HW5

Junyu Wang

First inspection of the data

```
source("~/personal/school/stats133/stat133/HW5/data_clean_preprocess.R")
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
str(raw_weather_data)
```

```
## 'data.frame':    11192 obs. of  21 variables:
##  $ id          : Factor w/ 11192 levels "196009230ram",..: 1 2 4 6 3 5 7
11 8 9 ...
##  $ home_team   : Factor w/ 40 levels "Arizona Cardinals",..: 21 11 14 33 3
29 11 14 10 3 ...
##  $ home_score  : int  21 28 14 19 20 24 25 28 28 42 ...
##  $ away_team   : Factor w/ 40 levels "Arizona Cardinals",..: 35 31 8 26 40
10 29 13 31 8 ...
##  $ away_score  : int  43 35 17 21 0 41 27 9 20 7 ...
##  $ temperature : int  66 72 60 72 62 61 77 53 54 54 ...
##  $ wind_chill  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ humidity    : num  78 80 76 44 80 77 50 78 78 76 ...
##  $ wind_mph    : int  8 16 13 10 9 9 11 16 15 9 ...
##  $ weather     : Factor w/ 8364 levels "-0 degrees- relative humidity 58%-
wind 8 mph- wind chill -15",..: 6489 7332 5338 7239 5786 5561 7770 3872 4066
4056 ...
##  $ date        : Factor w/ 2099 levels "1/1/1967","1/1/1978",..: 1918 1934
1951 1951 1951 1951 2026 360 360 360 ...
##  $ temperature2: num  66 72 60 72 62 61 77 53 54 54 ...
##  $ humidity2   : num  78 80 76 44 80 77 50 78 78 76 ...
##  $ wind2       : num  8 16 13 10 9 9 11 16 15 9 ...
##  $ year        : num  1960 1960 1960 1960 1960 1960 1960 1960 1960 1960
...
##  $ monthnum    : num  9 9 9 9 9 9 9 10 10 10 ...
##  $ month       : Factor w/ 7 levels "August","December",..: 7 7 7 7 7 7 7
6 6 6 ...
##  $ decade      : Factor w/ 6 levels "1960s","1970s",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ total_score : int  64 63 31 40 20 65 52 37 48 49 ...
##  $ diff_score  : int  -22 -7 -3 -2 20 -17 -2 19 8 35 ...
##  $ home_win    : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
```

```
head(raw_weather_data)
```

```
##             id           home_team home_score           away_team
## 1 196009230ram    Los Angeles Rams         21 St. Louis Cardinals
## 2 196009240dal      Dallas Cowboys         28 Pittsburgh Steelers
## 3 196009250gnb   Green Bay Packers         14       Chicago Bears
## 4 196009250sfo San Francisco 49ers         19     New York Giants
## 5 196009250clt     Baltimore Colts         20 Washington Redskins
## 6 196009250phi Philadelphia Eagles         24    Cleveland Browns
##   away_score temperature wind_chill humidity wind_mph
## 1         43          66         NA       78        8
## 2         35          72         NA       80       16
## 3         17          60         NA       76       13
## 4         21          72         NA       44       10
## 5          0          62         NA       80        9
## 6         41          61         NA       77        9
##                                         weather      date temperature2
## 1  66 degrees- relative humidity 78%- wind 8 mph 9/23/1960           66
## 2 72 degrees- relative humidity 80%- wind 16 mph 9/24/1960           72
## 3 60 degrees- relative humidity 76%- wind 13 mph 9/25/1960           60
## 4 72 degrees- relative humidity 44%- wind 10 mph 9/25/1960           72
## 5  62 degrees- relative humidity 80%- wind 9 mph 9/25/1960           62
## 6  61 degrees- relative humidity 77%- wind 9 mph 9/25/1960           61
##   humidity2 wind2 year monthnum     month decade total_score diff_score
## 1        78     8 1960        9 September  1960s          64        -22
## 2        80    16 1960        9 September  1960s          63         -7
## 3        76    13 1960        9 September  1960s          31         -3
## 4        44    10 1960        9 September  1960s          40         -2
## 5        80     9 1960        9 September  1960s          20         20
## 6        77     9 1960        9 September  1960s          65        -17
##   home_win
## 1    FALSE
## 2    FALSE
## 3    FALSE
## 4    FALSE
## 5     TRUE
## 6    FALSE
```

Weather Information

Remove % in column humidity and convert such values to numeric format

```
head(raw_weather_data$humidity)
```

```
## [1] 78 80 76 44 80 77
```

Extract the temperature from column weather, and create a column temperature2 with these values.

```
head(raw_weather_data$temperature2)
```

```
## [1] 66 72 60 72 62 61
```

Extract the humidity values from column weather, and create a column humidity2 with these values.

```
head(raw_weather_data$humidity2)
```

```
## [1] 78 80 76 44 80 77
```

Extract the wind speed values from column weather, and create a column wind2 with these values.

```
head(raw_weather_data$wind2)
```

```
## [1]  8 16 13 10  9  9
```

Check new columns coincide with the pre-existing ones:

```
summary(raw_weather_data$temperature)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -7.00   46.00   59.00   56.69   72.00   96.00
```

```
summary(raw_weather_data$temperature2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -7.00   46.00   59.00   56.69   72.00   96.00
```

```
summary(raw_weather_data$humidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   57.00   69.00   67.21   79.00  100.00    1907
```

```
summary(raw_weather_data$humidity2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   57.00   69.00   67.21   79.00  100.00    1907
```

```
summary(raw_weather_data$wind_mph)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    7.00   10.00   10.21   13.00   32.00    1845
```

```
summary(raw_weather_data$wind2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    7.00   10.00   10.21   13.00   32.00    1845
```

Data information

Create a column year that contains the number of the year.

```
head(raw_weather_data$year)
```

```
## [1] 1960 1960 1960 1960 1960 1960
```

Create a column monthnum that contains the number of the month.

```
head(raw_weather_data$monthnum)
```

```
## [1] 9 9 9 9 9 9
```

Create a column month that contains the name of the corresponding month as factor.

```
head(raw_weather_data$month)
```

```
## [1] September September September September September September
## Levels: August December Feburary January November October September
```

Create a column decade that indicates the corresponding decade of each played game.

```
head(raw_weather_data$decade)
```

```
## [1] 1960s 1960s 1960s 1960s 1960s 1960s
## Levels: 1960s 1970s 1980s 1990s 2000s 2010s
```

Scores Information

Create a column total_score that contains the total number of scored points in each game.

```
head(raw_weather_data$total_score)
```

```
## [1] 64 63 31 40 20 65
```

Create a column diff_score that indicates the difference of home_score and away_score.

```
head(raw_weather_data$diff_score)
```

```
## [1] -22  -7  -3  -2  20 -17
```

Create a column home_win that shows whether home_score is greater than away_score.

```
head(raw_weather_data$home_win)
```

```
## [1] FALSE FALSE FALSE FALSE  TRUE FALSE
```

Basic Exploration

Inspect variables home_score, away_score, temperature, wind_mpg

```
summary(raw_weather_data$home_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    14.0    21.0    22.1    28.0    72.0
```

```
summary(raw_weather_data$away_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.00   19.00   19.35   27.00   62.00
```

```
summary(raw_weather_data$temperature)
```

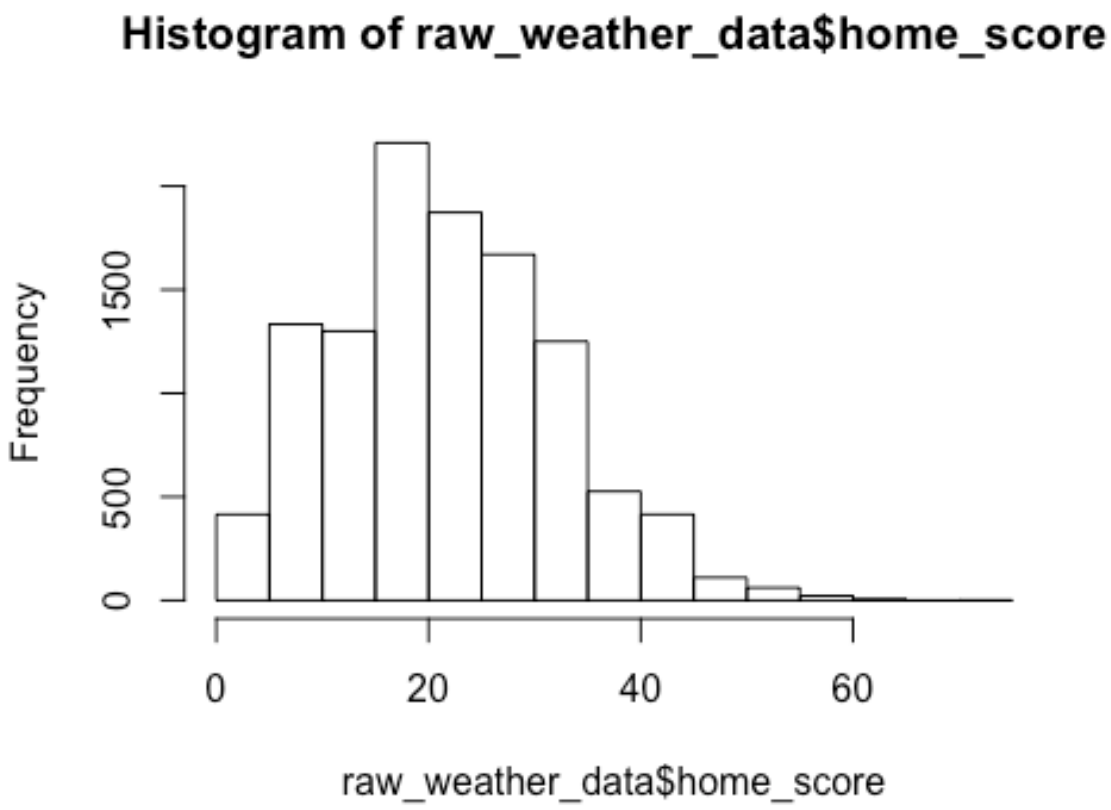
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -7.00   46.00   59.00   56.69   72.00   96.00
```

```
summary(raw_weather_data$wind_mph)
```

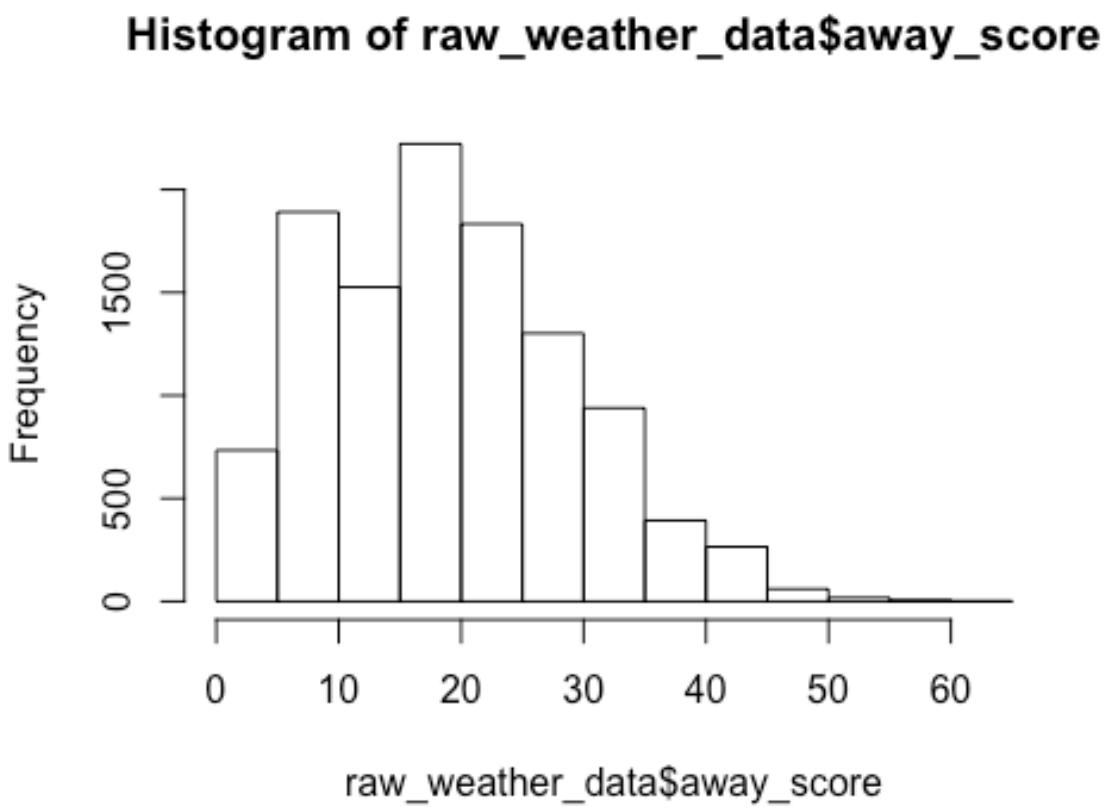
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    7.00   10.00   10.21   13.00   32.00    1845
```

Visually inspect variables home_score, away_score, temperature, humidity and wind_mph.

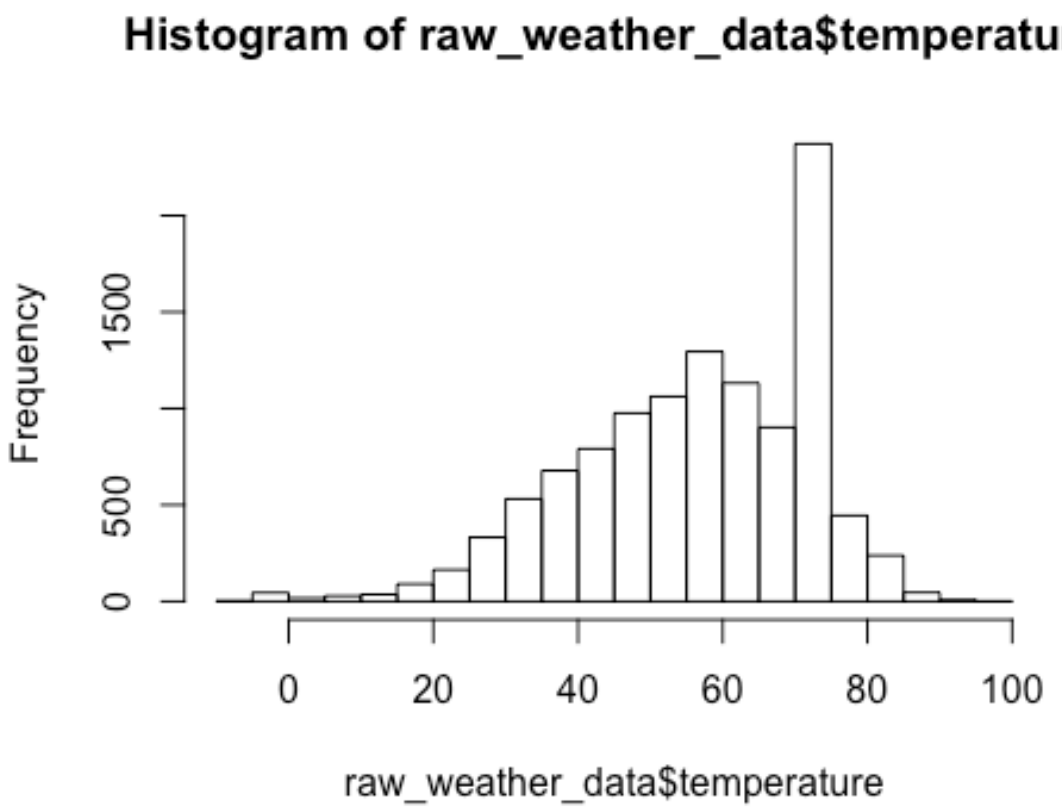
```
hist(raw_weather_data$home_score)
```



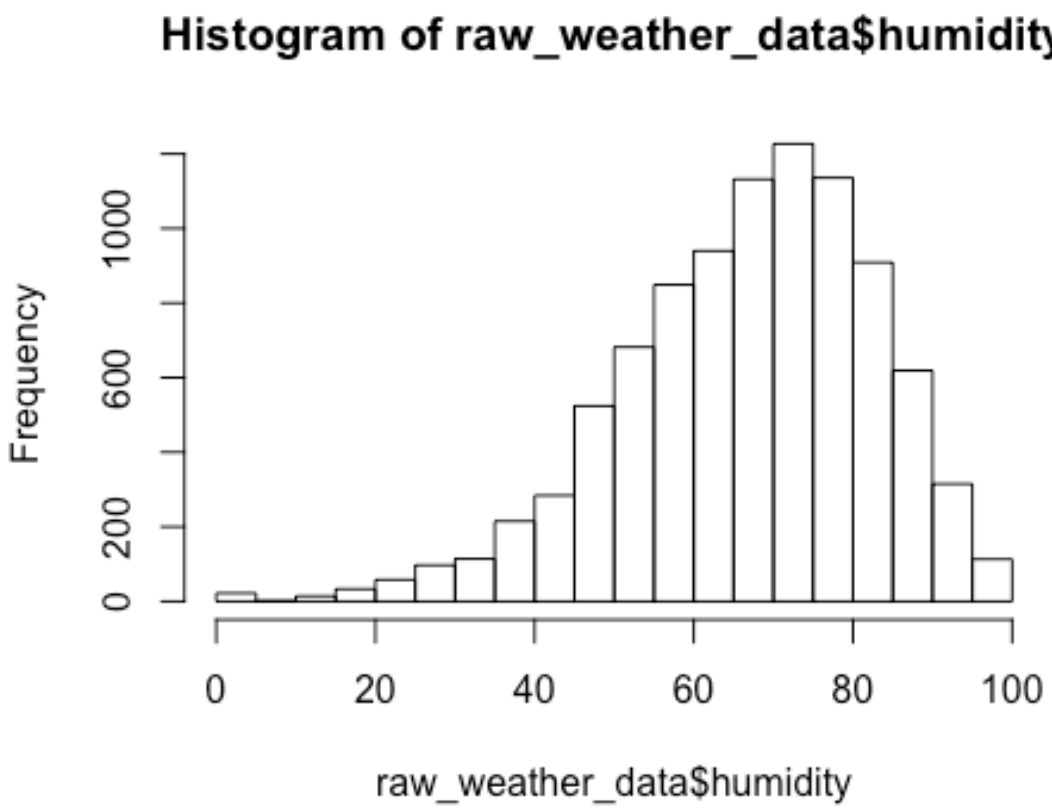

```
hist(raw_weather_data$away_score)
```



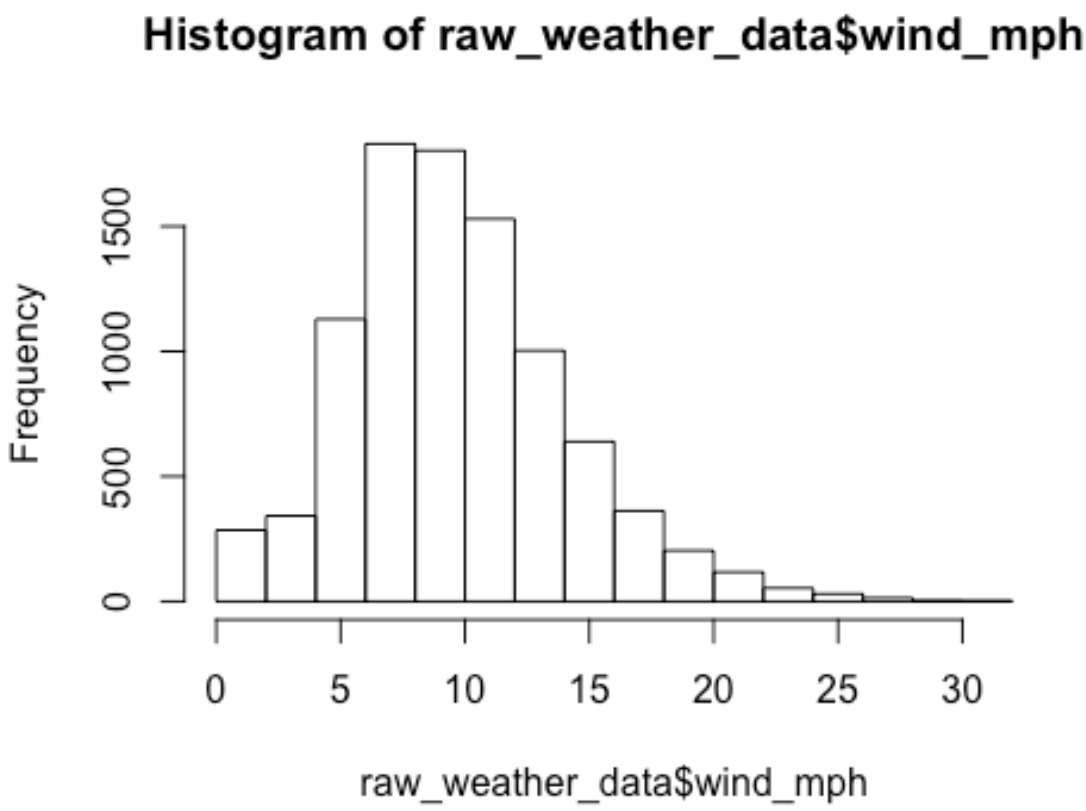

```
hist(raw_weather_data$temperature)
```




```r
hist(raw_weather_data$humidity)
```




```
hist(raw_weather_data$wind_mph)
```



What team has the maximum home score?

```
team_with_max_home_score
```

```
## [1] Washington Redskins
## 40 Levels: Arizona Cardinals Atlanta Falcons ... Washington Redskins
```

What team has the maximus away score?

```
team_with_max_away_score
```

```
## [1] Atlanta Falcons
## 40 Levels: Arizona Cardinals Atlanta Falcons ... Washington Redskins
```

What is the most common home score?

```
m_common_home_score
```

```
##  17
## 816
```

What is the most common away score?

```
m_common_away_score
```

```
##  17
## 917
```

What has been the maximum temperature in a game?

```
max_temp
```

```
## [1] 96
```

What was the date of the maximum temperature?

```
date_max_temp
```

```
## [1] 9/8/2013
## 2099 Levels: 1/1/1967 1/1/1978 1/1/1984 1/1/1989 1/1/1995 ... 9/9/2013
```

What has been the minimum temperature in a game?

```
min_temp
```

```
## [1] -7
```

What was the date of the minimum temperature?

```
date_min_temp
```

```
## [1] 1/20/2008
## 2099 Levels: 1/1/1967 1/1/1978 1/1/1984 1/1/1989 1/1/1995 ... 9/9/2013
```

How many games have been played with a temperature of 90 degrees or more?

```
games_with_high_temp
```

```
## [1] 15
```

How many games have been played with a temperature below 0 degrees (do not include 0)?

```
games_with_low_temp
```

```
## [1] 7
```

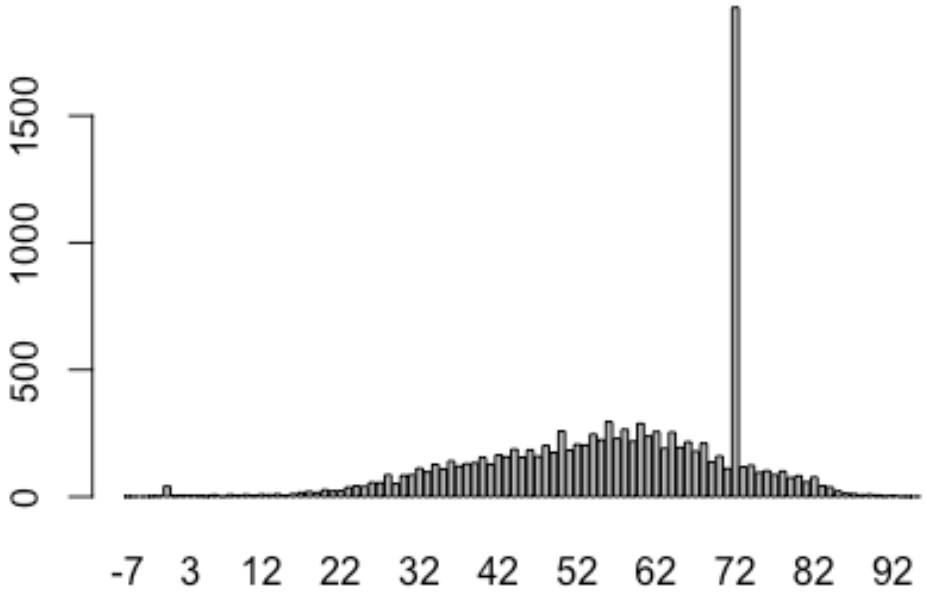
What is the most common temperature?

```
m_common_temp
```

```
##   72
## 1926
```

Make a bar chart with the frequency table of temperatures. Is there anything that catches your attention?

```
barplot(temp_freq)
```



The frequency of temperature 72 is much higher than the rest of temperatures.

Data Files

Export data for all games with selected columns.

```
head(cleaned_data)
```

```
##             id           home_team home_score           away_team
## 1 196009230ram    Los Angeles Rams         21 St. Louis Cardinals
## 2 196009240dal      Dallas Cowboys         28 Pittsburgh Steelers
## 3 196009250gnb   Green Bay Packers         14       Chicago Bears
## 4 196009250sfo San Francisco 49ers         19     New York Giants
## 5 196009250clt     Baltimore Colts         20 Washington Redskins
## 6 196009250phi Philadelphia Eagles         24    Cleveland Browns
##   away_score total_score diff_score home_win      date year     month
## 1         43          64        -22    FALSE 9/23/1960 1960 September
## 2         35          63         -7    FALSE 9/24/1960 1960 September
## 3         17          31         -3    FALSE 9/25/1960 1960 September
## 4         21          40         -2    FALSE 9/25/1960 1960 September
## 5          0          20         20     TRUE 9/25/1960 1960 September
## 6         41          65        -17    FALSE 9/25/1960 1960 September
##   decade temperature humidity wind_mph
```

```
## 1  1960s           66       78        8
## 2  1960s           72       80       16
## 3  1960s           60       76       13
## 4  1960s           72       44       10
## 5  1960s           62       80        9
## 6  1960s           61       77        9
```

Export data for games during each decade with selected columns.

```
print("files are correctly generated")
```

```
## [1] "files are correctly generated"
```

Data Analysis

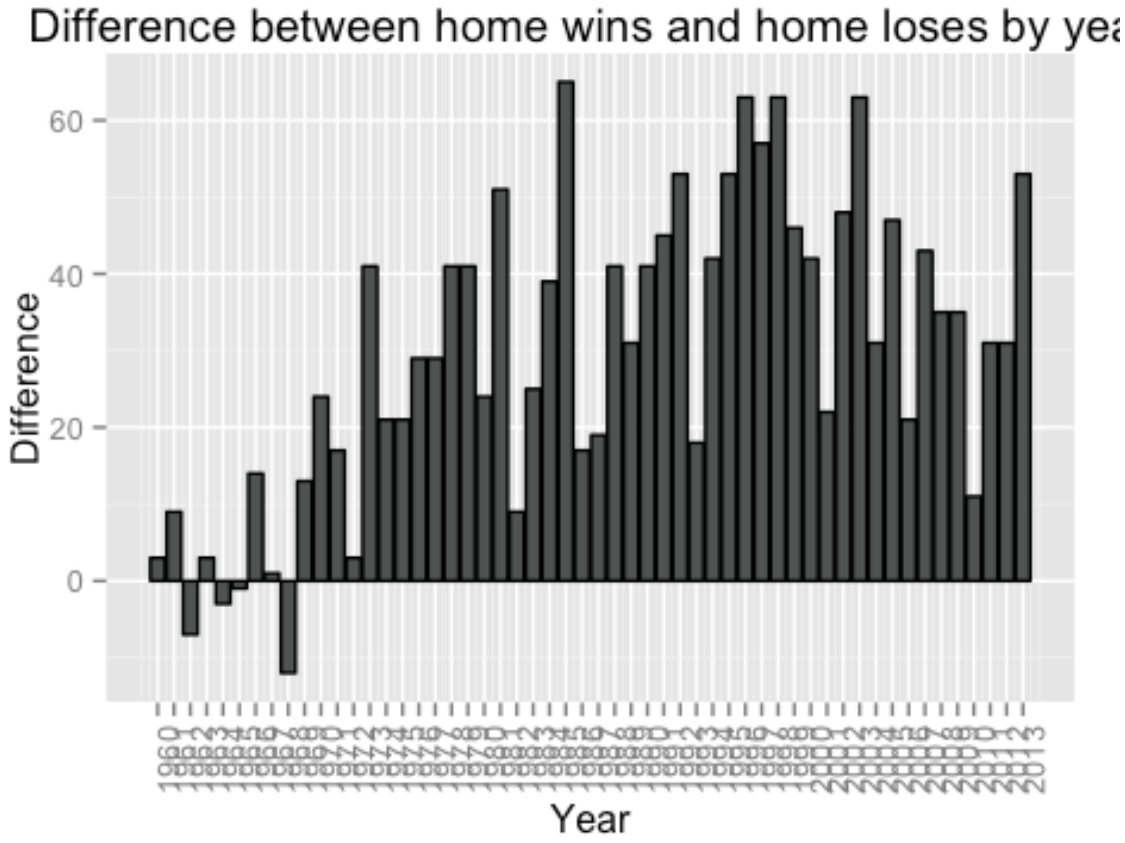
```
source("~/personal/school/stats133/stat133/HW5/data_analysis.R")
```

1. Does play at home really have an advantage for the home team?

Number of More Home victories per year

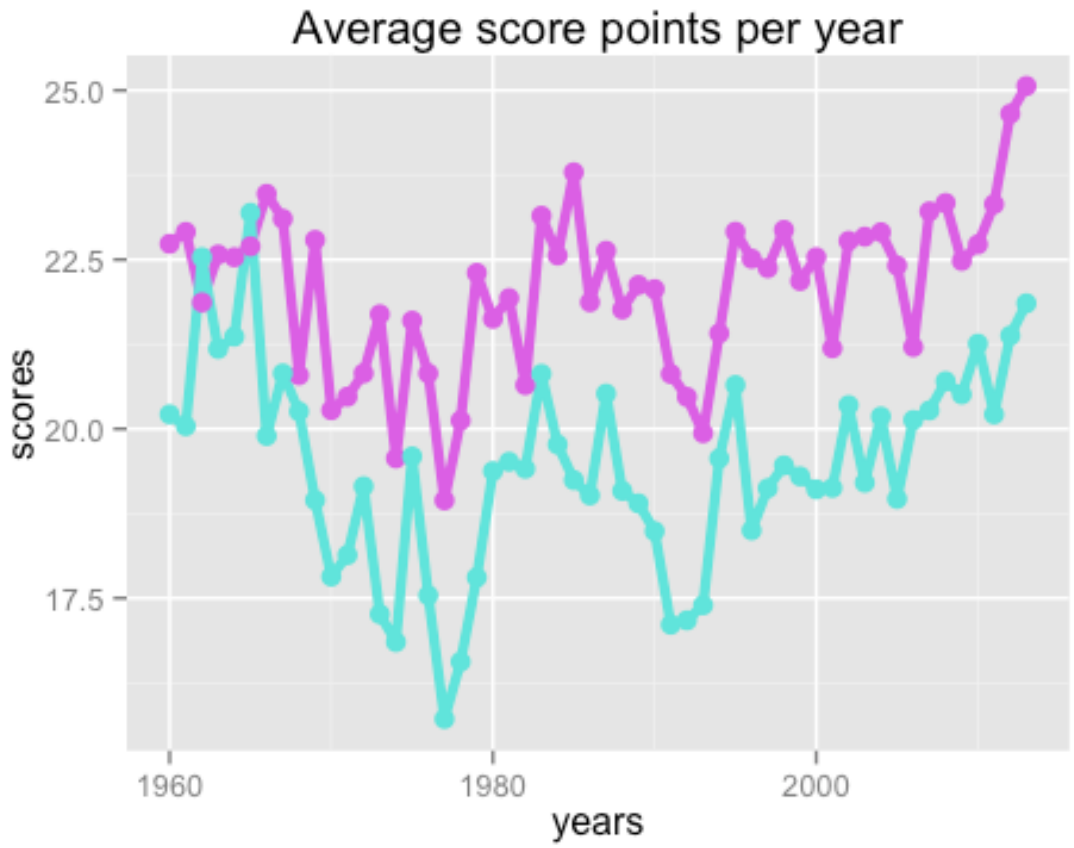
```
ggplot(data = result, aes(years, diffs)) +
  geom_bar(color = "#000000", fill = "#515252", stat = "identity") +
  ggtitle("Difference between home wins and home loses by year") +
  ylab("Difference") +
  scale_x_continuous(name = "Year",
                     breaks = years) +
  theme(axis.text.x = element_text(angle= 90))
```

```
## Warning: Stacking not well defined when ymin != 0
```



Average score points per year

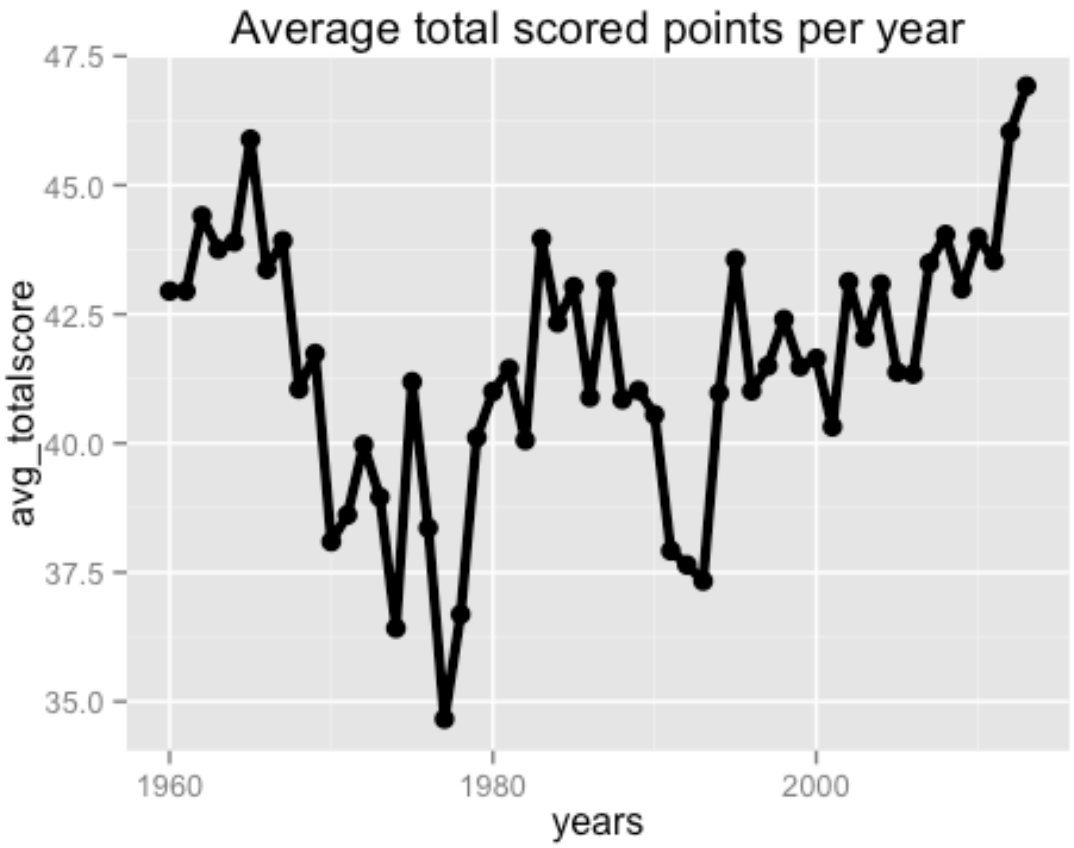
```
ggplot(data = avgs) +
  geom_line(aes(x = years, y = avg_homescore), color = "#DB5FE4", size = 1.5)
+
  geom_line(aes(x = years, y = avg_awayscore), color = "#5FE4DB", size = 1.5)
+
  geom_point(aes(x = years, y = avg_homescore), color = "#DB5FE4", size =
3.0) +
  geom_point(aes(x = years, y = avg_awayscore), color = "#5FE4DB", size =
3.0) +
  ggtitle("Average score points per year") +
  ylab("scores")
```



Other type of analysis for this conclusion: We can also conduct a t-test for null hypothesis stating that difference in home victories and away team victories = 0 each year. We can set up a one-side t-test and reject the null if p-value is smaller than the significance level.

2. Has the total number of scored points per game changed over time?

```
ggplot(data = avg_scores) +
  geom_line(aes(x = years, y = avg_totalscore), color = "#000000", size =
1.5) +
  geom_point(aes(x = years, y = avg_totalscore), color = "#000000", size =
3.0) +
  ggtitle("Average total scored points per year")
```



Other type of analysis for this conclusion: We can conduct a significance test for null hypothesis stating that scores each year is the same so expected value for each year's score = total scores of all years / number of years. We can set up a chi-square test and reject the null if test statistic is larger than corresponding critical score.