Data Analysis Project - Baseball Data

Henry Daniel

4/23/2023

Loading packages

library(tidyverse) library(openintro) library(infer) library(MASS) library(infer) library(GGally)

Loading dataset

BaseballTeamPayroll = BaseballTeamPayroll.csv

Introduction

For this project, we will making use of BaseballTeamPayroll dataset from the MASS package. BaseballTeam-Payroll is a data set about baseball teams in the USA from year 1985 to 2016. the data set has 918 rows and 50 columns. It contains information such as payroll, wins, losses, runs, hits, home runs, win percentage and so on. The data set contains 348 missing values. All of the missing values come from the division win column and the world series win column. The rest of the columns have 0 missing values. The goal of this project is to understand the relationship between Payroll and various other variables within the data set through hypothesis tests and predictive modeling.

Research Question

Based on the introduction, we would like to study the Payroll variable. We would like to ask the following questions:

Do the teams with higher payrolls win their divison more on average?

Can a teams payroll be explained or predicted by HR, W, ERA, WinPercentage, DivID and H?

Hypothesis

We will test out the first research question using hypothesis testing with a 95% confidence level. Our hypothesis is that the teams a higher payroll also win their divison more on average. In mathematical form it would be

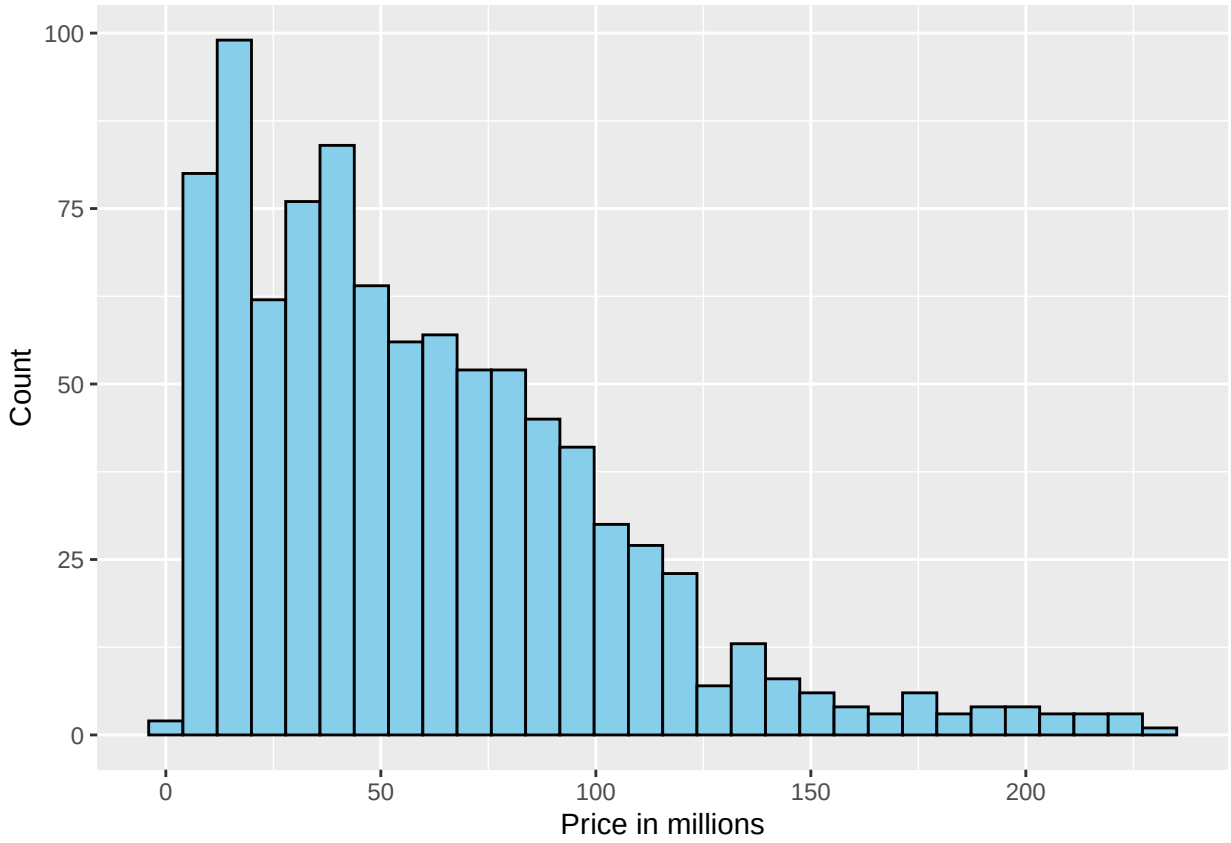
Ho: payroll win - payroll lose = 0, Ha: payroll win - payroll lose 0

Exploratory Data Analysis

First, we need to look at the distribution of our main variable before we begin any sort of analysis. We will make use of a histogram plot

```
BaseballTeamPayroll %>%
 ggplot(aes(x = payroll))+
 geom_histogram(color = "black", fill = "skyblue")+
 labs(x = "Price in millions",
 y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It looks like the teams payrolls are skewed to the right. This is evident because of the presence of the outliers and tail towards the right. Let's look at some summary statistics of our main variable.

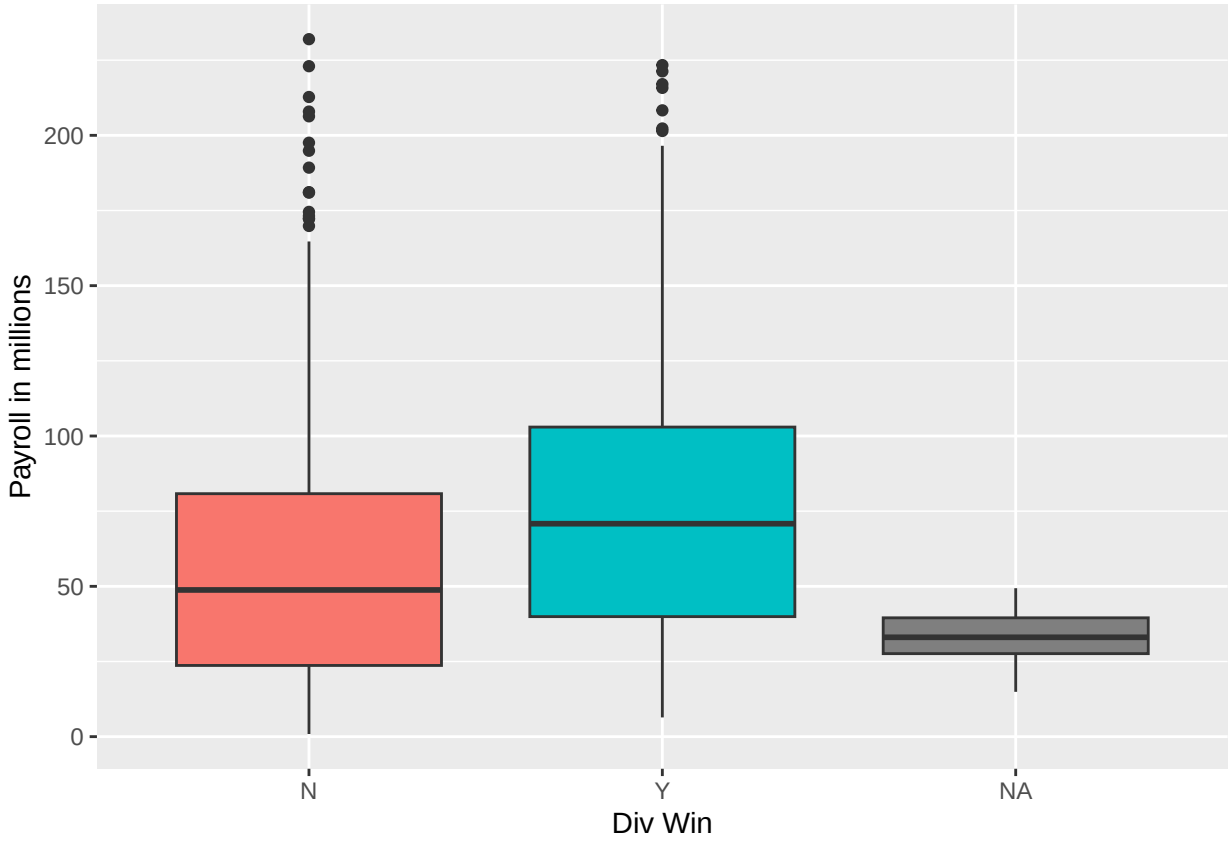
```
BaseballTeamPayroll %>%
summarise(Mean = mean(payroll),
SD = sd(payroll),
Min = min(payroll),
Q1 = quantile(payroll, 0.25),
Median = median(payroll),
Q3 = quantile(payroll,.25),
Max = max(payroll))
```

```
##       Mean       SD  Min       Q1   Median       Q3      Max
## 1 60.04263 43.30992 0.88 25.43571 50.53732 25.43571 231.9789
```

The right skewness is supported by the fact that the mean payroll is quite higher than the median. Payroll has a large variability with a standard deviation of 43.3. The minimum payroll is .88 (880,000) and the maximum payroll was 231.9789 (231,978,900).

Let's look at DivWin, the variable of interest for our first question and price.

```
BaseballTeamPayroll%>%
 ggplot(aes(x = DivWin, y = payroll, fill = DivWin))+
 geom_boxplot(show.legend = F)+
 labs(x = "Div Win",
 y = "Payroll in millions")
```

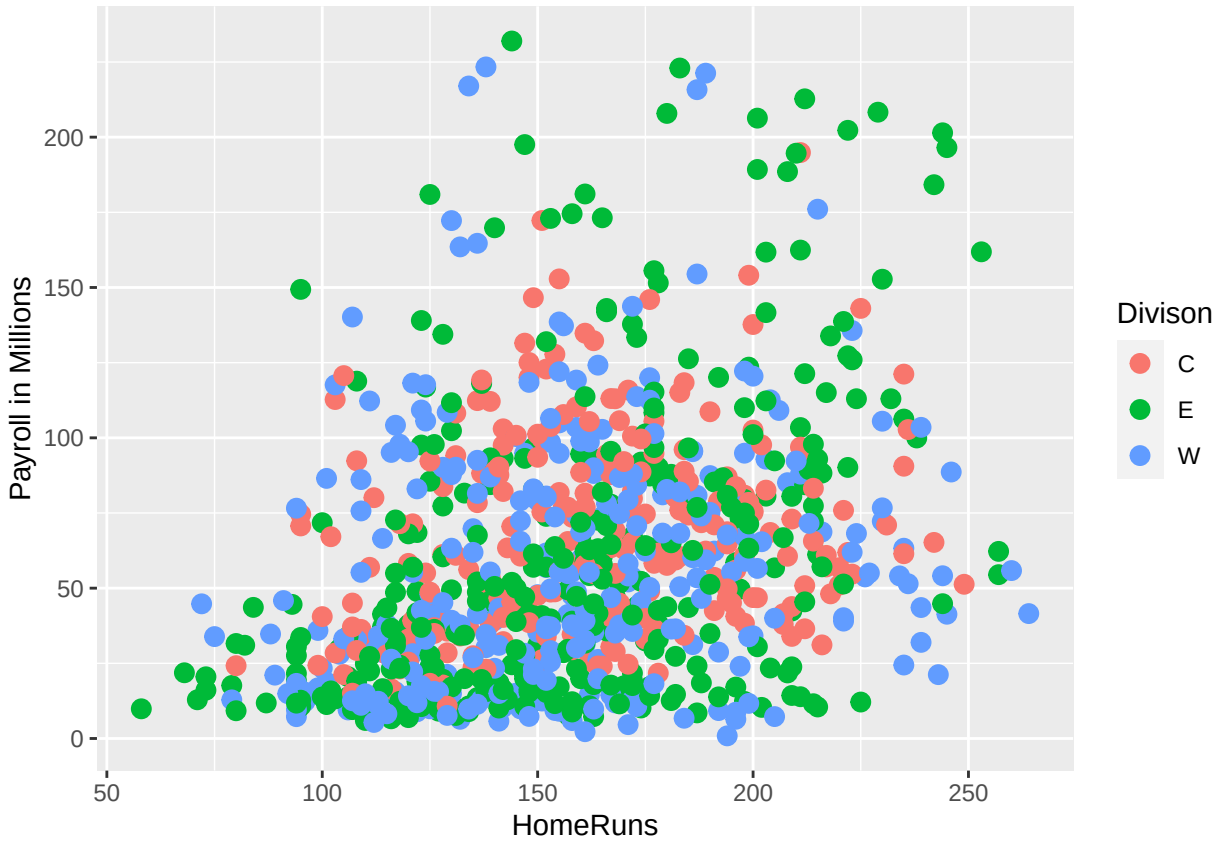


We can see that there is a clear difference between the two box plot. Teams that win their division seem to have a higher payroll.But this plot isn't enough evidence for us to deem that the difference is significant enough.

Before we can answer the second research question using predictive modeling. We need to understand the relationship between the main response variable (Payroll) and the explanatory variables (HR, W, ERA, WinPercentage, DivID and H). We can understand their relationship making use of ggplot() function to make plots and find their correlation

Payroll vs Home runs

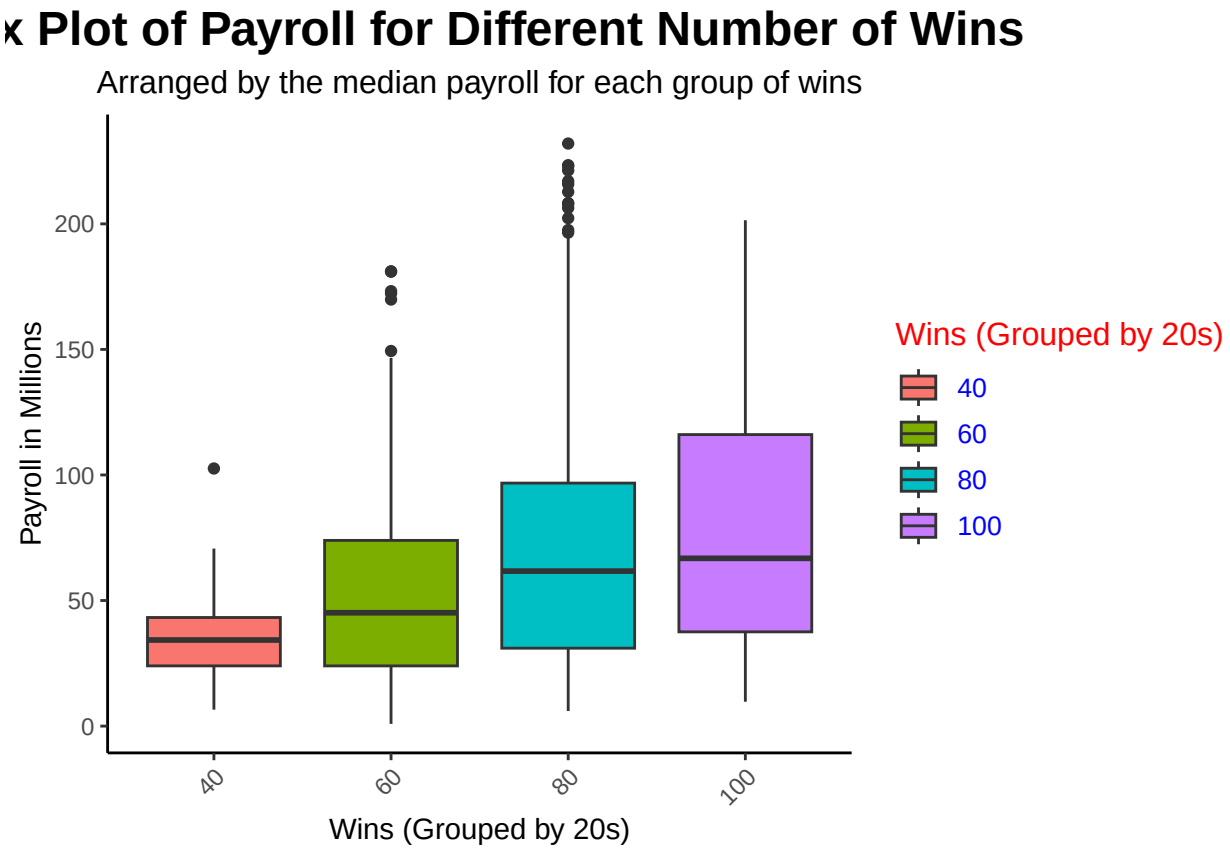
```
BaseballTeamPayroll %>%
 ggplot(aes(x = HR, y = payroll, col = divID ))+
 geom_point(size = 3)+
 labs(x = "HomeRuns",
 y = "Payroll in Millions",
 col = "Divison")
```



In the plot above we plotted home runs against payroll. From this chart we cannot really see a relationship between the number of home runs and a teams payroll.

payroll vs wins

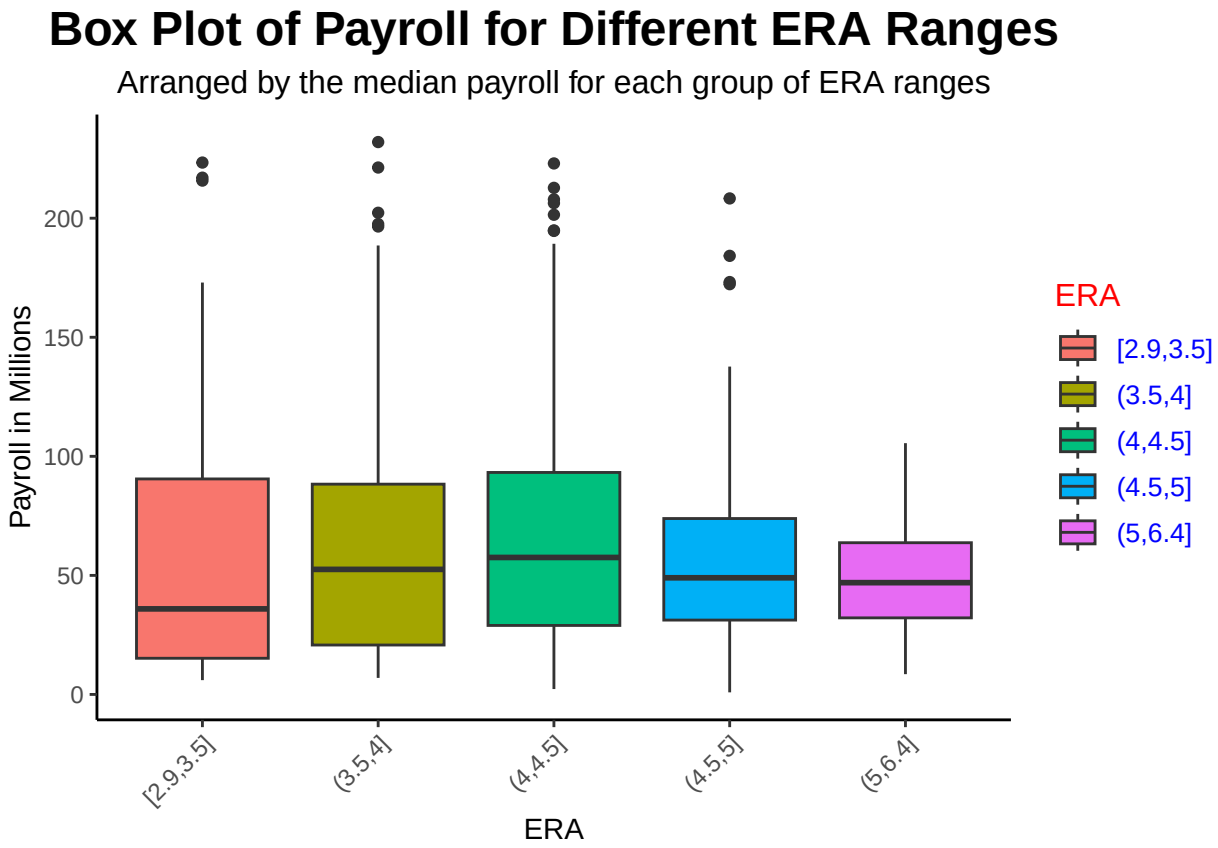
```
BaseballTeamPayroll %>%
  mutate(W_group = cut(W, breaks = seq(40, 120, 20), include.lowest = TRUE, labels = seq(40, 100, 20)))
  ggplot(aes(x = W_group, y = payroll, fill = as.factor(W_group))) +
  geom_boxplot() +
  labs(x = "Wins (Grouped by 20s)",
       y = "Payroll in Millions",
       title = "Box Plot of Payroll for Different Number of Wins",
       subtitle = "Arranged by the median payroll for each group of wins") +
  scale_x_discrete(limits = as.character(seq(40, 100, 20))) +
  scale_fill_discrete(name = "Wins (Grouped by 20s)") +
  theme_classic() +
  theme(legend.title = element_text(color = "red", size = 12),
        legend.text = element_text(color = "blue", size = 10),
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", size = 18, hjust = 0.5),
        plot.subtitle = element_text(size = 12, hjust = 0.5),
        strip.text.x = element_text(face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



For the above side-by-side box plot, we arranged the number of wins in groups of 20s. From the plot you can see that teams with around 100 wins have the highest median payroll and teams that have around 40 wins have the lowest median payroll. Teams that have around 80 wins ave the second highest median payroll but also the highest variability. This could indicate that there are various other factors that would have affect on the number of wins besides payroll.

Payroll vs ERA

```
BaseballTeamPayroll %>%
  mutate(ERA_group = cut(ERA, breaks = c(2.9, 3.5, 4, 4.5, 5, 6.4), include.lowest = TRUE)) %>%
  ggplot(aes(x = ERA_group, y = payroll, fill = as.factor(ERA_group))) +
  geom_boxplot() +
  labs(x = "ERA",
       y = "Payroll in Millions",
       title = "Box Plot of Payroll for Different ERA Ranges",
       subtitle = "Arranged by the median payroll for each group of ERA ranges") +
  scale_fill_discrete(name = "ERA") +
  theme_classic() +
  theme(legend.title = element_text(color = "red", size = 12),
        legend.text = element_text(color = "blue", size = 10),
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", size = 18, hjust = 0.5),
        plot.subtitle = element_text(size = 12, hjust = 0.5),
        strip.text.x = element_text(face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



This plot shows Payroll vs ERA. There are 5 different era ranges. The era range with the highest median payroll is 4 to 4.5 ERA. the ERA range with the lowest median payroll is 2.9 to 3.5. This is interesting because the lower the ERA the better. From this graph I the take away is that era dosent really affect payroll.

Compute and report summary statistics

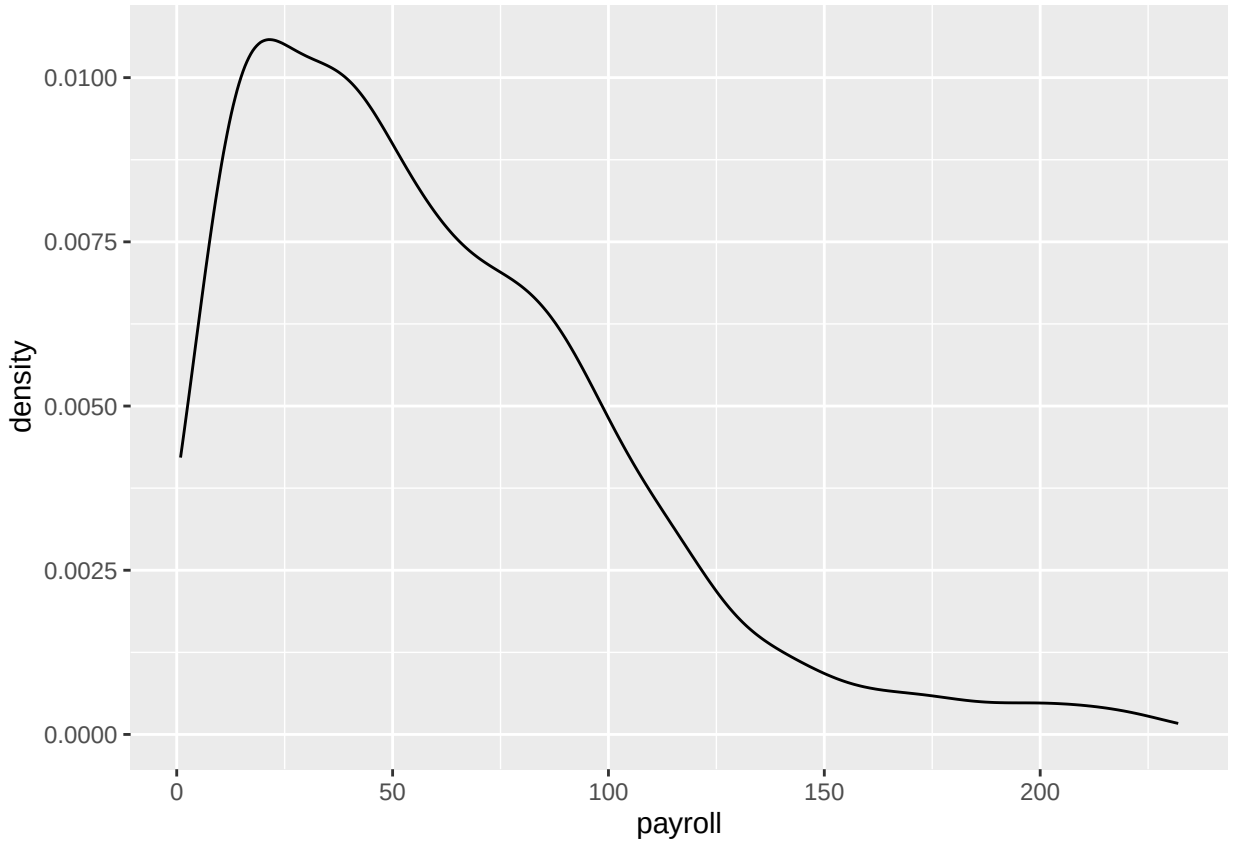
```
library(knitr)

BaseballTeamPayroll %>%
  summarise(Mean = format(mean(payroll), nsmall = 2, big.mark = ","),
            SD = format(sd(payroll), nsmall = 2),
            Min = format(min(payroll), big.mark = ","),
            Q1 = format(quantile(payroll, 0.25), big.mark = ","),
            Median = format(median(payroll), nsmall = 2, big.mark = ","),
            Q3 = format(quantile(payroll,.25), big.mark = ","),
            Max = format(max(payroll), big.mark = ",")) %>%
  kable(col.names = c("Mean", "SD", "Min", "Q1", "Median", "Q3", "Max"))
```

Mean	SD	Min	Q1	Median	Q3	Max
60.04263	43.30992	0.88	25.43571	50.53732	25.43571	231.9789

Construct confidence interval for estimating the population mean of the main response variable

```
BaseballTeamPayroll %>%
  filter(!is.na(payroll))%>%
  ggplot(aes(x=payroll)) +
  geom_density()
```



The distribution does not look like normal distribution, but the sample size is sufficiently large (n = 918) for the application of central limit theorem normal approximation to hold.

$H0 : \mu = 60.04263 \; Vs \; Ha : \mu \neq 60.04263$

```
BaseballTeamPayroll %>%
  filter(!is.na(payroll))%>%
  t_test(
    response = payroll,
    conf_int = TRUE,
    conf_level = 0.95,
    mu = 60.04263,
    alternative = "two-sided"
  )
```

```
## # A tibble: 1 x 7
```

```
##    statistic  t_df p_value alternative estimate lower_ci upper_ci
##        <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1 0.00000184   917    1.00 two.sided       60.0     57.2     62.8
```

We are 95% confident that the average payroll for MLB teams is contained before the interval \57.23728 and \62.84799.

Two that display the association between the response variable and two explanatory I think might correlate.

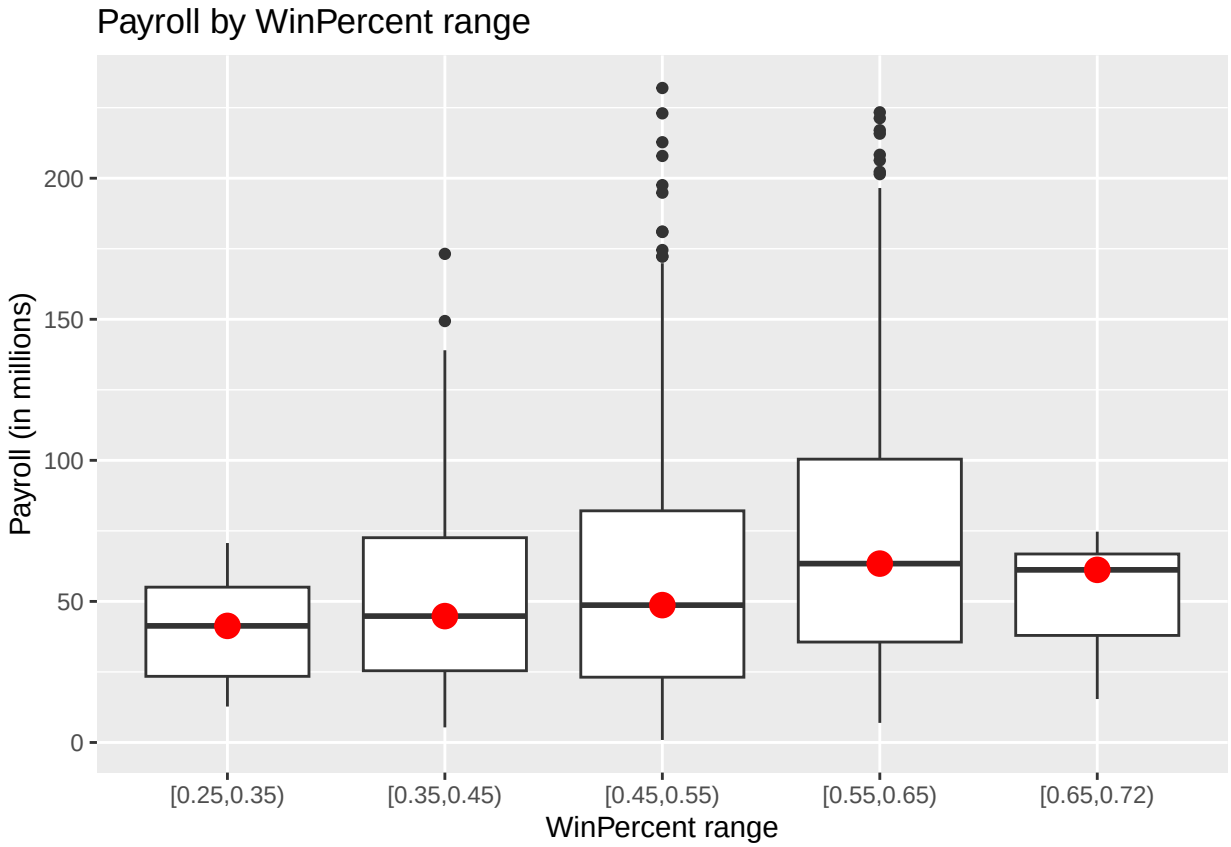
Graph 1

```
library(dplyr)
library(ggplot2)

BaseballTeamPayroll <- BaseballTeamPayroll %>%
  mutate(WinPercent_range = cut(WinPercent,
breaks = c(0.25, 0.35, 0.45, 0.55, 0.65, 0.72), right = FALSE))

payroll_median <- BaseballTeamPayroll %>%
  filter(!is.na(payroll)) %>%
  group_by(WinPercent_range) %>%
  summarize(median_payroll = median(payroll))

ggplot(BaseballTeamPayroll, aes(x = WinPercent_range, y = payroll)) +
  geom_boxplot() +
  geom_point(data = payroll_median, aes(x = WinPercent_range, y = median_payroll), color = "red", size =
  xlab("WinPercent range") +
  ylab("Payroll (in millions)") +
  ggtitle("Payroll by WinPercent range")
```



This graph can be is used to analyze the relationship between win percentage and payroll. From this graph we can determine that there is a correlation between the two variables and higher payroll translates to higher win percentage.

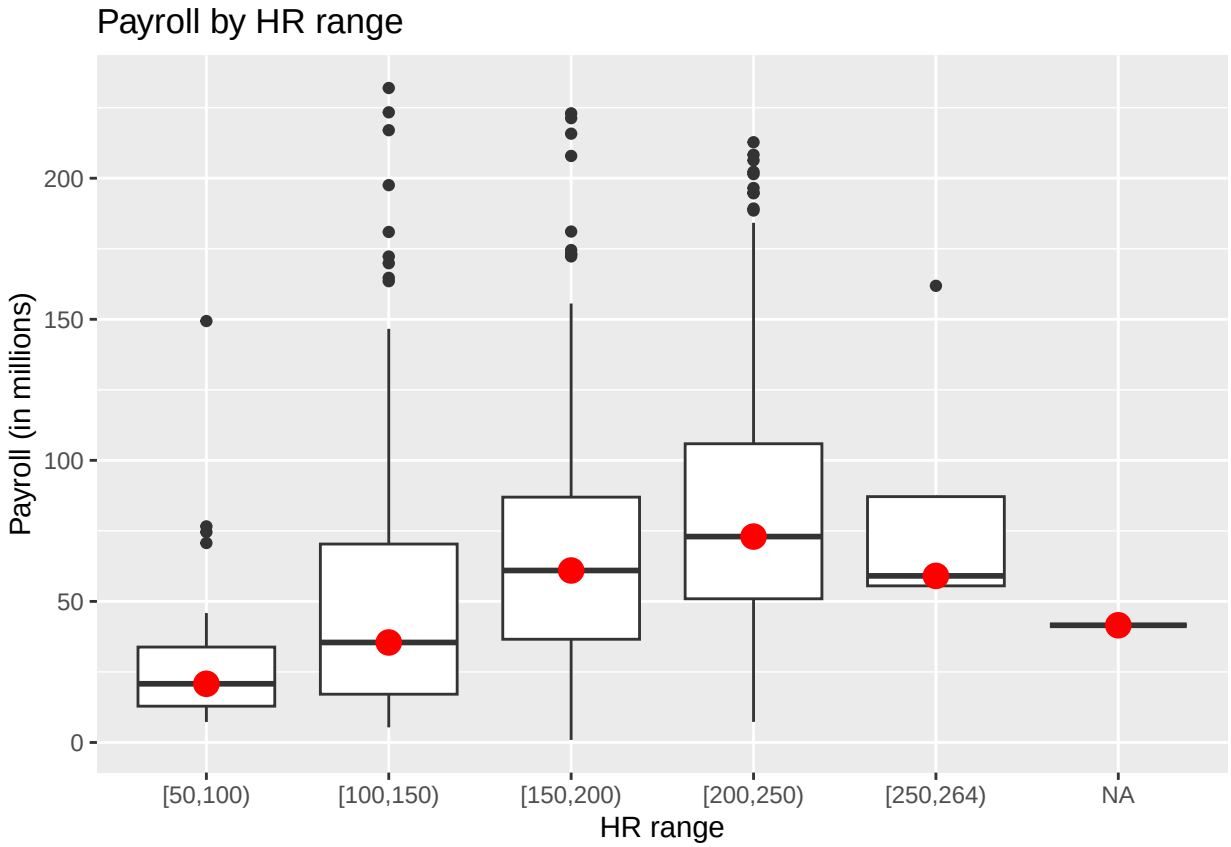
Graph 2

```
library(dplyr)
library(ggplot2)

BaseballTeamPayroll <- BaseballTeamPayroll %>%
  mutate(HR_range = cut(HR, breaks = c(50, 100, 150, 200, 250, 264), right = FALSE))

payroll_median <- BaseballTeamPayroll %>%
  filter(!is.na(payroll)) %>%
  group_by(HR_range) %>%
  summarize(median_payroll = median(payroll))

ggplot(BaseballTeamPayroll, aes(x = HR_range, y = payroll)) +
  geom_boxplot() +
  geom_point(data = payroll_median, aes(x = HR_range, y = median_payroll), color = "red", size = 4) +
  xlab("HR range") +
  ylab("Payroll (in millions)") +
  ggtitle("Payroll by HR range")
```



This is a visual representation of the median payroll value for various ranges of home runs. From this graph we can see that as the median payroll increases the amount of home runs is also increasing.

Compute summary statistics (mean and standard deviation) to summarize the response variable in your data by groups defined by the levels of one categorical variable from your dataset.

```
library(dplyr)

payroll_summary <- BaseballTeamPayroll %>%
  group_by(DivWin) %>%
  summarize(mean_payroll = mean(payroll, na.rm = TRUE),
            sd_payroll = sd(payroll, na.rm = TRUE))

print(payroll_summary)
```

```
## # A tibble: 3 x 3
##   DivWin mean_payroll sd_payroll
##   <chr>         <dbl>      <dbl>
## 1 N              56.9      40.6
## 2 Y              77.8      51.8
## 3 <NA>           33.1       8.53
```

The categorical variable I chose was if the team won their division or not in the particular season.

Compute and report the 90% confidence interval for the difference in population means.

```
BaseballTeamPayroll %>%
  filter(!is.na(payroll), !is.na(DivWin))%>%
  t_test(
    response = payroll,
    explanatory= DivWin,
    order = c("N","Y"),
    conf_int = TRUE,
    conf_level = 0.90)
```

```
## # A tibble: 1 x 7
##   statistic  t_df    p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>      <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1     -4.88  217. 0.00000201 two.sided      -20.9    -27.9    -13.8
```

I can conclude that the mean payroll for teams that wins their division is different than the mean payroll of teams a team who loses their division. Teams that win their divison tend to have a higher payroll.

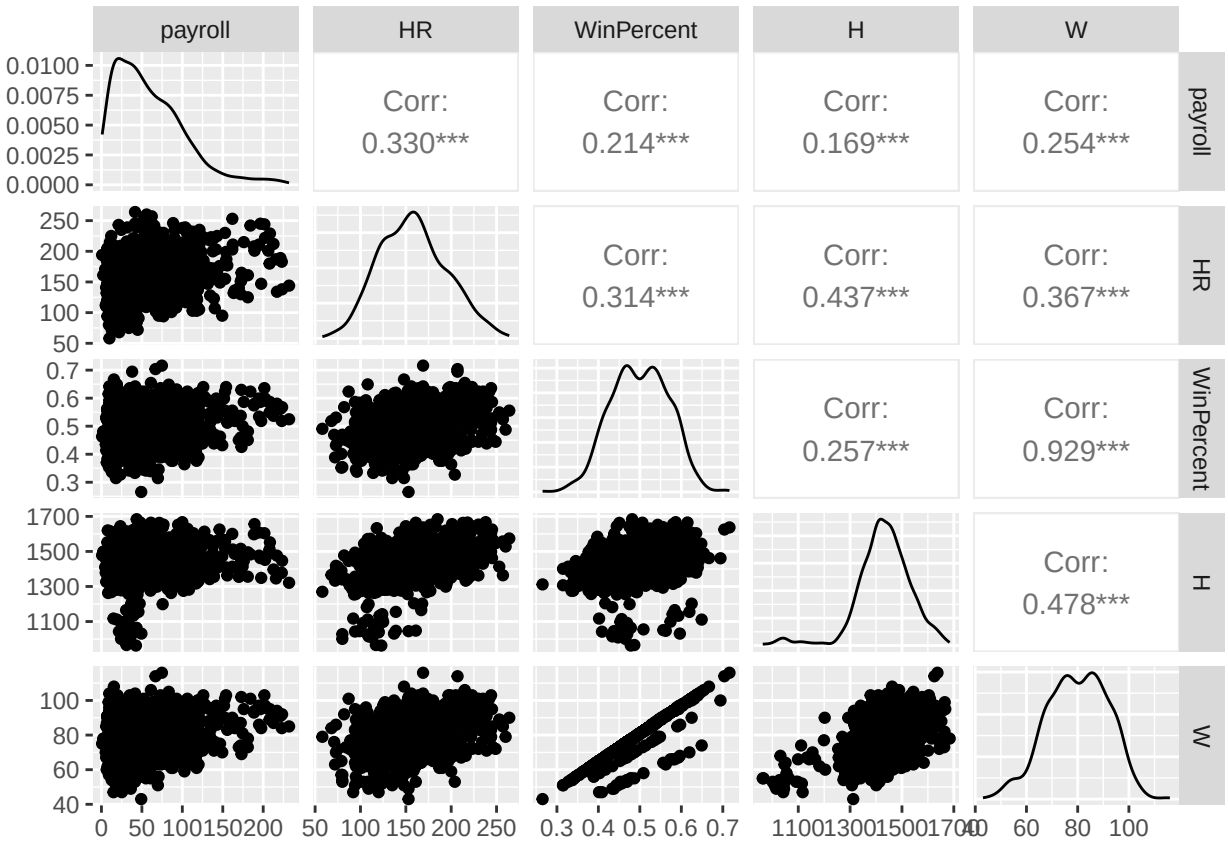
Dap Part III

Correlation refers to the statistical relationship between two random variables or bivariate data, regardless of whether the relationship is causal or not. This relationship can be measured quantitatively, using a range of values from -1 to 1. If the correlation is positive, it means that as one variable increases, the other variable tends to increase as well. Conversely, a negative correlation means that as one variable increases, the other variable tends to decrease. The strength of the correlation determines how close it is to -1 or 1, and the closer it is, the more likely it is that the relationship between the variables is a straight line with no variation. To calculate and plot the correlation between response and explanatory variables, we can use the ggpairs() function from the GGally package, which not only calculates the correlation but also provides a visual representation of the relationship. The code for this function is shown below.

1a

```
library(GGally)
BaseballTeamPayroll%>%
   dplyr::select(payroll,HR,WinPercent,H,W)%>%

 ggpairs()
```



We can see that Payroll and Homeruns have the highest positive correlation of .330. This is the highest correlation from our response variable. As you can see there are other variables with higher correlation such as the number of hits a team had and the number of wins.

1b

```
full_model = lm(payroll ~ HR + W +
 ERA + WinPercent + H, data = BaseballTeamPayroll)
summary(full_model)
```

```
##
## Call:
## lm(formula = payroll ~ HR + W + ERA + WinPercent + H, data = BaseballTeamPayroll)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.639 -29.297  -5.758  19.999 173.925
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.970979  26.306966   1.595   0.1110
## HR            0.419409   0.050615   8.286 4.15e-16 ***
## W             0.544919   0.547622   0.995   0.3200
## ERA         -12.656439   4.941969  -2.561   0.0106 *
## WinPercent  -81.839783  68.882946  -1.188   0.2351
## H             0.001648   0.024264   0.068   0.9459
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.25 on 912 degrees of freedom
## Multiple R-squared:  0.1412, Adjusted R-squared:  0.1365
## F-statistic: 29.98 on 5 and 912 DF,  p-value: < 2.2e-16
```

estimated regression equation

payroll = 41.97 + 0.419 x HR + 0.5449 x W -12.656439 x ERA - 81.84 x WinPercent + .0016 x H

1c

From the summary table above, we can see that the HR variable has a positive estimate of 0.419, which means that a unit increase in HR is associated with an increase in payroll by $419, on average, holding all other variables constant. ERA has a negative estimate of -12.656, indicating that a unit increase in ERA is associated with a decrease in payroll by $12,656, on average, holding all other variables constant. The variable W has an estimate of 0.545 and a p-value of 0.32, which suggests that it is not statistically significant and may not have a significant impact on payroll. WinPercent has an estimate of -81.84 and a p-value of 0.24, indicating that it is also not statistically significant. H has an estimate of 0.00165 and a p-value of 0.946, which suggests that it is not statistically significant and may not have a significant impact on payroll.

1d

Overall, the model has an adjusted R-squared of 0.1365, which suggests that only about 13.65% of the variation in payroll can be explained by the explanatory variables collectively. The F-statistic of 29.98 on 5 and 912 degrees of freedom indicates that the model as a whole is statistically significant, but the individual variables may not be. Further analysis may be necessary to determine which variables are truly significant in explaining the variation in payroll.

1e

From the summary output of the model, we see that the variable "W" has a p-value of 0.3200, which is greater than 0.05. Therefore, we can remove this variable from the model.

```
model_reduced <- lm(formula = payroll ~ HR + ERA + WinPercent + H, data = BaseballTeamPayroll)
summary(model_reduced)
```

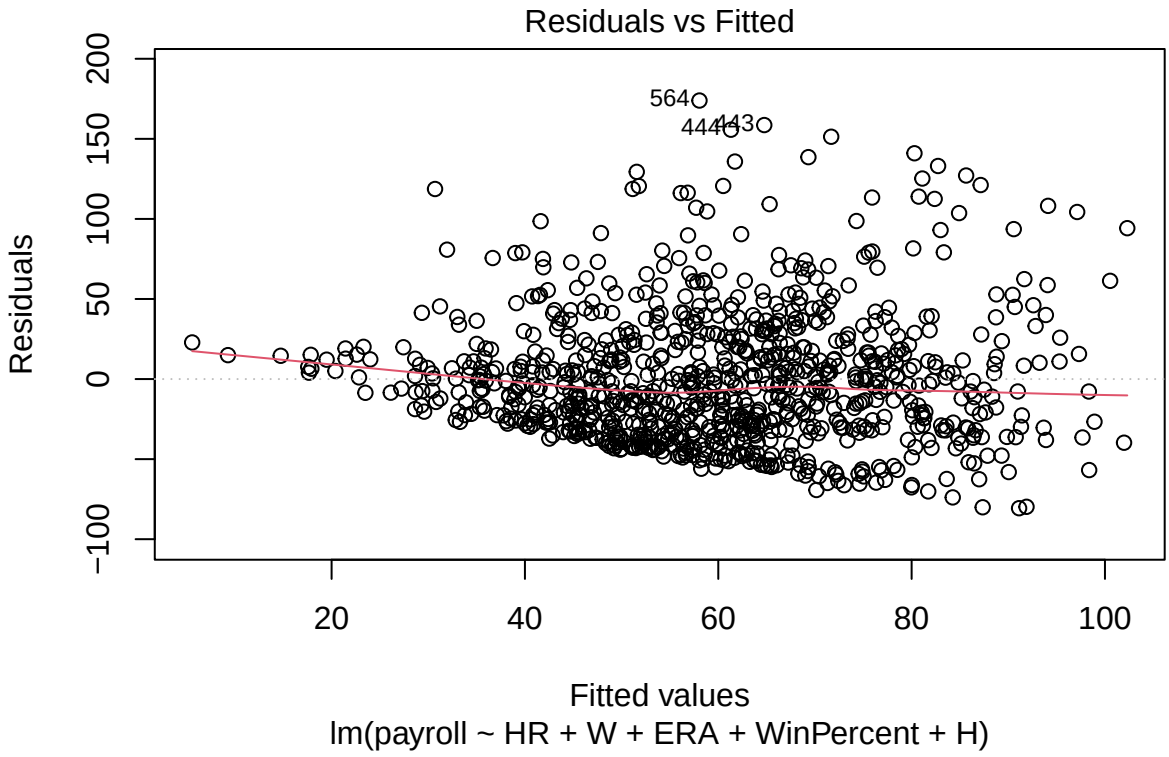
```
##
## Call:
## lm(formula = payroll ~ HR + ERA + WinPercent + H, data = BaseballTeamPayroll)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.749 -29.147  -5.603  20.274 176.649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.13174   26.02235   1.465    0.143
## HR            0.43720    0.04735   9.233  < 2e-16 ***
## ERA         -15.94133    3.67765  -4.335 1.62e-05 ***
## WinPercent  -20.29702   30.32625  -0.669    0.503
```

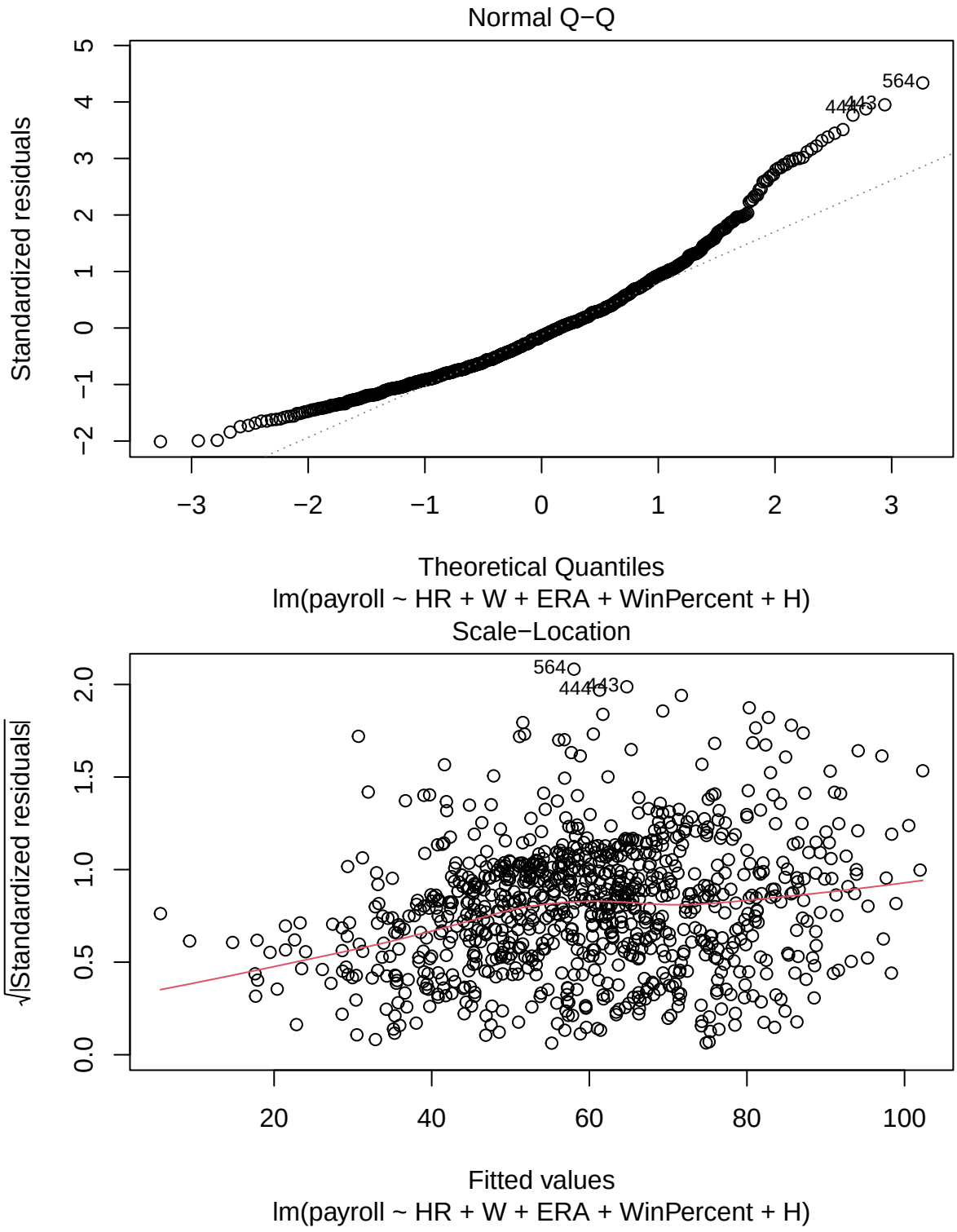
```
## H             0.02097    0.01455   1.441    0.150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.25 on 913 degrees of freedom
## Multiple R-squared:  0.1402, Adjusted R-squared:  0.1365
## F-statistic: 37.23 on 4 and 913 DF,  p-value: < 2.2e-16
```

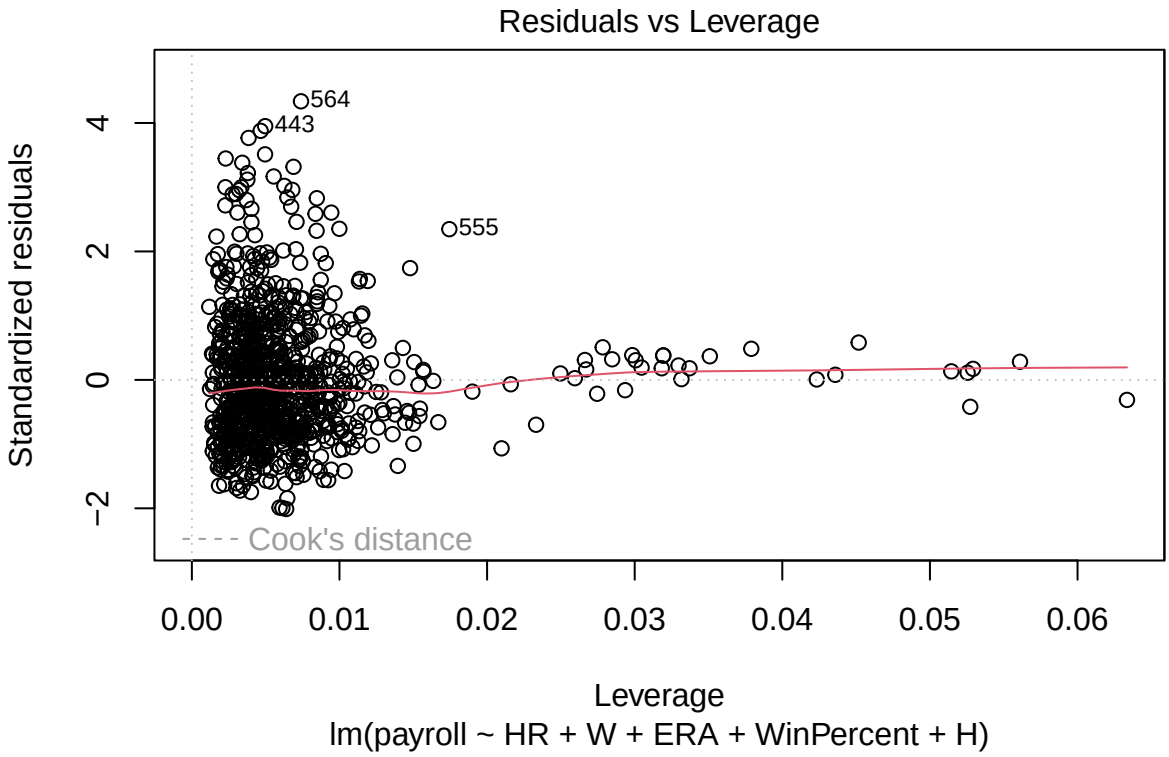
We can see that after removing the "W" variable, the Adjusted R-squared value is still .1365.The reduced model still explains approximately 13.6% of the variability in the response variable. The remaining variables in the model are not statistically significant (have p-values higher than .05).

1f

```
full_model = lm(payroll ~ HR + W +
 ERA + WinPercent + H, data = BaseballTeamPayroll)
plot(full_model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(payroll ~ HR + W + ERA + WinPercent + H)
Scale−Location
√|Standardized residuals|
Fitted values
lm(payroll ~ HR + W + ERA + WinPercent + H)



Interpretation of residual plots:

Residuals vs Fitted: The plot shows a relatively random scatter of residuals around zero, indicating that the assumption of linearity is reasonable. However, there is some non-constant variance in the residuals, as indicated by the widening of the spread of residuals at higher fitted values.

Normal Q-Q: The plot shows the residuals roughly following a straight line, which indicates that the assumption of normality is reasonable. However, there is some deviation from normality, particularly in the tails of the distribution.

Scale-Location: The plot shows the residuals spreading out pretty equally across the range of fitted values, this means that the assumption of constant variance is reasonable.

Residuals vs Leverage: The plot shows no points with particularly high leverage, indicating that there are no extreme values that are affecting the results.

Conclusion

In conclusion, this data analysis project has provided insights into the relationship between baseball team payroll and various performance metrics, including homeruns, wins, earned run average, win percentage, divisional standings, and hits. The findings reveal that there is no significant correlation between team payroll and these performance metrics. These results suggest that other factors, such as team management, player skills, and strategic planning, may play a more critical role in determining a team's success. These findings have implications for team owners, coaches, and players, who may need to rethink the traditional notion that high team payroll translates into better performance. By leveraging these insights, teams can optimize their performance and achieve greater success on the baseball field.

The main issue with my data analysis project was the lack of correlation between a MLB team's payroll and the other variables being analyzed. To improve the project, I could have considered different variables or factors that could potentially influence a team's success, such as player performance metrics or team

management strategies. It is encouraging to see that there was a correlation between hits, home runs, and wins, as well as between ERA and win percentage, and I could have explored these relationships in more depth to gain a better understanding of their impact on a team's overall success.This was my first time using Postit Cloud and coding in this way, so it was quite a challenge at first. But I actually appreciated being exposed to something different. Like any new tool or technique, there was a lot to learn, but I managed to adapt and get the hang of it. In future projects, I can use this experience to tackle future projects.