Pradeep Sahoo Midterm

Pradeep Sahoo

7/5/2018

My Github repository for my assignments can be found at this URL: My Github

```
library(mdsr)
library(tidyverse)
library(tibble)
```

The tidyverse packages

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

a. Plotting - ggplot2
b. Data munging/wrangling - dplyr
c. Reshaping (speading and gathering) data - tidyr
d. Importing/exporting data - readr

2. Now can you name two functions that you've used from each package that you listed above for these tasks?

a. Plotting - ggplot(), geom_boxplot()
b. Data munging/wrangling - summarize(), filter()
c. Reshaping data - gather(), spread(), separate(), unite()
d. Importing/exporting data - read_csv(), read_tsv(), read_delim()

R Basics

```
My_data.name___is.too0OooLong <- c( 1 , 2   , 3 )
My_data.name___is.too0OooLong
```

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

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

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

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

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

my_vector is converted to string.

Data import/export

```
file_path = ('/Users/pradeepsahoo/Downloads/rail_trail.txt')
text_data <- read_delim(file_path, delim = "|")
```

```
## Parsed with column specification:
## cols(
##   hightemp = col_integer(),
##   lowtemp = col_integer(),
##   avgtemp = col_double(),
##   spring = col_integer(),
##   summer = col_integer(),
##   fall = col_integer(),
##   cloudcover = col_double(),
##   precip = col_double(),
##   volume = col_integer(),
##   weekday = col_integer()
## )
```

```
glimpse(text_data)
```

```
## Observations: 90
## Variables: 10
## $ hightemp   <int> 83, 73, 74, 95, 44, 69, 66, 66, 80, 79, 78, 65, 41,...
## $ lowtemp    <int> 50, 49, 52, 61, 52, 54, 39, 38, 55, 45, 55, 48, 49,...
## $ avgtemp    <dbl> 66.5, 61.0, 63.0, 78.0, 48.0, 61.5, 52.5, 52.0, 67....
## $ spring     <int> 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, ...
## $ summer     <int> 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, ...
## $ fall       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, ...
## $ cloudcover <dbl> 7.6, 6.3, 7.5, 2.6, 10.0, 6.6, 2.4, 0.0, 3.8, 4.1, ...
## $ precip     <dbl> 0.00, 0.29, 0.32, 0.00, 0.14, 0.02, 0.00, 0.00, 0.0...
## $ volume     <int> 501, 419, 397, 385, 200, 375, 417, 629, 533, 547, 4...
## $ weekday    <int> 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, ...
```

```
write_csv(text_data,path = '/Users/pradeepsahoo/R-Assignments/R_Project1/rail_trail.csv')
csv_data <- read_csv( '/Users/pradeepsahoo/R-Assignments/R_Project1/rail_trail.csv')
```

```
## Parsed with column specification:
## cols(
##   hightemp = col_integer(),
##   lowtemp = col_integer(),
##   avgtemp = col_double(),
##   spring = col_integer(),
##   summer = col_integer(),
##   fall = col_integer(),
##   cloudcover = col_double(),
##   precip = col_double(),
##   volume = col_integer(),
##   weekday = col_integer()
## )
```

```
glimpse(csv_data)
```

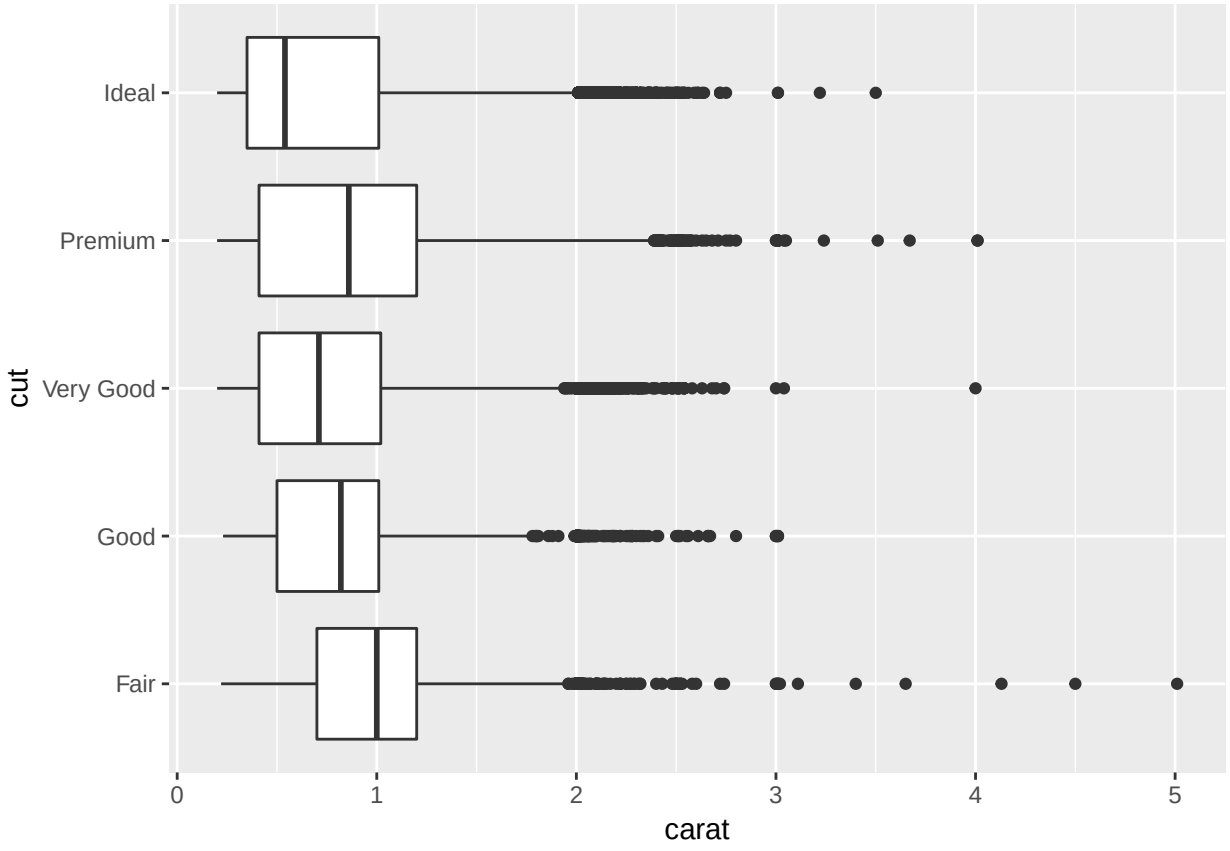
```
## Observations: 90
## Variables: 10
```

```
## $ hightemp   <int> 83, 73, 74, 95, 44, 69, 66, 66, 80, 79, 78, 65, 41,...
## $ lowtemp    <int> 50, 49, 52, 61, 52, 54, 39, 38, 55, 45, 55, 48, 49,...
## $ avgtemp    <dbl> 66.5, 61.0, 63.0, 78.0, 48.0, 61.5, 52.5, 52.0, 67....
## $ spring     <int> 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, ...
## $ summer     <int> 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, ...
## $ fall       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, ...
## $ cloudcover <dbl> 7.6, 6.3, 7.5, 2.6, 10.0, 6.6, 2.4, 0.0, 3.8, 4.1, ...
## $ precip     <dbl> 0.00, 0.29, 0.32, 0.00, 0.14, 0.02, 0.00, 0.00, 0.0...
## $ volume     <int> 501, 419, 397, 385, 200, 375, 417, 629, 533, 547, 4...
## $ weekday    <int> 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, ...
```

Visualization

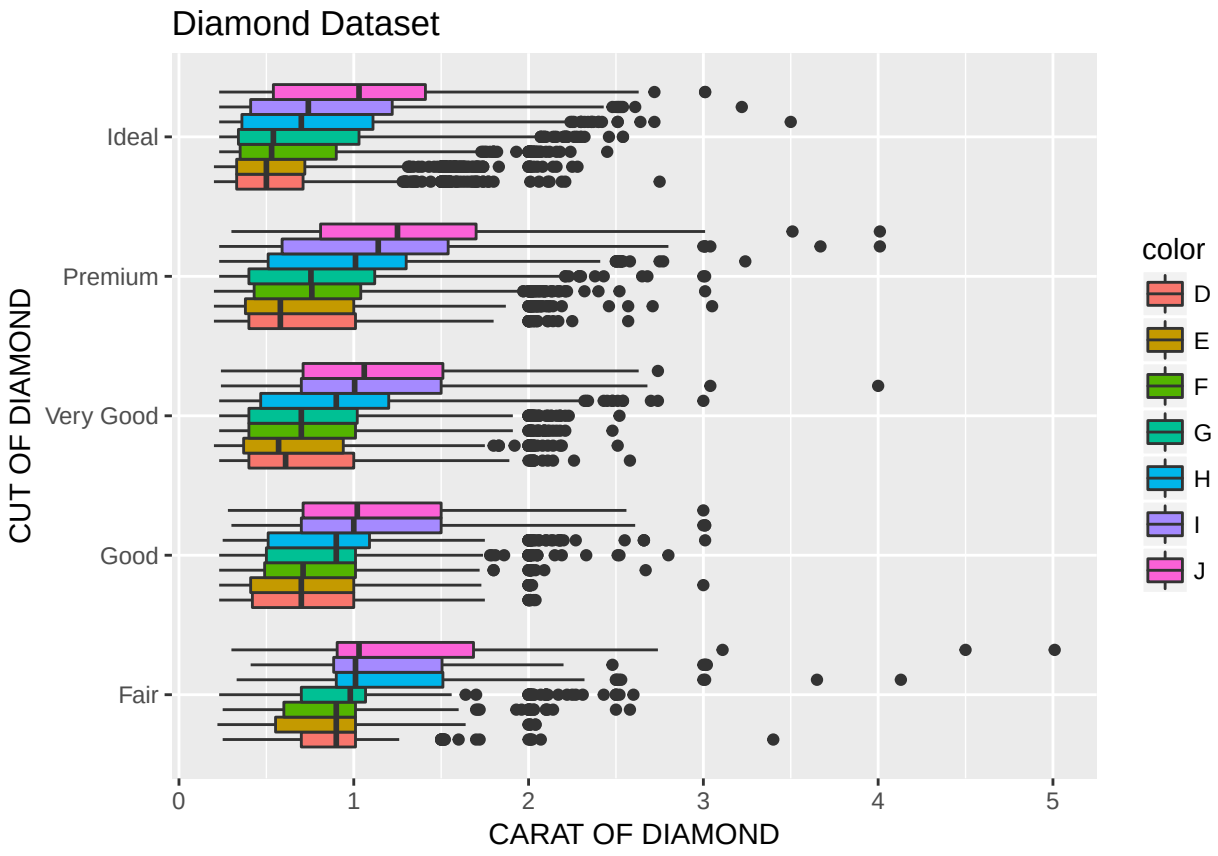
Mrs. President: - There is no %age symbol so it is difficult to under what the number suggest. - Then the sum is not summing up to 100 in one variable/bucket. - There is no signinificance of the color and it is creating confusion.

```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x= cut, y = carat)) + coord_flip()
```

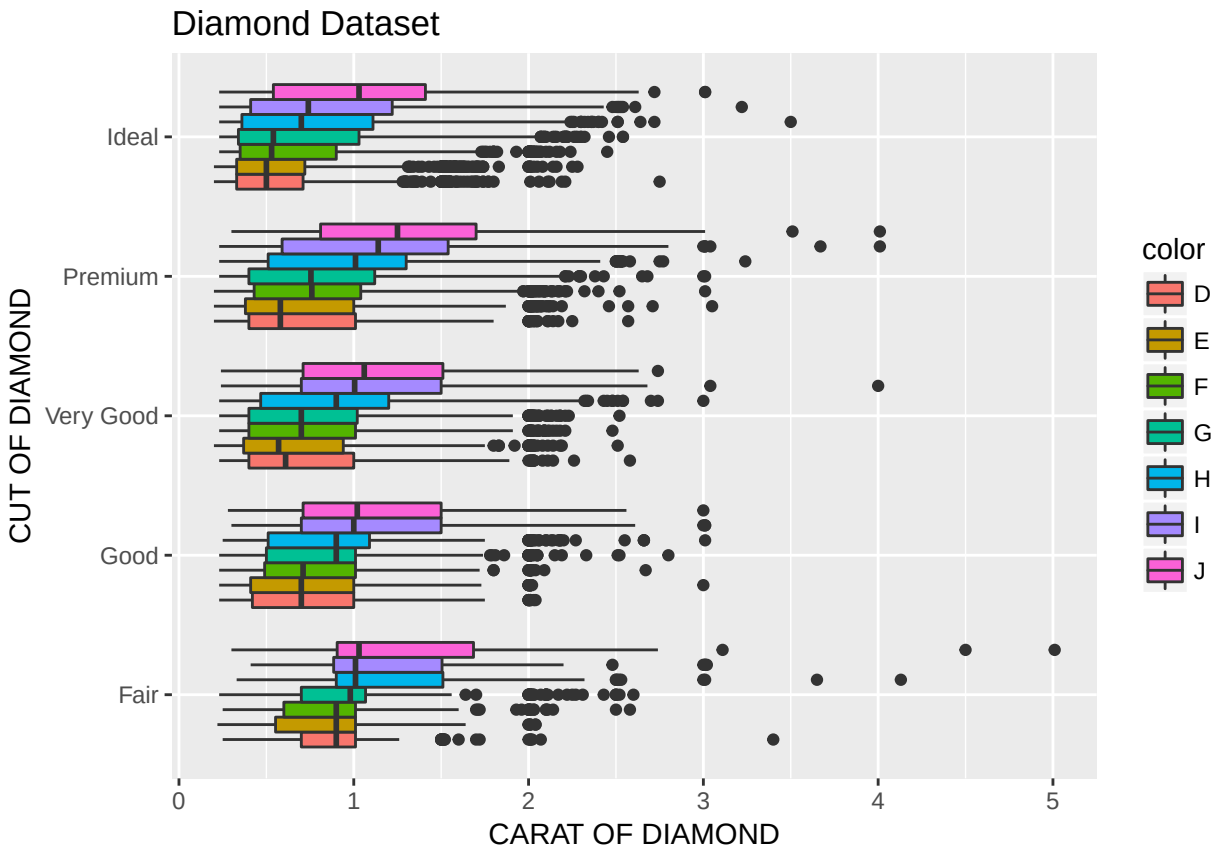



```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = cut,y = carat, fill = color, position = "dodge") ) + ggtitle ("Diamond
```

```
## Warning: Ignoring unknown aesthetics: position
```




```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = cut,y = carat, fill = color)) + ggtitle ("Diamond Dataset") + xlab( "CU
```



Data munging and wrangling

```
glimpse(table2)
```

```
## Observations: 12
## Variables: 4
## $ country <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanis...
## $ year    <int> 1999, 1999, 2000, 2000, 1999, 1999, 2000, 2000, 1999, ...
## $ type    <chr> "cases", "population", "cases", "population", "cases",...
## $ count   <int> 745, 19987071, 2666, 20595360, 37737, 172006362, 80488...
```

```
diamonds1 <- diamonds%>% mutate(price_per_carat = price / carat)
glimpse(diamonds1$price_per_carat)
```

```
##  num [1:53940] 1417 1552 1422 1152 1081 ...
```

```
  by_cut  <- filter(group_by(diamonds,cut), carat < 1.5 & price >10000)
    summarise(by_cut,
              count = n())
```

```
## # A tibble: 5 x 2
##   cut       count
##   <ord>     <int>
## 1 Fair          4
## 2 Good         17
## 3 Very Good   155
```

```
## 4 Premium     173
## 5 Ideal       485
```

• Do the results make sense? Why? The result make sense as more ideal cut diamonds are more expensive though they have less carats. • Do we need to be wary of any of these numbers? Why?

EDA

1. During what time period is this data from?

```
library(ggplot2)
data("txhousing")
glimpse(txhousing)
```

```
## Observations: 8,602
## Variables: 9
## $ city      <chr> "Abilene", "Abilene", "Abilene", "Abilene", "Abilene...
## $ year      <int> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000...
## $ month     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5...
## $ sales     <dbl> 72, 98, 130, 98, 141, 156, 152, 131, 104, 101, 100, ...
## $ volume    <dbl> 5380000, 6505000, 9285000, 9730000, 10590000, 139100...
## $ median    <dbl> 71400, 58700, 58100, 68600, 67300, 66900, 73500, 750...
## $ listings  <dbl> 701, 746, 784, 785, 794, 780, 742, 765, 771, 764, 72...
## $ inventory <dbl> 6.3, 6.6, 6.8, 6.9, 6.8, 6.6, 6.2, 6.4, 6.5, 6.6, 6....
## $ date      <dbl> 2000.000, 2000.083, 2000.167, 2000.250, 2000.333, 20...
```

```
min(txhousing$year)
```

```
## [1] 2000
```

```
max(txhousing[,2])
```

```
## [1] 2015
```

The data is from 2000 to 2015 2. How many cities are represented?

```
count(unique(txhousing[c("city")]))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    46
```

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

```
tx_sales <- group_by(txhousing,city,year,month)
max(tx_sales$sales, na.rm = TRUE)
```

```
## [1] 8945
```

```
        #summarise(tx_sales,max = max(sales))
```

4. What kind of relationship do you think exists between the number of listings and the number of sales? Check your assumption and show your work.

Listing is always greater than Sales.

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

```
tx_sales_city <- group_by(txhousing,city)
tx_sales_city_notna <- filter(tx_sales_city,  sales =='NA')
          summarise(tx_sales_city,count=n())
```

```
## # A tibble: 46 x 2
##    city                  count
##    <chr>                 <int>
##  1 Abilene                 187
##  2 Amarillo                187
##  3 Arlington               187
##  4 Austin                  187
##  5 Bay Area                187
##  6 Beaumont                187
##  7 Brazoria County         187
##  8 Brownsville             187
##  9 Bryan-College Station   187
## 10 Collin County           187
## # ... with 36 more rows
```

```
    summarise(tx_sales_city_notna,count=n())
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: city <chr>, count <int>
```

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.

• Any cities that stand out that you'd want to investigate further? • Why might we want to filter out all cities and months with sales less than 500?