Assignment 1

1. What will the following piece of code do?
   for i in range(3):
   print(i)

   _____

2. What is the difference between range(1, 3, 10) and range (1,10, 3)

   _____

   _____

3. What is the difference between indexing a dataframe using *iloc*, *loc* and *ix*, in Pandas

   _____

   _____

   _____

4. I have a dataframe **df1** with the following 3 columns:

   student_id: This is a unique ID assigned to each student
   first_name: Student's first name
   last_name: Student's last name
   email: Student's email address

   I have a second dataframe **df2** with the following 2 columns:

   student_id: Same unique ID as above
   grade: final letter grade given to the student in this class offered last year

   Students who did not complete the class will have an entry in df1 but not in df2. I want to merge the two dataframes, with the following command, such that only students who completed the class end up in the merged dataframe, **merged_df**.

   merged_df = df1.merge(df2, on='student_id', how='_____')

   What value should I pass in for the parameter *how*?

   _____

5. Try out the following two operations at a python command line and report what the resulting value is:

   ```
   None == None
   ```

   _____

   ```
   np.nan == np.nan
   ```

   _____

6. A zip file containing stock price data has been uploaded to Brightspace. This zip file contains CSV files for 100 stock symbols (named <symbol>.csv e.g. AAPL.csv, LYFT.csv, etc. The zip file also contains a directory “spy”, which contains SPY.csv (a comma separated file containing price history for the SPY ETF) and SPY.txt (the same file, but tab-separated).

   Use pandas to load data for these 100 stocks (but not SPY) into dataframes, store the dataframes in a dictionary using symbol as the key. Print out the number of rows for each, the start date for each symbol and the minimum date (across symbols), as you iterate through the data. Do not do this (or any subsequent tasks) by copying and pasting your code 100 times!

7. Within each of the dataframes in your dictionary, create a return column and assign the stock’s ticker symbol as name for this new column i.e. the dataframe containing the data loaded in from AAPL.csv should have a return column named “AAPL”, etc.

8. Merge the above 100 dataframes into a single dataframe (assigned to a variable named mdf_intersect) such that the final dataframe has 101 columns, a Date column and one column for each of the 100 symbols, this column will contain the return value you computed for that symbol. Do the merge such that it only contains the intersection of dates across symbols. Print out the following stats:

   a. Number of rows in the merged dataframe

      _____

9. Create a new column called “sum_returns” that is the sum of the returns of all your symbols. Now load SPY.csv into a separate dataframe, create a return column with the name “SPY”, and merge it with your large dataframe (using the intersection of dates). What is the correlation (use pandas!) between the “sumreturn” column and the newly merged “SPY” column (which contains returns for SPY)?

10. Assign the correlation value to a variable and use a **print()** statement with an ***f string*** to print out the following (note that I have used 50% as the correlation here just to show you the format I want you to print) with the correct correlation, start date and end date:

    The correlation between SPY and LYFT is 50.0% (start date: YYYYMMDD, end date: YYYYMMDD)

11. Is this correlation high or low? What is an explanation for why this correlation is high or low?

12. Copy your code used to create mdf_intersect and repeat the merge, this time creating a dataframe assigned to a variable named mdf_union, containing the union of dates.
    a. What is the number of rows in this new dataframe?

    b. Calculate the number of days where 90 or more symbols have a positive return

13. Now call the method dropna() on mdf_union and assign the result to a new variable named mdf_union_dropna. What are the number of rows and how do they compare to the number of rows in mdf_intersect **from step 8**? Explain what we just did here.

14. Now call the method fillna(0.0) on mdf_union and assign the result to a new variable named mdf_union_fillna.What are the number of rows and how do they compare to the number of rows in mdf_union from step 8? Explain what we just did here.

15. Save dataframe mdf_union_fillna to a new CSV file. We will use this file in class!