Group 3 - Analysis of Salaries in Major League Soccer

Dylan Hull, Jueyao Liu, Colin Zhang

Introduction

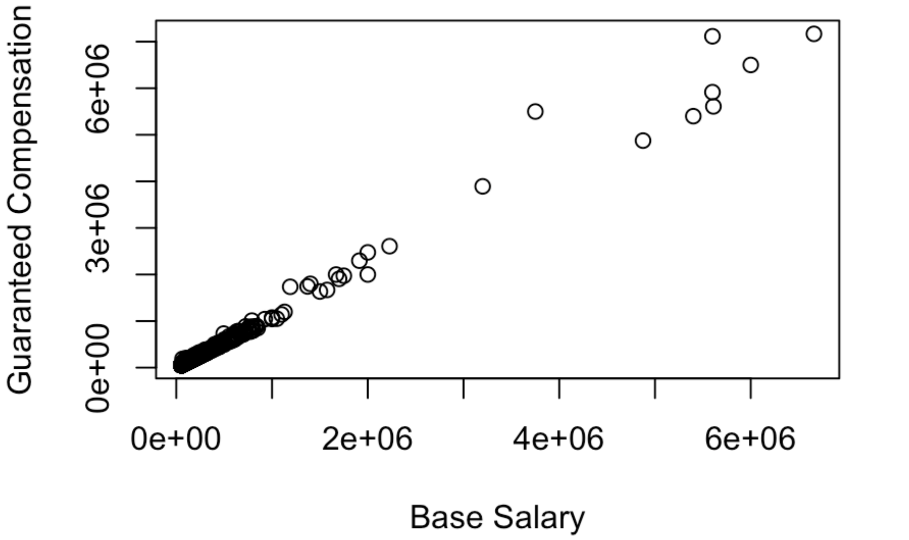
Major League Soccer is the top tier of American professional soccer, bringing in players from all around the world. Although a relatively new league (it was only founded in 1996), the league has over 25 clubs and 600 players. However, it is often viewed as a step below soccer leagues in other countries. The United States' most popular sports, football and baseball, take up much of the country's attention while the MLS is left to scrounge for fans and popularity. That being said, some players still make millions of dollars and the athletes are well compensated for their efforts. It is also one of the fastest growing leagues in the world, with 7 expansion teams since 2016 and 3 more to come by 2023 (MLS).

The MLS has an interesting dynamic to its players and who gets paid. A lot of the time, when a world class player finds the competition in the European leagues to be too tough, they will "retire" to the MLS. Oftentimes, these players command exceptionally large salaries when compared to homegrown players in the MLS system, as they are quite famous players tha bring a lot of publicity and success to the club. This causes a few exceptional outliers in the MLS salary distribution. With this in mind, the salaries of the MLS become an interesting topic to discuss. What exactly motivates or allows a team to pay a certain player more?

In this paper, we examine several variables that relate to salaries of players in the MLS and determine how these variables will affect how players get paid. The dataset we are working with provides salaries, positions, and teams of every player in the MLS in 2017. We are going to focus on how team and position affects their salaries. First, we want to determine if there is a difference between how much certain teams will pay their players. With the new expansion teams in mind, we want to look at the difference between their salaries and more established clubs. For example, will the expansion teams pay their players at a different price level than some of the original clubs that have been around for almost 25 years? In order to answer this question we will look at the two expansion teams in 2017 (Atlanta and Minnesota) and two original clubs (Dallas and Washington, DC). Also, does the size of the market affect the salary of the players? To determine this we will look at two big cities compared to two smaller cities. We compared Colorado (Denver) to Chicago and Los Angeles to Salt Lake CIty, two pairs of cities in similar regions with different market sizes. Since we also want to determine if position affects the player salary, we took a look at how much forwards were paid compared to how much defenders were paid as well as goalkeepers vs non goalkeepers. Essentially, were goal scorers more heavily compensated than those who prevented goals, and were field players paid more than a goalkeeper? An analysis of these salaries can help determine how a certain team may make decisions about how many players of a certain position to sign if they need to account for how much money they plan to spend. All of these questions were assessed using permutation tests. Finally, we looked at how the salaries of the players across the entire MLS were distributed. We checked the distribution against several types of distributions (exponential, gamma, normal) using the Kolmogorov-Smirnov test.

Data Processing

For certain problems, we decided to remove the outliers in salary. To do this, we plotted 'base salary' against 'guaranteed compensation'.



From a simple visual inspection of the scatterplot, we determine the 9 points closest to the upper-right corner of the plot are outliers. This was expected, as there are usually star players who sign very lucrative contracts compared to most members in the league. These are players that have a base salary higher than 3 million dollars.

From the four positions, there were some players who had a couple positions instead of just one. To simplify things, the first position listed was assumed to be the primary position and secondary positions were removed.

Question 1 - Does Clubs Affect Salary?

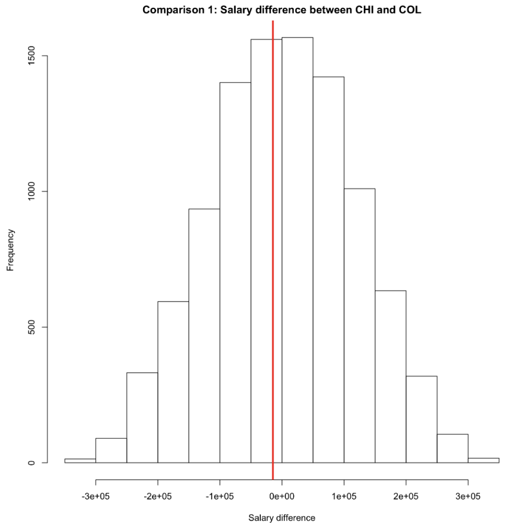
Part 1 -- Club Location Effect: Large Market vs. Small Market

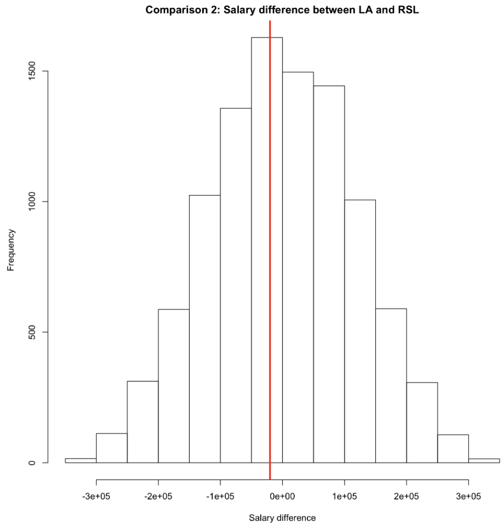
As the soccer industry grows rapidly in the United States, professional MLS players become high-paying jobs and each club pays their players millions of dollars each year to compete for the championship. Given that salary occupies a significant proportion of a club's expenditure, we are curious about the salary differences between MLS clubs. Firstly, we investigate the location effect.

We hypothesize that MLS clubs located in a large market, such as California and Chicago, spend more on player salaries than clubs located in a small market.

Two comparisons are carried out for this hypothesis. The first one is in the middle west between Chicago Fire FC(CHI) and Colorado Rapids(COL), and the second comparison is on the west coast between LA Galaxy(LA) and Real Salt Lake(RSL). Chicago Fire FC and LA Galaxy represent the big market whereas Colorado Rapids and Real Salt Lake represent the small market. After removing outliers, we performed the permutation tests under the null hypothesis of equal average salary. The test statistic is the difference between subtracting the average salary from clubs located in the small market from clubs located in the large market.

The permutation test result (Permutation size = 10000)				
Comparison	Difference (large market - small market)	P value	The 95% CI (lower bound)	The 95% CI (upper bound)
CHI vs COL	-14697.1	0.912	-222762.4	195457.8
LA vs RSL	-19955.1	0.872	-225844.2	193133.9

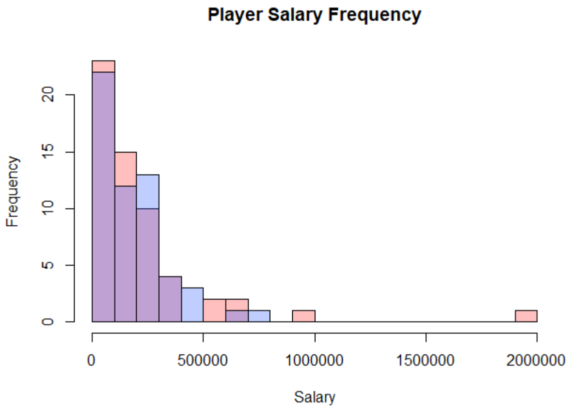

Comparison 1: Salary difference between CHI and COL

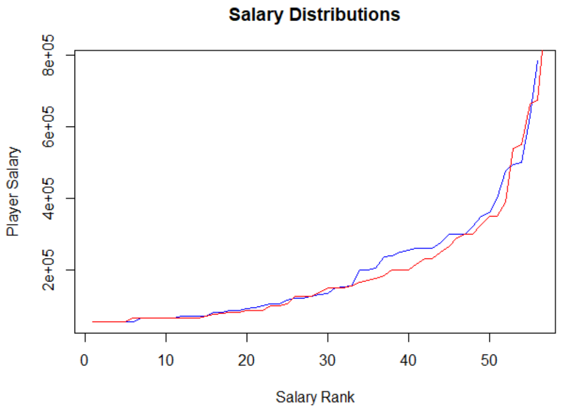

Comparison 2: Salary difference between LA and RSL

From the chart and the plot above, we cannot reject either of our hypotheses since both p values, one is 0.912 and the other is 0.872, are extremely high. Interestingly, the average salary of clubs located in a small market is higher than the average salary of clubs located in a large market, when excluding outliers. Thus, the result indicates there is no significant effect of location on the player salary.

Part 2 -- Club Age Effect: Old vs. New Clubs

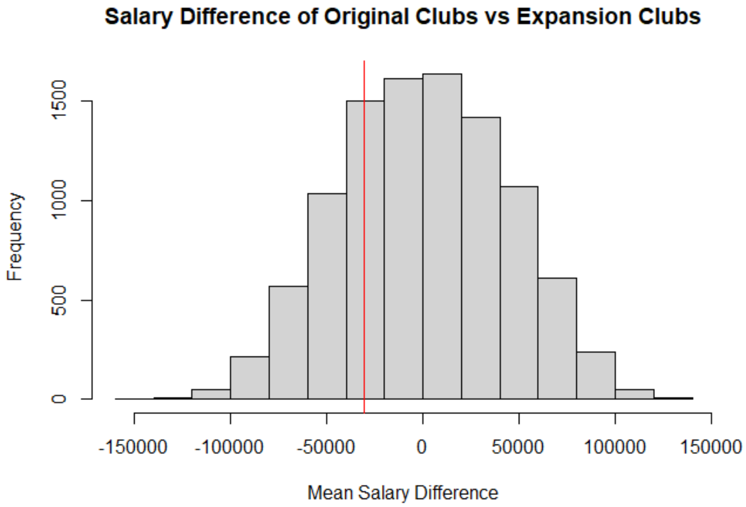
Another attempt to distinguish salary distributions between clubs was to look at the difference in salaries between old and new clubs. There were two expansion teams in 2017, Atlanta and Minnesota, so we used two original teams' (Dallas and DC) salaries to compare similar sample sizes. Since the league was founded in 1996, the original teams were in their 22nd year while expansion clubs were only in their first year. The distributions of old teams' salaries versus new teams' salaries are shown in the following plots.





In these graphs, red represents expansion club salaries and blue represents original club salaries. The two visually appear to have similar distributions but we proceed to run permutation tests on the combined samples to determine statistical significance.

After 10000 iterations of a permutation sample, the following distribution of the difference of means (mean(original) - mean(expansion) was determined.



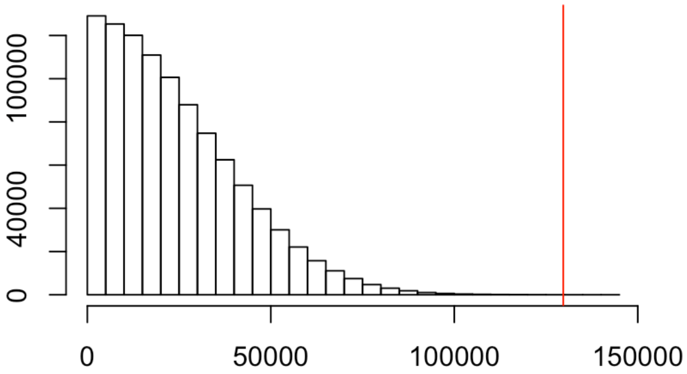
The actual salary difference is shown by the red line, showing that the new clubs actually have a higher average salary than the older clubs. After the permutation test was run, it was determined that the two sided p-value was equal to .52, much higher than a .05 significance level. This indicates that there is likely not a difference in how clubs pay their players based on club age.

Question 2 - Does Position Affect Salary?

Part 1 -- Forwards vs. Defenders

Forwards are players who attempt to score, while defenders are players who stop the other team's forwards from scoring. We were interested in comparing the two positions and wanted to know if the positions had different mean salaries.

To do this, we used a permutation test. We set B = 1,000,000 permutations, and the statistic used was the absolute value of the difference between the means.



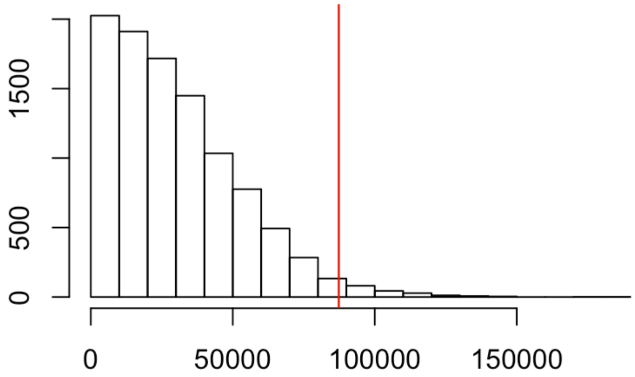
The above figure is the histogram of the test statistic. The red line indicates the statistic of the original dataset, which is clearly far away from most of the data. The p-value is 1.2999e-05, indicating a very high significance, and that it was likely that the two means were different from each other.

While we did remove some outliers, it's possible that using another statistic that is more resistant to outliers (ex. median) might indicate that the salaries between the two positions are not very different.

Part 2 -- Goalkeeper vs. Others

Goalkeeper is a relatively uncommon position, as each team only has one on the field at a time. We were interested in comparing the position with all other positions and wanted to know if it had a different mean salary from the others.

To do this, we used a permutation test. We set B = 10,000 permutations, and the statistic used was the absolute value of the difference between the means.



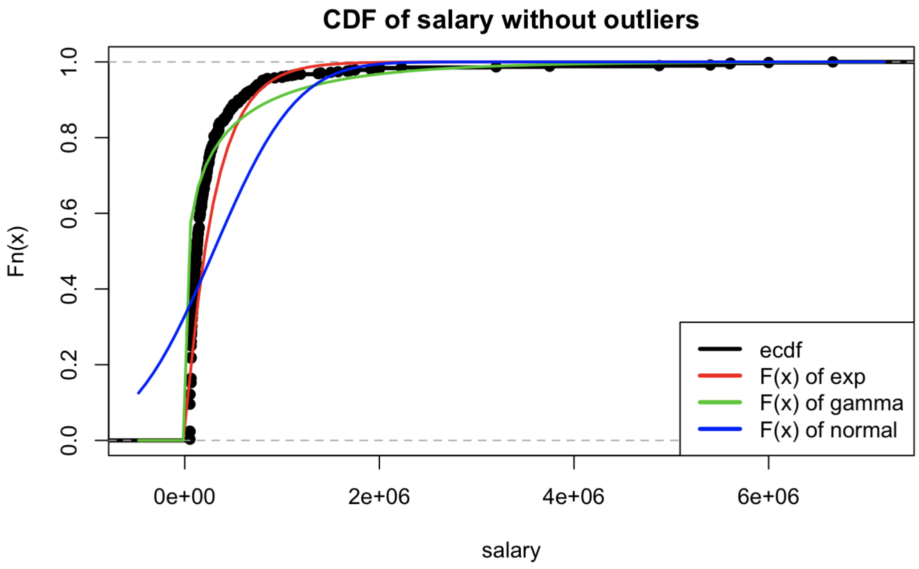
The above figure is the histogram of the test statistic. The red line indicates the statistic of the original dataset, which is closer to the higher end of the data. The p-value is 0.021, indicating a somewhat high significance, and that it was likely that the two means were different from each other.

Note that the sample size for goalkeepers was 65, compared to 551 players of other positions. This should not have affected the results as the permutation test makes no assumptions on the size of the samples.

Question 3 -- Analysis of Salary Distribution

Finally, we analyzed the overall MLS salary distribution excluding outliers. We fitted three distributions, normal, exponential, and gamma. We then tested the similarities based on the Kolmogorov-Smirnov statistic (D). Maximum likelihood estimators (MLE) were calculated in the Kolmogorov-Smirnov test (KS test).

The Kolmogorov-Smirnov test result		
Distribution	MLE	The KS statistic (D)
Normal	$\hat{\mu}=224453.5$, $\hat{\sigma}=285366.\ 1$	0.318
Exponential	$\hat{\lambda}=4.455*10^{-6}$	0.211
Gamma	$\hat{\alpha}=1.133$, $\hat{\beta}=5.046*10^{-6}$	0.203



Based on the resulting chart and the plot above, the normal distribution has the worst fitness. It has the highest Kolmogorov-Smirnov statistic (D=0.318) and its CDF curve doesn't show consistency with the ECDF curve of salary as well. The exponential and gamma distribution's CDFs look quite consistent with the original ECDF. Also, given the alpha hat of the gamma distribution is close to 1 and gamma(1,β) is an exponential distribution, the gamma distribution under MLE is close to an exponential distribution. Their Kolmogorov-Smirnov statistics are close to each other (0.211 and 0.203) as well. Therefore, although the gamma distribution has a slightly lower Kolmogorov-Smirnov statistic, we decided to use the exponential distribution for interpretation since it has better interpretability.

As a result, the MLS salary distribution is close to an exponential distribution with λ equals $4.455*10^{-6}$.

Discussion and Conclusion

We investigated the club location, club age, position effect on MLS salary and its distribution. The tests showed that the two club effects don't have significant impacts on salary. There are several possible explanations for this. In regards to location, one possible explanation is that a soccer club plays both home and away games, the club's fanbases may not limit to its home state. TV broadcast and live streaming can bring new fans and incomes to clubs even located in a small town. Therefore, if an MLS club has fans all over the states, investment in clubs located on the small market can be as profitable as the clubs located on the large market. A possible explanation for the lack of significance when discussing the age of the clubs is that there are conflicting thoughts on how money is spent by each club. Older clubs may have more money to spend but they may also have a more reliable farm system that can bring in cheaper players so they do not need to spend it. Meanwhile, newer clubs may not have the money that older clubs do, but in order to have success they have to spend more to bring in flashy, skilled players. Also, clubs are limited by a salary cap so they will not be able to spend above a certain threshold, preventing a wider distribution that could lead to larger differences.

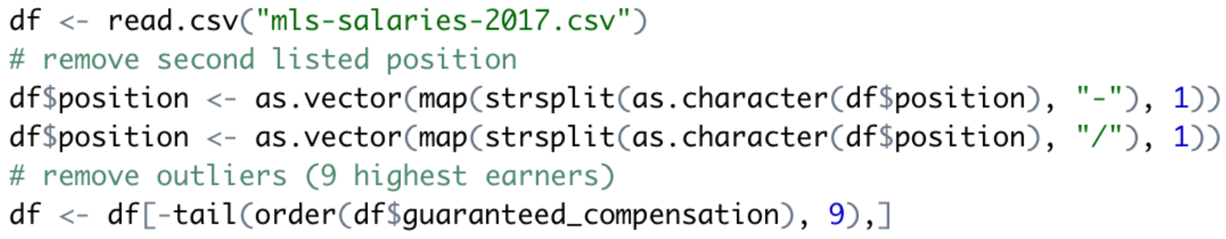
In contrast, the position effect is significantly influential. Forwards get paid higher than defenders and non goalkeepers get paid higher than goalkeepers. Generally, fans prefer watching attacking rather than defending plays, because shootings and goals bring more excitement. That may be the reason why clubs are willing to pay more to non-goalkeepers and forwards.

The salary distribution has a exponential pattern with λ equals 4.455*10^{-6}. It indicates that the salaries increase rapidly for top-tier players, and the top-tier players get paid significantly higher than average players.

For future study, perhaps the outliers that we removed could be included in the analysis. We can try performing log transformations to reduce the skewness of the data, so these outliers could be included, as the data is currently being used untransformed.

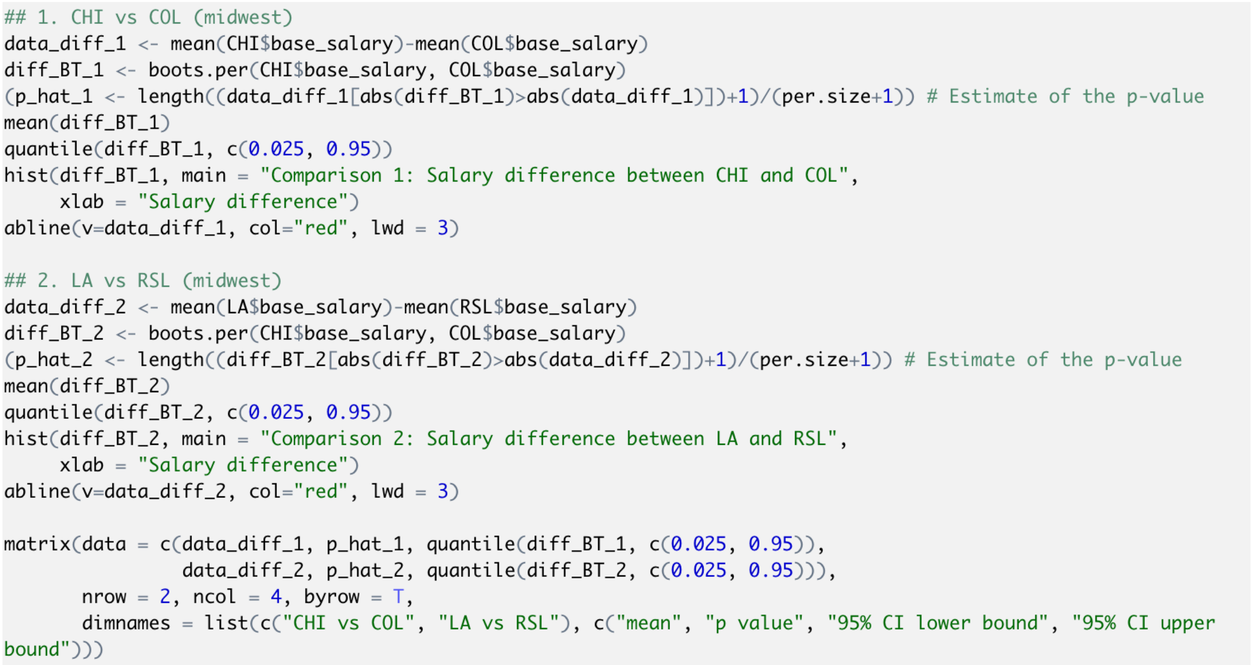
Appendix

1. Data Processing

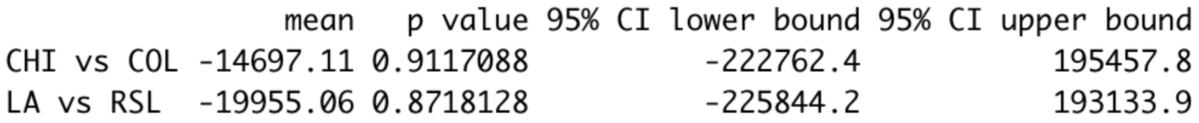


```
df <- read.csv("mls-salaries-2017.csv")
# remove second listed position
df$position <- as.vector(map(strsplit(as.character(df$position), "-"), 1))
df$position <- as.vector(map(strsplit(as.character(df$position), "/"), 1))
# remove outliers (9 highest earners)
df <- df[-tail(order(df$guaranteed_compensation), 9),]
```

2. Question 1 part 1

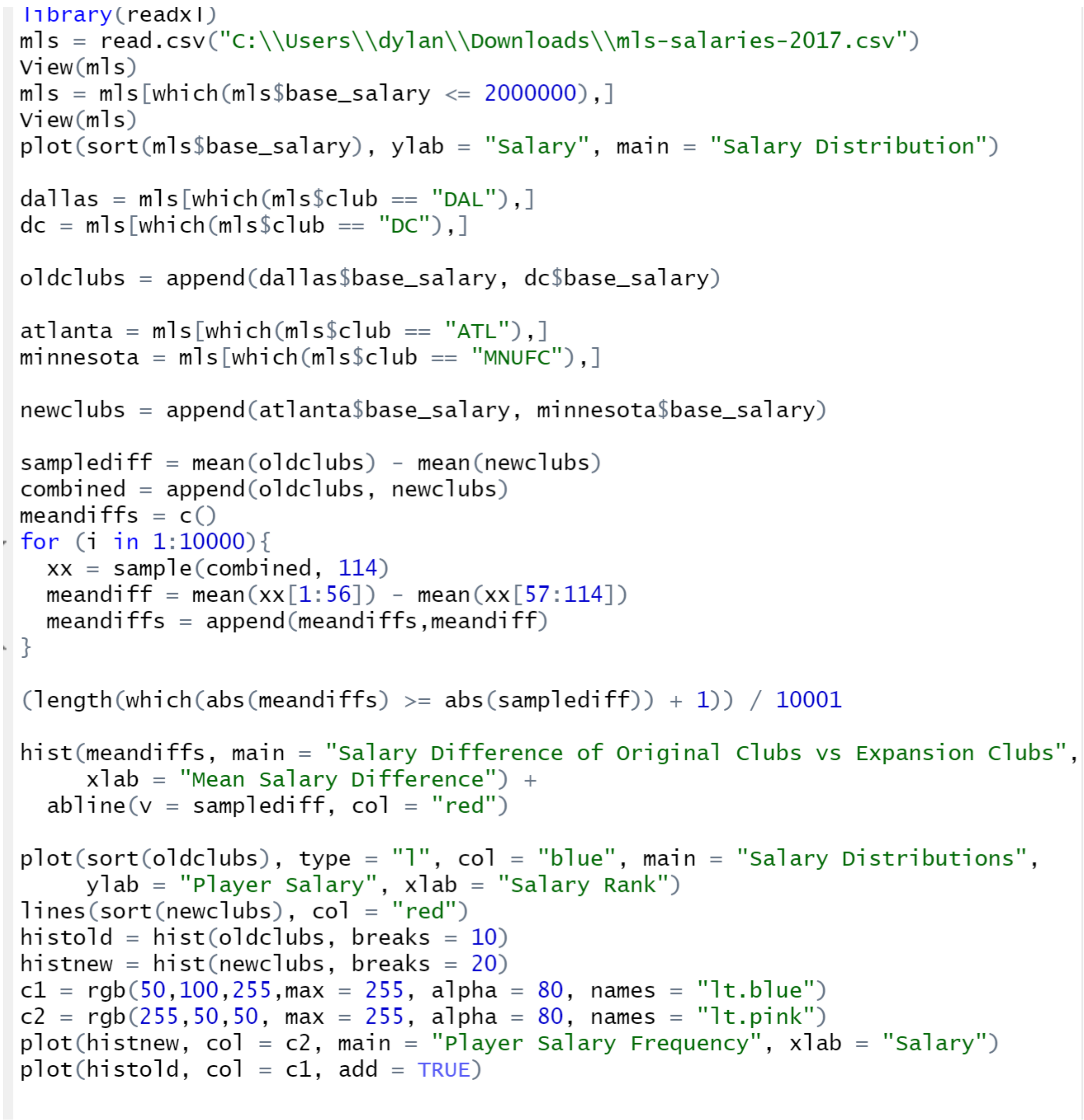


```
## 1. CHI vs COL (midwest)
data_diff_1 <- mean(CHI$base_salary)-mean(COL$base_salary)
diff_BT_1 <- boots.per(CHI$base_salary, COL$base_salary)
(p_hat_1 <- length((data_diff_1[abs(diff_BT_1)>abs(data_diff_1)])+1)/(per.size+1)) # Estimate of the p-value
mean(diff_BT_1)
quantile(diff_BT_1, c(0.025, 0.95))
hist(diff_BT_1, main = "Comparison 1: Salary difference between CHI and COL",
     xlab = "Salary difference")
abline(v=data_diff_1, col="red", lwd = 3)

## 2. LA vs RSL (midwest)
data_diff_2 <- mean(LA$base_salary)-mean(RSL$base_salary)
diff_BT_2 <- boots.per(CHI$base_salary, COL$base_salary)
(p_hat_2 <- length((diff_BT_2[abs(diff_BT_2)>abs(data_diff_2)])+1)/(per.size+1)) # Estimate of the p-value
mean(diff_BT_2)
quantile(diff_BT_2, c(0.025, 0.95))
hist(diff_BT_2, main = "Comparison 2: Salary difference between LA and RSL",
     xlab = "Salary difference")
abline(v=data_diff_2, col="red", lwd = 3)

matrix(data = c(data_diff_1, p_hat_1, quantile(diff_BT_1, c(0.025, 0.95)),
                data_diff_2, p_hat_2, quantile(diff_BT_2, c(0.025, 0.95))),
       nrow = 2, ncol = 4, byrow = T,
       dimnames = list(c("CHI vs COL", "LA vs RSL"), c("mean", "p value", "95% CI lower bound", "95% CI upper
bound")))
```



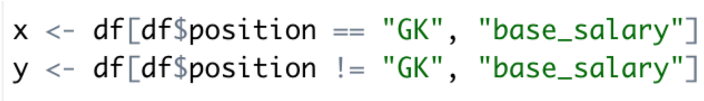
```
                mean   p value 95% CI lower bound 95% CI upper bound
CHI vs COL -14697.11 0.9117088          -222762.4           195457.8
LA vs RSL  -19955.06 0.8718128          -225844.2           193133.9
```

3. Question 1 Part 2

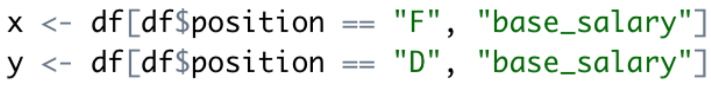


```
library(readxl)
mls = read.csv("C:\\Users\\dylan\\Downloads\\mls-salaries-2017.csv")
View(mls)
mls = mls[which(mls$base_salary <= 2000000),]
View(mls)
plot(sort(mls$base_salary), ylab = "Salary", main = "Salary Distribution")

dallas = mls[which(mls$club == "DAL"),]
dc = mls[which(mls$club == "DC"),]

oldclubs = append(dallas$base_salary, dc$base_salary)

atlanta = mls[which(mls$club == "ATL"),]
minnesota = mls[which(mls$club == "MNUFC"),]

newclubs = append(atlanta$base_salary, minnesota$base_salary)

samplediff = mean(oldclubs) - mean(newclubs)
combined = append(oldclubs, newclubs)
meandiffs = c()
for (i in 1:10000){
  xx = sample(combined, 114)
  meandiff = mean(xx[1:56]) - mean(xx[57:114])
  meandiffs = append(meandiffs,meandiff)
}

(length(which(abs(meandiffs) >= abs(samplediff)) + 1)) / 10001

hist(meandiffs, main = "Salary Difference of Original Clubs vs Expansion Clubs",
     xlab = "Mean Salary Difference") +
  abline(v = samplediff, col = "red")

plot(sort(oldclubs), type = "l", col = "blue", main = "Salary Distributions",
     ylab = "Player Salary", xlab = "Salary Rank")
lines(sort(newclubs), col = "red")
histold = hist(oldclubs, breaks = 10)
histnew = hist(newclubs, breaks = 20)
c1 = rgb(50,100,255,max = 255, alpha = 80, names = "lt.blue")
c2 = rgb(255,50,50, max = 255, alpha = 80, names = "lt.pink")
plot(histnew, col = c2, main = "Player Salary Frequency", xlab = "Salary")
plot(histold, col = c1, add = TRUE)
```

4. Question 2
   a. Separate the data into groups as necessary for a permutation test

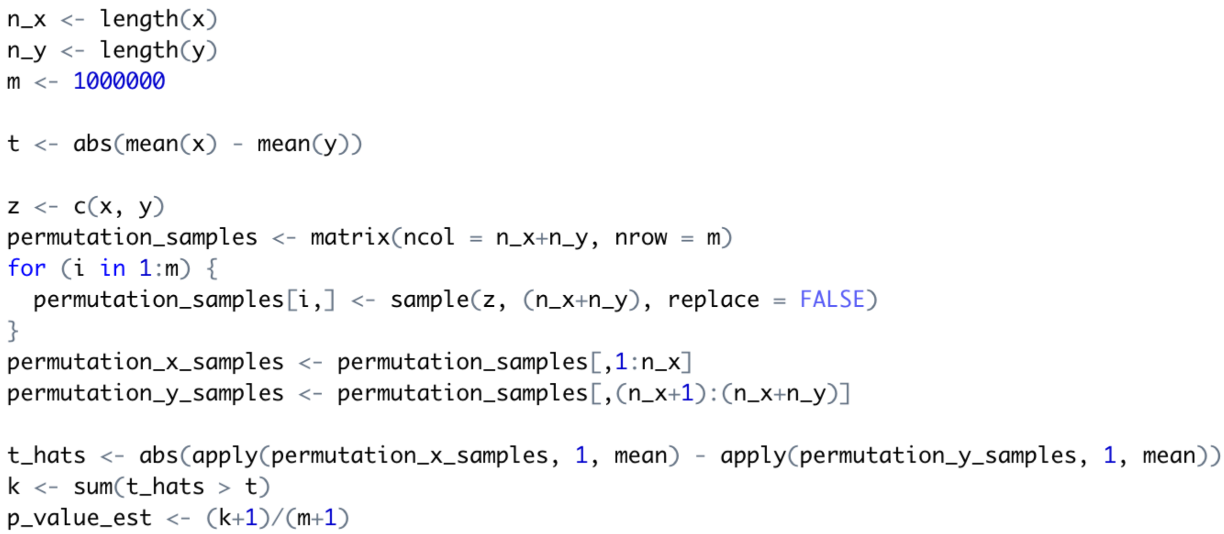


```
x <- df[df$position == "GK", "base_salary"]
y <- df[df$position != "GK", "base_salary"]
```



```
x <- df[df$position == "F", "base_salary"]
y <- df[df$position == "D", "base_salary"]
```

   b. Perform permutation test



```r
n_x <- length(x)
n_y <- length(y)
m <- 1000000

t <- abs(mean(x) - mean(y))

z <- c(x, y)
permutation_samples <- matrix(ncol = n_x+n_y, nrow = m)
for (i in 1:m) {
  permutation_samples[i,] <- sample(z, (n_x+n_y), replace = FALSE)
}
permutation_x_samples <- permutation_samples[,1:n_x]
permutation_y_samples <- permutation_samples[,(n_x+1):(n_x+n_y)]

t_hats <- abs(apply(permutation_x_samples, 1, mean) - apply(permutation_y_samples, 1, mean))
k <- sum(t_hats > t)
p_value_est <- (k+1)/(m+1)
```

5. Question 3

```r
salary.adjust <- mls$base_salary
n <- length(salary.adjust)
xbar <- mean(salary.adjust)
s <- sd(salary.adjust)
logxbar <- mean(log(salary.adjust))
## 1 lambda of rexp()
lambda <- 1/xbar
## 2 alpha and beta of rgamma()
a <- 0.5 / (log(xbar) - logxbar)
b <- a / xbar
## 3 mu and sd of rnorm()
mu <- xbar
sd <- s
## plotting
ecdf(salary)
plot(ecdf(salary), col = 1, lwd = 3, main = "CDF of salary without outliers", xlab = "salary")
curve(pexp(x,lambda), add = T,col = 2,  lwd = 2)
curve(pgamma(x,a,b), add = T,col = 3,  lwd = 2)
curve(pnorm(x,mu,sd), add = T,col = 4,  lwd = 2)
legend("bottomright", legend = c("ecdf", "F(x) of exp", "F(x) of gamma", "F(x) of normal"),
       lwd = 3, col = 1:4)

exp.data <- rexp(n,lambda)
gamma.data <- rgamma(n,a,b)
normal.data <- rnorm(n, mu, sd)
ks.test(salary.adjust, exp.data)
ks.test(salary.adjust, gamma.data)
ks.test(salary.adjust, normal.data)
```

```
p-value will be approximate in the presence of ties
        Two-sample Kolmogorov-Smirnov test

data:  salary.adjust and exp.data
D = 0.23888, p-value = 1.776e-15
alternative hypothesis: two-sided

p-value will be approximate in the presence of ties
        Two-sample Kolmogorov-Smirnov test

data:  salary.adjust and gamma.data
D = 0.16639, p-value = 1.006e-07
alternative hypothesis: two-sided

p-value will be approximate in the presence of ties
        Two-sample Kolmogorov-Smirnov test

data:  salary.adjust and normal.data
D = 0.28007, p-value < 2.2e-16
alternative hypothesis: two-sided
```

References

Crawford, Chris. “U.S. Major League Soccer Salaries.” *Kaggle*, 13 July 2017, www.kaggle.com/crawford/us-major-league-soccer-salaries?select=mls-salaries-2017.csv.

staff, MLSsoccer. “MLS Expansion: New Timeline Released for Inaugural Season of Newest Clubs.” *Mlssoccer*, 13 Aug. 2020, www.mlssoccer.com/news/mls-expansion-new-timeline-released-inaugural-season-newest-clubs.

1. *Descriptive Title*

   Analysis of the Salary Distribution Across Teams and Positions in Major League Soccer in 2017

2. *Group Members*

   Dylan Hull, Jueyao Liu, Colin Zhang

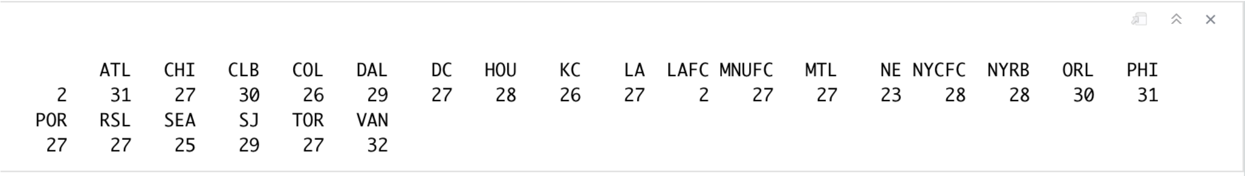
3. *If your dataset, or similar data sets, have been analyzed elsewhere, give a full citation of this companion paper/analysis. Or maybe a Wikipedia reference, to a page that prompted your interest and questions.*

   Not applicable.

4. *A clearly defined explanation of the scientific problem and questions of interest.*

   We proposed three scientific problems, all based around the salary of players in the MLS.

   1. We want to look at the difference in salaries between players of different clubs. It would be interesting to determine if certain clubs spend more on their players and if the financial differences are significant.
      a. Pick maybe a couple pairs of clubs?
      b. Atlanta (ATL) and Minnesota (MNUFC) vs DC and DAL
      c. Big market vs small market:
         i. CHI vs COL (midwest)
         ii. LA vs RSL



```
        ATL   CHI   CLB   COL   DAL    DC   HOU    KC    LA  LAFC MNUFC   MTL    NE NYCFC  NYRB   ORL   PHI
    2    31    27    30    26    29    27    28    26    27     2    27    27    23    28    28    30    31
  POR   RSL   SEA    SJ   TOR   VAN
   27    27    25    29    27    32
```

   2. Another question that would be of interest is if there is a difference in salary for players of different positions. There are four main positions in soccer (forward, midfield, defender, goalkeeper), and it would be interesting to see which, if any, positions are valued higher than the others when it comes to salary.
      a. Goalkeeper vs non-goalkeeper
      b. Forward vs defender

3. We want to determine the distribution of the salaries, i.e., we want to know whether salaries come from a normal, exponential, gamma, or some other distribution.

5. *The source of your data.*

https://www.kaggle.com/crawford/us-major-league-soccer-salaries?select=mls-salaries-2017.csv

6. *A description of your data.*

Dataset Size: Around *n* = 600

Number of Variables: 6

Variables of Interest: Club, position, base salary, guaranteed compensation. Club and position are categorical variables, while base salary and guaranteed compensations are numerical variables.

Unusual Challenges: There are a few rows that contain null values, but not much of the data is missing. The data for base salary and guaranteed compensation both have high outliers, as some players receive notably more pay than the vast majority. Even after removing certain outliers, the data is heavily skewed to the right.

7. *Proposed Analysis: Broadly, what approach will you use to analyze the data in order to address your questions of interest? How will you use some form of Monte Carlo sampling or resampling (e.g. bootstrap) in performing, or in assessing, your statistical analysis?*

To reduce the impact of extreme outliers, we are planning to remove the 9 players whose salaries are higher than 2*10^6.

Then, for question 1 and question 2, we will investigate using permutation and parametric bootstrap.

For question 2 particularly, we need to modify the position variable before the test. There are 4 basic positions: GK(goalkeeper), D(defender), M(midfield), F(forward). However, some players are assigned to multiple positions (eg. M-F and D-M). So for sake of conciseness, we will assign M-F to F and D-M to M.

For the last question, we will use Monte Carlo to test the distribution of salaries. If the result shows that the 2017 MLS salary distribution comes from some normal, exponential, gamma, or some other distribution, we will then compute the bootstrap confidence interval and other statistics.

Question:

1. The difference between salaries between clubs. (Permutation bootstrap)
2. The difference between salaries between positions (Defender, Midfield, vs Forward). (bootstrap)
    a. There are 4 main positions: GK(goalkeeper), D(defender), M(midfield), F(forward).
    b. Some players are assign to multiple positions (eg. M-F), she suggested we can assign one from the 4 main positions to those players. For example, always assign M-F to M.

3. Interested in if the overall salaries come from a normal/exponential/ gamma distribution (Monte Carlo)?
    a. She suggested that we could also try to fit a gamma distribution. And we can compute a (bootstrap) confidence interval if it does come from an exponential distribution.

Other advice:

4. Possible outliers: we should remove the outlier for our project and have to mention it in our proposal.

https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017
Soccer match dataset home/away scores… neutral sites, location, tournament…

https://www.kaggle.com/vardan95ghazaryan/top-250-football-transfers-from-2000-to-2018
The columns contain the following information: the name of a football player, selling team and league, the league and team where a player is sold, an estimated market value of a player, an actual value of a transfer, the position of a player and season when a transfer took place.

https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

This dataset contains about 10 years of daily weather observations from many locations across Australia.

RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

This dataset is huge so maybe not but still interesting stuff lots of variables to work with

https://www.kaggle.com/uciml/forest-cover-type-dataset

- Can you build a model that predicts what types of trees grow in an area based on the surrounding characteristics? A past Kaggle competition project on this topic can be found here.
- What kinds of trees are most common in the Roosevelt National Forest?
- Which tree types can grow in more diverse environments? Are there certain tree types that are sensitive to an environmental factor, such as elevation or soil type?

https://www.kaggle.com/residentmario/ramen-ratings
Ramen ratings

https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

This data set contains every soccer player available in FIFA 15, 16, 17, 18, 19, and also FIFA 20. The data set includes more than 100 attributes of players, such as player overall score, player positions, player attribute, and player personal data. Below are some details of the variables.

- Overall score and potential score:
- Player positions, with the role in the club and in the national team.
- Player attributes with statistics as Attacking, Skills, Defense, Mentality, GK Skills, etc.
- Player personal data like Nationality, Club, DateOfBirth, Wage, Salary, etc.

The data set is **massive**. So, only investigation of data from a single year is still profound. FIFA 20 is preferred since it is the closest year. **This data set was used for my STAT 423 last**

quarter, so I am pretty familiar with it but it is massive and always used in some linear regression tests. So, it may not fitted with this project, but i still list it here

https://www.kaggle.com/kimjihoo/coronavirusdataset?select=TimeGender.csv
COVID-19 has infected more than 10,000 people in South Korea.
KCDC (Korea Centers for Disease Control & Prevention) announces the information of COVID-19 quickly and transparently.
We make a structured dataset based on the report materials of KCDC and local governments.
Also, we analyze and visualize the data using various data mining or visualization techniques.

This data set is some and has many missing data, so may be a good candidate for Monte Carlo simulation.

https://www.kaggle.com/crawford/us-major-league-soccer-salaries?select=mls-salaries-2017.csv

This data set contains the Major League Soccer Union salaries of every MLS player each year. This is a collection of salaries from 2007 to 2017.

Since the data set is not super large (n=600), I think it is a good data set to apply the bootstrap method. Eg. We can apply the bootstrap sampling to test if the salaries between two clubs are differ from each other. (since the original distribution of salaries is likely not to be normal)