lab 7.2

AUTHOR
Dina Al Jibori

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.3     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(Lahman)
library(broom)
data(Teams)
library(patchwork)
```

Question 1

```
NewTeams <- Teams %>%
  filter(yearID >= 2000)
glimpse(NewTeams)
```

```
Rows: 690
Columns: 48
$ yearID         <int> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2…
$ lgID           <fct> AL, NL, NL, AL, AL, AL, NL, NL, AL, NL, AL, NL, NL, AL,…
$ teamID         <fct> ANA, ARI, ATL, BAL, BOS, CHA, CHN, CIN, CLE, COL, DET, …
$ franchID       <fct> ANA, ARI, ATL, BAL, BOS, CHW, CHC, CIN, CLE, COL, DET, …
$ divID          <chr> "W", "W", "E", "E", "E", "C", "C", "C", "C", "W", "C", …
```

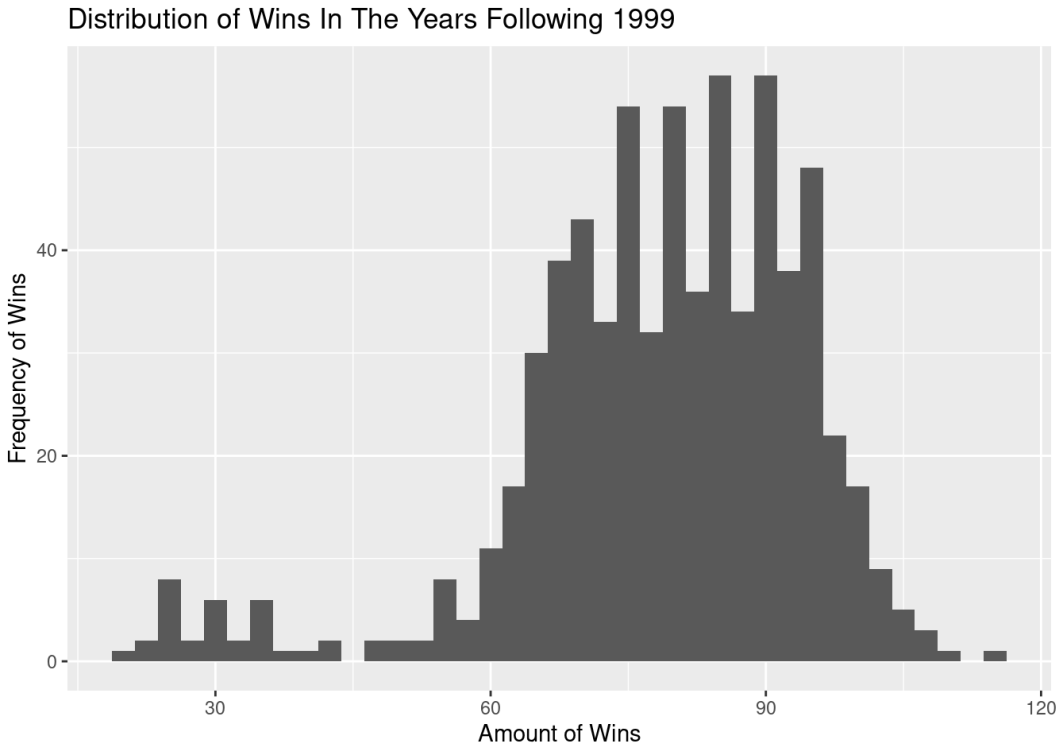
```
$ Rank          <int> 3, 3, 1, 4, 2, 1, 6, 2, 2, 4, 3, 3, 4,
4, 2, 3, 5, 4, 1…
$ G             <int> 162, 162, 162, 162, 162, 162, 162, 163,
162, 162, 162, …
$ Ghome         <int> 81, 81, 81, 81, 81, 81, 81, 82, 81, 81,
81, 81, 81, 81,…
$ W             <int> 82, 85, 95, 74, 85, 95, 65, 85, 90, 82,
79, 79, 72, 77,…
$ L             <int> 80, 77, 67, 88, 77, 67, 97, 77, 72, 80,
83, 82, 90, 85,…
$ DivWin        <chr> "N", "N", "Y", "N", "N", "Y", "N", "N",
"N", "N", "N", …
$ WCWin         <chr> "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", …
$ LgWin         <chr> "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", …
$ WSWin         <chr> "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", …
$ R             <int> 864, 792, 810, 794, 792, 978, 764, 825,
950, 968, 823, …
$ AB            <int> 5628, 5527, 5489, 5549, 5630, 5646,
5577, 5635, 5683, 5…
$ H             <int> 1574, 1466, 1490, 1508, 1503, 1615,
1426, 1545, 1639, 1…
$ X2B           <int> 309, 282, 274, 310, 316, 325, 272, 302,
310, 320, 307, …
$ X3B           <int> 34, 44, 26, 22, 32, 33, 23, 36, 30, 53,
41, 29, 36, 27,…
$ HR            <int> 236, 179, 179, 184, 167, 216, 183, 200,
221, 161, 177, …
$ BB            <int> 608, 535, 595, 558, 611, 591, 632, 559,
685, 601, 562, …
$ SO            <int> 1024, 975, 1010, 900, 1019, 960, 1120,
995, 1057, 907, …
$ SB            <int> 93, 97, 148, 126, 43, 119, 93, 100, 113,
131, 83, 168, …
$ CS            <int> 52, 44, 56, 65, 30, 42, 37, 38, 34, 61,
38, 55, 52, 35,…
$ HBP           <int> 47, 59, 59, 49, 42, 53, 54, 64, 51, 42,
43, 60, 83, 48,…
$ SF            <int> 43, 58, 45, 54, 48, 61, 45, 58, 52, 75,
49, 51, 61, 70,…
$ RA            <int> 869, 754, 714, 913, 745, 839, 904, 765,
816, 897, 827, …
$ ER            <int> 805, 698, 648, 855, 683, 751, 849, 700,
775, 835, 755, …
$ ERA           <dbl> 5.00, 4.35, 4.05, 5.37, 4.23, 4.66,
5.25, 4.33, 4.84, 5…
```

```
$ CG             <int> 5, 16, 13, 14, 7, 5, 10, 8, 6, 7, 6, 5,
8, 10, 9, 2, 6,…
$ SHO            <int> 3, 8, 9, 6, 12, 7, 5, 7, 5, 2, 6, 4, 2,
6, 11, 7, 4, 7,…
$ SV             <int> 46, 38, 53, 33, 46, 43, 39, 42, 34, 33,
44, 48, 30, 29,…
$ IPouts         <int> 4344, 4331, 4321, 4300, 4358, 4351,
4364, 4369, 4327, 4…
$ HA             <int> 1534, 1441, 1428, 1547, 1433, 1509,
1505, 1446, 1511, 1…
$ HRA            <int> 228, 190, 165, 202, 173, 195, 231, 190,
173, 221, 177, …
$ BBA            <int> 662, 500, 484, 665, 498, 614, 658, 659,
666, 588, 496, …
$ SOA            <int> 846, 1220, 1093, 1017, 1121, 1037, 1143,
1015, 1213, 10…
$ E              <int> 134, 107, 129, 116, 109, 133, 100, 111,
72, 94, 105, 12…
$ DP             <int> 182, 138, 138, 151, 120, 190, 139, 156,
147, 176, 171, …
$ FP             <dbl> 0.978, 0.982, 0.979, 0.981, 0.982,
0.978, 0.983, 0.982,…
$ name           <chr> "Anaheim Angels", "Arizona
Diamondbacks", "Atlanta Brav…
$ park           <chr> "Edison International Field", "Bank One
Ballpark", "Tur…
$ attendance     <int> 2066982, 2942251, 3234304, 3297031,
2585895, 1947799, 2…
$ BPF            <int> 102, 105, 101, 95, 104, 102, 97, 102,
101, 125, 95, 94,…
$ PPF            <int> 103, 103, 99, 96, 103, 102, 98, 102,
100, 125, 95, 95, …
$ teamIDBR       <chr> "ANA", "ARI", "ATL", "BAL", "BOS",
"CHW", "CHC", "CIN",…
$ teamIDlahman45 <chr> "ANA", "ARI", "ATL", "BAL", "BOS",
"CHA", "CHN", "CIN",…
$ teamIDretro    <chr> "ANA", "ARI", "ATL", "BAL", "BOS",
"CHA", "CHN", "CIN",…
```

The dimensions of this new filtered data set are 690 by 48.

Question 2

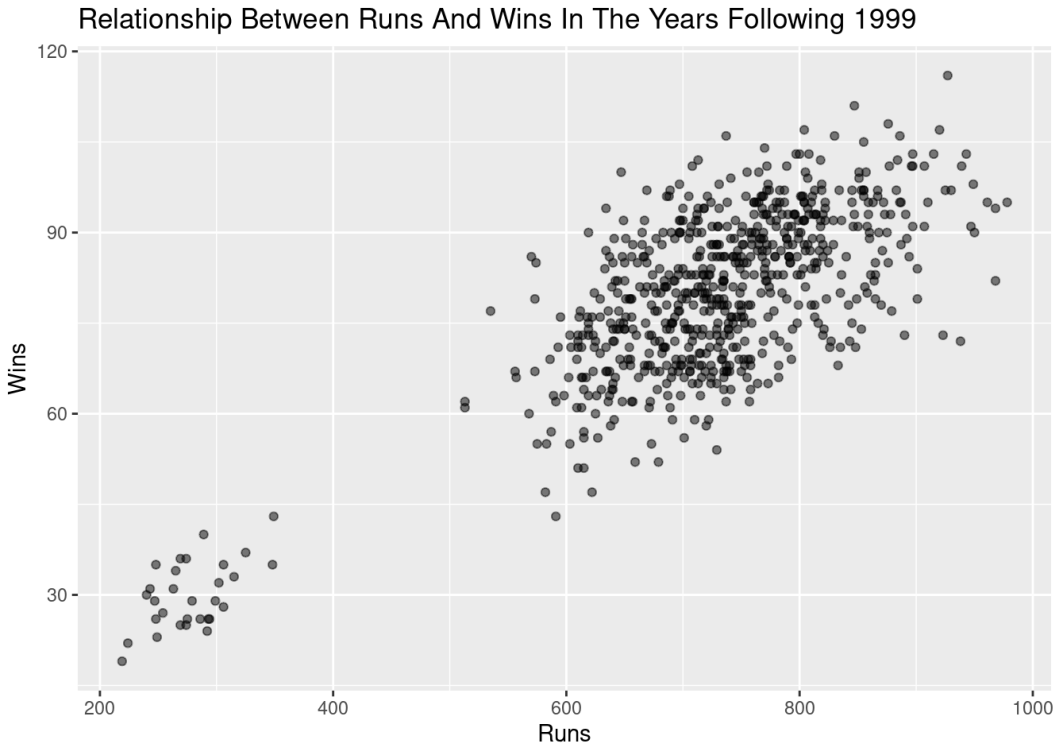
```
NewTeams %>%
  ggplot(aes(x = W)) +
  geom_histogram(binwidth = 2.5) +
  labs(title = "Distribution of Wins In The Years Following 199
```



The shape of this graph is fairly unimodal with small variation and it's right leaning. It shows that if a Team has around 90 wins then they will most likely have the most frequency of wins. I drew something similair for the plot in Lab 7.1 as I assumed that the teams who win more often have a higher likelyhood of having a higher frequency of wins.

Question 3

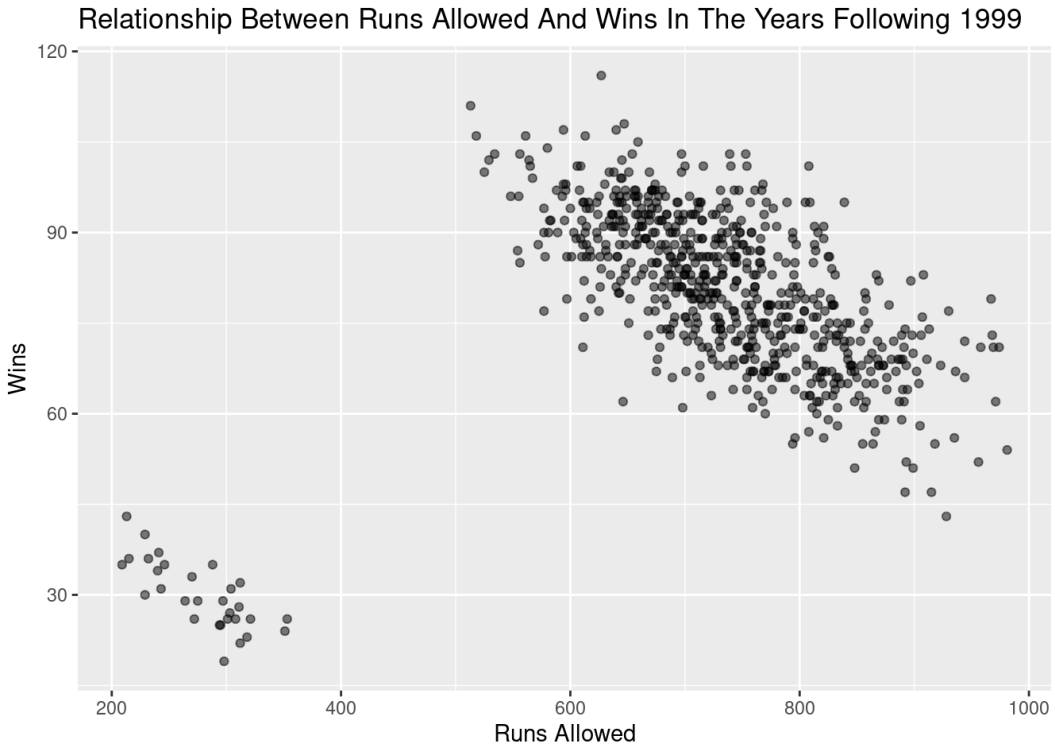
```
NewTeams %>%
  ggplot(aes(x = R, y = W)) +
  geom_point(alpha=0.5) +
  labs(title = "Relationship Between Runs And Wins In The Years
```



This scatterplot shows a strong, positive, linear association between Wins and Runs of accidents. There is no outlier but I would like to note the lack of data between 400 and 500 and I assume that's due to the lack of games in the Covid Years (2019-2021). My plot for the first lab looks similair to this as I assumed that there would be a positive relationship between Homeruns and Wins however I didnt account for the years in quarentine.

Question 4

```
NewTeams %>%
  ggplot(aes(x = RA, y = W)) +
  geom_point(alpha=0.5) +
  labs(title = "Relationship Between Runs Allowed And Wins In Th
```



There is a strong, negative, linear association between the Runs allowed and Wins with a some outliers around 30 wins. Similarly to the previous graph, theres lack of data between 400 and 500 Runs allowed and that's due to the lack of games in the Covid Years (2019-2021). However, this has a negative relationship betweeen x and y while the previous graph has a positive relationship between x and y.

Question 5

```
set.seed(123)
set_type <- sample(x = c('train', 'test'),
                   size = 690,
                   replace = TRUE,
                   prob = c(0.8, 0.2))
NewTeams <- NewTeams %>%
  mutate(set_type = set_type)

teams_train <- NewTeams %>%
  filter(set_type == 'train')

teams_test <- NewTeams %>%
  filter(set_type == 'test')
```

Question 6

```
set.seed(123)
```

```
model_1 <- lm(W ~ R, data = teams_train)

glance(model_1) %>%
    select(r.squared)
```

```
# A tibble: 1 × 1
  r.squared
      <dbl>
1     0.594
```

```
W_pred_linear <- predict(model_1, newdata = teams_test)

teams_test %>%
    mutate(W_pred_linear = W_pred_linear,
           resid_sq_linear = (W - W_pred_linear)^2) %>%
    summarize(TSS = sum((W - mean(W))^2),
              RSS_linear = sum(resid_sq_linear)) %>%
    mutate(Rsq_linear = 1 - RSS_linear/TSS) %>%
    select(Rsq_linear)
```

```
  Rsq_linear
1  0.6491418
```

```
print(coef(model_1))
```

```
(Intercept)           R
 8.60600027  0.09797272
```

The equation is W = 8.60600027 + 0.09797272*R. The R squared value for the training data is 0.5942856. The R squared value for the testing data is 0.6491418.

Question 7

```
average_runs <- NewTeams %>%
  summarise(mean_runs = mean(R))
average_runs
```

```
  mean_runs
1  718.2203
```

```
average_wins <- NewTeams %>%
  summarise(mean_wins = mean(W))
average_wins
```

```
  mean_wins
1  78.75217
```

```
RunsPredicted <- c(average_runs$mean_runs, 600, 850)

DataPredicted <- data.frame(R = RunsPredicted)

WinsPredicted <- predict(object = model_1 , newdata = DataPredi
WinsPredicted
```

```
       1        2        3
78.97199 67.38963 91.88281
```

The average number of season runs is 719 and an average of 78 wins. Based on the previous model, I predict a team that scored the average number of runs would win 78 games. While a team that scored 600 runs would win 67 games and a team that scored 850 runs would win 91 games.

Question 8

```
set.seed(123)

model_2 <- lm(W ~ R + RA, data = teams_train)

glance(model_2) %>%
    select(r.squared)
```

```
# A tibble: 1 × 1
  r.squared
      <dbl>
1     0.781
```

```
W_pred_dbl <- predict(model_2, newdata = teams_test)

teams_test %>%
    mutate(W_pred_dbl = W_pred_dbl,
           resid_sq_dbl = (W - W_pred_dbl)^2) %>%
    summarize(TSS = sum((W - mean(W))^2),
              RSS_dbl = sum(resid_sq_dbl)) %>%
    mutate(Rsq_dbl = 1 - RSS_dbl/TSS) %>%
    select(Rsq_dbl)
```

```
    Rsq_dbl
```

```
1 0.8041205
```

```
print(coef(model_2))
```

```
(Intercept)           R          RA
 27.6109097   0.1373287  -0.0658608
```

The equation is W = 27.6109097 + 0.1373287 * R - 0.0658608 * RA .The R squared value for the training data is 0.7808217. The R squared value for the testing data is 0.8041205. This model evaluates two of the variables (R and RA) rather than one of the variables in the previous one in relation to the wins.

Question 9

```
set.seed(123)
model_3 <- lm(W ~ R + BB + poly(SO, degree = 2, raw = TRUE), da
model_3
```

```
Call:
lm(formula = W ~ R + BB + poly(SO, degree = 2, raw = TRUE),
data = teams_train)

Coefficients:
                      (Intercept)
R
                       -2.146e+01
6.710e-02
                                BB  poly(SO, degree = 2, raw =
TRUE)1
                        2.852e-02
6.535e-02
poly(SO, degree = 2, raw = TRUE)2
                       -2.733e-05
```

```
glance (model_3) %>%
  select(r.squared)
```

```
# A tibble: 1 × 1
  r.squared
      <dbl>
1     0.624
```

```
pred_poly <- predict (model_3, newdata = teams_test)
```

```
teams_test %>%
  mutate(pred_poly = pred_poly,
         resid_sq_poly = (W - pred_poly)^2) %>%
  summarize(TSS = sum((W - mean(W))^2),
RSS_poly = sum(resid_sq_poly)) %>%
  mutate(Rsq_poly = 1 - RSS_poly/TSS) %>%
  select(Rsq_poly)
```

```
   Rsq_poly
1 0.6583934
```

The equation is W= −21.46 + 0.0671 * R + 0.0285 * BB + 0.06535 * SO − 0.00002733 * SO. The R squared value for the training data is 0.6237166. The R squared value for the testing data is 0.6583934.

Question 10

Through all three model the values of the the training R squared were fairly close to the values of the testing R squared differing by a few decimal points. The better metric for deciding how well a model will perform on new data is higher R squared value. By that metric, model 2 has the best proformance for new data.

Question 11

A predictive model having a positive coefficient between one of the predictors is not direct evidence of causation. There are many more factors involve and you cant surely conclude that two variables cause each other rather than just happen to move in the same direction.