STA404/504 HW5

Pei Wang

08/21/2022

Learning Objectives:

- Data cleaning and manipulation with **dplyr** (%>%)

The items below will be considered for grading:

- The data manipulation is done in **dplyr**
- The plots are correct, with professionalism. Axis labels and titles are correct and complete. Units are clearly labeled.
- Proper grammar in the write-up.
- The discussion and the story told is interesting and appropriate.

Please use dplyr to work on any data manipulation related tasks.

Question 1

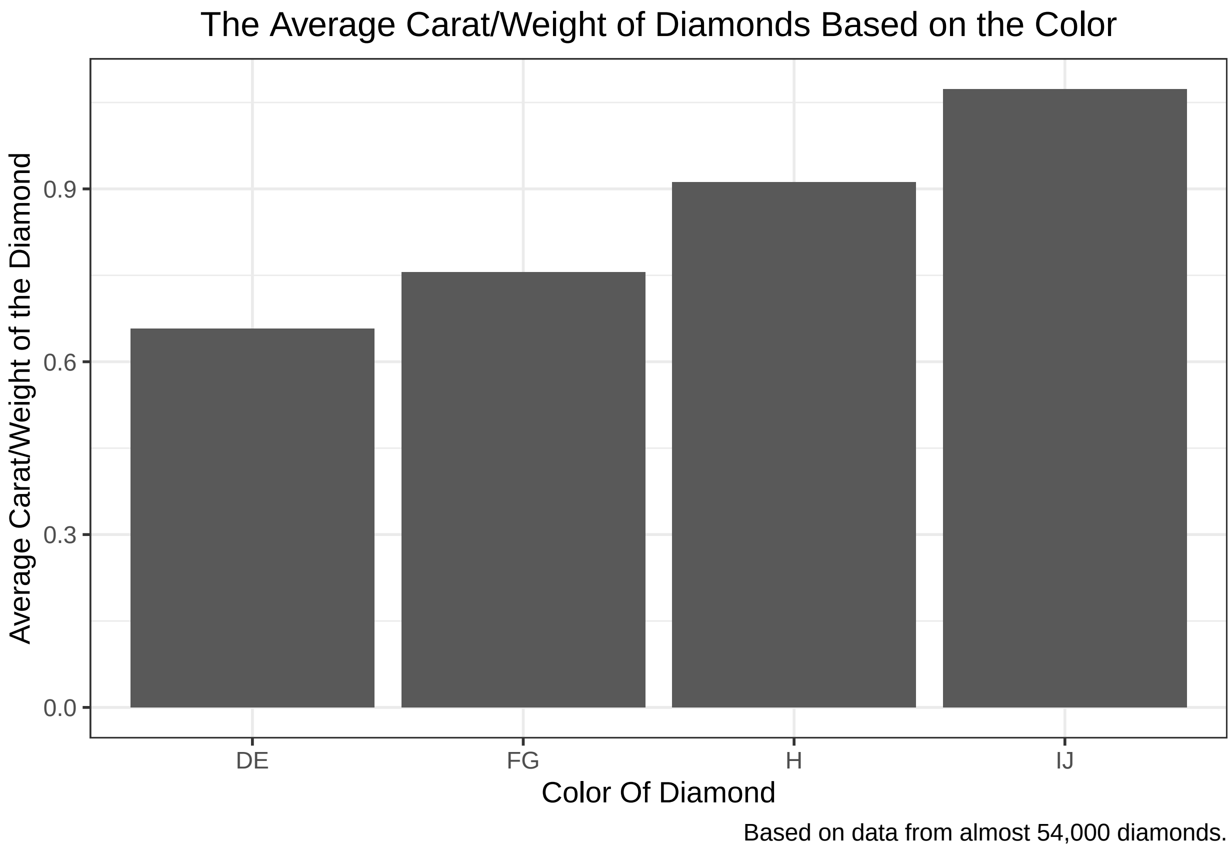
In day10 class, we practiced combining categories "SI1" and "SI2" together as a new category "SI" in the "diamonds" data. We also discussed the general logic to combine the other categories using a nested **ifelse** statement. Please following the rule below to combine smaller categories into bigger once. Make sure the order of the new categories is similar to the original data. Then compute the mean carat for each new **color** type and show it in a graph. **Note:** the graph here is different to the one we made in the class.

- Combine the categories "D" and "E" to a new category "DE";
- Combine the categories "F" and "G" to a new category "FG";
- Combine the categories "I" and "J" to a new category "IJ";
- For all other categories, keep the original category setting.

```
# Create a new data frame called q1Data
  # we mutate the data and combine the groups of colors
  # after that we group the data by those groups
  # then summarize the data by the mean minutes
q1Data <- diamonds %>%
  mutate(q1Category = ordered(ifelse(color=="D"|color=="E", "DE",
                  ifelse(color=="F"| color=="G","FG",
                  ifelse(color=="I" | color =="J","IJ", as.character(color)))),
                  levels=c("DE", "FG", "H", "IJ"))) %>%
  group_by(q1Category) %>%
  summarize(q1meanCarat = mean(carat))


# create a plot of the data
```

```
ggplot(data = q1Data) +
  geom_histogram(aes(x = q1Category, y = q1meanCarat),  stat="identity") +
  labs(x = "Color Of Diamond", y = "Average Carat/Weight of the Diamond",
       caption = "Based on data from almost 54,000 diamonds.") +
  ggtitle("The Average Carat/Weight of Diamonds Based on the Color") +
  theme_bw() +
  theme(plot.title = element_text(hjust=0.5))
```



Question 2

During the lecture, we analyzed a tennis dataset ("atp_matches_2021.csv", from https://github.com/JeffSackmann/tennis_atp). Use the same data and answer the following questions:

1. In year 2021, find the player that had the greatest number of wins on each type of surface. Report the player's name and the number of wins for each type of surface in a table.

```
tennis <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_matches_2021.c
```

```
## Rows: 2733 Columns: 49
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): tourney_id, tourney_name, surface, tourney_level, winner_entry, wi...
## dbl (35): draw_size, tourney_date, match_num, winner_id, winner_seed, winner...
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
ts <- select(tennis, tourney_id:surface, tourney_date,
             match_num, winner_id,
             winner_name, winner_hand,winner_age,
             loser_id,loser_name, loser_hand,loser_age,
             minutes,winner_rank,winner_rank_points,
             loser_rank,loser_rank_points)


q1p1Data <- ts %>%
  group_by(surface, winner_name) %>%
  summarize(totalWins=n()) %>%
  arrange(desc(totalWins)) %>%
  top_n(1,totalWins)
```

```
## ‘summarise()‘ has grouped output by ’surface’. You can override using the
## ‘.groups‘ argument.
```

```
q1p1Data
```

```
## # A tibble: 3 x 3
## # Groups:   surface [3]
##   surface winner_name       totalWins
##   <chr>   <chr>                 <int>
## 1 Hard    Daniil Medvedev          51
## 2 Clay    Casper Ruud              29
## 3 Grass   Matteo Berrettini        11
```

Daniil Medvedev had the most wins on the Hard surface, Casper Ruud had the most wins on the Clay surface, and Matteo Berrettini had the most wins on the Grass surface.

2. For each of the players you detected in question 2(1), find the tourney (tourneys) that they had the greatest number of wins? Report the information in a table. (In this question, it’s fine to use the question 2, part (1) result, and “hard code” the specific location of the player in that dataset to extract the player’s name.)

```
q1p2Data <- ts %>%
  filter(winner_name == "Daniil Medvedev" |
           winner_name == "Casper Ruud" |
           winner_name == "Matteo Berrettini") %>%
  group_by(winner_name, tourney_name) %>%
  summarize(totalWins=n()) %>%
  arrange(desc(totalWins)) %>%
  top_n(1,totalWins)
```

```
## ‘summarise()‘ has grouped output by ’winner_name’. You can override using the
## ‘.groups‘ argument.
```

```
q1p2Data
```

```
## # A tibble: 9 x 3
## # Groups:   winner_name [3]
##   winner_name        tourney_name         totalWins
##   <chr>              <chr>                    <int>
## 1 Daniil Medvedev    Us Open                      7
## 2 Matteo Berrettini  Wimbledon                    6
## 3 Casper Ruud        Bastad                       4
## 4 Casper Ruud        Geneva                       4
## 5 Casper Ruud        Gstaad                       4
## 6 Casper Ruud        Kitzbuhel                    4
## 7 Casper Ruud        Madrid Masters               4
## 8 Casper Ruud        Monte Carlo Masters          4
## 9 Casper Ruud        San Diego                    4
```

Daniil Medvedev won the Us Open with 7 wins, Matteo Berrettini won the Wimbledon with 6 wins, and Casper Ruud won the Bastad with 4 wins.