PS4 Andy Fan Will Sigal

PS4: Due Sat Nov 2 at 5:00PM Central. Worth 100 points.

Style Points (10 pts)

Submission Steps (10 pts)

1. This problem set is a paired problem set.
2. Play paper, scissors, rock to determine who goes first. Call that person *Partner 1*.
    - Partner 1 (name and cnet ID): Andy Fan, fanx
    - Partner 2 (name and cnet ID):
3. Partner 1 will accept the `ps4` and then share the link it creates with their partner. You can only share it with one partner so you will not be able to change it after your partner has accepted.
4. "This submission is our work alone and complies with the 30538 integrity policy." Add your initials to indicate your agreement:

AF WS

5. "I have uploaded the names of anyone else other than my partner and I worked with on the problem set **here**" (1 point)
6. Late coins used this pset: (Andy:0); (will:0) Late coins left after submission: (Andy: 3) ; (will: 4)
7. Knit your `ps4.qmd` to an PDF file to make `ps4.pdf`,
    - The PDF should not be more than 25 pages. Use `head()` and re-size figures when appropriate.
8. (Partner 1): push `ps4.qmd` and `ps4.pdf` to your github repo.
9. (Partner 1): submit `ps4.pdf` via Gradescope. Add your partner on Gradescope.
10. (Partner 1): tag your submission in Gradescope

Important: Repositories are for tracking code. **Do not commit the data or shapefiles to your repo.** The best way to do this is with `.gitignore`, which we have covered in class. If you do accidentally commit the data, Github has a guide. The best course of action depends on whether you have pushed yet. This also means that both partners will have to download the initial raw data and any data cleaning code will need to be re-run on both partners' computers.

Download and explore the Provider of Services (POS) file (10 pts)

(This file records all providers certified to bill Medicare and Medicaid. Each provider has a unique CMS certification number, consistent across different years)

(The data dictionary is located at the bottom of the page as "Provider of Services File- Hospital & Non-Hospital Facilities Data Dictionary)

```
### SETUP
import pandas as pd
import altair as alt
import time
import os
import warnings
warnings.filterwarnings('ignore')
```

1. (Partner 1) This is a fairly large dataset and we won't be using most of the variables. Read through the rest of the problem set and look through the data dictionary to identify which variables you will need to complete the exercise, and use the tool on data.cms.gov into restrict to those variables ("Manage Columns") before exporting ("Export"). Download this for 2016 and call it pos2016.csv. What are the variables you pulled?

required variables:

provider type code: PRVDR_CTGRY_CD

subtype code: PRVDR_CTGRY_SBTYP_CD

CMS certification number: PRVDR_NUM

Termination code: PGM_TRMNTN_CD

zip code: ZIP_CD

Date of termination: TRMNTN_EXPRTN_DT

Facility name: FAC_NAME

```
### import data (default directory is just the repo)

#os.chdir('/Users/willsigal/Documents/GitHub/problem-set-4-andy-fan-will-sigal')
 ↪  #will wd

os.chdir('d:\\UChicago\\Classes\\2024Qfall\\Programming
 ↪  Python\\problem-set-4-andy-fan-will-sigal') #andy wd

df_2016 = pd.read_csv('pos2016.csv', encoding="latin1", on_bad_lines="skip")
```

2. (Partner 1) Import your pos2016.csv file. We want to focus on short-term hospitals. These are identified as facilities with provider type code 01 and subtype code 01. Subset your data to these facilities. How many hospitals are reported in this data? Does this number make sense? Cross-reference with other sources and cite the number you compared it to. If it differs, why do you think it could differ?

```
###subset to short term hositals
df_2016 = df_2016[(df_2016['PRVDR_CTGRY_SBTYP_CD'] == 1) &
 ↪  (df_2016['PRVDR_CTGRY_CD'] == 1)]
```

a.There are 7275 short term hospitals reported in the data. This number at a glance makes sense nationwide.

b.Looking at the Article by Jane Wishner, Patricia Solleveld, Robin Rudowitz, Julia Paradise, and Larisa Antonisse, they found "nearly 5,000 short-term, acute care hospitals in the United States". This is less than the data we have. It is possibly because some of the short-term hospitals in our dataset closed (they are temporary after all, and could have only been setup for less than a year)

3. (Partner 1) Repeat the previous 3 steps with 2017Q4, 2018Q4, and 2019Q4 and then append them together. Plot the number of observations in your dataset by year.

```
### import data
df_2017 = pd.read_csv('pos2017.csv', encoding="latin1", on_bad_lines="skip")

df_2018 = pd.read_csv('pos2018.csv', encoding="latin1", on_bad_lines="skip")

df_2019 = pd.read_csv('pos2019.csv', encoding="latin1", on_bad_lines="skip")
```

```
###subset to short term hositals
df_2017 = df_2017[(df_2017['PRVDR_CTGRY_SBTYP_CD'] == 1) &
 ↪  (df_2017['PRVDR_CTGRY_CD'] == 1)]
df_2018 = df_2018[(df_2018['PRVDR_CTGRY_SBTYP_CD'] == 1) &
 ↪  (df_2018['PRVDR_CTGRY_CD'] == 1)]
df_2019 = df_2019[(df_2019['PRVDR_CTGRY_SBTYP_CD'] == 1) &
 ↪  (df_2019['PRVDR_CTGRY_CD'] == 1)]
```

The number of short-term hospitals in each of the years:

2017: 7260 hospitals

2018: 7277 hospitals

2019: 7303 hospitals

```
### create new year columns
df_2016['year'] = 2016
df_2017['year'] = 2017
df_2018['year'] = 2018
df_2019['year'] = 2019

### append
df = pd.concat([df_2016, df_2017, df_2018, df_2019], ignore_index=True)
```

```
### altair plot
df1_3 = df.groupby('year').size().reset_index(name='count')

chartQ1_3 = alt.Chart(df1_3).mark_bar().encode(
    x= alt.X('year:O', axis=alt.Axis(labelAngle=90)),
    y= alt.Y('count:Q', title='number of hospitals',
 ↪  scale=alt.Scale(domain=[7000, 7500], clamp=True))
).properties(
    width=400,
    height=200,
    title='Observations of Hospitals by Year'
)
chartQ1_3.save("chartQ1_3.png")
chartQ1_3
```

```
alt.Chart(...)
```

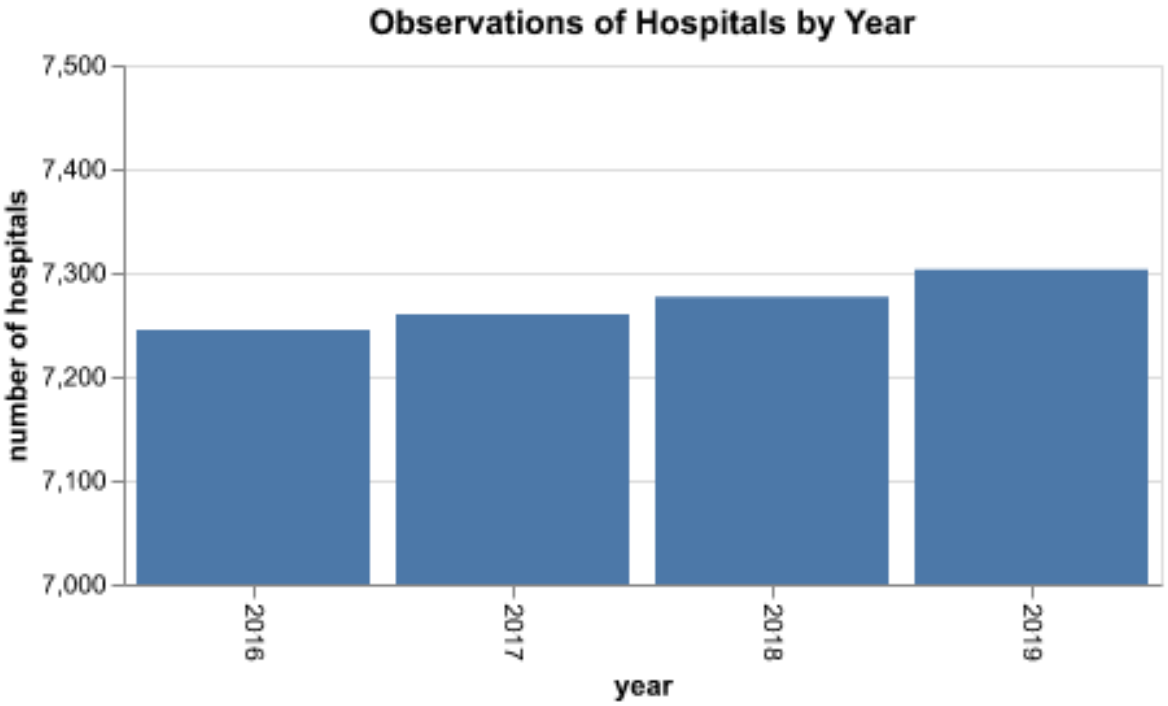

Figure 1: Chart Q1.3

4. (Partner 1) Each hospital is identified by its CMS certification number. Plot the number of unique hospitals in your dataset per year. Compare this to your plot in the previous step. What does this tell you about the structure of the data?

a.plot unique hospitals

```
### plot
df1_4 = df.groupby('year')['PRVDR_NUM'].nunique().reset_index()

chartQ1_4 = alt.Chart(df1_4).mark_bar().encode(
    x= alt.X('year:O', axis=alt.Axis(labelAngle=90)),
    y= alt.Y('PRVDR_NUM:Q', title='number of unique hospitals',
 ↪  scale=alt.Scale(domain=[7000, 7500], clamp=True))
).properties(
    width=400,
    height=200,
    title='Unique Hospitals by Year'
)
```

```
chartQ1_4.save("chartQ1_4.png")
chartQ1_4
```

```
alt.Chart(...)
```

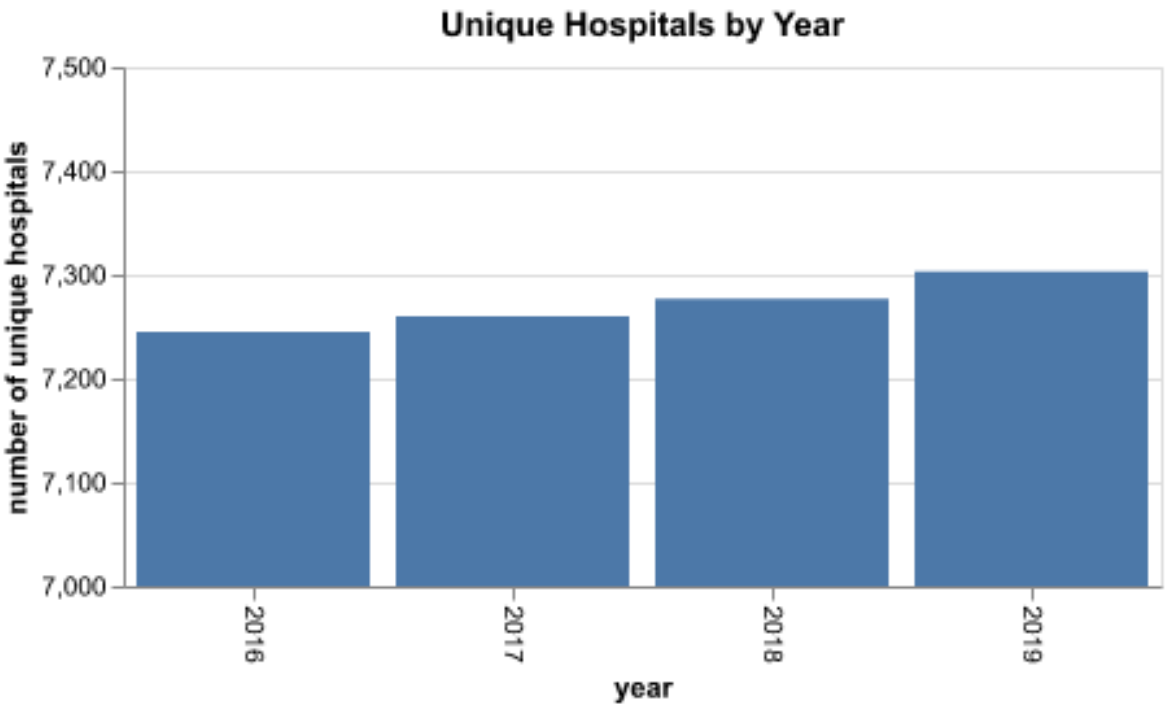

Figure 2: Chart Q1.4

```
b.compare to previous step, what does this tell about data structure:
```

The plots are identical. Which means each hospital is a unique observation, tallied each year.

Identify hospital closures in POS file (15 pts) (*)

We will now use the 2016-2019 files to identify hospital closures nationally. A hospital issuspected to have closed if its Termination Code in the POS file lists them as an "Active Provider" in 2016 and then they are either not active or do not appear in the data at all in a subsequent year.

1. (Partner 2) Use this definition to create a list of all hospitals that were active in 2016 that were suspected to have closed by 2019. Record the facility name and zip of each hospital as well as the year of suspected closure (when they become terminated or disappear from the data). How many hospitals are there that fit this definition?

2. (Partner 2) Sort this list of hospitals by name and report the names and year of suspected closure for the first 10 rows.

3. (Partner 2) However, not all suspected hospital closures are true closures. For example, in the case of a merger, a CMS certification number will be appear to be "terminated," but then the hospital re-appear under a similar name/address with a new CMS certifi cation number in the next year. As a first pass to address this, remove any suspected hospital closures that are in zip codes where the number of active hospitals does not decrease in the year after the suspected closure.

a. Among the suspected closures, how many hospitals fit this definition of potentially being a merger/acquisition?

b. After correcting for this, how many hospitals do you have left?

c. Sort this list of corrected hospital closures by name and report the first 10 rows.

Download Census zip code shapefile (10 pt)

Navigate to the Census shapefiles (link), select "gz_2010_us_860_00_500k.zip" and download the resulting shapefile. Note: If you have difficulty downloading or working with the file because the size is too large for your computer, reach out to the instruction team on Ed so that we can give them a smaller initial shapefile to work with.

1. (Partner 1) This is non-tabular data. 1a. What are the five file types and what type of information is in each file? 1b. It will be useful going forward to have a sense going forward of which files are big versus small. After unzipping, how big is each of the datasets?

a.

b.

2. (Partner 1) Load the zip code shapefile and restrict to Texas zip codes. (Hint: you can identify which state a zip code is in using the first 2-3 numbers in the zip code (Wikipedia link). Then calculate the number of hospitals per zip code in 2016 based on the cleaned POS file from the previous step. Plot a choropleth of the number of hospitals by zip code in Texas.

Calculate zip code's distance to the nearest hospital (20 pts) (*)

1. (Partner 2) Create a GeoDataFrame for the centroid of each zip code nationally: zips_all_centroids. What are the dimensions of the resulting GeoDataFrame and what do each of the columns mean?

2. (Partner 2) Create two GeoDataFrames as subsets of zips_all_centroids. First, create all zip codes in Texas: zips_texas_centroids. Then, create all zip codes in Texas or a bordering state: zips_texas_borderstates_centroids, using the zip code prefixes to make these subsets. How many unique zip codes are in each of these subsets?

3.(Partner 2) Then create a subset of zips_texas_borderstates_centroids that contains only the zip codes with at least 1 hospital in 2016. Call the resulting GeoDataFrame zips_withhospital_centroids What kind of merge did you decide to do, and what variable are you merging on?

4. (Partner 2) For each zip code in zips_texas_centroids, calculate the distance to the nearest zip code with at least one hospital in zips_withhospital_centroids.

a. This is a computationally-intensive join. Before attempting to do the entire join, subset to 10 zip codes in zips_texas_centroids and try the join. How long did it take? Approximately how long do you estimate the entire procedure will take?

b. Now try doing the full calculation and time how long it takes. How close is it to your estimation?

c. Look into the .prj file and report which unit it is in. Convert the given unit to miles, using an appropriate conversion you find online (estimates are okay).

5.(Partner 2) Calculate the average distance to the nearest hospital for each zip code in Texas.

a. What unit is this in

b. Report the average distance in miles. Does this value make sense?

c. Map the value for each zip code

Effects of closures on access in Texas (15 pts)

1. (Partner 1) Using the corrected hospital closures dataset from the first section, create a list of directly affected zip codes in Texas– that is, those with at least one closure in 2016-2019. Display a table of the number of zip codes vs. the number of closures they experienced.

2.(Partner 1) Plot a choropleth of which Texas zip codes were directly affected by a closure in 2016-2019– there was at least one closure within the zip code. How many directly affected zip codes are there in Texas?

3. (Partner 1) Then identify all the indirectly affected zip codes: Texas zip codes within a 10-mile radius of the directly affected zip codes. To do so, first create a GeoDataFrame of the directly affected zip codes. Then create a 10-mile buffer around them. Then, do a spatial join with the overall Texas zip code shapefile. How many indirectly affected zip codes are there in Texas?

4.(Partner 1) Make a choropleth plot of the Texas zip codes with a different color for each of the 3 categories: directly affected by a closure, within 10 miles of closure but not directly affected, or not affected.

Reflecting on the exercise (10 pts)

(Partner 1) The "first-pass" method we're using to address incorrectly identified closures in the data is imperfect. Can you think of some potential issues that could arise still and ways to do a better job at confirming hospital closures?

(Partner 2) Consider the way we are identifying zip codes affected by closures. How well does this reflect changes in zip-code-level access to hospitals? Can you think of some ways to improve this measure?