DATA 101 Exam 1

Shreehar Joshi

Due: Monday 10/26 at 11:59pm

Academic Honesty Statement (fill in your name)

I, Shreehar Joshi, 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

```
library(tidyverse)
```

```
nba <- read_csv("data/nba_salaries.csv")
```

Questions

Question 1

The highest paid players for the NBA 2015-2016 season are as follows:

```
nba %>%
  arrange(desc(salary))
```

```
## # A tibble: 417 x 4
##    player          position team                  salary
##    <chr>           <chr>    <chr>                  <dbl>
##  1 Kobe Bryant     SF       Los Angeles Lakers      25
##  2 Joe Johnson     SF       Brooklyn Nets           24.9
##  3 LeBron James    SF       Cleveland Cavaliers     23.0
##  4 Carmelo Anthony SF       New York Knicks         22.9
##  5 Dwight Howard   C        Houston Rockets         22.4
##  6 Chris Bosh      PF       Miami Heat              22.2
##  7 Chris Paul      PG       Los Angeles Clippers    21.5
##  8 Kevin Durant    SF       Oklahoma City Thunder   20.2
##  9 Derrick Rose    PG       Chicago Bulls           20.1
## 10 Dwyane Wade     SG       Miami Heat              20
## # ... with 407 more rows
```

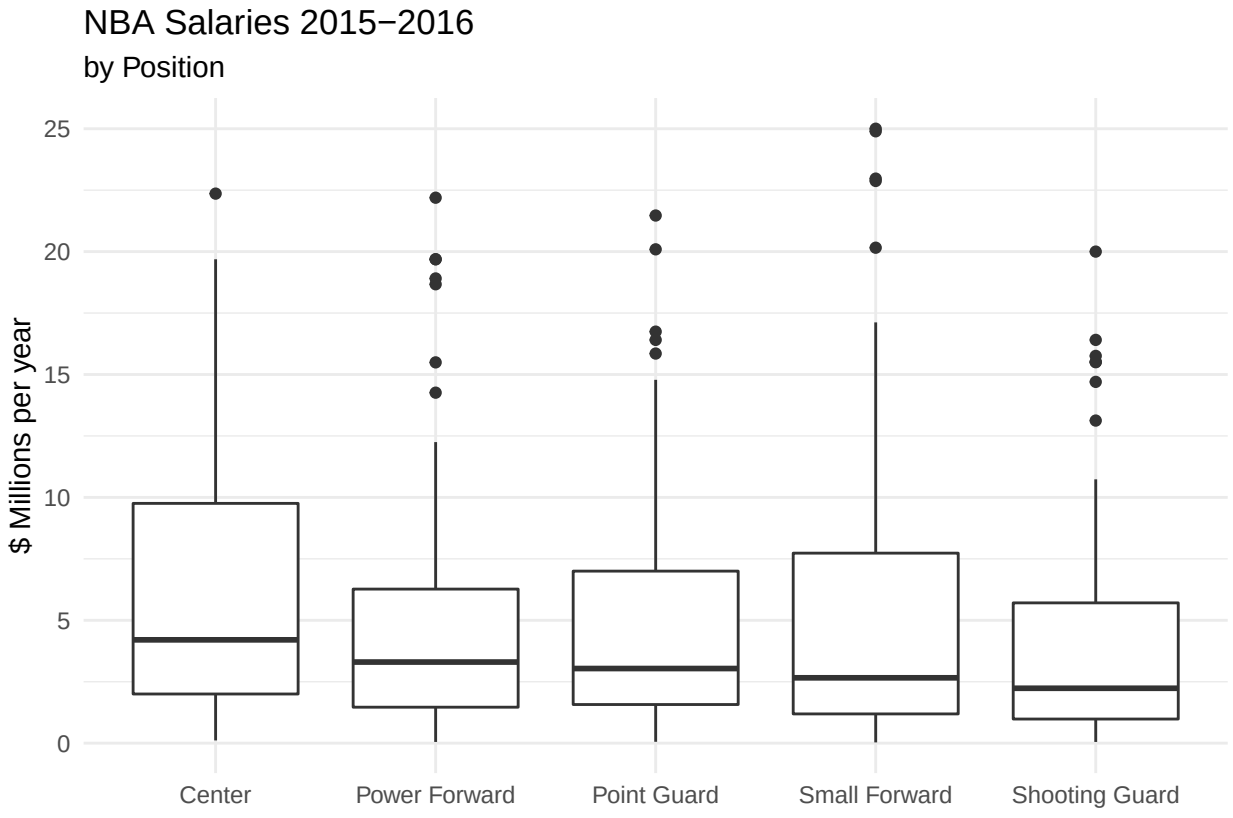
Kobe Bryant had the highest salary of 25.0 million USD per year in the NBA season 2015-2016. He was followed by Joe Johnson and LeBron James each with the salary of 24.9 and 23.0 million USD per year respectively. All the top three highest paid players played in the "Small Forward" position.

Question 2

Now, lets visualize the distribution of salaries based on position in the NBA 2015-2016.

```
nba %>%
  ggplot(aes(x = position, y = salary)) +
  geom_boxplot() +
  theme_minimal() +

  scale_x_discrete(labels = c("C" = "Center",
                              "PG" = "Point Guard",
                              "SG" = "Shooting Guard",
                              "PF" = "Power Forward",
                              "SF" = "Small Forward")) +
  #Reference (https://ggplot2.tidyverse.org/reference/scale_discrete.html)
  labs(
    x = " ",
    y = "$ Millions per year",
    title = "NBA Salaries 2015-2016",
    subtitle = "by Position"
  )
```



The position "Center" has the highest median salary and the position "Shooting Guard" has the least median salary. The position "Center" also has the highest range and interquartile range in salaries. "Point Guard" and "Shooting Guard" have the highest number of outliers while as "Small Forward" has outliers with the highest salaries among all the positions.

Question 3

Now lets find out the number of players that play in each position.

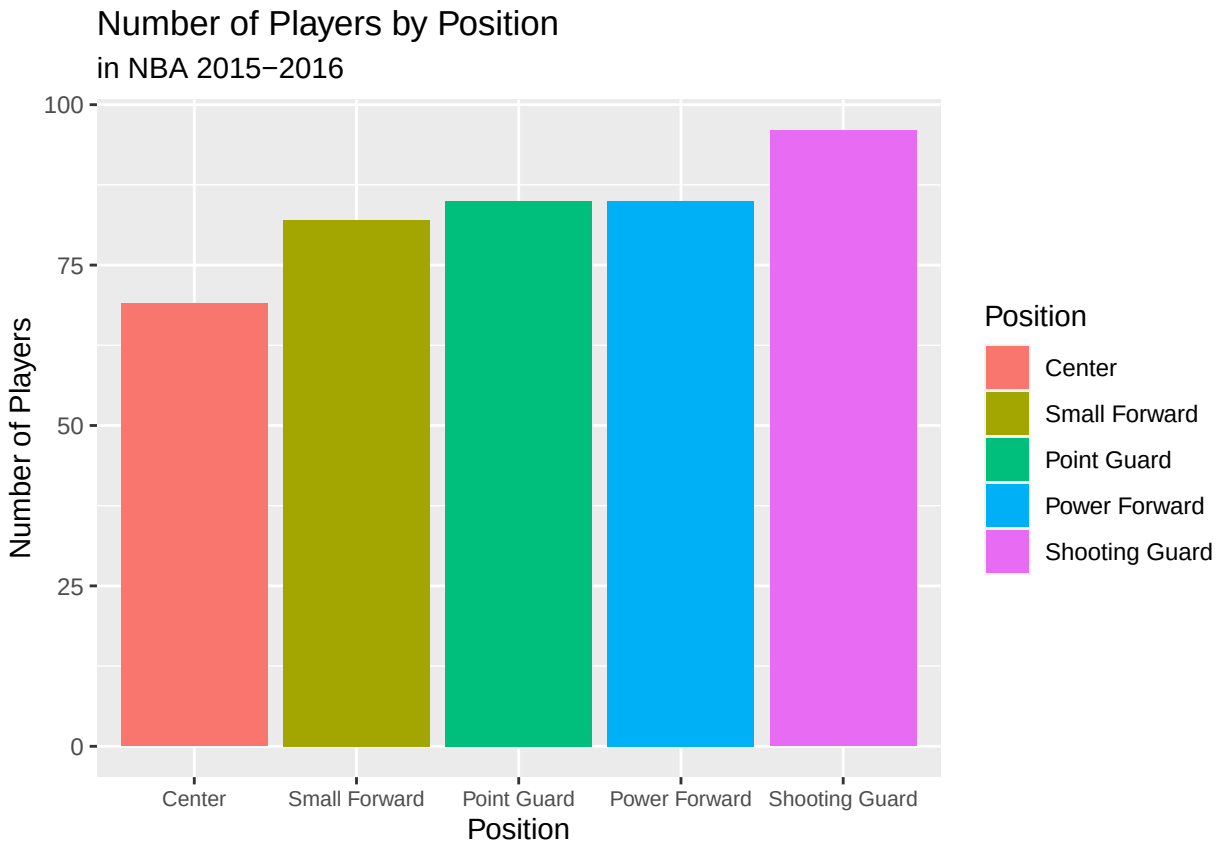
```
nba %>%
  count(position)
```

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

The table above shows that the number of players in Center, Power Forward, Point Guard, Small Forward, and Shooting Guard are 69, 85, 85, 82, and 96 respectively.

Finally, lets visualize the number of players in each position.

```
nba %>%
  ggplot(aes(x = fct_rev(fct_infreq(position)),
             fill = fct_rev(fct_infreq(position)))) +
  geom_bar() +
  theme(axis.text.x = element_text(size = 8)) +
  #Reference (http://www.cookbook-r.com/Graphs/Axes_(ggplot2)/)
  scale_x_discrete(labels = c("C" = "Center",
                              "PG" = "Point Guard",
                              "SG" = "Shooting Guard",
                              "PF" = "Power Forward",
                              "SF" = "Small Forward")) +
  #Reference (https://ggplot2.tidyverse.org/reference/scale_discrete.html)
  scale_fill_discrete(labels = c("C" = "Center",
                                 "PG" = "Point Guard",
                                 "SG" = "Shooting Guard",
                                 "PF" = "Power Forward",
                                 "SF" = "Small Forward")) +
  #Reference (http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/)
  labs(
    x = "Position",
    y = "Number of Players",
    title = "Number of Players by Position",
    subtitle = "in NBA 2015-2016",
    fill = "Position"
  )
```

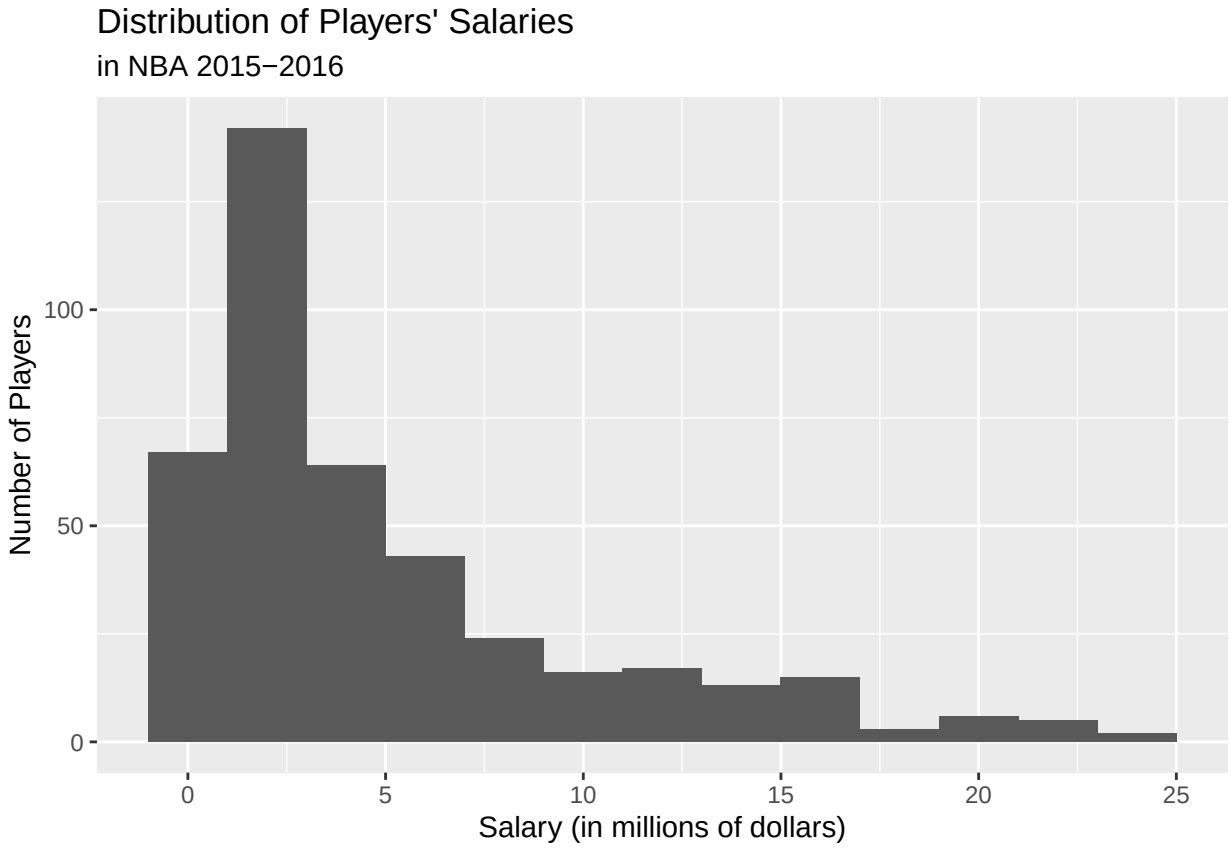


It is evident from the graph above that the position "Center" had the least number of players and the position "Shooting Guard" had the highest number of players in the NBA 2015-2016. The other three positions (Small Forward, Point Guard, and Power Forward) had the number of players that were in between the number of players in "Center" and "Shooting Guard". Additionally, "Point Guard" and "Power Forward" had the number of players that were in between the number of players in "Center" and "Shooting Guard". Additionally, "Point Guard" and "Power Forward" had the same number of players, both of which were greater than the number of players in "Small Forward".

Question 4

Now, lets visualize the distribution of players' salaries.

```
nba %>%
  ggplot(aes(x = salary)) +
  geom_histogram(binwidth = 2) +
  labs(
    x = "Salary (in millions of dollars)",
    y = "Number of Players",
    title = "Distribution of Players' Salaries",
    subtitle = "in NBA 2015-2016"
  )
```



The distribution of salaries is unimodal and right skewed. The histogram above suggests that majority of NBA players had an annual salary of around 3 million USD per year for the season 2015-2016. It also suggests that some players were able to have salaries far greater than the average with the maximum being 25 million USD per year.

Question 5

Now lets find the average player salary for the top 10 highest paying teams.

```
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
```

Cleveland Cavaliers had the average salary of 10.2 million USD per year, which was the highest for a team. It was followed by Houston Rockets and Miami Heats in the second and third position with average salaries of 7.11 million and 6.79 million USD per year respectively.

Question 6

Now, lets classify the salaries of the players into "Low", "Moderate", and "High".

```
nba_salary <- nba %>%
  mutate(salary_level = case_when(
    salary < 8                ~ "Low",
    salary >= 8 & salary < 16 ~ "Moderate",
    salary >= 16              ~ "High"
  ))
```

After this, lets calculate the proportion of players at each salary level.

```
nba_salary %>%
  count(salary_level, sort = TRUE) %>%
  mutate(prop_players = n / sum(n))
```

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

Majority of players had a "Low" salary as their proportion is 0.782, which is the highest. Only a few players had a "High" salary as their proportion is 0.0528, which is the least. The proportion of players whose salary is classified as "Medium" is 0.165 and it falls in between the proportion of players that have "High" and "Low" salaries.

Question 7

Now lets create a dataframe to store the starting lineup salaries of each team.

```
starters <- nba %>%
  select (-player) %>%
  group_by(team, position) %>%
  filter(salary == max(salary)) %>%
  ungroup() %>%
  distinct() %>%
  arrange(team, position)

starters
```

```
## # A tibble: 147 x 3
##    position team           salary
##    <chr>    <chr>           <dbl>
##  1 C        Atlanta Hawks   12
```

```
##  2 PF       Atlanta Hawks   18.7
##  3 PG       Atlanta Hawks    8
##  4 SF       Atlanta Hawks    4
##  5 SG       Atlanta Hawks    5.75
##  6 C        Boston Celtics   2.62
##  7 PF       Boston Celtics   5
##  8 PG       Boston Celtics   7.73
##  9 SF       Boston Celtics   6.80
## 10 SG       Boston Celtics   3.43
## # ... with 137 more rows
```

To create the dataframe above, at first, we removed the "player" column by using select function in the nba dataframe. Then, the dataframe was piped to be grouped on the basis of team and position. And then, we used filter function to find the maximum salary. Had we just grouped on the basis of team, then the maximum salary on the basis of teams would have been filtered instead of filtering the maximum salary on the basis of each positions in each teams. After the filter function, ungroup function is used to remove the groups in the resulting dataframe. It is followed by distinct function to avoid any repetition of the highest salaries by selecting only the unique rows. The distinct function is followed by arrange function to sort the output alphabetically first by team name and then by position.

Question 8

Now, adding a column with appropriate player names to the starters dataframe by left joining it with nba dataframe.

```
starters <- left_join(starters, nba)
```

```
## Joining, by = c("position", "team", "salary")
```

```
starters
```

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

To create the dataframe above, we used left join to join the two dataframes (starters and nba). In this type of join, all rows from starters dataframe are returned along with the matching rows from nba dataframe.

Question 9

We can count the number of players in each of position in each of the teams and if we find any position with more than 1 value in any team, then we can conclude that the team has multiple starters in a single position.

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

```
## # A tibble: 1 x 3
##   team           position     n
##   <chr>          <chr>    <int>
## 1 Indiana Pacers C            2
```

In the team Indiana Pacers, the Center position had two players with the same highest salary.

Question 10

Finally, lets view the teams with the name of their players with highest salaries in each position.

```
starters_unique <- starters %>%
  filter(player != "Ian Mahinmi") %>%
  pivot_wider(-salary, names_from = position, values_from = player)

knitr::kable(starters_unique,
             col.names = c("Team",
                           "Center",
                           "Power Forward",
                           "Point Guard",
                           "Small Forward",
                           "Shooting Guard"))
```

Team	Center	Power Forward	Point Guard	Small Forward	Shooting Guard
Atlanta Hawks	Al Horford	Paul Millsap	Jeff Teague	Thabo Sefolosha	Kyle Korver
Boston Celtics	Tyler Zeller	Jonas Jerebko	Avery Bradley	Jae Crowder	Evan Turner
Brooklyn Nets	Andrea Bargnani	Thaddeus Young	Jarrett Jack	Joe Johnson	Bojan Bogdanovic
Charlotte Hornets	Al Jefferson	Marvin Williams	Kemba Walker	Michael Kidd-Gilchrist	Nicolas Batum
Chicago Bulls	Joakim Noah	Nikola Mirotic	Derrick Rose	Doug McDermott	Jimmy Butler
Cleveland Cavaliers	Tristan Thompson	Kevin Love	Kyrie Irving	LeBron James	Iman Shumpert
Dallas Mavericks	Zaza Pachulia	David Lee	Deron Williams	Chandler Parsons	Justin Anderson
Denver Nuggets	JJ Hickson	Kenneth Faried	Jameer Nelson	Danilo Gallinari	Gary Harris
Detroit Pistons	Aron Baynes	NA	Reggie Jackson	Stanley Johnson	Jodie Meeks
Golden State Warriors	Andrew Bogut	Draymond Green	Stephen Curry	Andre Iguodala	Klay Thompson
Houston Rockets	Dwight Howard	Terrence Jones	Ty Lawson	Trevor Ariza	James Harden
Indiana Pacers	Jordan Hill	Lavoy Allen	Rodney Stuckey	Paul George	Monta Ellis
Los Angeles Clippers	Cole Aldrich	Blake Griffin	Chris Paul	Paul Pierce	J.J. Redick

Team	Center	Power Forward	Point Guard	Small Forward	Shooting Guard
Los Angeles Lakers	Roy Hibbert	Julius Randle	D'Angelo Russell	Kobe Bryant	Louis Williams
Memphis Grizzlies	Marc Gasol	Zach Randolph	Mike Conley	Jeff Green	Tony Allen
Miami Heat	NA	Chris Bosh	Goran Dragic	Luol Deng	Dwyane Wade
Milwaukee Bucks	Miles Plumlee	Jabari Parker	Greivis Vasquez	Giannis Antetokounmpo	Khris Middleton
Minnesota Timberwolves	Nikola Pekovic	Kevin Garnett	Ricky Rubio	Shabazz Muhammad	Kevin Martin
New Orleans Pelicans	Omer Asik	Ryan Anderson	Jrue Holiday	Quincy Pondexter	Eric Gordon
New York Knicks	Robin Lopez	Kristaps Porzingis	Jose Calderon	Carmelo Anthony	Arron Afflalo
Oklahoma City Thunder	Enes Kanter	Serge Ibaka	Russell Westbrook	Kevin Durant	Dion Waiters
Orlando Magic	Nikola Vucevic	Channing Frye	Brandon Jennings	Tobias Harris	Victor Oladipo
Philadelphia 76ers	Joel Embiid	Carl Landry	Kendall Marshall	Gerald Wallace	Nik Stauskas
Phoenix Suns	Tyson Chandler	Mirza Teletovic	Eric Bledsoe	P.J. Tucker	Devin Booker
Portland Trail Blazers	Ed Davis	Meyers Leonard	Damian Lillard	Al-Farouq Aminu	Gerald Henderson
Sacramento Kings	DeMarcus Cousins	NA	Rajon Rondo	Rudy Gay	Marco Belinelli
San Antonio Spurs	Boris Diaw	LaMarcus Aldridge	Tony Parker	Kawhi Leonard	Danny Green
Toronto Raptors	Jonas Valanciunas	Patrick Patterson	Kyle Lowry	DeMarre Carroll	DeMar DeRozan
Utah Jazz	Tibor Pleiss	Trevor Booker	Dante Exum	Gordon Hayward	Alec Burks
Washington Wizards	Nene Hilario	Markieff Morris	John Wall	Martell Webster	Bradley Beal

```
#Reference (https://bookdown.org/yihui/rmarkdown-cookbook/kable.html)
```

To create the dataframe above, at first, the starters dataframe is piped into a filter function which removes the record for the player "Ian Mahinmi". Then the resulting dataframe is pivoted using pivot_wider function. The first argument passed in this function removes the salary column. The second and third arguments will create columns with names from position and the values for those newly created columns taken from players' names. In the end, knitr:kable function is used to print the entire table.