Midterm

STA 141B | Fall 2021

Instructions

You will find in the Box folder containing this file

- `tastdb-2010.csv`, containing a dataset
- `TAST_codebook.pdf`, containing a data dictionary
- `solutions.ipynb`, a notebook in which you will report your work.

Answer the questions below, adding one or more cells in your solution notebook between the problem numbers in which to do so. Don't change the cells in which the problem numbers have been written.

Once you are finished, print your solution notebook to pdf (make sure that it is not unnecessarily long due to long output) and submit this pdf to gradescope. Also submit a zipped folder containing your solution notebook, dataset, codebook, and pdf to the Canvas assignment 'midterm.' (IMPORTANT: notice you will be submitting work twice, once on Canvas, once on Gradescope!)

Remember, this must be your work, and your work alone. You can post questions on piazza, but do NOT talk about your work or otherwise collaborate with others, inside or outside of the class.

Exercise 1: Ethics statement.

For this problem, all that you have to do is to read and then type your full name in acknowledgment of the statement below:

As a student at UC Davis, I hold myself to a high standard of integrity, and by signing/accepting this statement I reaffirm my pledge to act ethically by honoring the UC Davis Code of Academic Conduct. I will also encourage other students to avoid academic misconduct.

I will not attempt to gain any unfair advantage over my fellow students during this exam.

I understand that suspected misconduct will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through dismissal from the University and a grade penalty up to a grade of "F" for the course.

I understand that if I fail to acknowledge or sign this statement, my instructor may not grade this work and may assign a grade of "0" or "F".

Trans-Atlantic Slave Trade

In this homework, we will uncover some of the numbers behind the Trans-atlantic slave trade (TAST), also known as the middle passage, that brought African slaves to the Americas. The middle passage is reported to have forcibly migrated over 10 million Africans to the Americas over a roughly 3 century time span. Many aspects of the TAST is little known by most people, such as the countries that constituted this network of slave ships, the regions from which the slaves were taken, and the number of slaves captured from Africa.

This last number is especially important since the number of slaves taken from Africa can impact other estimates that result from this. For example, when estimating the population of Africa in a given decade, demographers will use population growth models and more recent census data. For example, there are roughly X number of people in Africa and such populations tend to grow at rate M. Then if we want to calculate the population one century ahead then we just apply a simple formula that assumes that the population grows at this rate. But if the population is being drained by the slave trade, then this number will tend to be underestimated because the growth rate is overestimated. To account for this models need to take into account this drain on the population.

Throughout this homework you will need to follow the principles of graphical excellence and the grammar of graphics. Use only Plotnine for your graphics, do not use Pyplot, Seaborn, or Plotly since they do not follow closely the grammar of graphics. Be sure to include titles and necessary contextual captions.

Attention: The Trans-Atlantic Slave Trade remains one of the most horrific abuses of human rights in history. This homework deals with the numbers behind this forced migration, please be aware that this is a sensitive topic for possibly yourself and others. A suitable amount of respect and seriousness is required when dealing with this data.

Exercise 2: The data.

a. Read in the Trans-Atlantic Slave Trade database with Pandas. Hint: if you use the unix tool `file` you can find that this CSV is encoded with iso-8859-1 character set. Make sure that all missing values are encoded as NaN.

b. Use `.info()` to print out the names of the columns, their types, etc.

c. Open up the pdf file: `TAST_codebook.pdf` which is the data dictionary for this and other related datasets. Many of the variables in the code book are not in this dataset because it is describing an updated dataset. Create a list where you describe the meaning of the columns of your imported dataframe. You can group similar columns together when describing their rough meaning, such as ownera,...,ownerp are owners of the slave ships.

Throughout we will disregard all time variables other than year since they are not reliable.

Exercise 3: Estimating the total number of captives.

a. We will try to estimate the number of people captured into slavery and forced through the middle passage. What variable would you use to estimate the total number of captives taken from Africa? Let me call this Variable A in this problem statement. How much of the data for Var A is missing?

b. Create a preliminary estimate of the total number of captives taken from Africa by assuming that Var A is Missing Completely at Random (whether the variable is missing or not is independent of the variable and all other variables). You can simply divide the total count for the non-missing entries by the proportion of non-missing entries.

c. What other variables do you expect to be associated with Var A and why? Select 2 top possibilities. Visualize these associations using an appropriate plot. Do you trust the answer to 3.b? Why or why not?

Exercise 4: The national flag that the ships flew.

a. We want to understand the trends of the country of the slave ships (the national flag, identifying the country, is identified by the national variable). Subselect the values of national that have more than 300 voyages with that value.

b. Create a DataFrame that filters out the voyages where national does not have one of these values. You should be retaining voyages with only these most common values.

c. Create a variable, nationality, that is a string of easily readable names for these values by looking them up in the pdf codebook.

d. Using Plotnine, plot the counts of the voyages by nationality as a function of voyage year. Think about how best to display the count of a voyage by year and then how should you be including the nationality variable.

e. In this plot, what are the geometric elements and aesthetic mappings? What other components of the grammar of graphics are you using?

f. Do you observe any abrupt changes in the patterns of these counts for a given nationality? Investigate the cause for this change (using Google, etc.).

Exercise 5: Looking at some of these ships.

a. Search for the slave ship mentioned in the following wikipedia article: `https://en.wikipedia.org/wiki/Brookes_(ship)` *Hint:* Look at all records of ships with 'Brook' in the name and try to match the characteristics to those described. How many voyages for this ship are in the data (try to exclude ships with the same name)?

b. Create a variable that is True if there was a resistance (like a slave revolt) on the ship and False if not. Plot the density of ships as a function of year with and without revolts, and compare these distributions.

c. The movie Amistad was based on a real life slave ship and slave uprising. Read about it here: `https://en.wikipedia.org/wiki/La_Amistad` Try to find this ship by searching for it by name and also by searching for ships in the same 10 year period as this event with a slave resistance. If you think you found it describe it, otherwise describe the events of another voyage that you did find.

Exercise 6: Other patterns.

a. The departure and arrival locations are quite detailed. Look in the appendix of the codebook for the location codes. Make a more coarse version of both arrival and departure port variables (select just the last departure and first arrival) so that for example,

30000 Caribbean 31600 Martinique 36101 Fort-Royale

is just encoded as '3' or Caribbean.

b. Plot the trend of voyages as a function of departure location. What trends do you see?

c. Do the same for arrival location.

d. Plot the proportion of captives that are men as a function of year. Include a smoother to describe the over all trend. Also include in the plot another possible confounding variable.

e. Describe the geometries, aesthetic mappings, and other aspects of the plot.