Gary Gregg
INFO 474, Assignment #2
26 January 2017

Selection of a Dataset

For this assignment, we choose to browse the catalog of U.S. Government data located at https://catalog.data.gov/. This catalog contains over 230,000 datasets compiled by the U.S. Government to include those provided at the federal, state and local government level. We choose to look for a larger dataset (greater than a megabyte), and available in a common format that is easily ingested by Tableau. The formats we choose to search for include CSV, Excel, JSON and XML formats. We reject a number of federal and state provided datasets because they contained insufficient data for an interesting exploration. For example, some federal or state data consist only of three measures for all fifty states, and some state data consist of a similar number of measures for each of the counties in that state. Datasets of these types are typically only a few kilobytes in size, consisting of less than a thousand data points. After some effort in filtering datasets by subject, region, origin or format, we happen upon the *NCHS - Leading Cause of Death* Dataset, published by the Centers for Disease Control and Prevention. See:

https://catalog.data.gov/dataset/age-adjusted-death-rates-for-the-top-10-leading-causes-of-death-united-states-2013).

The description of the dataset at the link as as follows:

> This dataset presents the age-adjusted death rates for the 10 leading causes of death in the United States beginning in 1999. Data are based on information from all resident death certificates filed in the 50 states and the District of Columbia using demographic and medical characteristics. Age-adjusted death rates (per 100,000 population) are based on the 2000 U.S. standard population. Populations used for computing death rates after 2010 are postcensal estimates based on the 2010 census, estimated as of July 1, 2010. Rates for census years are based on populations enumerated in the corresponding censuses. Rates for non-census years before 2010 are revised using updated intercensal population estimates and may differ from rates previously published. Causes of death classified by the International Classification of Diseases, Tenth Revision (ICDâ10) are ranked according to the number of deaths assigned to rankable causes. Cause of death statistics are based on the underlying cause of death.

This dataset is available as a 1.3 megabyte CSV file, and consists of four dimensions and two measures. As described in the excerpt for the source, the data are some quite interesting statistics regarding causes of death in each of the fifty states, and how they compare to each other. We chose to use this dataset, and import it into Tableau.

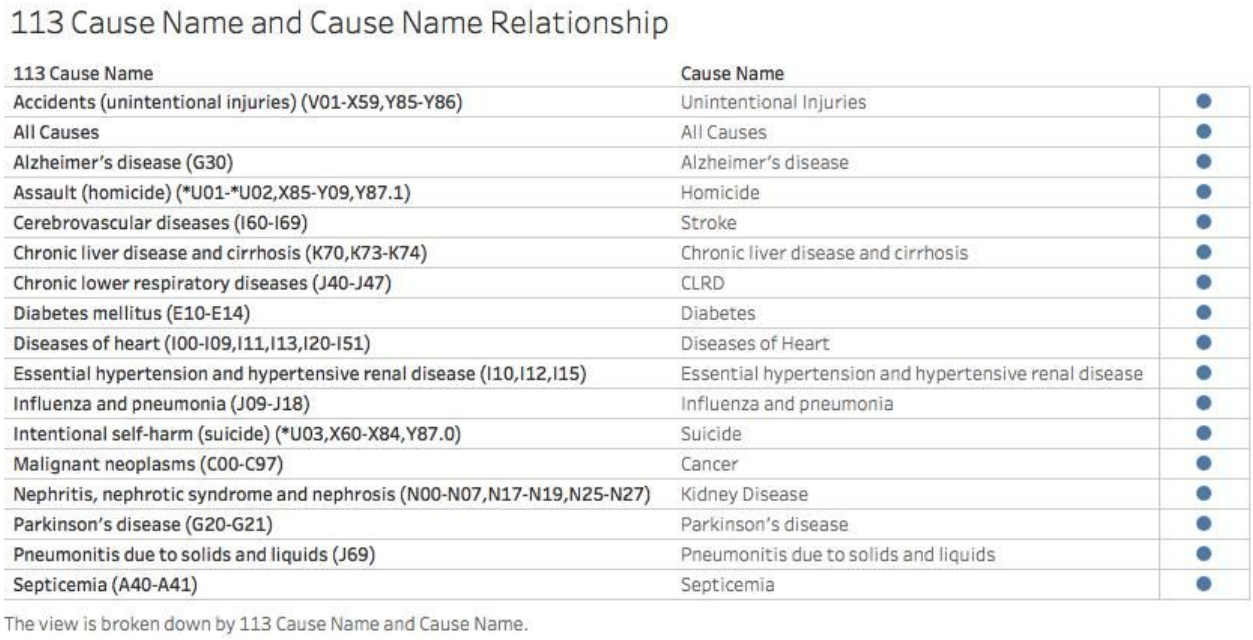
Data Description and Preparation

The fields provided by the dataset include the following dimensions: *113 Cause Name*, *Cause Name*, *State*, and *Year*. They also include the following measures: *Age-adjusted Death Rate*, and *Deaths*. The data appear to be quite clean, and ready for analysis. No special preparation is required for an initial exploration. However, we need to seek an interpretation of the meaning of some of the fields. We decide to let our inquiries about the semantics of the fields be among the questions that are answered in this assignment.

Questions

Question 1: What is the relationship between the fields *113 Cause Name* and *Cause*?

We drag both the *113 Cause Name* and *Cause Name* dimensions to the row tray in Tableau, and set the mark type to a circle. The following graphic results:



113 Cause Name and Cause Name Relationship

113 Cause Name	Cause Name	
Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional Injuries	●
All Causes	All Causes	●
Alzheimer's disease (G30)	Alzheimer's disease	●
Assault (homicide) (*U01-*U02,X85-Y09,Y87.1)	Homicide	●
Cerebrovascular diseases (I60-I69)	Stroke	●
Chronic liver disease and cirrhosis (K70,K73-K74)	Chronic liver disease and cirrhosis	●
Chronic lower respiratory diseases (J40-J47)	CLRD	●
Diabetes mellitus (E10-E14)	Diabetes	●
Diseases of heart (I00-I09,I11,I13,I20-I51)	Diseases of Heart	●
Essential hypertension and hypertensive renal disease (I10,I12,I15)	Essential hypertension and hypertensive renal disease	●
Influenza and pneumonia (J09-J18)	Influenza and pneumonia	●
Intentional self-harm (suicide) (*U03,X60-X84,Y87.0)	Suicide	●
Malignant neoplasms (C00-C97)	Cancer	●
Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27)	Kidney Disease	●
Parkinson's disease (G20-G21)	Parkinson's disease	●
Pneumonitis due to solids and liquids (J69)	Pneumonitis due to solids and liquids	●
Septicemia (A40-A41)	Septicemia	●

The view is broken down by 113 Cause Name and Cause Name.

From the resulting graphic, we see that all *113 Cause Name* values are related to precisely one *Cause Name* value. Scanning the values (and relying on some layman's knowledge of medical terms), we see that *Cause Name* is simply a restatement of *113 Cause Name*. It appears, then, that these dimensions can be used interchangeably in any visualization that uses cause of death. For this exploration, we choose to use either *113 Cause Name* or *Cause Name* when one or the other more understandably describes the cause of death.

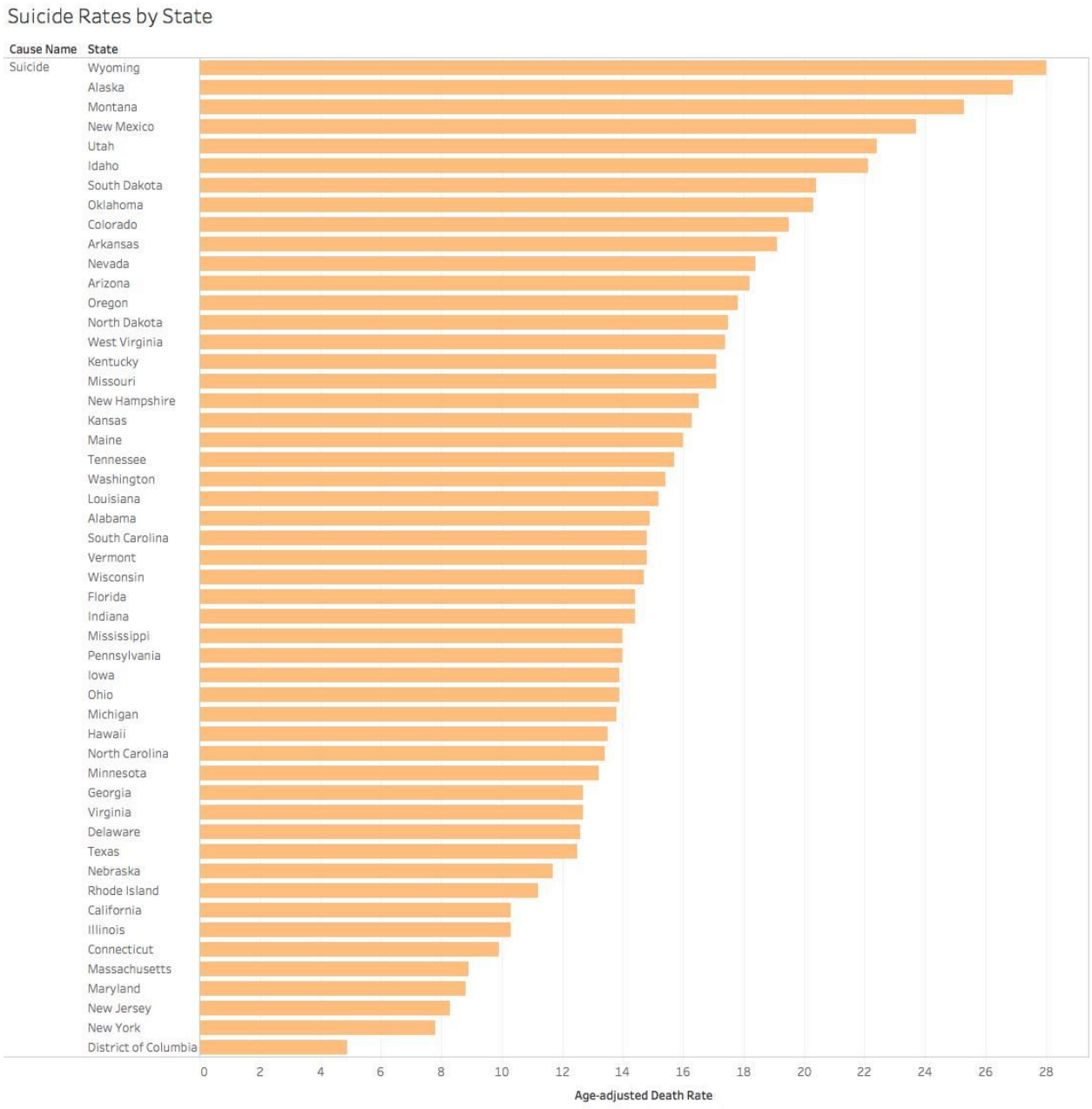
Question 2: What is the meaning of *Age-adjusted Date Rate*?

We let this be a question for the assignment, although answering it really does not require any visualization. It only requires a careful reading of the description of the dataset. It seems that the term “age adjusted” applies not to the age of the deceased, but to the age of the census used to determine the population of the state for the year in question. For this, it appears the creators of the dataset used the exact census numbers for the years 2000 and 2010 for the population of the state under consideration. For years before 2000, and between 2001 and 2009, and after 2010, it appears they used some sort of interpolative function to estimate the population, using the 2000 and 2010 census numbers as source. For the purposes of this research, let us assume that we do not need to question this further, and that population estimates are exact. The field *Age-adjusted Death Rate* might just as simply be called “Death Rate.” The description of the dataset notes that this death rate is given per 100,000 population. It would therefore appear to be possible to reconstruct the estimated population of a state for any given year by multiplying the number of deaths for any given cause by 100,000, then dividing by the death rate for that same state, year, and cause of death.

To ascertain the correctness of this interpretation, choose Alabama for the year 1999. Then choose any two values of *Cause Name*. For these two values of *Cause Name*, the ratio of *Death* should be equal to the ratio of *Age-adjusted Death Rate*. For Alabama in 1999, choose *Cause Names* of “Accidents,” and “All Causes.” For the ratio of *Deaths*, we come up with **2,313 / 44,806 = 0.0516**. For the ratio of *Age-adjusted Death Rate*, we come up with **52.20 / 1009.30 = 0.0517**. Let us assume that this result is close enough to declare as correct our interpretation of the fields.

Question 3: How do the states compare for the latest year in terms of suicide rates?

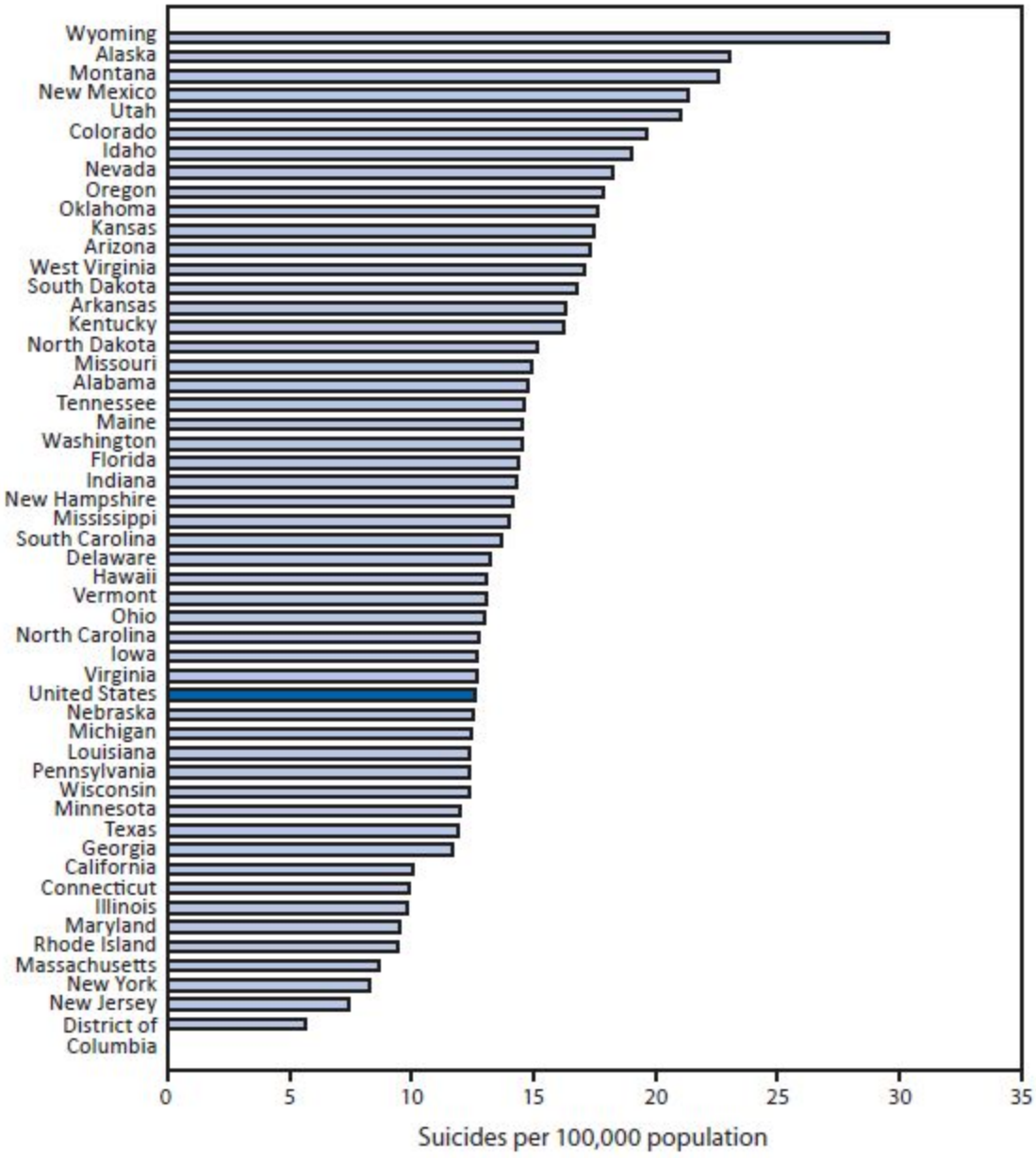
To answer this question, we set *Age-adjusted Death Rate* as a column, *Cause Name* for a row, and finally *State* for a row. We apply a filter on *State* because the data includes a row for the entire United States, and right now we are not interested in seeing that total. So we filter out United States. Next we choose *Cause Name* as a filter, and we choose to see only “Suicide.” Finally, we choose *Year* for a filter, and we choose to see only “2015,” which is the last year available. As a measure for *Age-adjusted Death Rate*, we temporarily choose “Count.” By doing this, we want to verify that there is only one row for each state. Seeing that there is indeed only one row for each state, we change the measure for *Age-adjusted Death Rate* to “Sum.” Since there is only one row under consideration, it does not matter if we choose “Sum,” “Average,” or “Median” for a measurement. For only one row, all these measures are the same. Now we choose a descending sort for *State* by *Age-adjusted Death Rate*. The following visualization results:



By checking online, we can determine if there is a similar per-state suicide death rate visualization. If what we find looks similar to what we have, above, we can be reasonably certain of two things: 1) That the data corresponds to known and accepted statistics; 2) That we have manipulated the data correctly in constructing our visualization.

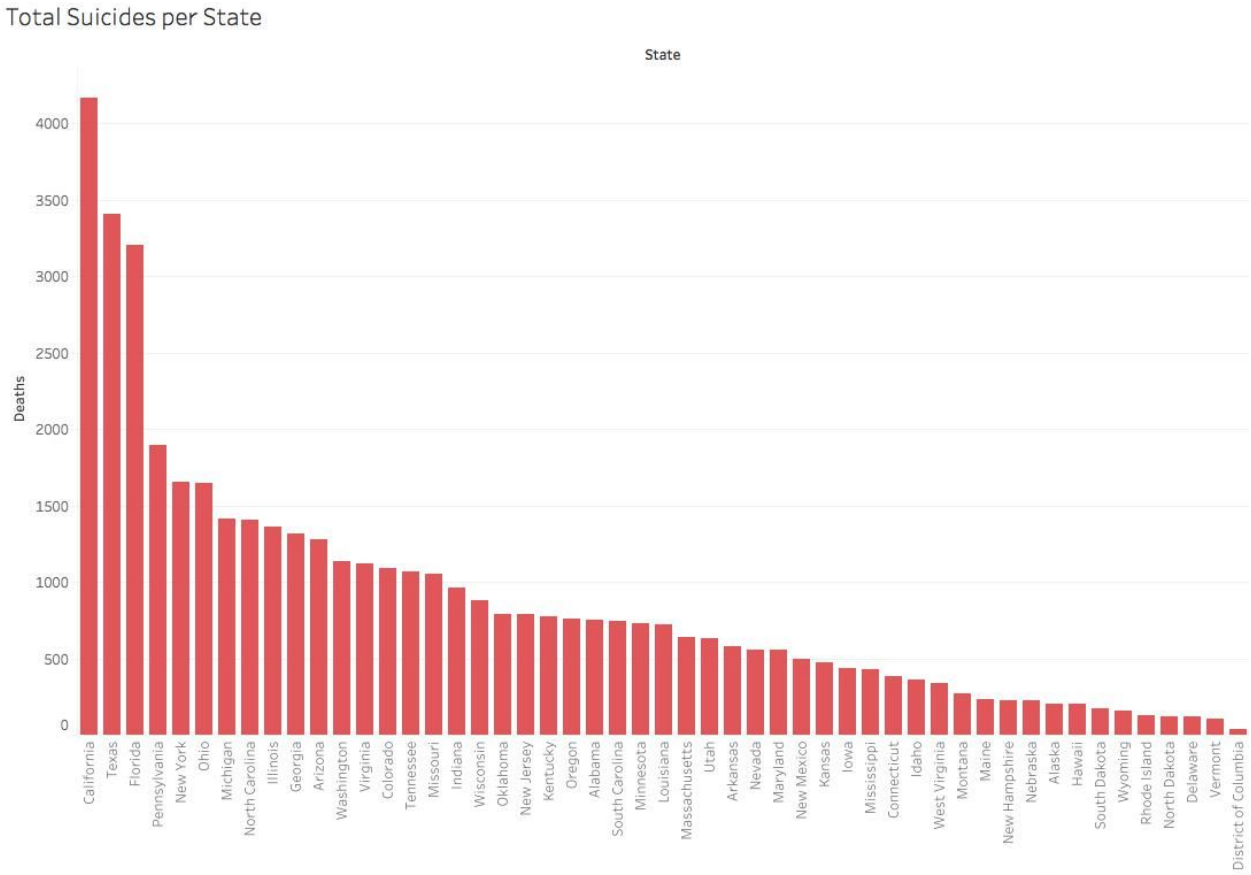
At https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6345a10.htm, we can find a similar visualization from the CDC for suicide by state for the year 2012. The similarity of the results is

such that we decide that we are on the right track, and that we do not need to redo the visualization for 2012 just to look for an exact match between what we have created, and what the CDC provides. Indeed, the visualization from the CDC might have come from the same data set that we use here! Below is the CDC visualization for 2012:



Question 4: How do the states compare for the latest year in terms of absolute number of suicides?

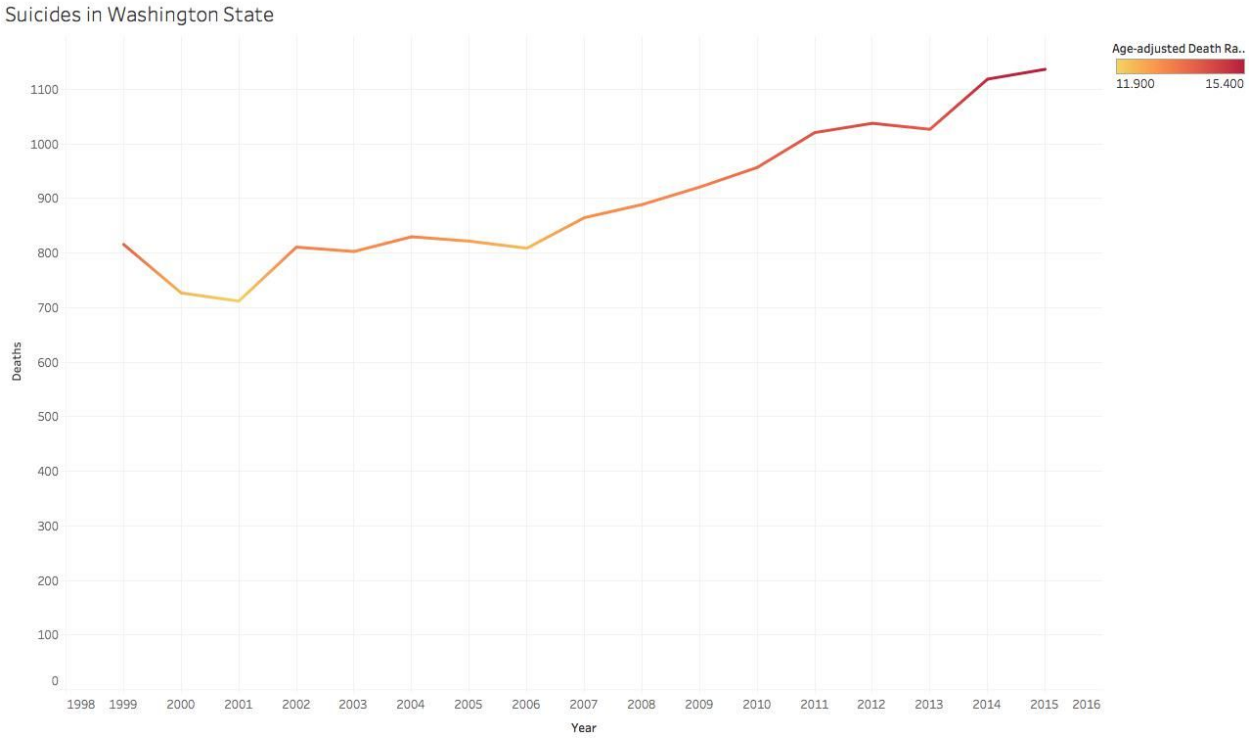
To answer this question, we set *State* as a column, and *Death* as a row. We next apply the same filters as in the previous question. We apply a filter on *State* because the data includes a row for the entire United States that we wish to exclude. We choose *Cause Name* as a filter, and we choose to see only “Suicide.” We choose *Year* for a filter, and we choose to see only “2015,” which is the last year available. Now we choose a descending sort for *State* by Deaths. Finally, for variety, we choose red for a color. We see in the resulting visualization the intuitive result that the states with the highest populations have the highest absolute number of suicides, namely California, Texas, and Florida:



Sum of Deaths for each State. The data is filtered on Cause Name and Year. The Cause Name filter keeps Suicide. The Year filter ranges from 2015 to 2015. The view is filtered on State, which excludes United States.

Question 5: How has the absolute number of suicides and the suicide rate in Washington State trended from 1999 through 2015?

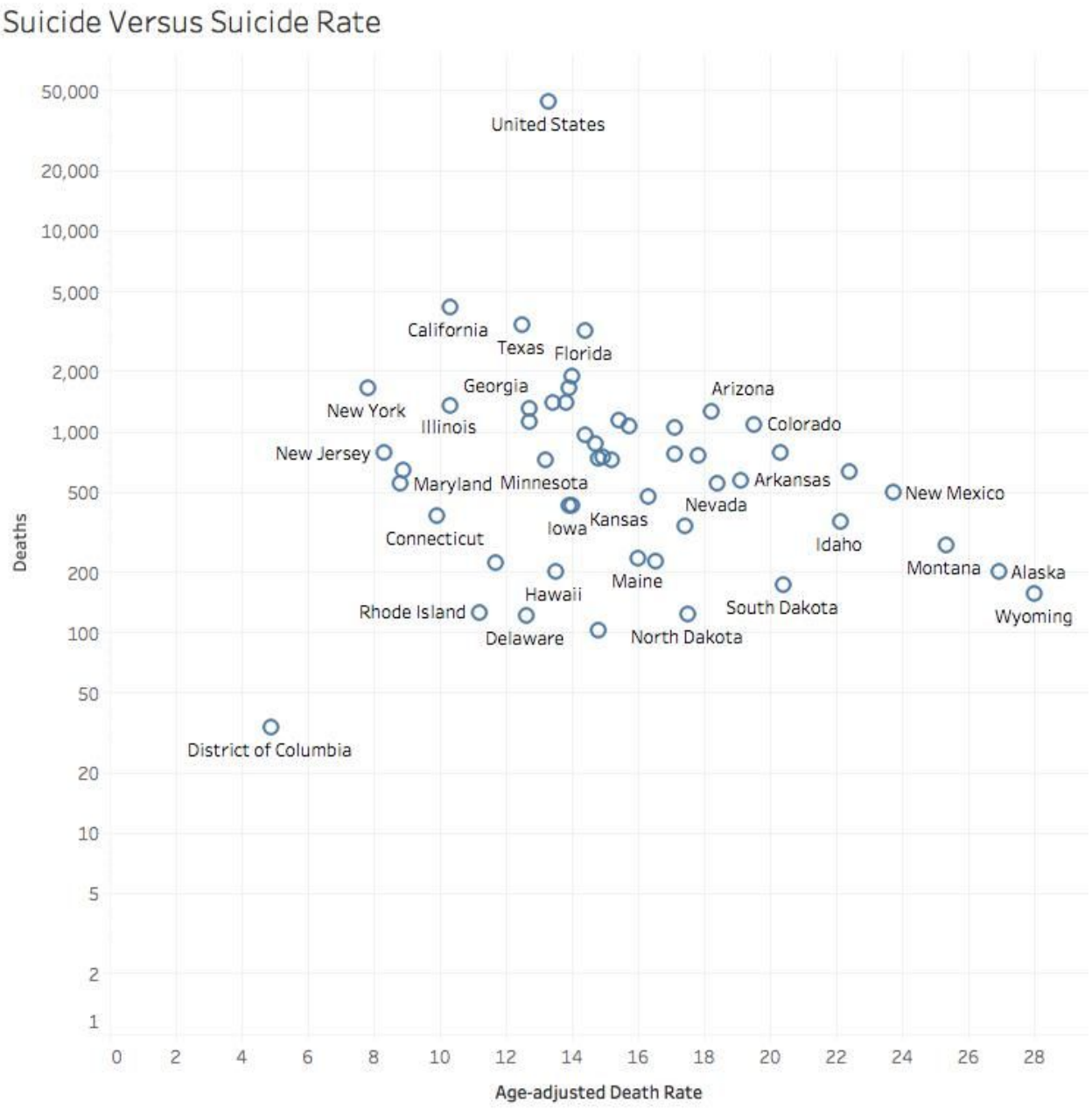
To answer this question, we set *Year* as a column, and *Death* as a row. We apply a *State* filter for Washington, and a *Cause Name* of “Suicide.” Next we set *Age-adjusted Death Rate* with a red-gold color coding. From the resulting visualization, we see the worrying trend that both the absolute number of suicides, and suicide rate are up over time:



The trend of sum of Deaths for Year. Color shows sum of Age-adjusted Death Rate. The data is filtered on State and Cause Name. The State filter keeps Washington. The Cause Name filter keeps Suicide.

Question 6: Is there a relationship between suicide rate, and absolute number of suicides in the latest year?

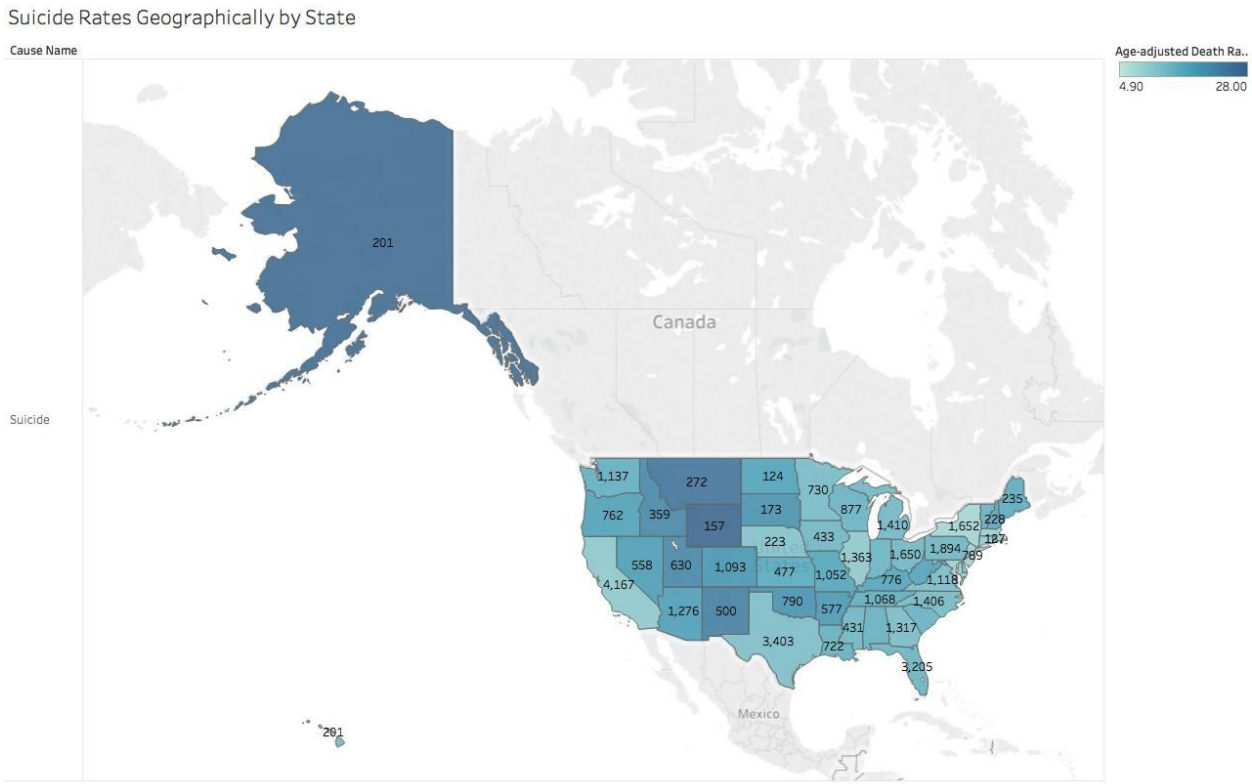
To answer this question, we set *Age-adjusted Death Rate* as a column, and *Death* as a row. We apply a *Cause Name* of “Suicide,” and a *Year* filter of “2015.” Many of the values for *Deaths* are clumped close to the low end of the scale, so we choose to click on the *Death* axis, and edit the axis to make it logarithmic. Next we drag *State* to the label icon. The resulting visualization shows no clear trend, neither linear, logarithmic, exponential, polynomial, nor power. We conclude, no, there is no obvious relationship between absolute number of suicides, and suicide rate. Knowing something about geography, though, we do notice, that many western (and less densely populated) states have higher death rates than there more densely populated eastern counterparts:



Age-adjusted Death Rate vs. Deaths. The marks are labeled by State. The data is filtered on Cause Name and Year. The Cause Name filter keeps Suicide. The Year filter ranges from 2015 to 2015.

Question 7: Can the results from question 3 be presented in such a way as to a show a possible geographic correlation (for example, that higher suicide rates exist in western states)?

To answer this question, we duplicate the worksheet produced in question 3. *Age-adjusted Death Rate* is represented here by shape, and we would prefer to see it represented as color. We therefore drag this measure to color in the marks tray. Next, we drag *Death* to label, just for reference. What we see here is a clear indication that western, less densely populated states have higher rates of suicide. Intuitively, we reject the idea that geographic location is causative of higher suicide rates. We suspect that the two measures are only correlated. We see, for example, that California has a relatively low suicide rate, despite being a western state. Our theory is bolstered by the observation that the least densely populated state, Alaska, has the highest suicide rate:



Map based on Longitude (generated) and Latitude (generated) broken down by Cause Name. Color shows sum of Age-adjusted Death Rate. The marks are labeled by sum of Deaths. Details are shown for State. The data is filtered on Year, which ranges from 2015 to 2015. The view is filtered on State and Cause Name. The State filter has multiple members selected. The Cause Name filter has multiple members selected.

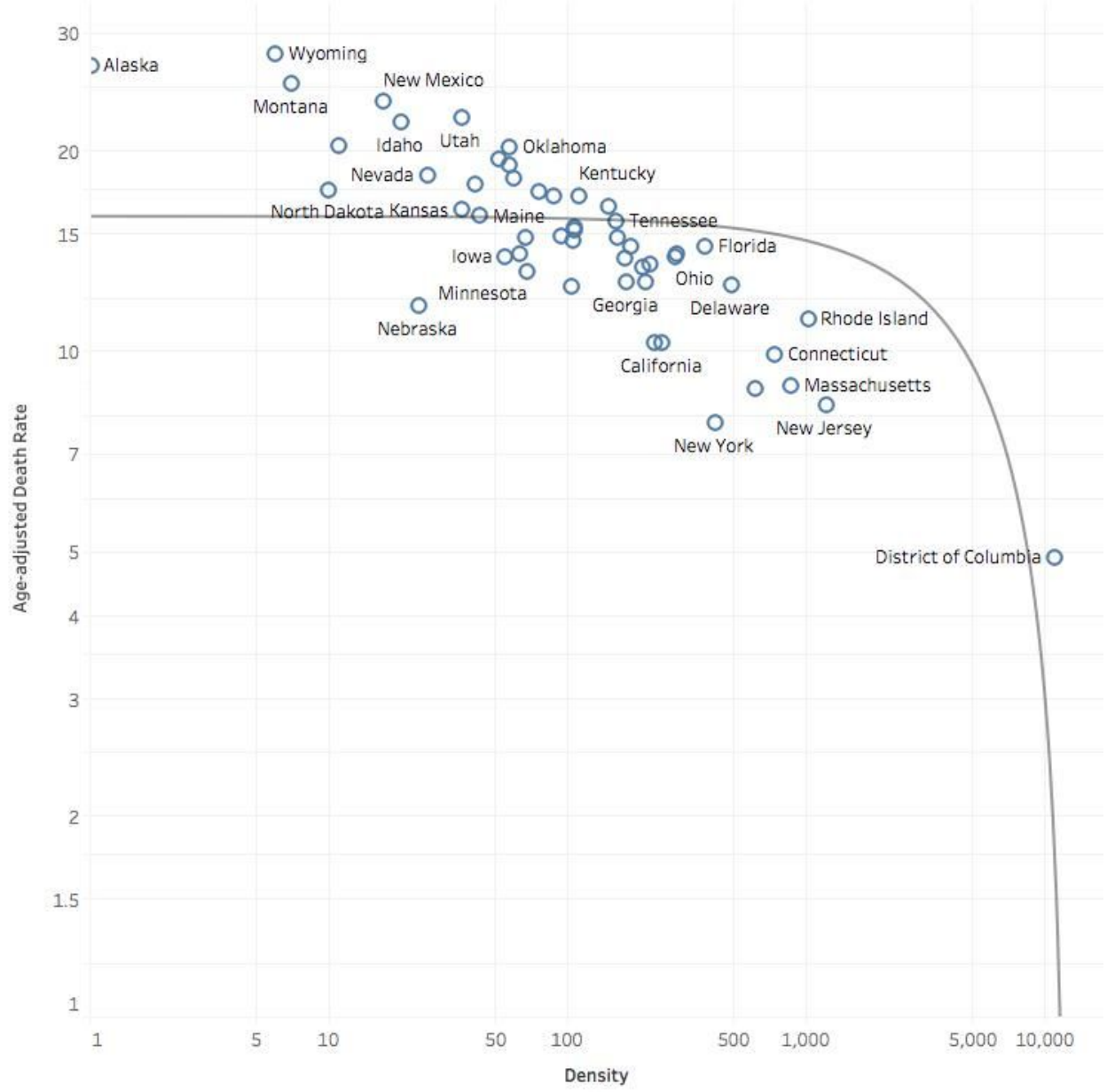
Question 8: Is there a relationship between population density, and suicide rate?

To answer this question, we first note that population density is not a statistic that is present in our data set. However, the data is readily obtainable online. For example, it is given in tabular format at:

https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population_density

Rather than look for the data a format ingestible by Tableau, we decide it will be quicker to simply copy the data by editing a CSV file. Although somewhat error prone, careful attention to data entry in the 53 resulting rows (header, 50 states, District of Columbia, and whole United States) should result in an accurate file. Next, we use the Data Source tab in Tableau to connect our new file with the existing NCHS file, linked by *State*. We use a left join to preserve all rows in the NCHS file, even if a corresponding *State* row is missing in our new file (which it should not be). Now we create a new workspace, add a column for *Density* (our new field) and a row of *Age-adjusted Death Rate*. Both axes are better represented on a logarithmic scale, so we change them from linear. We put in filters for *State* (remove the row for entire United States), *Cause* (limit to “Suicide”), and *Year* (limit to “2015”). We next add *State* for a label. We seem to see a linear trend, so we go to the Analytics tab, and add a trend line. Indeed the resulting trend line is significant. The R-Squared is **0.170412**, and the P-Value is **0.0026084**. We conclude that there is indeed a relationship between population density, and suicide rate. Without offering further proof, this author will theorize that isolation is a causative factor of suicide.

Suicide Rates by Population Density



Density vs. Age-adjusted Death Rate. The marks are labeled by State. The data is filtered on Cause Name and Year. The Cause Name filter keeps Suicide. The Year filter ranges from 2015 to 2015. The view is filtered on State, which excludes United States.