Exercise 03

This should be completed individually.

Section 1: Sample Size

1. Generate the following. A reference for how to produce random normal observations can be found here. Use seed(205) 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 2: Poisson Distribution

1. Using `numpy & seed(141)`, generate a Poisson distribution setting `lam=10` and `size=1000`
2. Compute mean and variance of your 1000 random Poisson values
3. Does the lambda = mean = variance?
4. Repeat 1-3 using `size = 5`. Does lambda=mean=variance?

Section 3: 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('2005-07-29' 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('2005-07-29' 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.

Use `psycopg2` and `pandas.io.sql` to query the data from your container and put it in a Pandas dataframe. Then follow the statistician's suggestions.

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