Homework 5

For this homework you will create a python notebook (`.ipynb` file) and output a `.html` file. Both of these files should then be uploaded to wolfware in the assignment link!

The purpose of this homework is to practice with SQL. Most homework assignments will have a part that pushes you beyond what was in the lectures! Learning to search for the right questions and browsing stackoverflow are really life skills that we should hone :) **Be sure to include markdown text describing what you are doing, even when not explicitly asked for!**

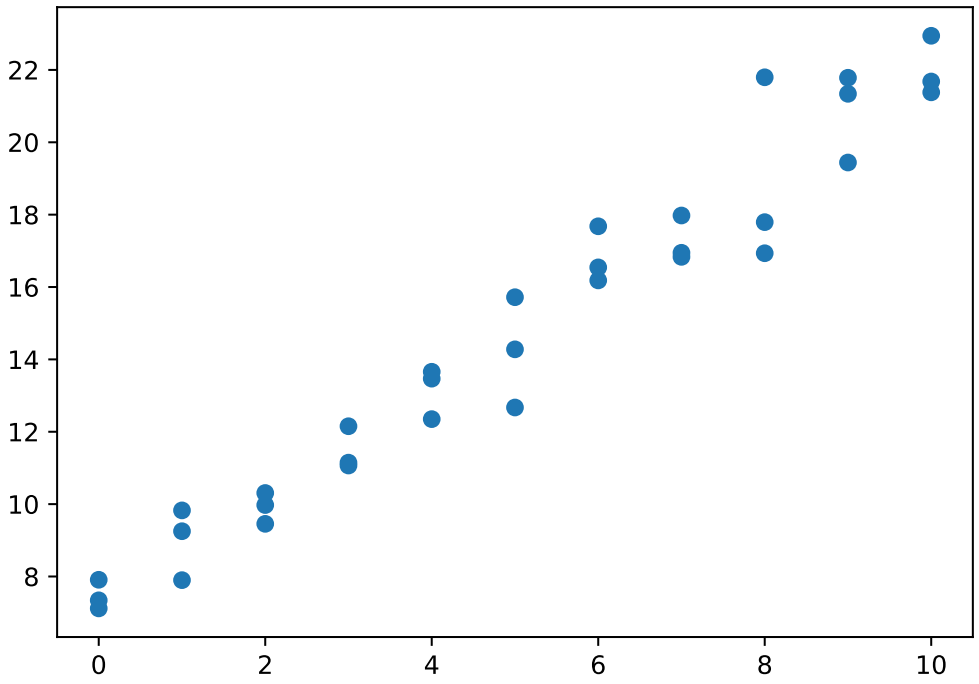
Part I Concept Questions

1. We discussed the "Five V's of Big Data". Give an example of a place where you've encountered big data or a topic you are interested in where big data would arise. Specifically address the five V's for the example and whether/how they apply to your example. (3 pts)
2. We looked at using simulation to investigate the sampling distribution of $\hat{p}$ in the notes. We'll now look at the sampling distribution of the sample slope from an SLR model using simulation.
   - Recall we assume the following model for SLR:

     $$Y_i = \beta_0 + \beta_1 x_i + E_i$$

     where the E_i are assumed to be indendent and identically distributed from a Normal distribution with mean 0 and variance σ^2. Let's assume $\sigma^2 = 1$ for simplicity.
   - We can generate data from this model by assuming values for β_0, β_1, and n, along with a sequence of x values via the following code:

```
#import some modules needed
import matplotlib.pyplot as plt
import numpy as np
from numpy.random import default_rng
rng = default_rng(32)
beta_0 = 7
beta_1 = 1.5
# get three 'values' of x at each integer from 0 to 10.
x = np.array(list(np.linspace(start = 0, stop = 10, num = 11))*3)
n = 33
#create the 'responses' modeled from the line plus a random deviation
y = beta_0 + beta_1*x + rng.standard_normal(n)
#visualize the data
plt.scatter(x = x, y = y)
plt.show()
```



- Now we can use `sklearn` to obtain the estimate for the slope (and save that value) as we did earlier in the course.
- Repeat the above process 5000 times (generating the y values, finding the 'best' slope, and saving that slope). We can use the many values of the sample slope as an approximation to the sampling distribution of the sample slope! (10 pts)
- Now, create a histogram of the sample slope values you found. (3 pts)
- Use your sampling distribution to approximate the probability of observing a sample slope larger than 1.65. Give an interpretation of this value and why it might be important for us in relation to a hypothesis test. (3 pts)

3. Read about the database we'll use below (part II). Give an example for each letter in the `CRUD` acronym in the context of this database. (3 pts)
4. What is the purpose of the `HAVING` clause when writing SQL code? (3 pts)

Part II - Querying a database (34 pts)

There is a database file on the assignment link called Lahman.db that is an sqlite database downloaded from here. This database has information on Major League Baseball.

1. Connect to the database and then look at all of the tables in the database (use `read_sql()` from `pandas` to have this returned as a data frame). (2 pts)
2. Write an SQL query using `pd.read_sql()` that returns all the teams that played in the year 2015 with all of the corresponding columns from the `Teams` table. (2 pts)
3. Write an SQL query using `pd.read_sql()` that returns all of the players in the hall of fame, the year they were voted into the hall of fame, and their category - see the `HallOfFame` table, the inducted variable is important here. (3 pts)

4. Write an SQL query using `pd.read_sql()` that return all unique managers of the Pittsburgh Pirates (`teamID` of `PIT`) and only that information from the `Managers` table. Hint: Check out `SELECT DISTINCT` (3 pts)

5. Use SQL code and the `HallOfFame` and `Managers` tables to return all of the `playerIDs` for the people that managed for a team that were inducted into the hall of fame. Also, programmatically report the number of such people - this can be done in `pandas` after returning the data from the call to `pd.read_sql()`. (4 pts)

6. Now use the same two tables (`HallOfFame` and `mANAGERS`) and an SQL query to return every season managed by each manager that made it to the hall of fame. You should return the `playerID` (manager ID), `G`, `W`, and `L` columns from the `Managers` table.
   Second, determine the overall win/loss records (sum of wins and sum of losses) for each of these hall of fame managers. Third, create a new variable that is the win/loss percentage (`W/(W+L)`). Lastly, sort the resulting data by the win/loss percentage variable (from largest to smallest). The last three parts can be done in `pandas` with the returned data or you can do it via SQL in your call to `pd.read_sql()`. (6 pts)

7. I'm going to give less guidance on this one but it will be similar to the above! Using SQL, construct a table of hall of fame pitchers (any hall of famer that pitched) that gives the `playerID` and their total (sum) for `GS`, `G`, `W`, `L`, `IPOuts`, `CG`, `SHO`, and `SV` columns. The summing can be done in `pandas` or in the SQL call. (6 pts)

8. For all of the hall of fame pitchers, use SQL to create a table of their batting statistics. Namely, the `playerID` and their total (sum) for `AB`, `R`, `H`, `HR`, `RBI`, `BB`, and `SO`. The summing can be done in `pandas` or in the SQL call. (4 pts)

9. Using `pandas` join the previous two tables together by pitcher. (If you want, try to do all of this via SQL! Not required though.) (4 pts)

Files to Submit

- Make sure all cells are run
- Click on File –> Save and Export Notebook As... –> HTML

That's it! Upload both the `.ipynb` and `.html` files to the assignment link.