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))
```

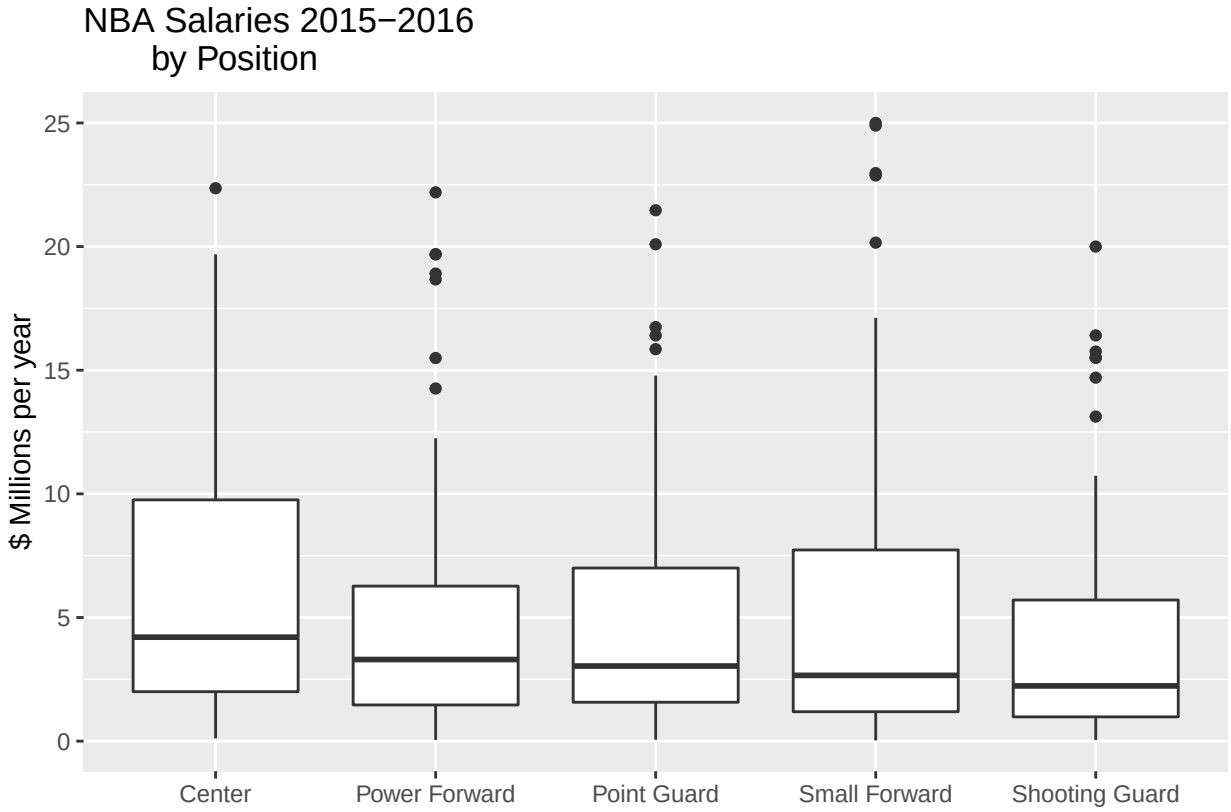
```
## # A tibble: 417 x 2
##    player          salary
##    <chr>            <dbl>
##  1 Kobe Bryant       25
##  2 Joe Johnson       24.9
##  3 LeBron James      23.0
##  4 Carmelo Anthony   22.9
##  5 Dwight Howard     22.4
##  6 Chris Bosh        22.2
##  7 Chris Paul        21.5
##  8 Kevin Durant      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 Guard
```



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.

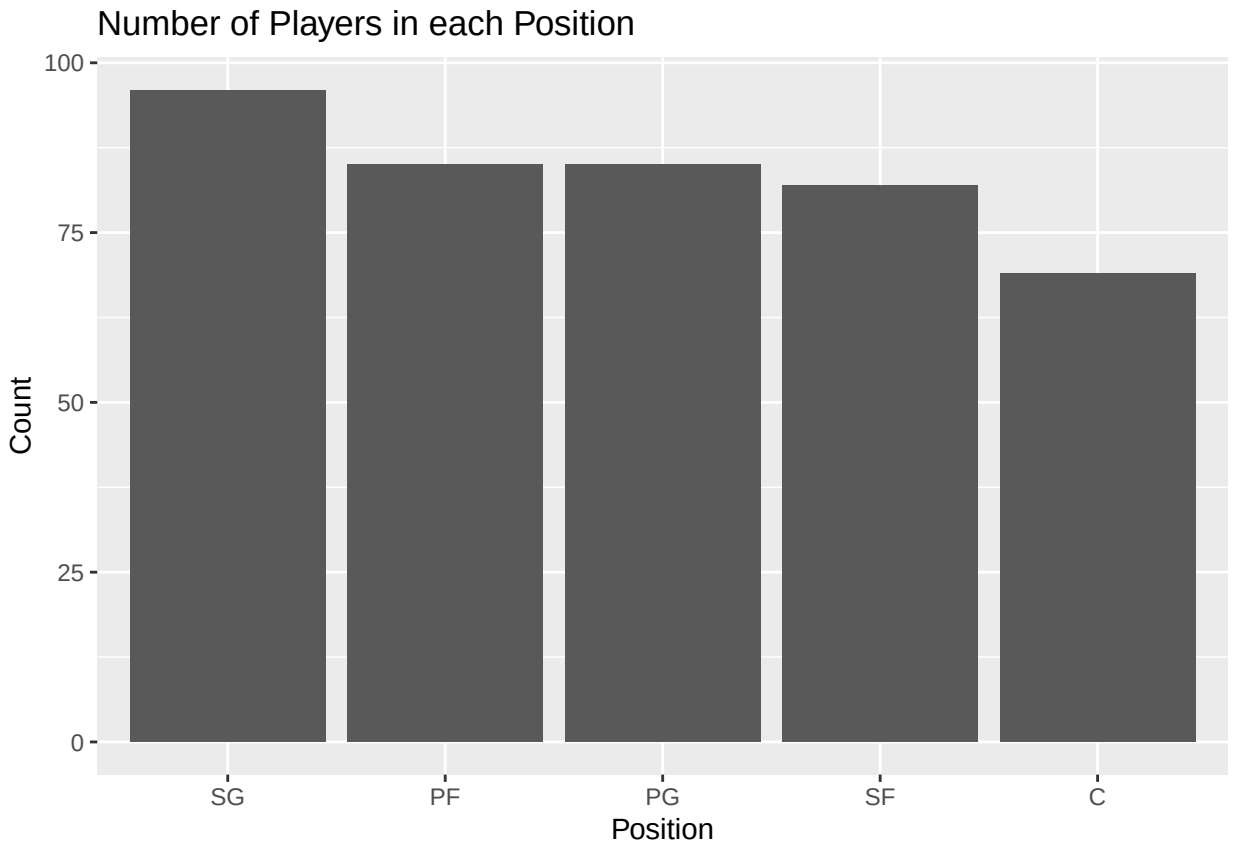
Question 3

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")
```



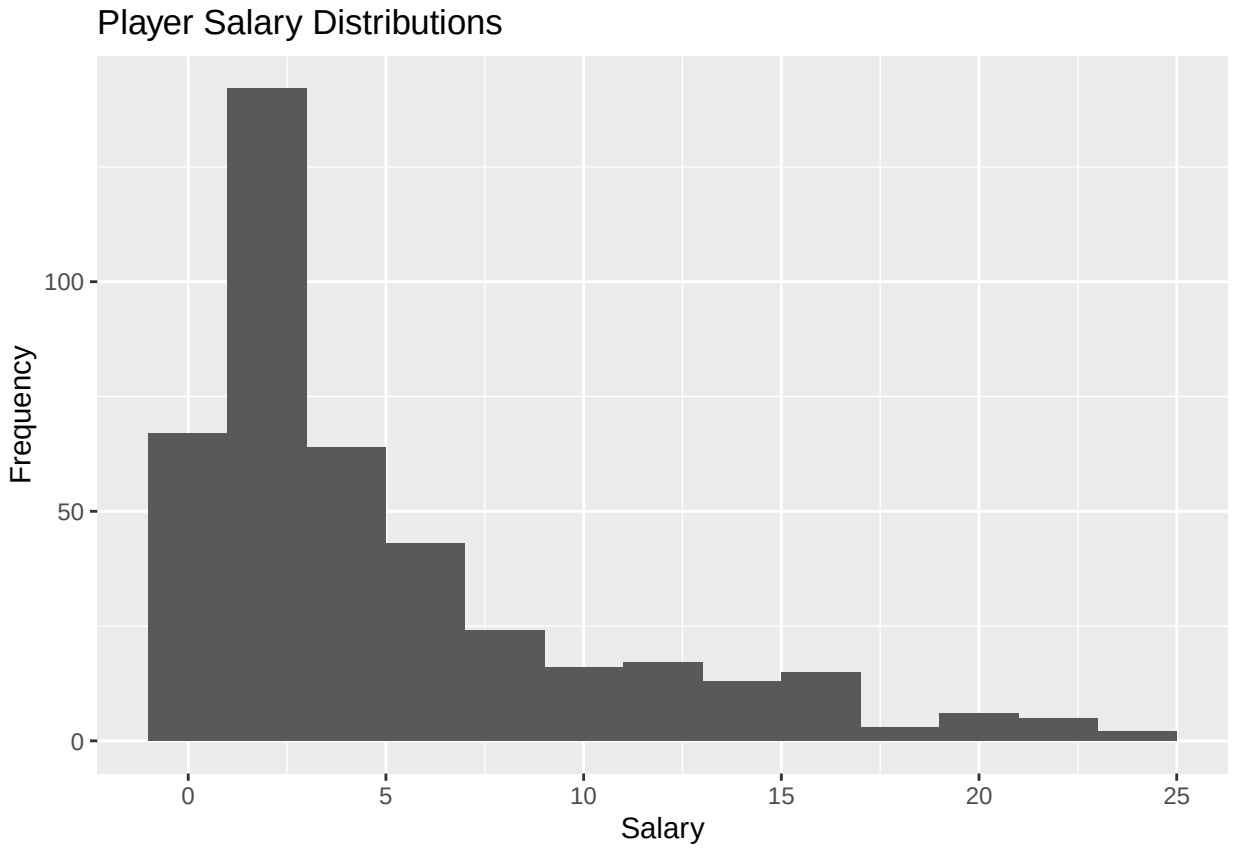
With this table and bar graph, it is visible that there are the most players in shooting guard at 96, while the center position has only 69, which is the least amount of players playing a certain position. Additionally, there are 82 playing small forward, 85 playing point guard, and 85 playing power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

print(starters)
```

```
## # A tibble: 148 x 4
## # Groups:   team [30]
##    team           position highest_salary player
##    <chr>          <chr>             <dbl> <chr>
##  1 Atlanta Hawks  C                 12    Al Horford
##  2 Atlanta Hawks  PF                18.7  Paul Millsap
##  3 Atlanta Hawks  PG                 8    Jeff Teague
##  4 Atlanta Hawks  SF                 4    Thabo Sefolosha
##  5 Atlanta Hawks  SG                 5.75 Kyle Korver
##  6 Boston Celtics C                  2.62 Tyler Zeller
##  7 Boston Celtics PF                 5    Jonas Jerebko
##  8 Boston Celtics PG                 7.73 Avery Bradley
##  9 Boston Celtics SF                 6.80 Jae Crowder
## 10 Boston Celtics SG                 3.43 Evan Turner
## # ... with 138 more rows
```

I approached this problem by recognizing that I wanted to keep all the rows from the starters dataframe, and only add the players' names that corresponded to the highest salary. Therefore, I used a left_join. I initially 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. In addition, to have the players' names correspond to the correct team and position that they are a part of, I also had to join by the "position" and "team".

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, position) %>%
  count(highest_salary) %>%
  print(n = Inf)
```

```
## # A tibble: 147 x 4
## # Groups:   team, position [147]
##     team                   position highest_salary     n
##     <chr>                  <chr>             <dbl> <int>
##   1 Atlanta Hawks          C                 12        1
##   2 Atlanta Hawks          PF                18.7      1
##   3 Atlanta Hawks          PG                 8        1
##   4 Atlanta Hawks          SF                 4        1
##   5 Atlanta Hawks          SG                 5.75     1
##   6 Boston Celtics         C                  2.62     1
##   7 Boston Celtics         PF                 5        1
##   8 Boston Celtics         PG                 7.73     1
##   9 Boston Celtics         SF                 6.80     1
##  10 Boston Celtics         SG                 3.43     1
##  11 Brooklyn Nets          C                  1.36     1
##  12 Brooklyn Nets          PF                11.2      1
##  13 Brooklyn Nets          PG                 6.3      1
##  14 Brooklyn Nets          SF                24.9      1
##  15 Brooklyn Nets          SG                 3.43     1
##  16 Charlotte Hornets      C                 13.5      1
##  17 Charlotte Hornets      PF                 7        1
##  18 Charlotte Hornets      PG                12        1
##  19 Charlotte Hornets      SF                 6.33     1
##  20 Charlotte Hornets      SG                13.1      1
##  21 Chicago Bulls          C                 13.4      1
##  22 Chicago Bulls          PF                 5.54     1
##  23 Chicago Bulls          PG                20.1      1
##  24 Chicago Bulls          SF                 2.38     1
##  25 Chicago Bulls          SG                16.4      1
##  26 Cleveland Cavaliers    C                 14.3      1
##  27 Cleveland Cavaliers    PF                19.7      1
##  28 Cleveland Cavaliers    PG                16.4      1
##  29 Cleveland Cavaliers    SF                23.0      1
##  30 Cleveland Cavaliers    SG                 8.99     1
##  31 Dallas Mavericks       C                  5.2      1
##  32 Dallas Mavericks       PF                15.5      1
##  33 Dallas Mavericks       PG                 5.38     1
##  34 Dallas Mavericks       SF                15.4      1
##  35 Dallas Mavericks       SG                 1.45     1
```

```
##  36 Denver Nuggets         C                5.61     1
##  37 Denver Nuggets         PF              11.2      1
##  38 Denver Nuggets         PG               4.34     1
##  39 Denver Nuggets         SF              14        1
##  40 Denver Nuggets         SG               1.58     1
##  41 Detroit Pistons        C                6.5      1
##  42 Detroit Pistons        PG              13.9      1
##  43 Detroit Pistons        SF               2.84     1
##  44 Detroit Pistons        SG               6.27     1
##  45 Golden State Warriors  C               13.8      1
##  46 Golden State Warriors  PF              14.3      1
##  47 Golden State Warriors  PG              11.4      1
##  48 Golden State Warriors  SF              11.7      1
##  49 Golden State Warriors  SG              15.5      1
##  50 Houston Rockets        C               22.4      1
##  51 Houston Rockets        PF               2.49     1
##  52 Houston Rockets        PG              12.4      1
##  53 Houston Rockets        SF               8.19     1
##  54 Houston Rockets        SG              15.8      1
##  55 Indiana Pacers         C                4        2
##  56 Indiana Pacers         PF               4.05     1
##  57 Indiana Pacers         PG               7        1
##  58 Indiana Pacers         SF              17.1      1
##  59 Indiana Pacers         SG              10.3      1
##  60 Los Angeles Clippers   C                1.10     1
##  61 Los Angeles Clippers   PF              18.9      1
##  62 Los Angeles Clippers   PG              21.5      1
##  63 Los Angeles Clippers   SF               3.38     1
##  64 Los Angeles Clippers   SG               7.08     1
##  65 Los Angeles Lakers     C               15.6      1
##  66 Los Angeles Lakers     PF               3.13     1
##  67 Los Angeles Lakers     PG               5.10     1
##  68 Los Angeles Lakers     SF              25        1
##  69 Los Angeles Lakers     SG               7        1
##  70 Memphis Grizzlies      C               19.7      1
##  71 Memphis Grizzlies      PF               9.64     1
##  72 Memphis Grizzlies      PG               9.59     1
##  73 Memphis Grizzlies      SF               9.45     1
##  74 Memphis Grizzlies      SG               5.16     1
##  75 Miami Heat             PF              22.2      1
##  76 Miami Heat             PG              14.8      1
##  77 Miami Heat             SF              10.2      1
##  78 Miami Heat             SG              20        1
##  79 Milwaukee Bucks        C                2.11     1
##  80 Milwaukee Bucks        PF               5.15     1
##  81 Milwaukee Bucks        PG               6.6      1
##  82 Milwaukee Bucks        SF               1.95     1
##  83 Milwaukee Bucks        SG              14.7      1
##  84 Minnesota Timberwolves C               12.1      1
##  85 Minnesota Timberwolves PF               8.5      1
##  86 Minnesota Timberwolves PG              12.7      1
##  87 Minnesota Timberwolves SF               2.06     1
##  88 Minnesota Timberwolves SG               7.08     1
##  89 New Orleans Pelicans   C                9.21     1
```

```
##  90 New Orleans Pelicans   PF               8.5      1
##  91 New Orleans Pelicans   PG              10.6      1
##  92 New Orleans Pelicans   SF               3.38     1
##  93 New Orleans Pelicans   SG              15.5      1
##  94 New York Knicks        C               12.6      1
##  95 New York Knicks        PF               4.13     1
##  96 New York Knicks        PG               7.40     1
##  97 New York Knicks        SF              22.9      1
##  98 New York Knicks        SG               8        1
##  99 Oklahoma City Thunder  C               16.4      1
## 100 Oklahoma City Thunder  PF              12.2      1
## 101 Oklahoma City Thunder  PG              16.7      1
## 102 Oklahoma City Thunder  SF              20.2      1
## 103 Oklahoma City Thunder  SG               5.14     1
## 104 Orlando Magic          C               11.2      1
## 105 Orlando Magic          PF               8.19     1
## 106 Orlando Magic          PG               8.34     1
## 107 Orlando Magic          SF              16        1
## 108 Orlando Magic          SG               5.19     1
## 109 Philadelphia 76ers     C                4.63     1
## 110 Philadelphia 76ers     PF               6.5      1
## 111 Philadelphia 76ers     PG               2.14     1
## 112 Philadelphia 76ers     SF              10.1      1
## 113 Philadelphia 76ers     SG               2.87     1
## 114 Phoenix Suns           C               13        1
## 115 Phoenix Suns           PF               5.5      1
## 116 Phoenix Suns           PG              13.5      1
## 117 Phoenix Suns           SF               5.5      1
## 118 Phoenix Suns           SG               2.13     1
## 119 Portland Trail Blazers C                6.98     1
## 120 Portland Trail Blazers PF               3.08     1
## 121 Portland Trail Blazers PG               4.24     1
## 122 Portland Trail Blazers SF               8.04     1
## 123 Portland Trail Blazers SG               6        1
## 124 Sacramento Kings       C               15.9      1
## 125 Sacramento Kings       PG               9.5      1
## 126 Sacramento Kings       SF              12.4      1
## 127 Sacramento Kings       SG               6.06     1
## 128 San Antonio Spurs      C                7.5      1
## 129 San Antonio Spurs      PF              19.7      1
## 130 San Antonio Spurs      PG              13.4      1
## 131 San Antonio Spurs      SF              16.4      1
## 132 San Antonio Spurs      SG              10        1
## 133 Toronto Raptors        C                4.66     1
## 134 Toronto Raptors        PF               6.27     1
## 135 Toronto Raptors        PG              12        1
## 136 Toronto Raptors        SF              13.6      1
## 137 Toronto Raptors        SG              10.0      1
## 138 Utah Jazz              C                2.9      1
## 139 Utah Jazz              PF               4.78     1
## 140 Utah Jazz              PG               3.78     1
## 141 Utah Jazz              SF              15.4      1
## 142 Utah Jazz              SG               9.46     1
## 143 Washington Wizards     C               13        1
```

```
## 144 Washington Wizards     PF                  8        1
## 145 Washington Wizards     PG                 15.9      1
## 146 Washington Wizards     SF                  5.61     1
## 147 Washington Wizards     SG                  5.69     1
```

```
starters %>%
  group_by(team, position) %>%
  count(highest_salary) %>%
  filter(n > 1)
```

```
## # A tibble: 1 x 4
## # Groups:   team, position [1]
##   team           position highest_salary     n
##   <chr>          <chr>             <dbl> <int>
## 1 Indiana Pacers C                     4     2
```

From displaying all 147 rows present in the table, we can quickly scan the "n" column and identify that there is only one observation where two players within a team earn the same highest salary in a given position. That is in the case of the Indiana Pacers in the center position, where there are two players that both earn the highest salary of $4 million for that position. This can be confirmed by just looking at the raw data, where indeed both Ian Mahinmi and Jordan Hill earn $4 million playing the center position for the Indiana Pacers, which is the highest salary for that team and position. Alternatively, we can filter to only show results where there are more than one highest_salary per position per team, if we do not want to look through all 147 rows. Therefore, the second table is much smaller and only shows the one observation with the Indiana Pacers C position.

Question 10

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

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

#values_fn = list
```