COMPSCIX 415.2 Homework 5 Midterm

Ganesh Saravanan

7/10/2018

My Github repository for all of my assignments can be found at this URL below https://github.com/gsaravanan1/rstudiodemo.git

```
library(mdsr)
library(tidyverse)
library(ggplot2)
library(nycflights13)
library(tibble)
```

Excercises

## The tidyverse packages (3 points)

1. Can you name which package is associated with each task below?

Plotting - ggplot

Data munging/wrangling - dplyr

Reshaping (speading and gathering) data - tidyr

Importing/exporting data - readr

2.Now can you name two functions that you've usedfrom each packagethat you listed above for thesetasks?

Plotting - geom_boxplot(), coord_flip()

Data munging/wrangling - count(), summarise ()

Reshaping (speading and gathering) data - select(),gather()

Importing/exporting data - read_csv(), write_delim()

## R Basics (1.5 points)

1. Fix this codewith the fewest number of changes possible so it works?

```
my_data <- c(1,2,3)
my_data
```

```
## [1] 1 2 3
```

2. Fix this code so it works:

Right code:

```
my_string <- c('has','an','error','in','it')
my_string
```

```
## [1] "has"   "an"    "error" "in"    "it"
```

3. Look at the code below and comment on what happened to the values in the vector.

Vector Type changed to 'Character'

```
my_vector <-c(1, 2,'3','4', 5)
my_vector
```

```
## [1] "1" "2" "3" "4" "5"
```

```
class(my_vector)
```

```
## [1] "character"
```

## Data import/export (3 points)

1.Download the rail_trail.txt file from Canvas (in the Midterm Exam section) and successfully import itinto R. Prove that it was imported successfully by including your import code and taking aglimpseofthe result.

2.Export the file into a comma-separated file and name it "rail_trail.csv". Make sure you define thepathcorrectly so that you know where it gets saved. Then reload the file. Include your export and importcode and take anotherglimpse.

```
my_data <- read.delim("~/Downloads/rail_trail.txt")
glimpse(my_data)
```

```
## Observations: 90
## Variables: 1
## $ hightemp.lowtemp.avgtemp.spring.summer.fall.cloudcover.precip.volume.weekday <fct> ...
```

```
write.csv(my_data, "~/Downloads/rail_trail.csv")
my_data1 <- read.csv("~/Downloads/rail_trail.csv")
glimpse(my_data1)
```

```
## Observations: 90
## Variables: 2
## $ X                                                                            <int> ...
## $ hightemp.lowtemp.avgtemp.spring.summer.fall.cloudcover.precip.volume.weekday <fct> ...
```

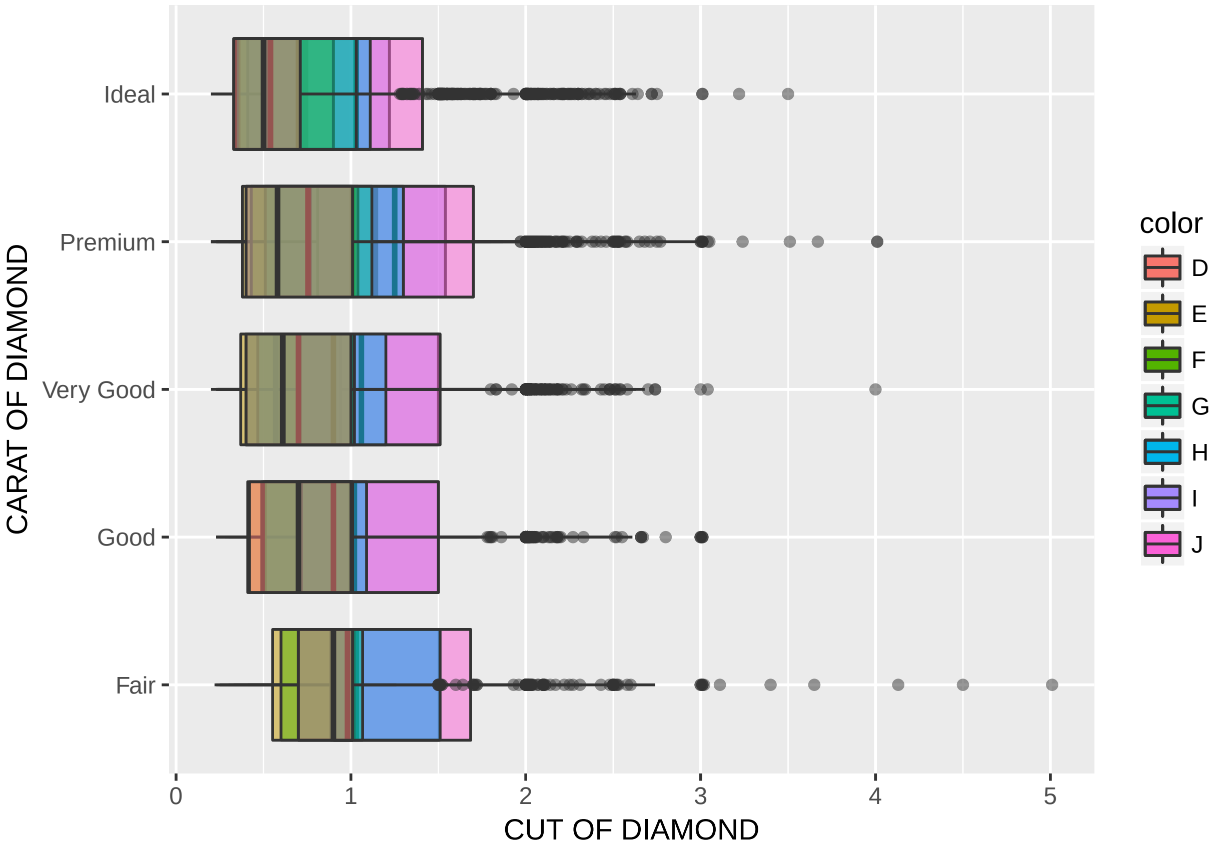
## Visualization (6 points)

1. Critique this graphic: giveonly three examples of what is wrong with this graphic. Be concise.

The chars could have been just one, with 2 legends , one of the legend represent male/female with color coding.The data density of a graph is the proportion of the total size of the graph that is dedicated displaying data.I prefers high data density graphs

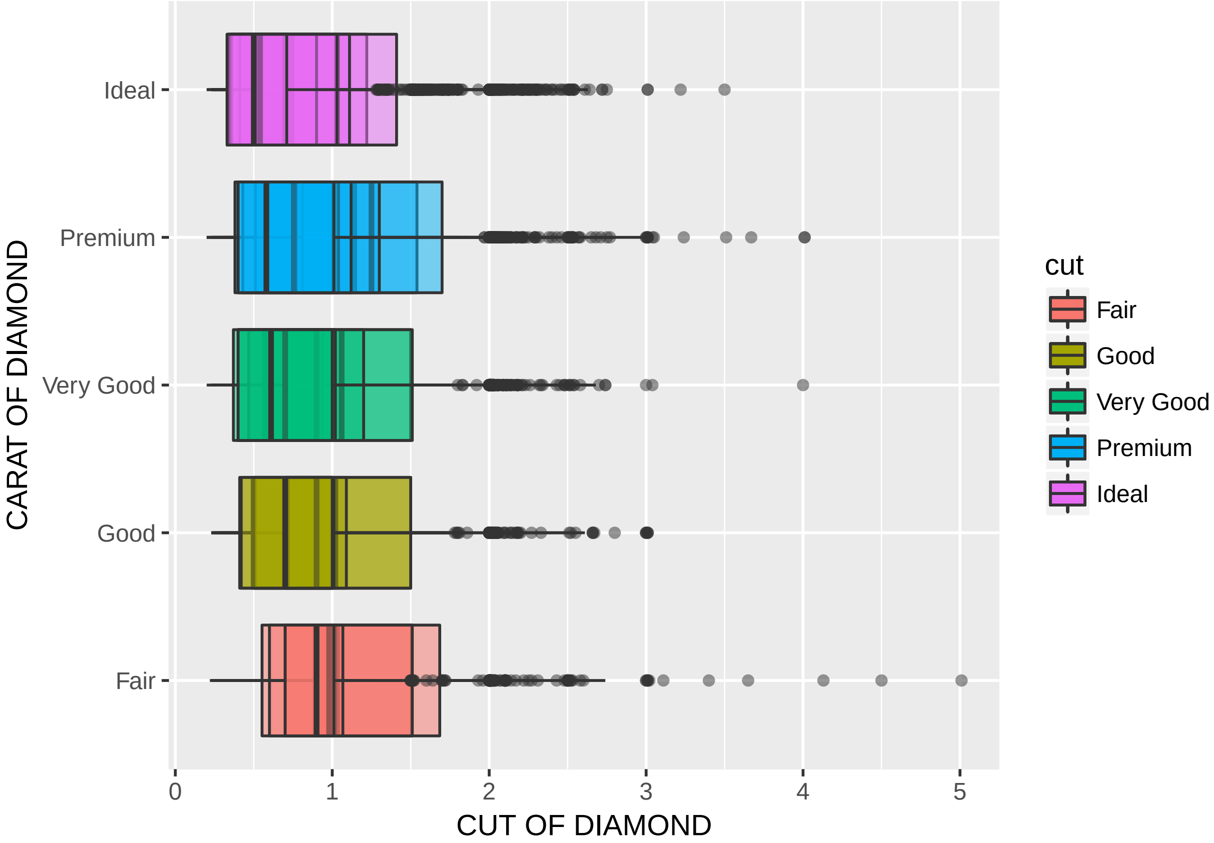
2. Reproduce this graphic using the diamondsdata set.

```r
gg <- ggplot()+ coord_flip()
mycolor <- as.vector(unique(diamonds$color))
for( i in mycolor){
 gg <- gg + geom_boxplot(data= diamonds[diamonds$color==i,],
                         aes(x = cut ,y = carat, fill= color),
                         alpha=0.5)

}
gg + xlab("CARAT OF DIAMOND") + ylab("CUT OF DIAMOND")
```



3.The previous graphic is not very useful. We can make it much more useful by changing one thing aboutit. Make the change and plot it again.

```
gg <- ggplot()+ coord_flip()
mycolor <- as.vector(unique(diamonds$color))
for( i in mycolor){
 gg <- gg + geom_boxplot(data= diamonds[diamonds$color==i,],
                         aes(x = cut ,y = carat, fill= cut),
                         alpha=0.5)

}
gg + xlab("CARAT OF DIAMOND") + ylab("CUT OF DIAMOND")
```



## Data munging and wrangling (6 points)

1.Is this data "tidy"?

May not be. Might be tidy if we order the year column as below:

```
dplyr::arrange(table2, year)
```

```
## # A tibble: 12 x 4
##     country      year type            count
##     <chr>       <int> <chr>           <int>
##   1 Afghanistan  1999 cases             745
##   2 Afghanistan  1999 population   19987071
##  3 Brazil       1999 cases          37737
##  4 Brazil       1999 population  172006362
##  5 China        1999 cases         212258
##  6 China        1999 population 1272915272
##  7 Afghanistan  2000 cases           2666
##  8 Afghanistan  2000 population   20595360
##  9 Brazil       2000 cases          80488
## 10 Brazil       2000 population  174504898
## 11 China        2000 cases         213766
## 12 China        2000 population 1280428583
```

2.Create a new column in the diamonds data set called price_per_carat that shows the price of each diamond per carat (hint: divide). Only show me the code, not the output

```
diamonds_cust <- diamonds
diamonds_cust <- transform(diamonds_cust, price_per_carat = price / carat)
```

3.For each cut of diamond in the diamonds data set, how many diamonds, and what proportion, have a price > 10000 and a carat < 1.5? There are several ways to get to an answer, but your solution must use the data wrangling verbs from the tidyverse in order to get credit.

```
diamonds_cust1 <- diamonds
diamonds_cust1 <- filter(diamonds, price>10000 ,carat <1.5)
diamonds_cust1 %>%
  group_by(cut) %>%
  summarise(counts = n() / nrow(diamonds_cust1))
```

```
## # A tibble: 5 x 2
##   cut        counts
##   <ord>       <dbl>
## 1 Fair      0.00480
## 2 Good      0.0204
## 3 Very Good 0.186
## 4 Premium   0.207
## 5 Ideal     0.582
```

Do the results make sense? Why?

There is a lot of variability in the distribution of carat sizes within each cut category. There is a slight negative relationship between carat and cut. Noticeably, the largest carat diamonds have a cut of “Fair” (the lowest).

Do we need to be wary of any of these numbers? Why?

This negative relationship can be due to the way in which diamonds are selected for sale. A larger diamond can be profitably sold with a lower quality cut, while a smaller diamond requires a better cut.

## EDA (6 points)

1. During what time period is this data from?

2000 - 2015

2. How many cties are represented?

46

3. which city, month and year had the highest number of sales?

city == Houston Year_Month == 201406

4. What kind of relationship do you think exists between the number of listings and the number of sales?

Listings and sales are relative. more the listings, more the sales.sales are much higher in the summer than in the winter.The range of sales varies over multiple orders of magnitude.

5. What proportion of sales is missing for each city?

0.44736842 sales is missing.

6. Looking at only the cities and months with greater than 500 sales:

Are the distributions of the median sales price (column name median), when grouped by city,different? The same? Show your work

The Distribution remains the same

```r
txhousing_cust1 <- filter(txhousing, sales>500 )
tx <- group_by(txhousing_cust1, city)
mean_sl <- mean(tx$sales)
pnorm(105000, mean=mean_sl, sd=150, lower.tail=FALSE)
```

```
## [1] 0
```

Any cities that stand out that you'd want to investigate further?

The biggest city, Houston, averages over ~4000 sales per month; the smallest city, San Marcos, only averages ~20 sales per month.

Why might we want to filter out all cities and months with sales less than 500?

This is to reduce noise in the data.