

Lab 4

Your Name and UID go here

2022-04-27

Contents

Examples	1
NFL 2021 Season Total Team Offense	1
Bivariate Association.	3
Context	5
All at Once	6
Your Work	8

```
## Date last run: 2022-04-27
```

```
## Hello World!
```

Examples

Requires library xtable.

NFL 2021 Season Total Team Offense

```
#### note ``` (3 backticks)
##### R code goes in here. Set code chunk environment options above

## We have an Excel file
library(gdata)

## Read in our data
xdf <- read.xls("NFL_offense_passing_2021.xlsx", sheet=1, header=TRUE)

head(xdf, n=6)
```

```
##      Team Att Cmp Cmp.. Yds.Att Pass.Yds TD INT Rate X1st X1st. X20 X40 Lng Sck SckY
## 1 49ers 514 343 66.7 8.6 4437 26 14 99.2 200 38.9 63 11 83 33 216
## 2 Bears 542 332 61.3 6.7 3635 16 20 75.5 180 33.2 40 7 64 58 428
## 3 Bengals 555 384 69.2 8.7 4806 36 14 106.9 208 37.5 63 16 82T 55 403
## 4 Bills 655 415 63.4 6.8 4450 36 16 91.3 236 36.0 51 8 61 27 166
## 5 Broncos 541 354 65.4 7.1 3856 20 9 91.7 179 33.1 46 7 64 40 263
## 6 Browns 520 320 61.5 7.0 3619 21 14 84.6 177 34.0 47 9 71T 49 299
## totalPoints
## 1 427
```

```
## 2      311
## 3      460
## 4      483
## 5      335
## 6      349
```

Rate: passer rating (NFL QB rating)

This data set was made by processing data obtained from NFL.com

Let's create the distribution of total season team points. `totalPoints` appears to be numerical; an integer. What does R think?

```
class(xdf[, "totalPoints"])
```

```
## [1] "integer"
```

```
### the par() function allows us to set parameters of subsequent graphic.
### here we set cex parameter, which controls the relative size of assets like title font
par(cex=0.65)
hist(xdf[, "totalPoints"], main="Total Team Points, NFL 2021 Season")
```

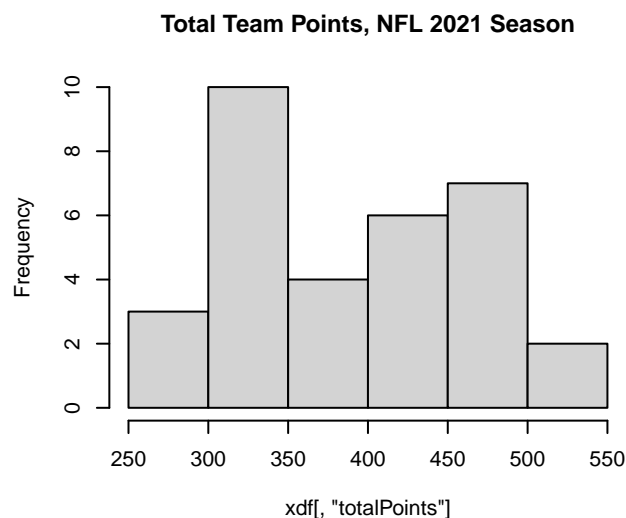


Figure 1: Distribution of Total Team Points

Let's create the distribution of season game average QB rating. `Rate` appears to be numerical. What does R think?

```
class(xdf[, "Rate"])
```

```
## [1] "numeric"
```

```
par(cex=0.65)
hist(xdf[, "Rate"], main="Team Game Avg Passer Rating, NFL 2021 Season")
```

Team Game Avg Passer Rating, NFL 2021 Season

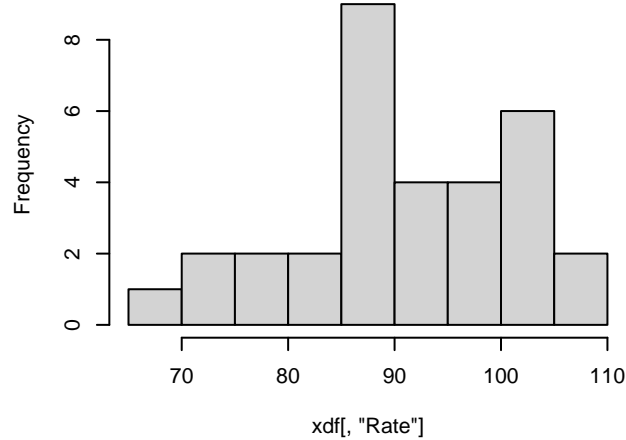


Figure 2: Distribution of Average Team Passer Rating

Bivariate Association.

By far the most popular way to graphically convey the bivariate relationship between two numeric attributes is the scatterplot.

```
par(cex=0.65)
plot(
  xdf[, "Rate"],
  xdf[, "totalPoints"],
  xlab="Game Avg QB Rating",
  ylab="Total Points Scored",
  main="Points Scored vs QB Rating, NFL 2021 Regular Season"
)
```

Points Scored vs QB Rating, NFL 2021 Regular Season

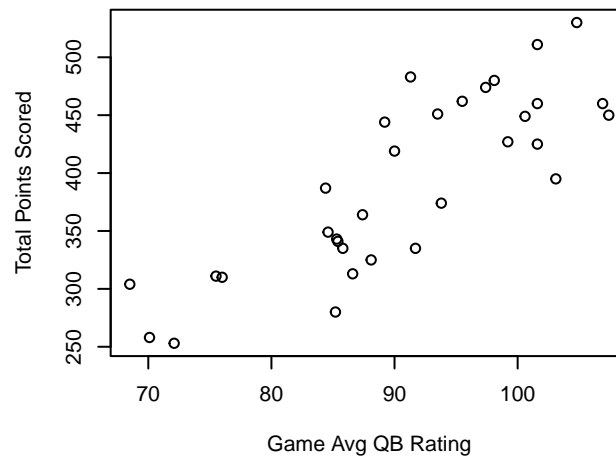


Figure 3: Scatterplot of total points scored vs QB rating, NFL 2021 regular season

Let's make this a little prettier.

```
par(cex=0.65)
plot(
  xdf[, "Rate"],
  xdf[, "totalPoints"],
  xlab="Game Avg QB Rating",
  ylab="Total Points Scored",
  pch=19,
  cex=2,
  col="#99339999",
  main="Points Scored vs QB Rating, NFL 2021 Regular Season"
)
```

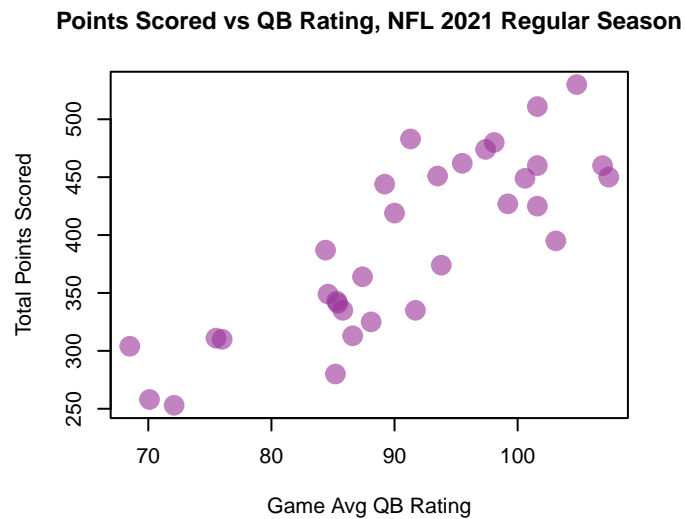


Figure 4: Scatterplot of total points scored vs QB rating, NFL 2021 regular season

Let's calculate some important univariate statistics.

```
xmeanPR <- mean(xdf[, "Rate"])
xmeanPR
```

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

```
xmeanPnts <- mean(xdf[, "totalPoints"])
xmeanPnts
```

```
## [1] 390.6875
```

```
##### sample standard deviation of rating
```

```
xsdPR <- sd(xdf[, "Rate"])
xsdPR
```

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

```
##### sample standard deviation of total points
```

```
xsdPnts <- sd(xdf[, "totalPoints"])
xsdPnts
```

```
## [1] 76.61358
```

Now let's perform a detailed calculation of some important bivariate statistics.

```

### sample size
n <- nrow(xdf)
n

## [1] 32

##### sample covariance
xcov <- sum( (xdf[ , "Rate"] - xmeanPR) * (xdf[ , "totalPoints"] - xmeanPnts) ) / (n - 1)
xcov

## [1] 674.5183

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

## [1] 0.8288188

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

## [1] 5.977739

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

## [1] -151.4747

```

Let's drop our LS line into our scatterplot.

```

par(cex=0.63)
plot(
  xdf[ , "Rate"],
  xdf[ , "totalPoints"],
  xlab="Game Avg QB Rating",
  ylab="Total Points Scored",
  pch=19,
  cex=2,
  col="#99339999",
  main="NFL 2021 Reg Season: Points Scored vs QB Rating, w/LS solution"
)

abline(a=xb0, b=xb1, lwd=3, col="#AA00AA99")

```

Context

Let's verbally put things into context:

Considering the 2021 NFL season, looking at team totals . . .

We have a total of 32 teams.

The average total team points scored over the season is 390.69 points.

The average of the game average team passer rating over the season is 90.7 rating points.

The covariance between total team points scored and game average passer rating is 674.52.

Pearson's correlation between total team points scored and game average passer rating is 0.8288.

NFL 2021 Reg Season: Points Scored vs QB Rating, w/LS solution

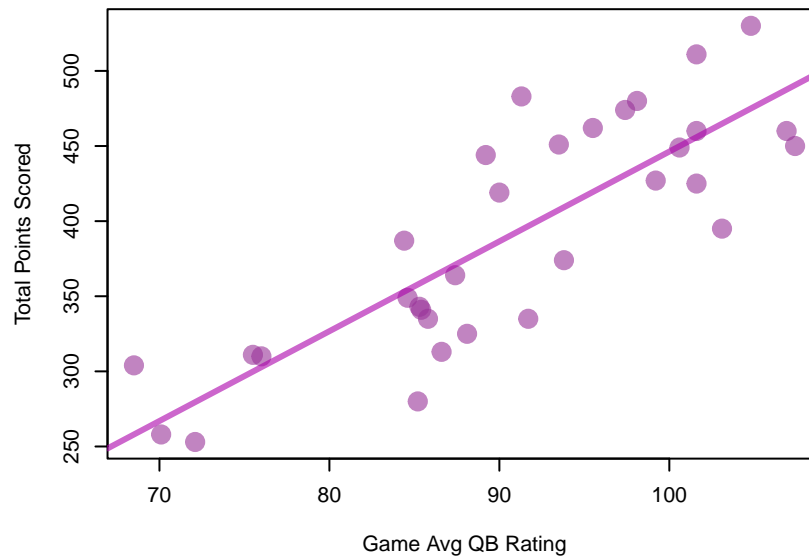


Figure 5: Scatterplot of points scored vs QB rating, NFL 2021 regular season with LS solution

If we fit an SLR (simple linear regression) line to total team points over game-average team passer rating, we obtain a slope of 5.978 and a y-intercept of -151.475.

One interpretation of this trend would be that a 10 unit increase in passer rating is associated with a 59.777 units increase — on average — in total team points scored over the season.

All at Once

R includes a high-level function that performs all the regression calculations.

The `lm()` function.

```
x <- xdf[ , "Rate"]
y <- xdf[ , "totalPoints"]
```

```
xlm <- lm(y ~ x)
```

```
summary(xlm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -77.829 -28.375  -6.111  42.761  88.707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -151.4747    67.2650  -2.252   0.0318 *
## x              5.9777     0.7368   8.113 4.67e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  
## Residual standard error: 43.58 on 30 degrees of freedom  
## Multiple R-squared:  0.6869, Adjusted R-squared:  0.6765  
## F-statistic: 65.83 on 1 and 30 DF,  p-value: 4.674e-09
```

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?

2 — Each team played 17 regular season games. Calculate the average points scored per game for each team. Create a scatterplot showing avg team points versus avg passer ratings with the least squares line ...

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?

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).

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

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.