Exercise 02

This should be completed individually.

Section 1: Querying a database in python

1. Connect to your local postgres instance from python (use the pyscopg2 library).
2. Write a query to return a dataset with film, rental, and payment information. Your dataset should have multiple rows per film, one for each time the film was rented and the amount spent on each rental. Create a dataframe with this information
3. Create a dataframe from the customer table

Section 2: Manipulating dataframes

1. Create a column for customer name that has the first name and last name in the same column.
2. Remove any inactive customers from the dataframe
   Hint: use the active field
3. Change the email addresses to be 'joe.person@wustl.edu', but only when their store_id is an even number
   Hint: use apply to run a function over the dataframe, don't forget to select the correct axis

Section 3: Visualizations

1. How much does each customer tend to spend in aggregate?
   Clarification: You want to first create total spend by customer, then you want to visualize that distribution, each customer being an observation. A box and whisker plot would be a good visualization
2. What does the distribution of film revenue look like?
   Clarification: You want to first calculate the revenue by film, you can sum the rental rate for each instance that the film was rented using the dataframe you created in section 1 part 2. A histogram would be a good visualization

Section 4: Analysis

Clarification: You do not need to use a statistical test to answer the following questions. Please use a visualization and interpret what see. Note the averages, counts, or sums where applicable, and the interpretation

1. On average, is the rental rate the same across movie ratings, treat each film as an observation?
2. Across the various film ratings, are we observing the same number of movies rented at store 2 and store 1?
3. On average, do films with a character of a 'robot' generate the same amount of revenue (use rental rate) as films that feature a 'teacher'?
   Hint: sum the rental rate by film across rentals to get revenue by film

Section 5: Sample Size

1. Generate the following. A reference for how to produce random normal observations can be found here. Use seed(200) to make sure the results are always the same.

   a. x1: 5 observations from a random normal distribution with a mean of 15 and standard deviation of 2
   b. x2: 25 observations from a random normal distribution with a mean of 15 and standard deviation of 2
   c. x3: 125 observations from a random normal distribution with a mean of 15 and standard deviation of 2
   d. x4: 625 observations from a random normal distribution with a mean of 15 and standard deviation of 2
2. Use subplots to display the histograms of all four sets of numbers. An example of how to do this is here. Set kde=True to see how well the histogram approximates a normal p.d.f.
3. Compute the sample means, standard deviations, and standard errors of x1, x2, x3, & x4.
4. Compare these to each other and the parameters of the distribution they come from. How do they differ?

Section 6: Poisson Distribution

1. Using `numpy & seed(100)`, generate a Poisson distribution setting `lam=12` and `size=1200`
2. Compute mean and variance of your 1200 random Poisson values
3. Does the lambda = mean = variance?
4. Repeat 1-3 using `size = 4`. Does lambda=mean=variance?

Section 7: Analysis

Its late 2005, and your boss at the DVD rental company wants to know how effective his customer promotion program was. He tells you, 'I want you to give me some descriptive information about how much the customers spent before and after the program started. Were the spending habits similar? Did they differ? Did the program help or make things worse?'

1. What is the outcome?
2. What is the main effect/predictor he wants to understand the impact of?
3. What is the hypothesis?

Lucky for you, your boss already asked Ted in Bethesda to give you a query for how to get the information.

Query:

```
with b4 as (
    select p.customer_id, sum(p.amount) as Payment_before
    from rental r
    left outer join payment p on p.rental_id = r.rental_id
    where rental_date < cast('2005-07-01' as timestamp) and
        amount is not null
    group by p.customer_id),
aft as (
    select p.customer_id, sum(p.amount) as Payment_after
    from rental r
    left outer join payment p on p.rental_id = r.rental_id
    where rental_date >= cast('2005-07-01' as timestamp) and
```

```
        amount is not null
    group by p.customer_id
)
select distinct c.customer_id, store_id, first_name, last_name,
    active, payment_before, payment_after
from customer c
left outer join b4 r on r.customer_id = c.customer_id
left outer join aft a on a.customer_id = c.customer_id
where payment_after is not null and payment_before is not null
```

Plus, the statistician you work with has some suggestions for how to give your boss what he wants. Query the data from your container and put it in a Pandas dataframe. Then follow the statistician's suggestions.

4. Compute summary statistics and create histograms of the payment_before and payment_after variables. (Try using `describe()` in pandas).
5. Compute the correlation between these two variables and create a scatterplot
6. Compute a variable which is the difference between the amounts spent before and after the program started: payment_after – payment_before.
7. Generate a histogram of the difference and conduct a one-sample t-test.
8. Interpret your results