EDP308: STATISTICAL LITERACY

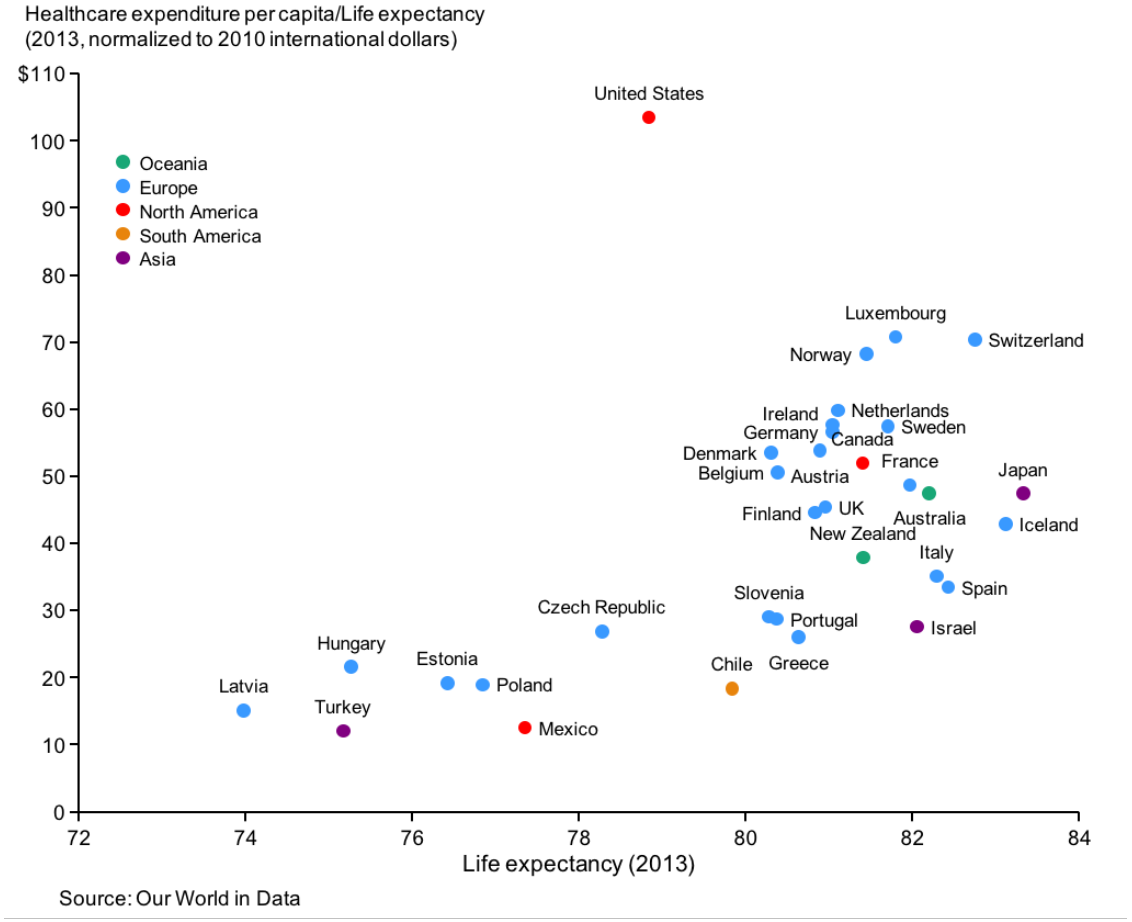
The University of Texas at Austin, Fall 2020
RAZ: Rebecca A. Zárate, MA

Overview

- Scatterplots
- Correlation
  - Linear Relationship
  - Strength
  - Direction
- Calculating the Pearson Correlation
- Effects on Correlation
  - Restricted Range
  - Outliers
- Spotify Top Tracks of 2000 Example
- Coefficient of Determination (R^2)
- Correlation in R

Quick Glance

- Do you think there is a relationship, an association, between Life Expectancy and Healthcare Expenditure?



Correlation

Correlation (Pearson's)

- Correlation measures the strength and direction of the *linear association/relationship* between two quantitative variables
- It answers the questions:
  - How do these two things relate to each other?
  - If one of the variables goes up, what happens to the other one?
  - If one of the variables goes down, what happens to the other one?

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Historical Moment: Karl Pearson

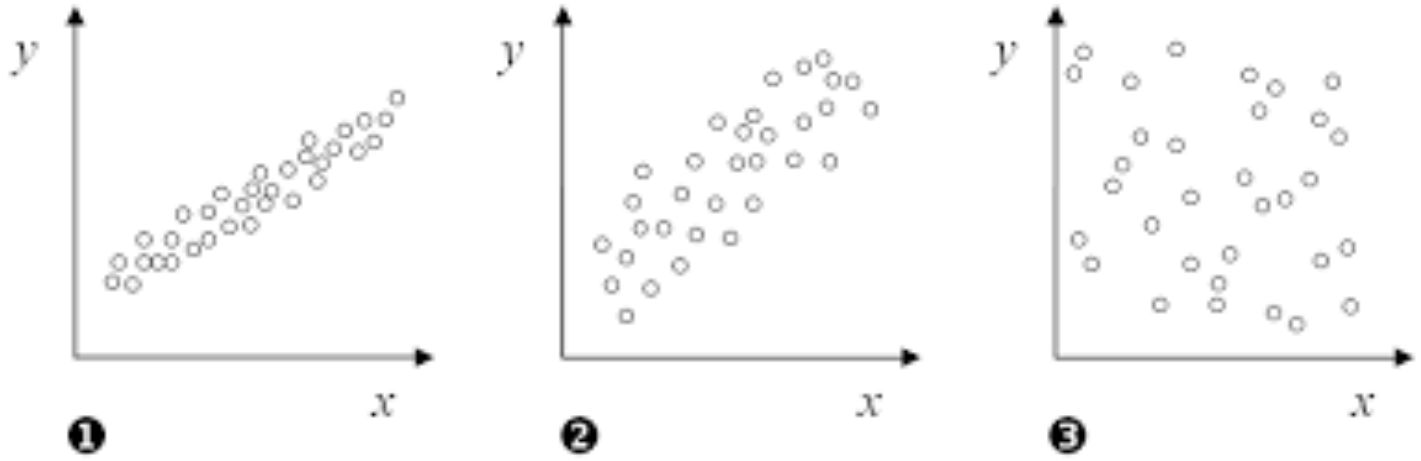
- **Karl Pearson** (March 27, 1857 – April 27, 1936)
- Some good stuff:
  - A mathematician and biostatistician.
  - Founded the first Statistics department at University College, London.
  - He developed some great statistical tools like the Pearson Product Moment Correlation… But…

Historical Moment: Karl Pearson

- **Karl Pearson** (March 27, 1857 – April 27, 1936)
- He was the protégé of Sir Francis Galton and like Galton was into "social Darwinism" and eugenics… (another one)
  - And as such has some racist thinking… Didn't really believe people could rise from the "lower levels"
  - He was quoted once as saying: "…if the bad stock be raised the good is lowered…"

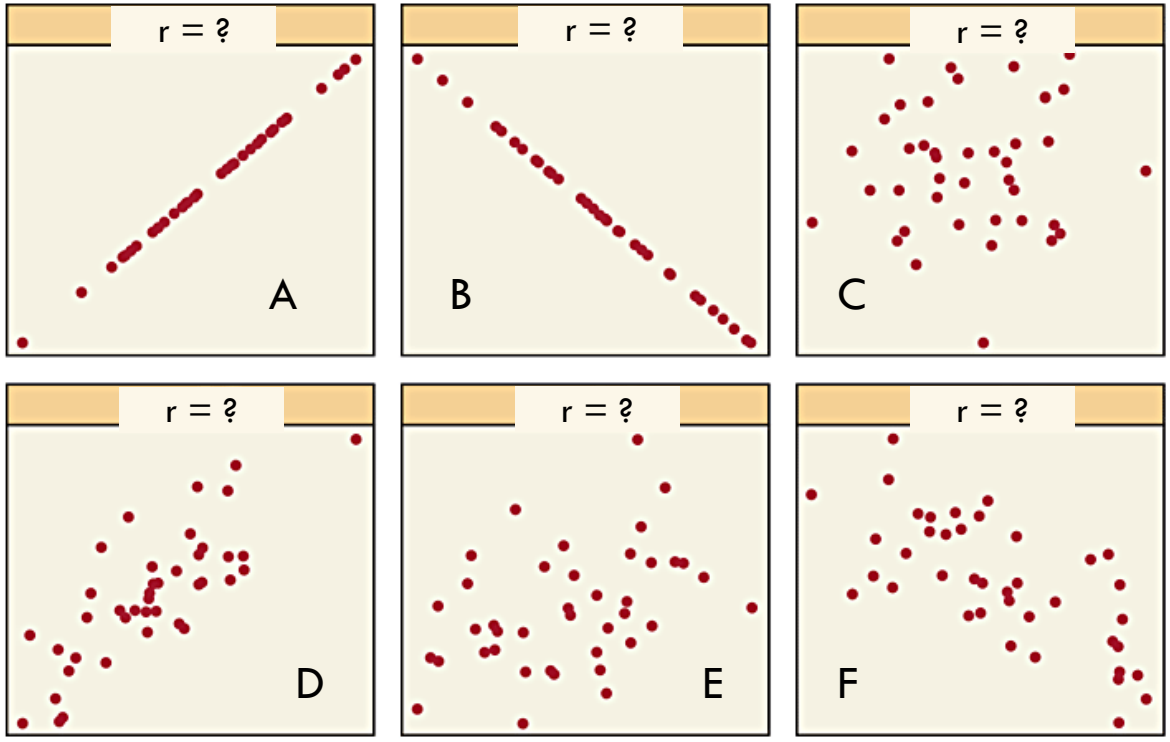
Main Characteristics

- Three essential characteristics of a correlation
  - ***(Linear) Trend***
    - Correlations apply ONLY to linear relationships!
    - Check scatterplots to ensure you have a linear relationship
  - ***Direction***
    - Positive or negative
  - ***Strength***
    - Magnitude of the correlation



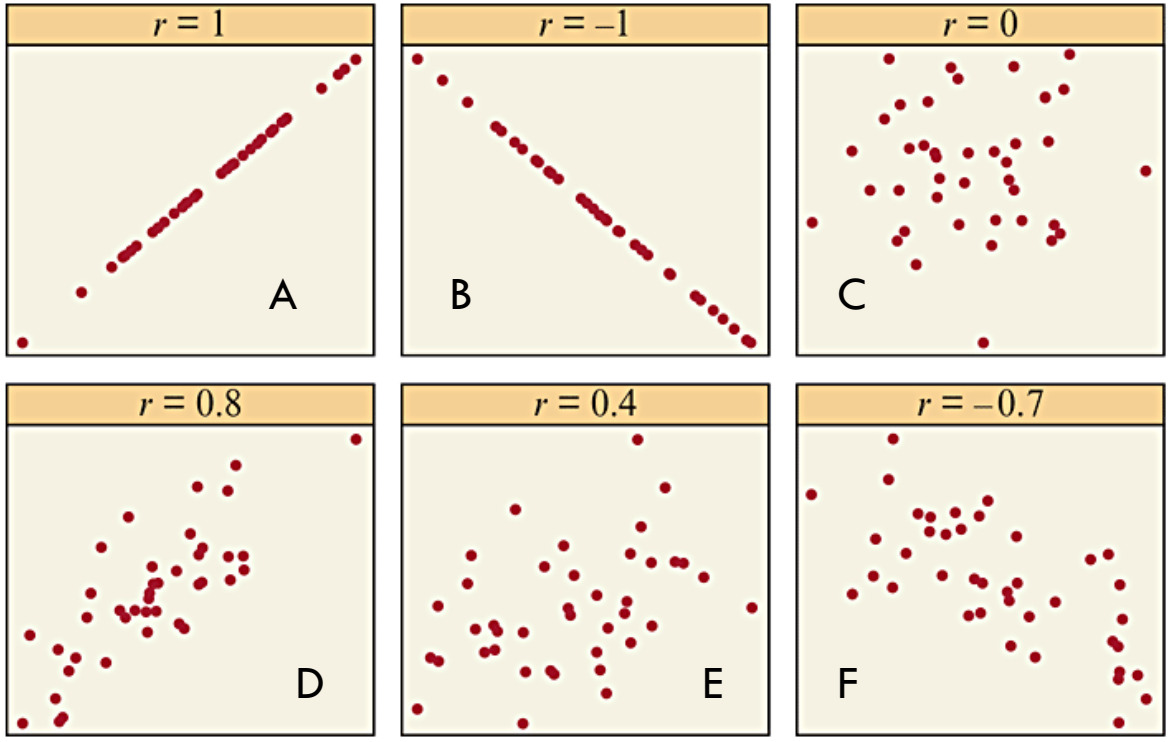
Scatter Plots

By eye, guess which correlation matches with each scatterplot.

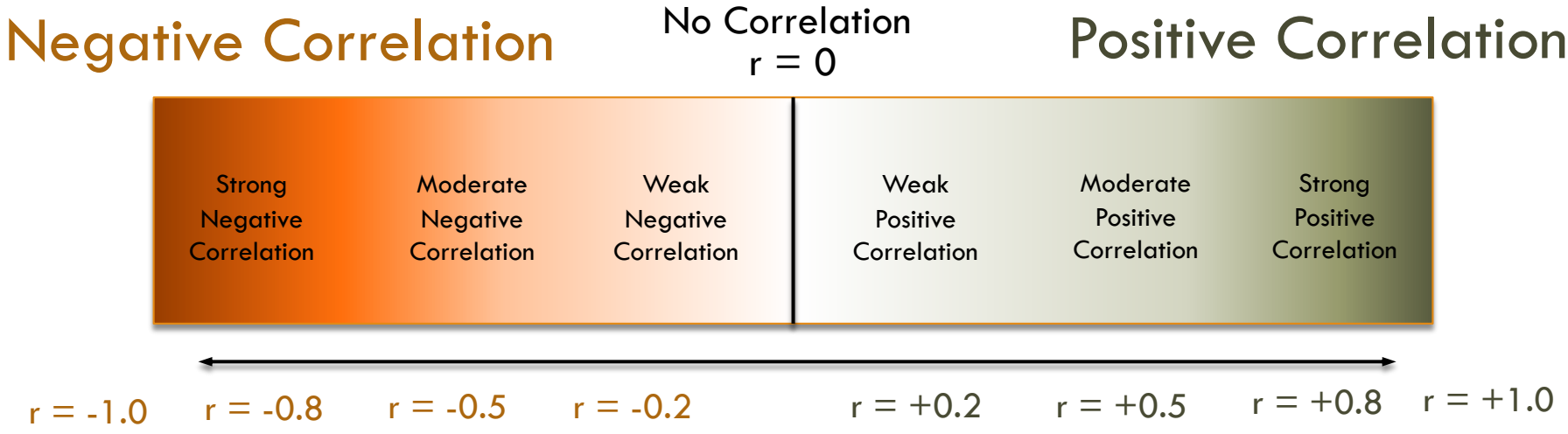


Scatter Plots

By eye, guess which correlation matches with each scatterplot.



Pearson Correlation Scale



- **No Association:** r = 0
- **Weak:** r = .1 to r = .3
- **Moderate**: r = .4 to r = .6
- **Strong**: r = .7 to r = .9
- **Perfect:** r = 1.0

Correlation Properties

- Always falls between -1 and +1.
- The sign of correlation denotes direction
  - (-) indicates negative linear association.
  - (+) indicates positive linear association.
- Correlation has a unit-less measure, it does not depend on the variables' units.
- Two variables have the same correlation no matter which is treated as the response variable.
- Correlation is not resistant to outliers.
- Correlation ***only*** measures strength of a **linear relationship.**
- Correlation does not imply causation!

Positive or Negative?

- **Correlation** refers to the degree to which two quantitative variables are associated.
  - Synonyms: association, relationship, covariance, dependence
- Positive Correlations
  - Cognitive ability and grades
  - Self-esteem and job success
  - Education and income
- Negative Correlations
  - Depression and self-esteem

Give me some more examples.

Spurious Correlations

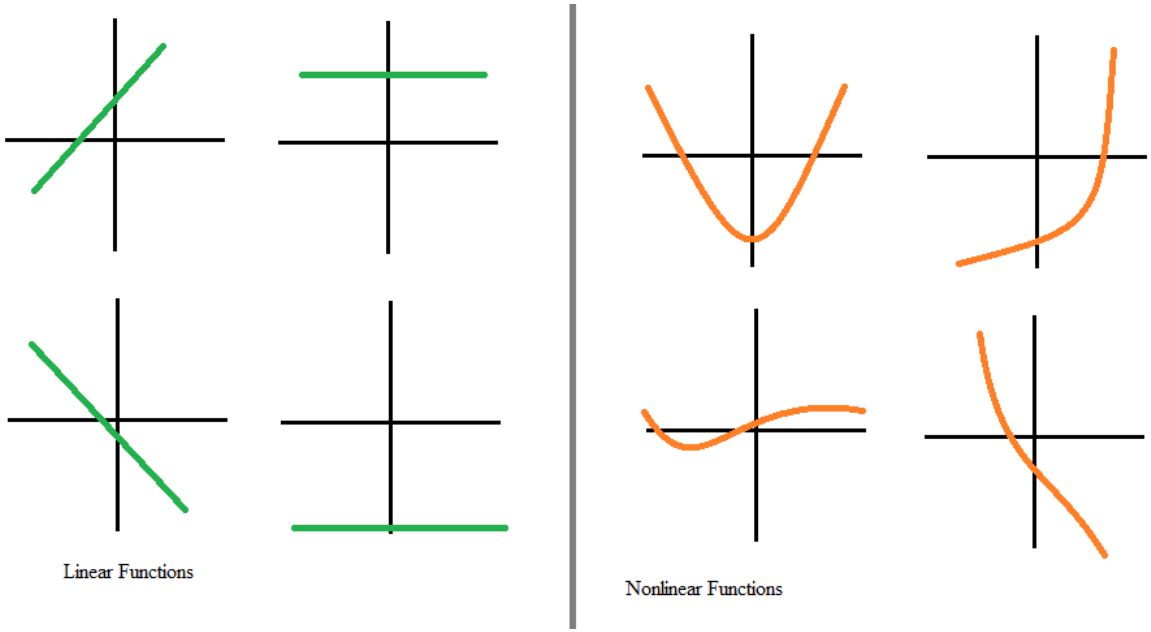
- I am sure most of you have heard the mantras:
  - The mitochondria is the power house of the cell.
  - Correlation does not equal causation.
    - Equally important, but we'll focus on the second one.

Spurious Correlation

Tyler Vigen

Linear vs. Non-Linear

- When we plot the relationship between two variables, the relationship can look either linear or nonlinear
- **Nonlinear relationships** are represented by one or more *curves* in the data
  - These are predictable, but with more advanced techniques
    - That's a different class…

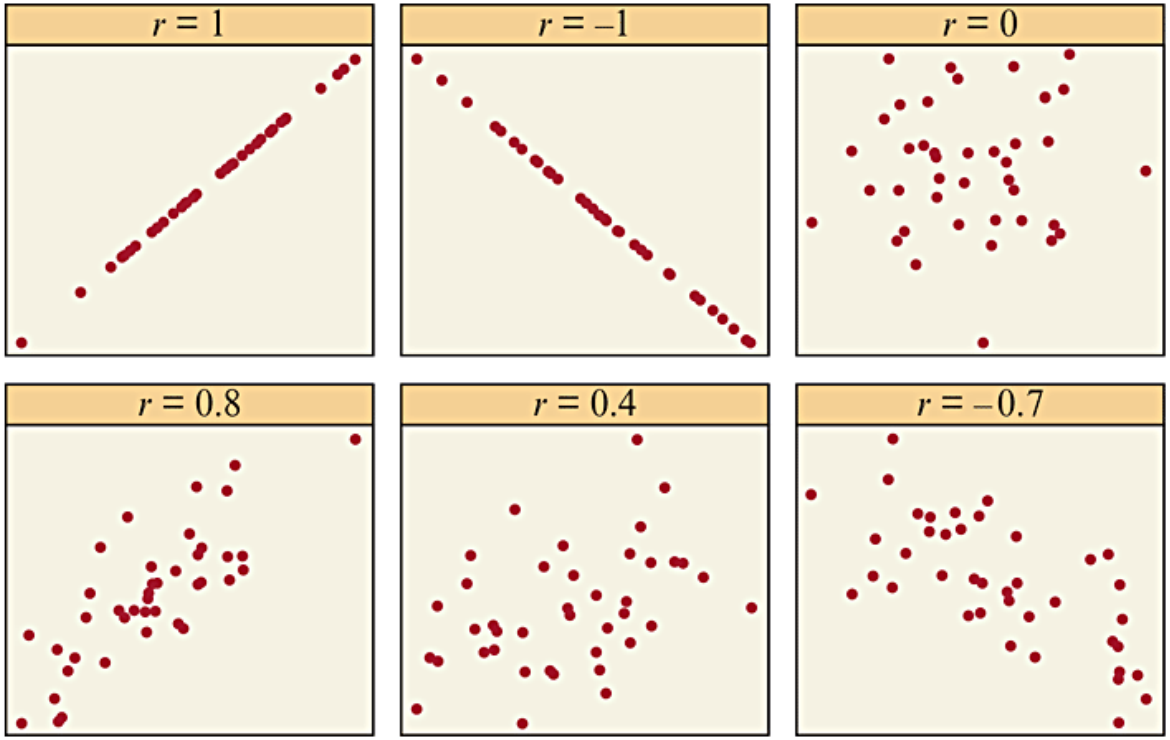


Pearson Correlation Coefficient

- We can calculate a test statistic (just like we've been doing) to tell us whether the correlation is strong
  - Always assuming the relationship is linear
  - This is known as the Pearson correlation coefficient
- To calculate Pearson correlation coefficient, we need:
  - A linear relationship between variables
  - Two quantitative variables
  - Data for each person in the sample on both variables
    - Ex. A measure of the level of stress (Variable 1) and the number of assignments due at the end of the semester (Variable 2)…

Stress and Work

- A measure of the level of stress (Variable 1) and the number of assignments due at the end of the semester (Variable 2)…

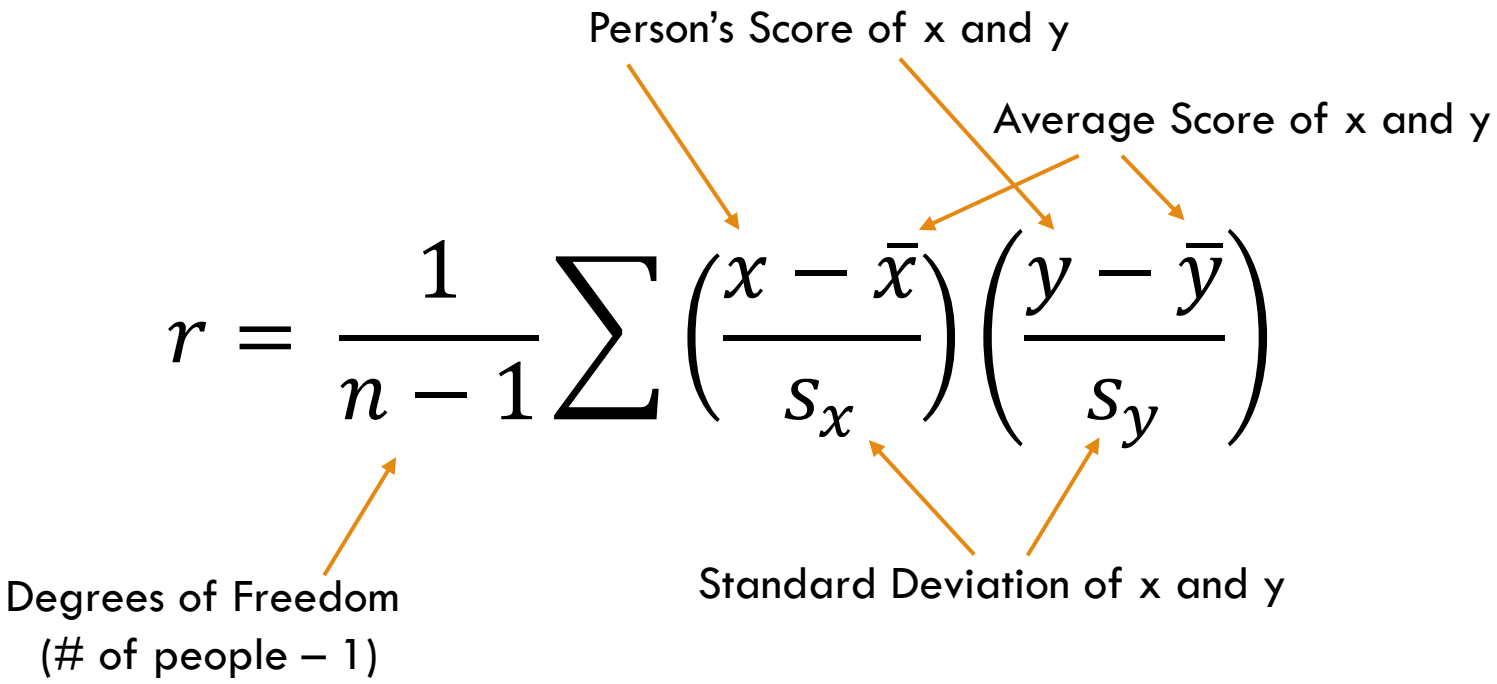


Which one of the scatter plots do you think would best model represent this correlation?

Calculating the Pearson Correlation

Calculating the Pearson Correlation

- The Pearson correlation coefficient is defined as:



Calculating the Pearson Correlation

- The Pearson correlation coefficient is defined as:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

(same thing)

$$r = \frac{1}{n-1}\sum z_x z_y$$

Translating each person's score on x and y into z-scores and multiplying them

Calculating the Pearson Correlation

- Your z-score on x multiplied by your z-score on y.
- Then add up everyone's ($z_x * z_y$)
- Lastly divide by the degrees of freedom

$$r = \frac{\sum z_x z_y}{df}$$

$$r = \frac{1}{n-1}\sum z_x z_y \qquad r = \frac{\sum z_x z_y}{n-1}$$

(all the same thing)

Try it. Studying and Correct Answers

The table below shows the number of hours spent studying, and the number of questions correct on a quiz. Compute the Pearson Correlation Coefficient.

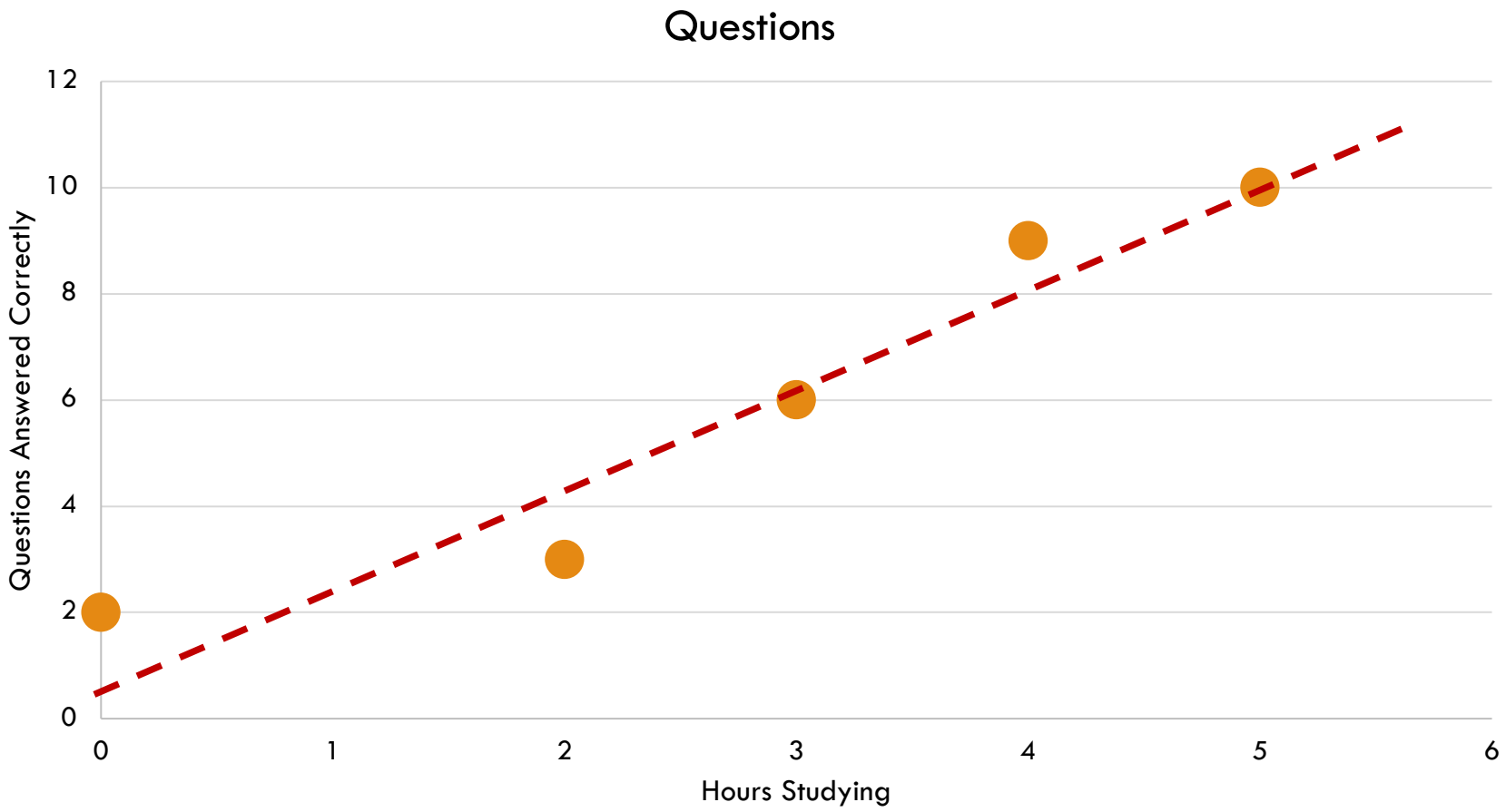
You are given the following:

$$\bar{X}_{Hours} = 2.8, s_{Hours} = 1.92,$$

$$\bar{X}_{Questions} = 6, s_{Questions} = 3.54$$

Hours	Questions
0	2
2	3
3	6
4	9
5	10

Studying and Questions Correct



Try it. Studying and Correct Answers

$$\bar{X}_{Hours} = 2.8, s_{Hours} = 1.92$$
$$\bar{X}_{Questions} = 6, s_{Questions} = 3.54$$

Hours	Questions	Z_{Hours}	$Z_{Questions}$	$Z_{Hours} * Z_{Questions}$
0	2			
2	3			
3	6			
4	9			
5	10			

$$\left(\frac{x - \bar{x}}{s_x}\right) = Z_{Hours} \qquad \left(\frac{y - \bar{y}}{s_y}\right) = Z_{Questions}$$

Try it. Studying and Correct Answers

$$\bar{X}_{Hours} = 2.8, s_{Hours} = 1.92$$
$$\bar{X}_{Questions} = 6, s_{Questions} = 3.54$$

Hours	Questions	Z_{Hours}	$Z_{Questions}$	$Z_{Hours} * Z_{Questions}$
0	2	-1.46	-1.13	
2	3	-0.42	-0.85	
3	6	0.10	0.00	
4	9	0.63	0.85	
5	10	1.14	1.13	

$$Z_{Hours} * Z_{Questions} = z_x z_y$$

Try it. Studying and Correct Answers

$$\bar{X}_{Hours} = 2.8, s_{Hours} = 1.92$$
$$\bar{X}_{Questions} = 6, s_{Questions} = 3.54$$

Hours	Questions	Z_{Hours}	$Z_{Questions}$	$Z_{Hours} * Z_{Questions}$
0	2	-1.46	-1.13	+1.65
2	3	-0.42	-0.85	+0.34
3	6	0.10	0.00	+0.00
4	9	0.63	0.85	+0.53
5	10	1.14	1.13	+1.29

$$r = \frac{1}{5-1}(1.64 + .36 + 0 + .53 + 1.29)$$
$$= \frac{1}{4}(3.82)$$
$$\approx .96$$

Degrees of Freedom: (n-1)

Sum of all the Product

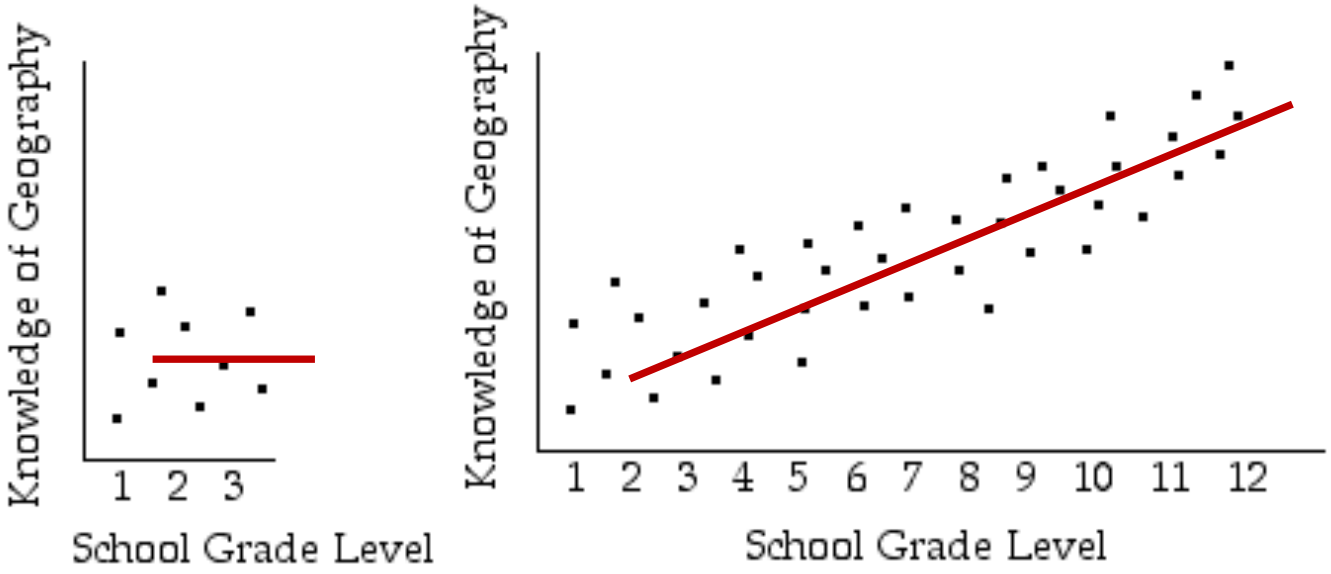
Effects on Correlation

Effects on Correlation

What sorts of factors might affect a correlation?

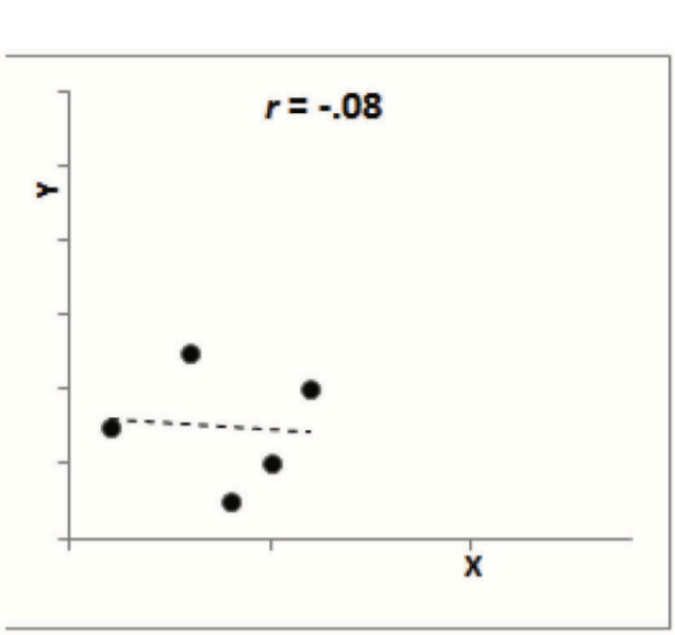
Restricted Range

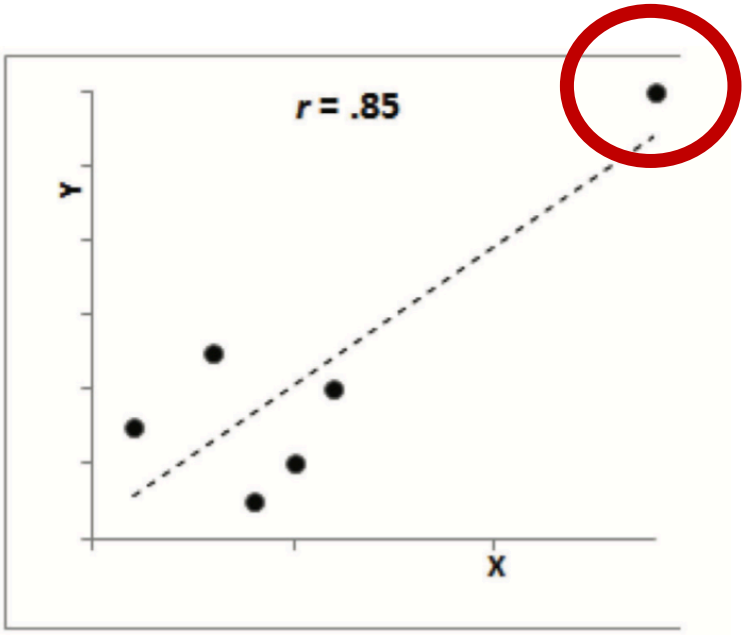
- Correlations will be manipulated when the entire range of a values is not represented
  - How much you zoom in matters



Outliers

- Outliers
  - Extreme values can distort and skew a correlation
    - They can pull things one way or the other
  - Outliers can make a correlation either weaker or stronger
  - Always check the scatterplot for outliers before computing a correlation





Spotify Top Tracks of 2000

Spotify Top Tracks of 2000

- This dataset contains audio statistics of the top 2000 tracks on Spotify and includes the following variables:
  - Year
  - BPM
  - Energy
  - Danceability
  - Loudness(bd)
  - Liveness
  - Valence
  - Length(Duration
  - Acousticness
  - Speechiness
  - Popularity

Which variables do you think will have high correlations?

A Correlation Matrix

	BeatsPerMinute_BPM	Energy	Danceability	Loudness_dB	Liveness	Valence	Duration	Acousticness	Speechiness	Popularity
BeatsPerMinute_BPM	1									
Energy	0.16	1								
Danceability	-0.14	0.14	1							
Loudness_dB	0.09	0.74	0.04	1						
Liveness	0.02	0.17	-0.1	0.1	1					
Valence	0.06	0.41	0.51	0.15	0.05	1				
Duration	0.02	0.04	-0.1	-0.04	0.01	-0.22	1			
Acousticness	-0.12	-0.67	-0.14	-0.45	-0.05	-0.24	-0.13	1		
Speechiness	0.09	0.21	0.13	0.13	0.09	0.11	-0.03	-0.1	1	
Popularity	0	0.1	0.14	0.17	-0.11	0.1	-0.04	-0.09	0.11	1

Which variables have the highest correlations?

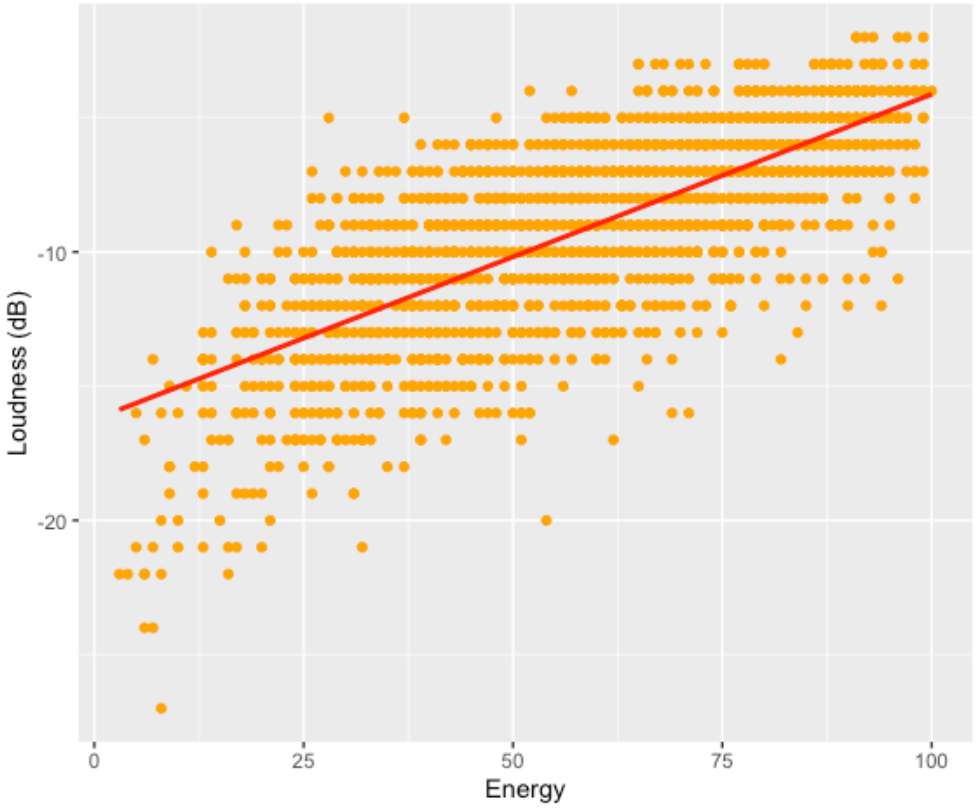
A Correlation Matrix

	BeatsPerMinute_BPM	Energy	Danceability	Loudness_dB	Liveness	Valence	Duration	Acousticness	Speechiness	Popularity
BeatsPerMinute_BPM	1									
Energy	0.16	1								
Danceability	-0.14	0.14	1							
Loudness_dB	0.09	0.74	0.04	1						
Liveness	0.02	0.17	-0.1	0.1	1					
Valence	0.06	0.41	0.51	0.15	0.05	1				
Duration	0.02	0.04	-0.1	-0.04	0.01	-0.22	1			
Acousticness	-0.12	-0.67	-0.14	-0.45	-0.05	-0.24	-0.13	1		
Speechiness	0.09	0.21	0.13	0.13	0.09	0.11	-0.03	-0.1	1	
Popularity	0	0.1	0.14	0.17	-0.11	0.1	-0.04	-0.09	0.11	1

- Energy and Loudness (.74)
- Energy and Acoustic-ness (-.67)
- Valance and Danceability (.51)
- Acoustic-ness and Loudness (-.45)
- Valance and Energy (.41)

Energy and Loudness (0.74)

- What is the correlation between Loudness dB (the actually decibel level) and Energy (a measure of intensity, ex death metal would have high energy)?
  - 0.74



Coefficient of Determination (R^2)

Coefficient of Determination (R^2)

- Tell us how much of the variability in one variable (ex. questions answered correctly) is explained by the variability in the other variable (ex. the number of hours you studied)
  - The proportion of variance in the response variable (ex. questions answered correctly) that is explained after accounting for the explanatory variable (ex. the number of hours you studied)
- $Coefficient\ of\ Determination\ (\mathrm{R}^2) = \mathrm{r}^2$
  - $1 - \mathrm{R}^2$ is the proportion of variance in the response variable that is left unexplained after accounting for the explanatory variable

Coefficient of Determination (R^2)

- How much of the variability in the sample IS explained?
  - R^2
- How much of the variance in the sample IS NOT explained?
  - $1 - \mathrm{R}^2$

Try it.

1. Compute and interpret the coefficient of determination (R^2) for the number of hours spent studying, and the number of questions correct on a quiz ($r = .96$).

2. Determine the proportion of variance in the number of quiz questions answered correctly that is left unexplained, after accounting for the number of hours spent studying.

Try it. #1

1. Compute and interpret the coefficient of determination for the number of hours spent studying, and the number of questions correct on a quiz ($r = .96$).

$$R^2 = .96^2 = .9216$$

Approximately 92.16% of the variance in the number of quiz questions answered correctly can be explained by the variance in the number of hours of studying.

Try it. #2

2. Determine the proportion of variance in the number of quiz questions answered correctly that is left unexplained, after accounting for the number of hours spent studying.

Approximately 92.16% of the variance in the number of quiz questions answered correctly can be explained by the variance in the number of hours of studying.

This means that $100\% - 92.16\% = 7.84\%$ of the variance in the number of quiz questions answered correctly is left unexplained, after accounting for the number of hours studied, perhaps due to some other variable like mood, interest in statistics, number of courses, etc

Up Next…

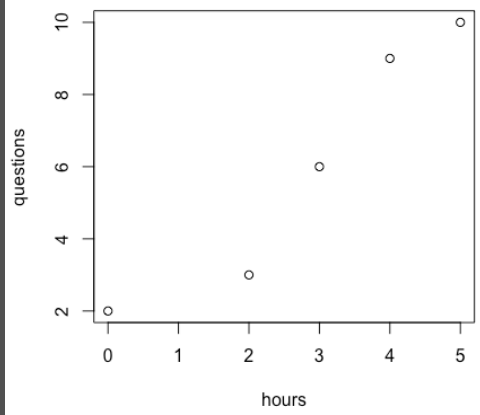
- It's all well and good to tell me how much two variables are correlated, but I want to be to predict someone's quiz grade based on the number of hours they study.
- For this we will need…

Regression

Correlation in R

Studying and Correct Answers in R

```r
# The data
hours <- c(0,2,3,4,5)
questions <- c(2,3,6,9,10)

# Correlations are quick with "cor()"
cor(hours, questions)

# We can also quickly look at a scatterplot
plot(hours, questions)

# Lastly, we can test the statistical significance of a correlation
# The null hypothesis is that the correlation is zero
cor.test(hours, questions)
```




```
        Pearson's product-moment correlation

data:  hours and questions
t = 5.6292, df = 3, p-value = 0.01109
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4689850 0.9971755
sample estimates:
     cor
0.955779
```

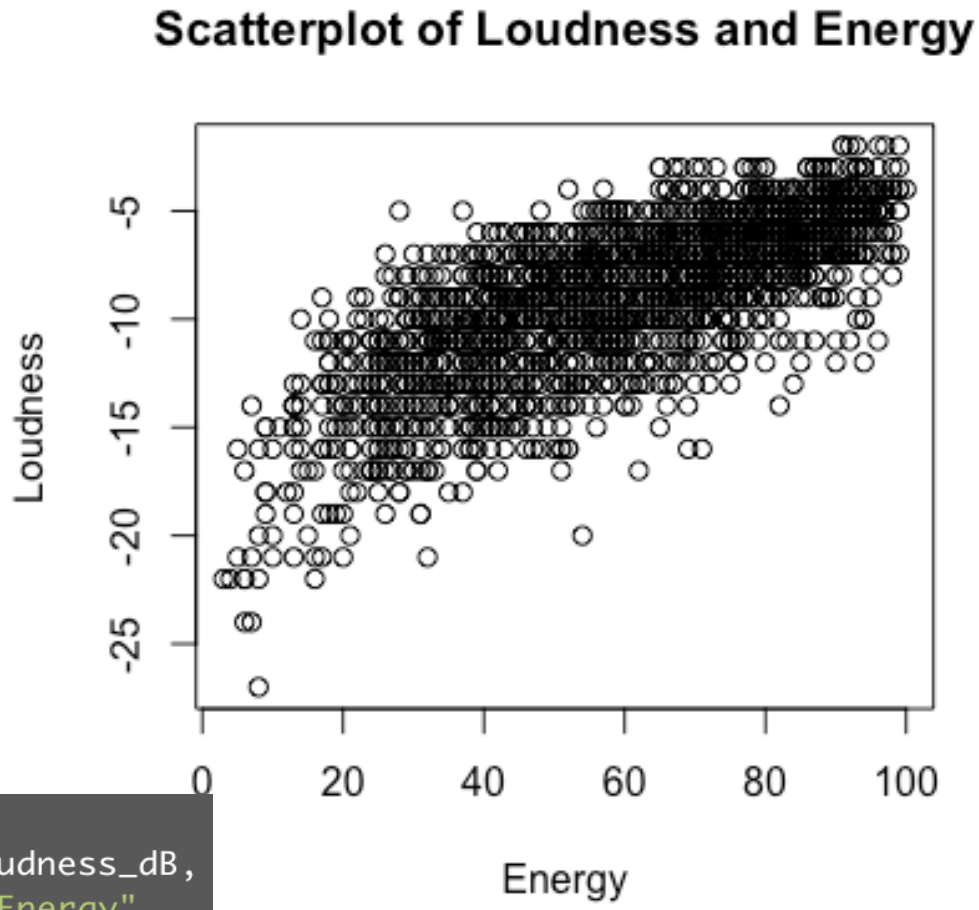
Spotify2000 Data in R

```r
# Read in the data
spotify_2000 <- read.csv("spotify_2000.csv")

# Sometimes R will misinterpret the class of a variable, so we reassign it
# R read the Duration of the song as a factor (categorical variable) rather than a number
# So we use the function "as.numeric" to change it
spotify_2000$Duration <- as.numeric(spotify_2000$Duration)

# We can easily get the correlation for two numeric variable using "cor()"
cor(spotify_2000$Energy, spotify_2000$Loudness_dB) # r = 0.74

# We can plot the data
plot(spotify_2000$Energy, spotify_2000$Loudness_dB,
     main = "Scatterplot of Loudness and Energy",
     xlab = "Energy",
     ylab = "Loudness")

# If we don't now which variable might correlate, we can use the correlation function
# on the numeric data to create a correlation matrix
spotify_corrs <- cor(spotify_2000[, 5:14])
```

Dataset: https://www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset
http://sortyourmusic.playlistmachinery.com/ by @plamere uses Spotify API

Plotting in R

- You can plot any two quantitative variables easily with “plot( )”




```
# We can plot the data
plot(spotify_2000$Energy, spotify_2000$Loudness_dB,
     main = "Scatterplot of Loudness and Energy",
     xlab = "Energy",
     ylab = "Loudness")
```

Spotify2000 Correlation Matrix

	BeatsPerMinute_BPM	Energy	Danceability	Loudness_dB	Liveness	Valence	Duration	Acousticness	Speechiness	Popularity
BeatsPerMinute_BPM	1									
Energy	0.16	1								
Danceability	-0.14	0.14	1							
Loudness_dB	0.09	0.74	0.04	1						
Liveness	0.02	0.17	-0.1	0.1	1					
Valence	0.06	0.41	0.51	0.15	0.05	1				
Duration	0.02	0.04	-0.1	-0.04	0.01	-0.22	1			
Acousticness	-0.12	-0.67	-0.14	-0.45	-0.05	-0.24	-0.13	1		
Speechiness	0.09	0.21	0.13	0.13	0.09	0.11	-0.03	-0.1	1	
Popularity	0	0.1	0.14	0.17	-0.11	0.1	-0.04	-0.09	0.11	1

```
# The rest of this code remove the upper half of the matrix because it is just a repeat of what is on the bottom
# You do not have to do this, it just makes it prettier
upper <- spotify_corrs
upper[upper.tri(spotify_corrs)] <- ""
upper <- as.data.frame(upper)
```