```
---
title: "Lab 4"
author:  Nandini Parmar 305561701
date: "2022-05-01"
output:
  pdf_document:
    toc: yes
    toc_depth: '3'
  html_document:
    theme: paper
    toc: yes
    toc_depth: 3
    toc_float: yes
---
\fontsize{10}{11}

  
```

– Attaching packages ——————————————— tidyverse 1.3.1 –

v ggplot2 3.3.5 v purrr 0.3.4

v tibble 3.1.6 v dplyr 1.0.7

v tidyr 1.1.4 v stringr 1.4.0

v readr 2.1.1 v forcats 0.5.1

– Conflicts ———————————————— tidyverse_conflicts() –

x dplyr::filter() masks stats::filter()

x dplyr::lag() masks stats::lag()

Date last run: 2022-05-01

Hello World!

\newpage

Your Work

Make sure to edit the "author" information in the YAML header near the top to include your name and UID

Complete/answer the following.

1 --- Does our NFL team data represent "stacked" data? Why or why not?

The NFL team data does not represent stacked data because the rows are independent from each other and

2 --- Each team played 17 regular season games. Calculate the average points scored per game for each

> 2a --- Does this plot visually look similar to its counterpart using season total points above? ...

> 2b --- Is the Pearson's correlation coefficient the same or different? Why?

```r
library(xtable)
library(tidyverse)
library(readxl)
library(ggthemes)

xdf <- read_excel("/Users/apurvashah/Documents/GitHub/stats10/lab4/NFL_offense_passing_2021.xlsx", shee
head(xdf, n=6)
```

```
## # A tibble: 6 x 17
##   Team      Att   Cmp `Cmp %` `Yds/Att` `Pass Yds`    TD   INT  Rate `1st` `1st%`  `20`
##   <chr>   <dbl> <dbl>   <dbl>     <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 49ers     514   343    66.7       8.6       4437    26    14  99.2   200   38.9    63
## 2 Bears     542   332    61.3       6.7       3635    16    20  75.5   180   33.2    40
## 3 Bengals   555   384    69.2       8.7       4806    36    14 107.    208   37.5    63
## 4 Bills     655   415    63.4       6.8       4450    36    16  91.3   236   36      51
## 5 Broncos   541   354    65.4       7.1       3856    20     9  91.7   179   33.1    46
## 6 Browns    520   320    61.5       7         3619    21    14  84.6   177   34      47
## # ... with 5 more variables: 40 <dbl>, Lng <chr>, Sck <dbl>, SckY <dbl>,
## #   totalPoints <dbl>
```

```r
x <- data.frame(Team = xdf$Rate, avg = xdf$totalPoints)
head(x)
```

```
##    Team avg
```

```
## 1  99.2 427
## 2  75.5 311
## 3 106.9 460
## 4  91.3 483
## 5  91.7 335
## 6  84.6 349
```

```
x$avg = round(x$avg/17, 2)

# head(x)

xmeanPR <- mean(x$Team)
xmeanPR
```

```
## [1] 90.69688
```

```
xmeanPnts <- mean(x$avg)
xmeanPnts
```

```
## [1] 22.9825
```

```
##### sample standard deviation of rating
xsdPR <- sd(x$Team)
xsdPR
```

```
## [1] 10.62254
```

```
##### sample standard deviation of total points
xsdPnts <- sd(x$avg)
xsdPnts
```

```
## [1] 4.50699
```

```
n <- nrow(xdf)
n
```

```
## [1] 32
```

```
###### sample covariance
xcov <- sum((x$Team - xmeanPR) * (x$avg - xmeanPnts) ) / (n - 1)
xcov
```

```
## [1] 39.68362
```

```
###### sample correlation
xcorr <- xcov / (xsdPR * xsdPnts)
xcorr
```

```
## [1] 0.8288888
```

```
###### sample regression line slope
xb1 <- xcorr * xsdPnts / xsdPR
xb1
```

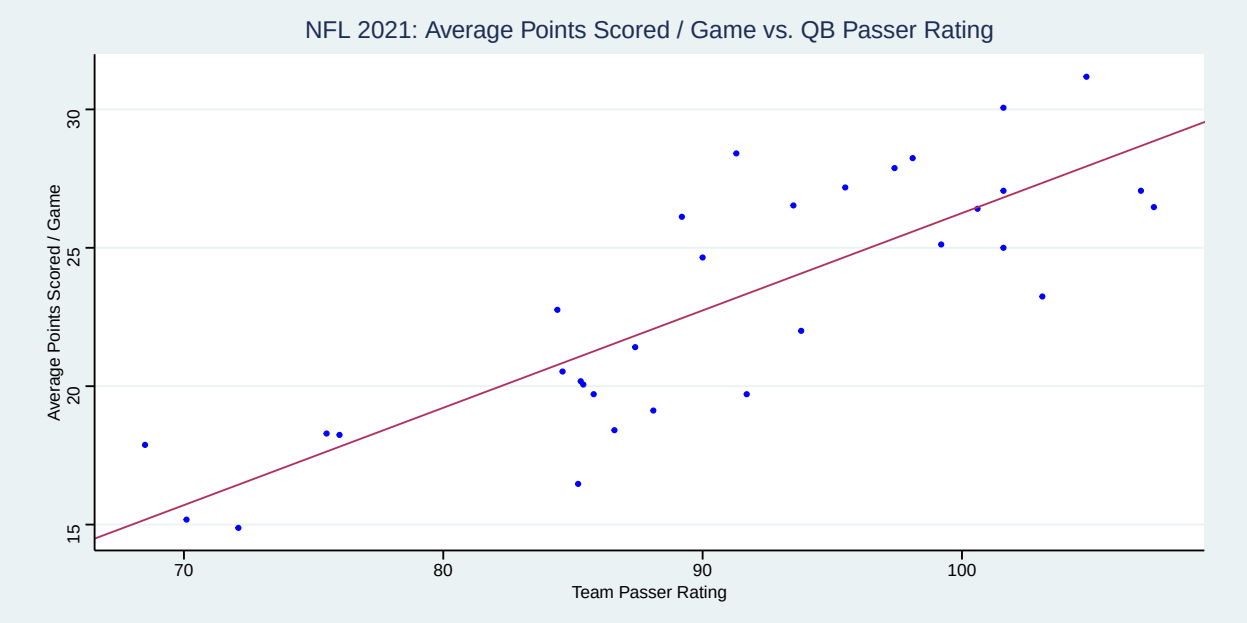
```
## [1] 0.3516855
```

```
###### sample regression line y-intercept
xb0 <- xmeanPnts - xb1 * xmeanPR
xb0
```

```
## [1] -8.914276
```

```
graph <- ggplot(data = x,  aes(Team, avg)) + geom_point(size = 1, color = "blue") +
  labs(title = "NFL 2021: Average Points Scored / Game vs. QB Passer Rating",
       x = "Team Passer Rating", y = "Average Points Scored / Game") + theme_stata() +
  geom_abline(slope=xb1, intercept=xb0, color = "maroon")

graph
```



The graph look very similar to the one above because the only difference would be in the how all of the data is scaled. The correlation coefficient is the same because the data is the same just scaled differently.

3 — Read in the NHL team-game data, and examine the relationship between assists and goals (for the regression part, goals are the outcome, assists are the predictor, i.e., regress goals on assists).

```
xdf <- read_tsv("/Users/apurvashah/Documents/GitHub/stats10/lab4/NHL_20202021_teamGame.tsv", col_names =
head(xdf)
```

```
## # A tibble: 6 x 23
##       date   season team           VorH  team_goals team_pim team_shots team_powerPlayGo~
```

```
##      <dbl>    <dbl> <chr>           <chr>      <dbl>    <dbl>      <dbl>             <dbl>
## 1 20210113 20202021 Pittsburgh Pen~ VT             3        6         34                 1
## 2 20210113 20202021 Philadelphia F~ HT             6        6         27                 2
## 3 20210113 20202021 Chicago Blackh~ VT             1        8         23                 1
## 4 20210113 20202021 Tampa Bay Ligh~ HT             5        6         33                 2
## 5 20210113 20202021 Montréal Canad~ VT             4       13         32                 2
## 6 20210113 20202021 Toronto Maple ~ HT             5       11         34                 2
## # ... with 15 more variables: team_powerPlayOpportunities <dbl>, team_blocked <dbl>,
## #   team_takeaways <dbl>, team_giveaways <dbl>, team_hits <dbl>, assists <dbl>,
## #   goals <dbl>, shots <dbl>, powerPlayGoals <dbl>, powerPlayAssists <dbl>,
## #   penaltyMinutes <dbl>, faceOffWins <dbl>, faceoffTaken <dbl>, shortHandedGoals <dbl>,
## #   shortHandedAssists <dbl>
```

```
x <- data.frame(Team = xdf$goals, avg = xdf$assists)
head(x)
```

```
##   Team avg
## 1    3   3
## 2    6  11
## 3    1   2
## 4    5  10
## 5    4   8
## 6    5   8
```

```
xmeanPR <- mean(x$Team)
xmeanPR
```

```
## [1] 2.898041
```

```
xmeanPnts <- mean(x$avg)
xmeanPnts
```

```
## [1] 4.87788
```

```
##### sample standard deviation of rating
xsdPR <- sd(x$Team)
xsdPR
```

```
## [1] 1.711624
```

```
##### sample standard deviation of total points
xsdPnts <- sd(x$avg)
xsdPnts
```

```
## [1] 3.009314
```

```
n <- nrow(xdf)
n
```

```
## [1] 1736
```

```r
###### sample covariance
xcov <- sum((x$Team - xmeanPR) * (x$avg - xmeanPnts) ) / (n - 1)
xcov
```

```
## [1] 4.846908
```

```r
###### sample correlation
xcorr <- xcov / (xsdPR * xsdPnts)
xcorr
```

```
## [1] 0.9409984
```

```r
###### sample regression line slope
xb1 <- xcorr * xsdPnts / xsdPR
xb1
```

```
## [1] 1.654429
```

```r
###### sample regression line y-intercept
xb0 <- xmeanPnts - xb1 * xmeanPR
xb0
```

```
## [1] 0.08327668
```

4 — With NHL team-game data, can we argue that increasing assists causes more goals?

Although when looking at the data, we can see that when there are more assists, there generally more goals, since we did not conduct and experiment and did not use random assignment we cannot assume causality here. We can say that there is a correlation.

5 — Back to the NFL data, can we say that improving passer rating causes more points to be scored?

Note that 4 and 5 are intended to test your reasoning. These questions may not be as simple as they appear.

Although we might see a trend that teams with better passer ratings cause more points to be scored we cannot say that this causes more points to be scored because we did not conduct a true experiment to test this. We cannot assume causality and at most can say that there might be a correlation between these two.