Instructions

- This homework assignment is worth 57 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: October 22, 2021 by 11:59 pm.**

Exercise 1

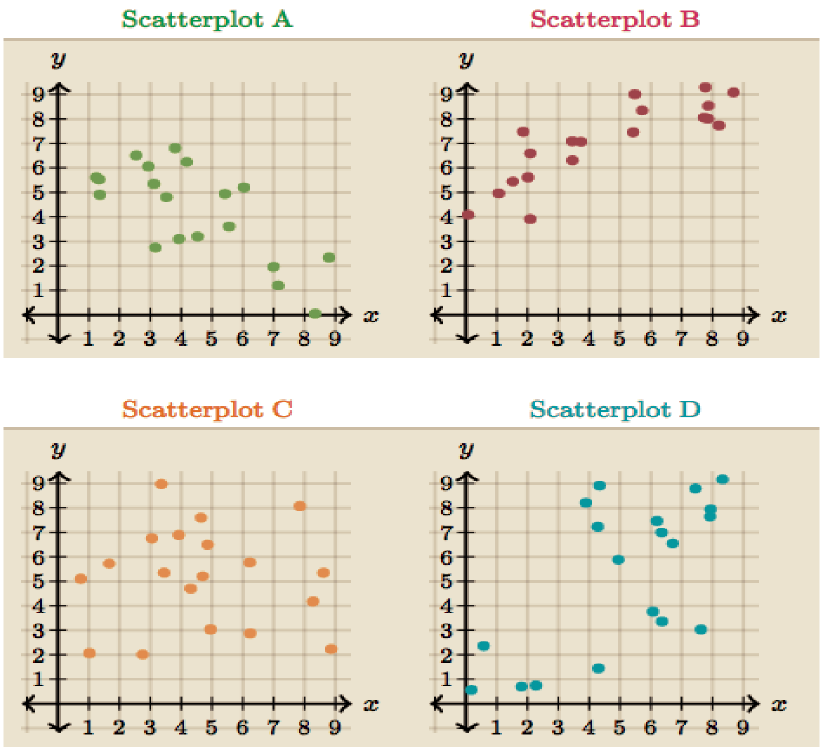
(4 points) What does correlation measure? Be specific.

Exercise 2

(4 points) Why is important to visualize the data using a scatter-plot before computing the correlation? Be specific.

Exercise 3

(5 points) Match the correlation coefficients with the scatter-plots shown below.



Scatterplot A

Scatterplot B

Scatterplot C

Scatterplot D

(a) $r = 0.65$

(b) $r = -0.02$

(c) $r = 0.84$

(d) $r = -0.72$

Exercise 4

What does Nadal do better on clay? Tennis player Rafael Nadal is considered by some to be the greatest clay-court player of all time, with 9 French Open titles among his 14 grand slam wins (as of June 2014). Nadal's performance on clay to his performance on other surfaces during the period 2008-2012 is shown in the below table.

Table 1: Nadal's performance on clay and non-clay surfaces

		Result		
		Loss	Win	Total
Surface	Non-Clay	3658	2715	6373
	Clay	1660	863	2523
	Total	5318	3578	8896

(a) (8 points) Compute the correlation coefficient. Interpret this number. Is this correlation significant?

(b) (5 points) Compute the α, the cross-product ratio. Interpret this number.

(c) (5 points) Compute the Q, Yule's Q. Interpret this number.

Exercise 5

Consider the `Teams.csv` data file. This is one of the files from `Lahman's baseball database`. The `Teams.csv` data file contains seasonal stats for major league teams going back to the first professional season in 1871. **In Python**, do the following:

(a) (4 points) Using pandas, read the csv file and create a data-frame called `mlb`.

(b) (8 points) Create two new variables: `RD` (run differential as `R - RA`) and `Wpct` (winning percentage as `W / (W + L)`).

(c) (4 points) We are interested in studying the relationship between `RD` and `Wpct` for recent seasons. Subset the on seasons since 2001.

(d) (5 points) Create a scatter plot between `RD` and `Wpct`. Describe this plot.

(e) (5 points) Compute the correlation between `RD` and `Wpct`. Describe this correlation.