Part 1

The first part of your assignments deals with understanding the concept of variation and covariation using visualization.

Variation is the tendency of the values of a variable to change from measurement to measurement. Variation describes the behavior between variables. Covariation is the tendency for the values of two or more variables to vary together in a related way. The best way to spot covariation is to visualize the relationship between two or more variables. How you do that should depend on the type of the variables involved. Let us try to understand this concept using the diamond dataset and going through the following steps:

1. Your first task is to summarize the dataset, what attributes it includes (what are types, etc).
2. Imagine we want to explore the distribution of the continuous variable "price" with respect to the categorical variable "cut". How would you visualize that?
3. Try your proposed visualization for each type of "cut". What if I want to see the distribution of the the price for all the "cut" in one plot (for instance, I want to compare the price based on the cut. How would you do that?
4. After you tried your suggestion, try to use visualizing boxplot of the price for each cut. You first need to make sure you understand what a boxplot shows.
5. Now start comparing the boxplot for various different types of cuts. What conclusion would you come to? Does the better cuts have a higher price (i.e., the intuitive conclusion)?
6. Now try to do similar analysis with clarity, what do you learn?
7. Now, let us see why low-quality diamonds seems to be more expensive?
8. Try to explore the other variables and see if you can find the answer before you go to the next step.
9. Hopefully, you have figured out what we were missing in the analysis already. If not, here is a hint: it "looks like" lower-quality diamonds have higher prices because there is an important confounding variable: the weight (carat) of the diamond. The weight of the diamond is the single most important factor determining the price of the diamond, and lower quality diamonds "tend" to be larger. How would you should that using

visualization? Notice that you need to show carat, price, and count effectively in a single visualization that supports the claim I made.

10. Now that you tried your answer to the previous question, try the following: draw a scatter plot where the x-axis is carat, the y-axis is price and the dots are color coded based on the count. Can you now see the claim I made.
11. Now you should be able to answer what variables are covarying in this dataset.

Source: Modified version of a subsection from R for data science by Wicham and Grolemund

Part 2

This part allows you to explore on your own using a new dataset:

There are 31 files named nyt1.csv, nyt2.csv, …, nyt31.csv in this dataset. Each one represent one (simulated) day's worth of ads shown and clicks recorded on the New York Times homepage in May 2012. Each row represents a single user. There are five columns: age, gender (female=0, male=1), number impressions, number clicks, and logged-in.

You'll be using Python to handle the data and you need to prepare a Jupyter notebook. Document all the step you go through in the notebook (including your code and the outcomes).
Once you load the data, follow the steps below to do some EDA:

1. Create a new variable, *age_group*, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64", and "65+".
2. For a single day:
    a. Plot the distribution of number impressions and click-through-rate (CTR=#clicks/#impressions) for these six age categories.
    b. Define a new variable to segment or categorize users based on their click behavior.
    c. Explore the data and make visual and quantitative comparisons across user segments/demographics (<18-year-old males versus <18-year-old females or logged-in versus not, for example).
    d. Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median,

variance, and max, and these can be calculated across the various user segments. Be selective. This about what will be important to track over time-what will compress the data, but still capture user behavior.

3. Now extend your analysis across days. Visualize some metrics and distributions over time.
4. Describe and interpret any patterns you find.

Source: Modified version of an exercise from “Doing Data Science” by O’Neil and Schutt