a boxplot. Explain the use of each graph and provide a short explanation for each produced graph.(9marks)

### 3.1 bar

**Explain usage**: The method of visualizing the distribution of variables depends on whether the variables are categorical or continuous. If the value is only in a small set, then this variable is a categorical variable. Categorical variables are usually saved as factors or character vectors in R. To check the distribution of categorical variables, you can use a bar chart.

**Note**: The height of the bar indicates the number of observations in each x value.

ggplot(data = diamonds) +  
geom\_bar(mapping = aes(x = cut))

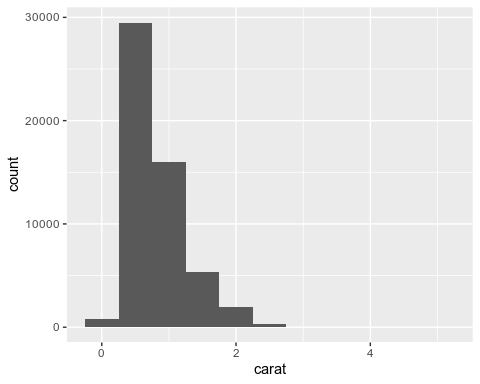


### 3.2 histogram

**Explain usage**: If it can be arbitrarily valued in an infinite ordered set, then this variable is a continuous variable. Numeric and date-time variables are two examples of continuous variables. To check the distribution of continuous variables, you can use a histogram.

**Note**: The histogram performs equal-width binning on the x-axis, and then uses the height of the bar to represent the number of observations that fall into each binning. In the figure below, the highest bar indicates that almost 30 000 observations have carat values between 0.25 and 0.75, which are the left and right values of the bar, respectively.

ggplot(data = diamonds) +  
geom\_histogram(mapping = aes(x = carat), binwidth = 0.5)



### 3.3 boxplot

**Explain usage**: One way to display the distribution of continuous variables by grouping of categorical variables is to use box plots. Box plots are a simple visual representation of the distribution of variable values. Such plots are very popular among statisticians. Each box plot includes the following content.

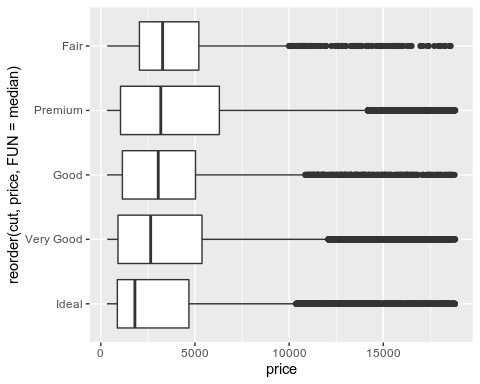
• For a rectangular box, the lower side represents the 25th percentile of the distribution, and the upper side represents the 75th percentile of the distribution. The distance between the upper and lower sides is called the interquartile range. There is a horizontal line in the middle of the box, indicating the median of the distribution, which is the 50th percentile of the distribution. These three lines can indicate the dispersion of the distribution, and can also help us to clarify whether the data is symmetric about the median or is biased to a certain side.

• The dots indicate observations that fall outside the upper and lower sides of the box by 1.5 times the interquartile range. These outliers are outliers, so they need to be plotted separately.

• The straight lines (or whiskers) extending from the top and bottom of the box can reach the farthest non-outlier points in the distribution.

**Note**: We can find counterintuitive phenomena from the box plot: the average price of better quality diamonds is lower!

ggplot(data = diamonds) +  
geom\_boxplot(  
mapping = aes(  
x = reorder(cut, price, FUN = median),  
y = price  
)  
) +  
coord\_flip()



### 3.4 scatterplot

**Explain usage**: For the visual representation of the relative change between two continuous variables, we have introduced a very good method: use geom\_point() to draw a scatterplot. You can think of related changes as a pattern of points.

**Note**: As the size of the data set continues to increase, the usefulness of the scatter plot becomes smaller and smaller, because the data points begin to appear drawn and pile up in a black area. We have already introduced a way to solve this problem, which is to add transparency using the alpha graphics attribute. you can see that there is an exponential relationship between the carat number and value of diamonds:

ggplot(data = diamonds) +  
geom\_point(  
mapping = aes(x = carat, y = price),  
alpha = 1 / 100  
)

## 

## (The template used in this project is original and has been submitted to [github](https://github.com/LK3024629/data_visualization) with permission “private”. After the exam, the project will be made public on June 11. If other privacy is involved, please contact me to delete the project immediately.)