DATA 101 Exam 1

Kamila Palys

Due: Monday 10/26 at 11:59pm

Academic Honesty Statement (fill in your name)

I, Kamila Palys, hereby affirm that I have not communicated with or gained information in any way from my classmates or anyone other than the Professor during this exam, that I have not assisted anyone else with this exam, and that all work is my own.

Load packages and data

```
# load required packages here
library(tidyverse)
```

```
# read in the data here
nba <- read_csv("data/nba_salaries.csv")
```

Questions

Question 1

First, we will make a table to view the salaries of the NBA players in descending order by using the "select" and "arrange" functions to display the variables we want, in the order we want.

```
nba %>%
  select(player, salary) %>%
  arrange(desc(salary))
```

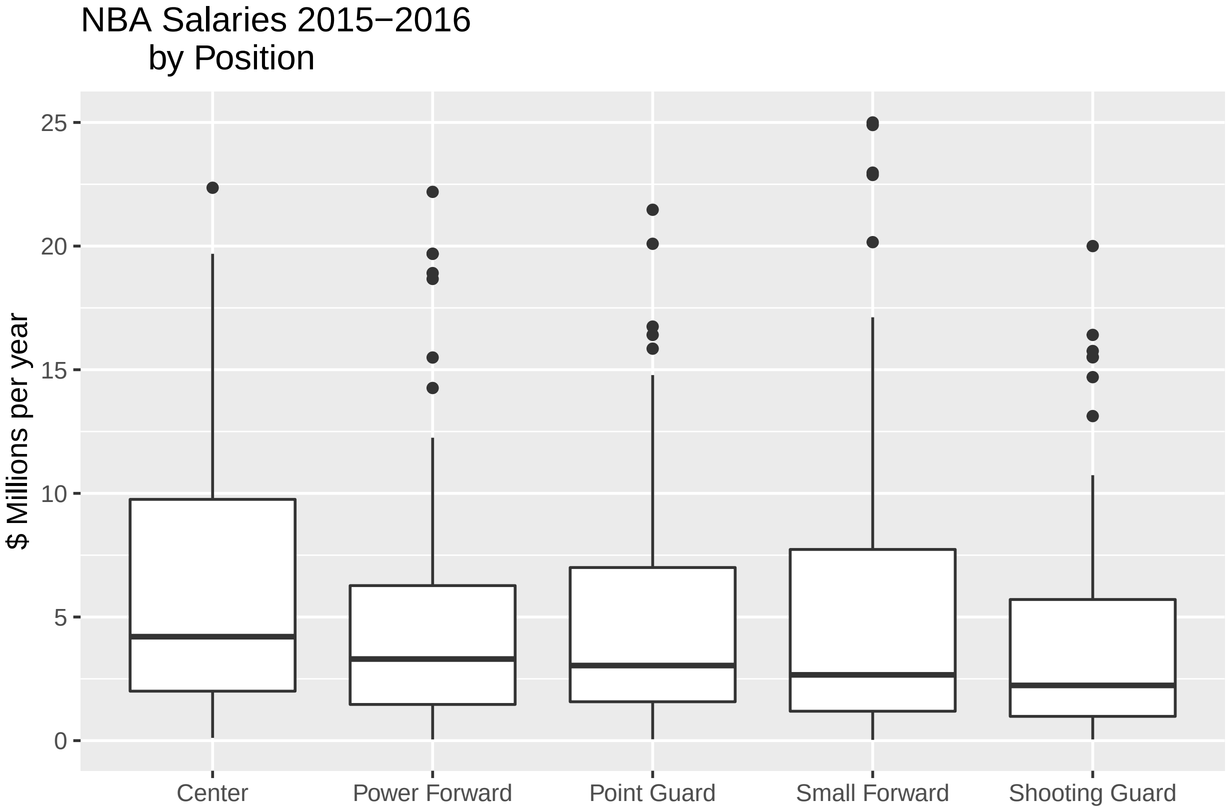
```
## # A tibble: 417 x 2
##    player          salary
##    <chr>            <dbl>
##  1 Kobe Bryant       25
##  2 Joe Johnson       24.9
##  3 LeBron James      23.0
##  4 Carmelo Anthony   22.9
##  5 Dwight Howard     22.4
##  6 Chris Bosh        22.2
##  7 Chris Paul        21.5
##  8 Kevin Durant      20.2
##  9 Derrick Rose      20.1
## 10 Dwyane Wade       20
## # ... with 407 more rows
```

From the table it is visible that Kobe Bryant has the highest salary of all the NBA players at $25 million.

Question 2

Now we will create a boxplot that compares the distribution of the players' salaries by position.

```
ggplot(data = nba) +
  geom_boxplot(mapping = aes(x = position, y = salary)) +
  labs(title = "NBA Salaries 2015-2016
       by Position",
       x = "",
       y = "$ Millions per year") +
  scale_x_discrete(labels = c("Center", "Power Forward", "Point Guard", "Small Forward", "Shooting Guar
```



From this graph, we are able to see that the highest median salary is earned by players with the center position, while the lowest median comes from the shooting guard players. The ones who play center, however, also have a more even spread of their salaries, seeing as their interquartile range, or the middle 50%, is the largest of all the positions. There are also very little outliers in the center players' salaries, and there are several high outliers in all four of the other positions, so that is not to say that players playing a position other than center cannot earn a high salary.

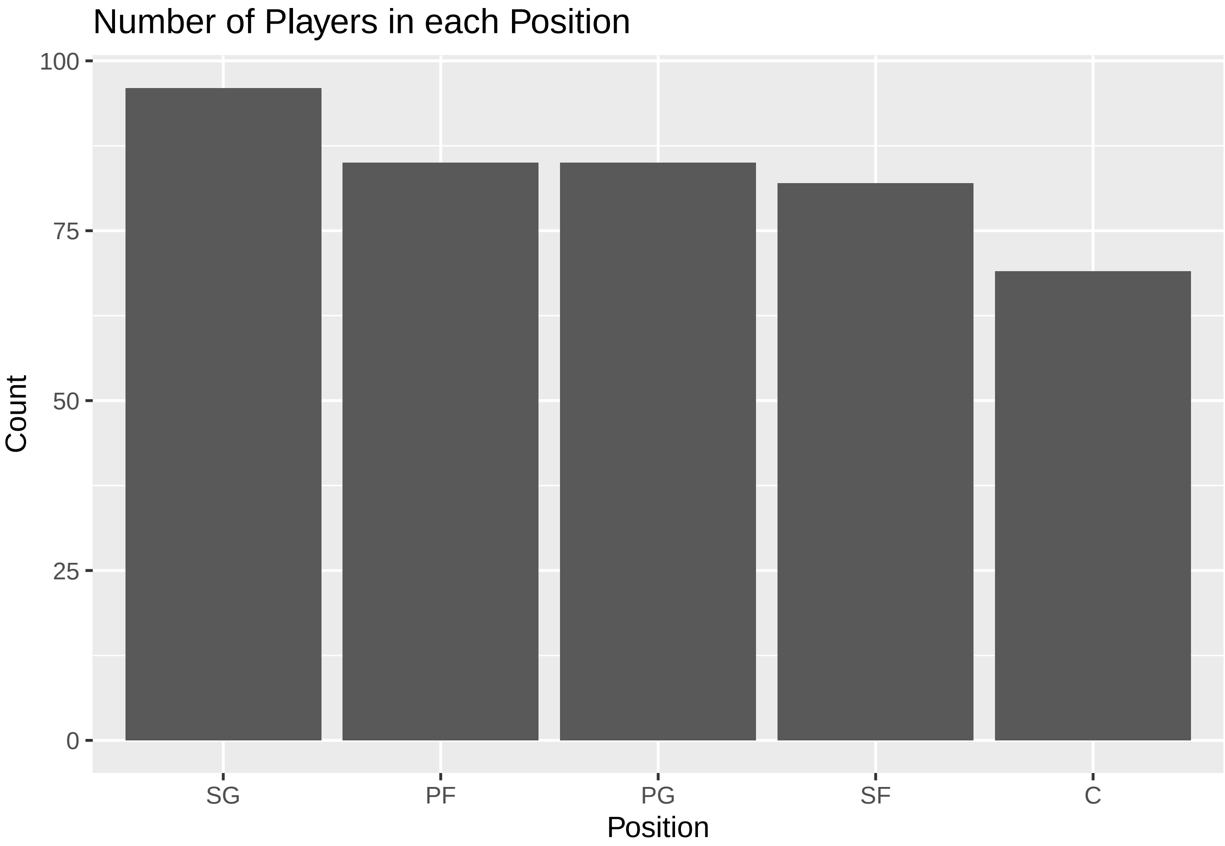
Question 3

Now we will take a look at how many players there are in each position with "group_by" and the "count" function.

```
nba %>%
  group_by(position) %>%
  count(position) %>%
  arrange(desc(position))
```

```
## # A tibble: 5 x 2
## # Groups:   position [5]
##   position     n
##   <chr>    <int>
## 1 SG          96
## 2 SF          82
## 3 PG          85
## 4 PF          85
## 5 C           69
```

```
ggplot(data = nba, mapping = aes(x = fct_infreq(position))) +
  geom_bar() +
  labs(title = "Number of Players in each Position",
       x = "Position",
       y = "Count")
```

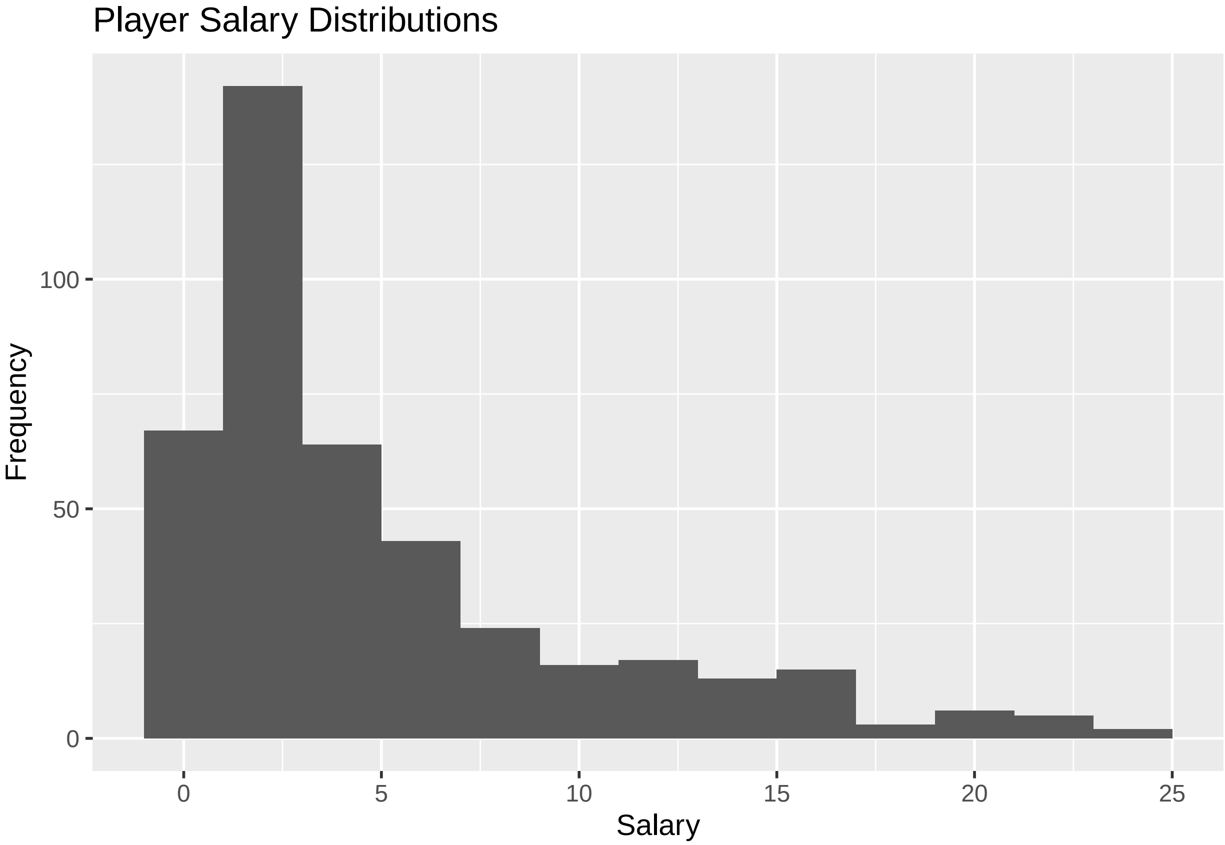


With this table and bar graph, it is visible that there are the most players in shooting guard at 96, while the center position has only 69, which is the least amount of players playing a certain position. Additionally, there are 82 playing small forward, 85 playing point guard, and 85 playing power forward.

Question 4

Here a display is created in the form of a histogram to show the distribution of the salaries of the NBA players.

```
ggplot(data = nba) +
  geom_histogram(mapping = aes(x = salary), binwidth = 2) +
  labs(title = "Player Salary Distributions",
       x = "Salary",
       y = "Frequency")
```



It is clear from this graph that a large number of players earned a salary between about $1 million to $3 million. It appears as though a large amount of players received a salary of up to $10 million, but a minority of them receive more than that and very few reach $20 million or more.

Question 5

Here, the average salaries per player in each team will be displayed.

```
nba %>%
  group_by(team) %>%
  summarise(avg_salary = mean(salary)) %>%
  arrange(desc(avg_salary)) %>%
  top_n(10)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## Selecting by avg_salary
```

```
## # A tibble: 10 x 2
##    team                  avg_salary
##    <chr>                      <dbl>
##  1 Cleveland Cavaliers        10.2
##  2 Houston Rockets             7.11
##  3 Miami Heat                  6.79
##  4 Golden State Warriors       6.72
##  5 Chicago Bulls                  6.57
##  6 San Antonio Spurs              6.51
##  7 Los Angeles Lakers             6.24
##  8 Sacramento Kings               6.22
##  9 Oklahoma City Thunder          6.05
## 10 Dallas Mavericks               5.98
```

As seen, the three top earning teams are the Cleveland Caveliers earning $10.2 million on average, Houston Rockets earning $7.11 million on average, and Miami heat with $6.79 million on average. The second and third highest earning teams do not have an extreme difference in average salaries per player, but the Cleveland Caveliers have a high jump, with each player earning almost $3 million more on average than the second highest earning team. It is possible that this may be due to the Cleveland Caveliers having more famous basketball players that earn a salary that is considered to be an outlier.

Question 6

Here we will create a new variable called "salary_level" with three different levels regarding the salary a player makes: Low, Moderate, and High. Then, in a column called "proportion", the proportion of players that fall within each salary level will be shown.

```
nba %>%
  mutate(salary_level = case_when(
    salary < 8 ~ "Low",
    salary >= 8 & salary <16 ~ "Moderate",
    salary >= 16 ~ "High"
  )) %>%
  count(salary_level) %>%
  mutate(proportion = n/sum(n)) %>%
  arrange(match(salary_level, "Low", "Moderate", "High"))
```

```
## Warning: Problem with `mutate()` input `^^--arrange_quosure_1`.
## i NAs introduced by coercion
## i Input `^^--arrange_quosure_1` is `match(salary_level, "Low", "Moderate", "High")`.

## Warning in match(salary_level, "Low", "Moderate", "High"): NAs introduced by
## coercion

## # A tibble: 3 x 3
##   salary_level     n proportion
##   <chr>        <int>      <dbl>
## 1 Low            326     0.782
## 2 High            22     0.0528
## 3 Moderate        69     0.165
```

As seen in the table above, there are 326 players that earn what we classify as a "low" salary, which is 0.782 of all players. Clearly, the majority of players (over 75%) earn what we call a "low" salary. There are 69 players that earn a "moderate" salary, making for 0.165 of all players, and only 22 players earn a "high" salary, which is 0.0528 of all players.

Question 7

The following table will show the highest salary earned for each position in each team.

```
starters <- nba %>%
  group_by(team, position) %>%
  summarise(highest_salary = (max(salary)))
```

```
## `summarise()` regrouping output by 'team' (override with `.groups` argument)
```

```
print(starters)
```

```
## # A tibble: 147 x 3
## # Groups:   team [30]
##    team           position highest_salary
##    <chr>          <chr>             <dbl>
##  1 Atlanta Hawks  C                 12
##  2 Atlanta Hawks  PF                18.7
##  3 Atlanta Hawks  PG                 8
##  4 Atlanta Hawks  SF                 4
##  5 Atlanta Hawks  SG                 5.75
##  6 Boston Celtics C                  2.62
##  7 Boston Celtics PF                 5
##  8 Boston Celtics PG                 7.73
##  9 Boston Celtics SF                 6.80
## 10 Boston Celtics SG                 3.43
## # ... with 137 more rows
```

I started this problem by creating a new datafram with the new name and piping the nba dataset into the following functions. At first I started to use the select() function knowing that we would only display certain variables, but I later realized it was unnecessary. Since we are applying the maximum function in summarise to each position in each time, I grouped by the team and position. This gave me the new column that we needed, which I called highest_salary. I had to use the print() function so that the table would actually display.

Question 8

Now, the names will be shown for each player that corresponds to the highest earning salary for each position in each team.

```
starters <- left_join(starters, nba, by = c("highest_salary" = "salary"))

starters$position.y <- NULL
starters$team.y <- NULL

print(starters)
```

```
## # A tibble: 218 x 4
##    team.x        position.x highest_salary player
##    <chr>         <chr>               <dbl> <chr>
##  1 Atlanta Hawks C                    12   Al Horford
##  2 Atlanta Hawks C                    12   Kemba Walker
##  3 Atlanta Hawks C                    12   Kyle Lowry
##  4 Atlanta Hawks PF                   18.7 Paul Millsap
##  5 Atlanta Hawks PG                    8   Jeff Teague
##  6 Atlanta Hawks PG                    8   O.J. Mayo
##  7 Atlanta Hawks PG                    8   Arron Afflalo
##  8 Atlanta Hawks PG                    8   Markieff Morris
##  9 Atlanta Hawks SF                    4   Thabo Sefolosha
## 10 Atlanta Hawks SF                    4   Jordan Hill
## # ... with 208 more rows
```

I approached this problem by thinking that I wanted to keep all data from the starters dataframe, and only add the players' names that corresponded to the highest salary. Therefore, I used a left_join to only add correspondin information from the nba data. I received many errors about the variable that I wanted to join by not existing, until I realized I needed to note that the "highest_salary" in the starters dataframe contains the same information as the "salary" in the nba dataframe.

Question 9

If we wanted to see in which instances there are multiple highest paid players within a position in a team, we can first group by the team and the position, since it is within those combinations that we are looking for multiple values. Then, we can simply count how many values of highest_salary there are within a position in every team.

```
starters %>%
  group_by(team.x, position.x) %>%
  count(highest_salary)
```

```
## # A tibble: 147 x 4
## # Groups:   team.x, position.x [147]
##    team.x         position.x highest_salary     n
##    <chr>          <chr>               <dbl> <int>
##  1 Atlanta Hawks  C                   12        3
##  2 Atlanta Hawks  PF                  18.7      1
##  3 Atlanta Hawks  PG                   8        4
##  4 Atlanta Hawks  SF                   4        5
##  5 Atlanta Hawks  SG                   5.75     1
##  6 Boston Celtics C                    2.62     1
##  7 Boston Celtics PF                   5        3
##  8 Boston Celtics PG                   7.73     1
##  9 Boston Celtics SF                   6.80     1
## 10 Boston Celtics SG                   3.43     2
## # ... with 137 more rows
```

From the table produced, we can see that for example, the center position in Atlanta Hawks has three players that all have the same highest salary. There are multiple occasions where more than one player playing a certain position in a team earns the highest salary simultaneously, however, there are also many instances where there is only one player that has the highest salary.

Question 10

```
starters_unique <- starters #[-c(Ian Mahinmi)]

  #pivot_wider(names_from = position.x, values_from = player)

#values_fn = list
```