A reflection on your data discovery process 10 points

The data discovery process was somewhat tedious for us. At first, we couldn't decide on a topic for Datafest. Since we couldn't figure out what topic we wanted, we decided to come up with random topics and look up data for the best random topic we could think of. The best random topics we thought of were earthquakes, ocean pollution, and power consumption. A method we used to find datasets was using Google and searching our topic along with the word shape file or GIS data at the end of the search. For example, for the topic of earthquakes, we searched up earthquake shapefiles or earthquake GIS data. We did this method for each topic and found the best and most interesting results with the topic of power consumption. The questions we came up with for earthquakes and ocean pollution were either too difficult to implement or the datasets were lackluster or overwhelming. When it came to finding data and discovering datasets, we found half of our desired data sets when using the California State Geoportal. For our topic about power consumption, the California State Geoportal provided us with data for our power outages and power lines. Additionally, the California State Geoportal made it easier for us to find and download data compared to other websites. They both were found by simply using the California State Geoportal Online search bar and typing in power outages in California and electrical transmission lines. The electrical transmission line dataset can also be found in the California Energy Commission Online by searching the electrical transmission lines dataset in the search bar of their website. For our county data, we got it from the Census website which was very difficult and confusing to navigate. The way we found the county data set was using the Census online search and searching TIGER/Line Shapefiles. Next, we clicked on FTP Archive by Layer and clicked the county folder link. We clicked on the 2020 folder and then chose the 6th county zip file. This process took multiple steps and was very confusing. Other dataset sites like California State Geoportal have a simpler and more efficient way of finding data when compared to Census. We also tried getting the population data from the Census website, but the website was very confusing and had zip files with confusing and cryptic names. Finding and downloading the correct files from Census was more difficult than downloading and finding files from the California State Geoportal. We often download from the Census website to find population data, but we ended up downloading and extracting a file that is related to something entirely different. This was very frustrating and time-consuming, but we eventually found population data from the National Mapping Pipeline System. We found this dataset through searching population density data shapefile on Google and clicking on the 8th result. We ended up using Google because the California State Geoportal Online didn't provide any useful population data for us and the Census website was too complex and complicated to navigate. Overall, I think this data discovery for DataFest helped me improve my data searching skills and if I needed to find and collect data from the internet in the future, I would be a lot more efficient.

A critical review of the data and sources used 10 points

The datasets I used for Datafest are the population density dataset, the power outages dataset, the powerlines dataset, and the counties in California dataset. For the power outages in the California dataset, the source is from the CA Governor's Office of

Emergency Services which makes it credible. This is credible because the Office of Emergency Services tracks where the power outages occur and are the ones that send people to go and fix the problem. This is also a primary source. This dataset is good for getting data about recent power outages. This is because the CA Governor's Office of Emergency Services updates it every 15 minutes, which makes the datasets very accurate and precise. However, it doesn't track long-term data which means that if you wanted to see long-term trends in power outages across California, using this dataset wouldn't achieve that objective. This dataset gets short-term data and resets the data collected after each week. This is a severe disadvantage and prevents users from seeing monthly or yearly data of outages. This makes it much more difficult to make correlations. We also wanted to find monthly power outage data online, but it was either blocked by a paywall or there was little to no documentation of it. My guess is that the data is private and not available to the public. The advantage of a short-term dataset is that it is recent and accurate data and doesn't clutter maps with data. The power outage data also shows the time it occurred, the cause of the power outage, and the estimated time it will be fixed. This data can be useful for seeing what caused the power outages like if they needed to repair a pole or if there were equipment problems. This could be used to track common and consistent issues in the electrical grids. Knowing these issues can make it easier to plan and strategize on how to improve the electrical grids and prevent these issues from occurring frequently. An ideal way of tracking long-term data using this dataset would be to wait until the end of each week to get yearly data. This would be very time-consuming and very inefficient. Yearly data can be used to possibly make a correlation between two variables or datasets. The size of the data points was too big which made it harder to see each layer when we imported it into ArcGIS Pro. We adjusted it to better fit the other layers for the map by making the data points smaller. In the context of my Datafest project, I think the dataset did a good enough job of showing how the short-term data possibly shows trends and correlations with our other datasets.

The population dataset showed areas in the United States that were the most populated. We needed to cut out the other states' populated areas so that California's populated areas were the only ones shown. The organization that made this dataset used data from the U.S. Census Bureau's TIGER Urban Areas to make this. The organization that made the dataset is called the National Pipeline Mapping System and is a government organization that maps where pipelines should be built across America. The data set considered 50,000 or more people with a population density of at least 1,000 people per square mile to be a populated area. Since the U.S. Census Bureau created the data being used, it makes the dataset credible. This is because the U.S. Census Bureau is a notable organization that collects data about a lot of things about the states in America like population, boundaries, race, ethnicity, and income. The problem with the data set is the age of the dataset. The data was collected in 2010 which shows the data is inaccurate and doesn't represent today's population. Our population has significantly grown since 2010 and areas that didn't meet the criteria of being a populated area can meet the criteria now. Another problem I have with the dataset is that it doesn't label the most populated areas. We needed to use the counties in the California dataset to approximate what specific cities and counties had populated areas. This dataset includes populated areas of other states, so we had to clip it. This is to exclude all the data in the United States except for California. It is also not a primary source and the National Pipeline Mapping System uses the data from the Census and alters the population density given to make a new and updated dataset. Since the dataset is edited and only uses parts of it, this is why it's a secondary source. In the

context of my Datafest project, the dataset did its job of showing what areas are populated in California, but we would've preferred using a more up-to-date and accurate dataset instead.

The dataset that shows the counties of California was taken from the U.S. Census Bureau. As I stated before, the Census is credible because they are known for collecting and documenting data about the states and they are a government-run organization. This is a primary source. The dataset was from 2020, so it might be a little inaccurate. This is because the boundaries of counties could have changed. If they didn't change, then the dataset is accurate. This dataset was useful to show what counties the power outages, powerlines, and populated areas were in. The dataset's position and font size of counties were difficult to see with other layers, so we had to adjust it and make the font size smaller. We positioned the names of counties so that it is more visible.Without this dataset, the person viewing the map would have a harder time understanding it.

Lastly, the dataset we used was the powerlines dataset and was made by the California Energy Commission. The California Energy Commission is a government-run organization that makes that data accurate and credible. The organization is known for documenting and collecting data regarding electricity and energy, which is why they are credible. This is also a primary source. The dataset is from 2017 which could indicate the data might be inaccurate. There is a 5-year gap and the mapping of current power lines could look different than the mapping of power lines from 2017. They could have built more power lines or torn down power lines in that 5-year gap. If the powerlines didn't change or be altered in California, then the data would still be accurate. However, I doubt this because disasters like earthquakes or fires could've damaged or taken down the powerlines. The dataset also shows that powerlines can be used to track which power lines are active or shut down and it also tracks the voltage of each power line, which can be useful to use in some scenarios. Some power lines went outside of California and we needed to clip them. We used the geoprocessing tool to clip it to keep the map concise and contained. The powerline color needed to be adjusted to make other layers more seeable. We made it less dark to help the colors not overlap with each other. In the context of my Datafest topic, I think this dataset was a good representation of the power lines in California and it showed which areas had an abundance of power lines. This was useful to see the correlation between power lines and population density, but it didn't show a correlation between power lines and power outages. Overall, it gave a good visualization of power lines to the people looking at the map.

A brief summary of your contribution to the team 15 points

One of my contributions to the team was getting all the data together and designing the production quality maps. I also made maps to showcase the specific datasets we found. This is seen in the Datasets slide of the presentation. I also contributed by searching and finding the datasets and those datasets were the powerlines dataset, the densely populated area dataset, and the counties in California dataset. For the presentation, I worked on the introductions slide, the background info and hypothesis slide, both dataset slides, and the references slide.

Appropriate APA style Citations 10 points

CA Governor's Office of Emergency Services. (2019, December 24). *Power Outage Incidents*. California State Geoportal. Retrieved June 19, 2022, from https://gis.data.ca.gov/datasets/CalEMA::power-outage-incidents/explore?location=36.501396%2C-120.398687%2C7.32

California Energy Commission. (2017, November). *California Electric Transmission Lines*. Retrieved June 19, 2022, from https://cecgis-caenergy.opendata.arcgis.com/datasets/260b4513acdb4a3a8e4d64e69fc84fee_0/explore?location=37.221216%2C-120.929034%2C9.27

National Pipeline Mapping System. (2018, January). *High Population Areas (HPA) Data*. Retrieved June 19, 2022, from https://www.npms.phmsa.dot.gov/PopulationData.aspx

US Census Bureau. (2021, December 16). *TIGER/Line Shapefiles*. Census. Retrieved June 20, 2022, from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html