Minh Nguyen - 01597092 602-2 Data Visualization

Project Proposal

1- Datasets

Natality Information (Births)

I have filtered and extracted the necessary variables and index of the datasets.
https://github.com/minhnaru/project

From section (URL):
Healthcare (https://github.com/caesar0301/awesome-public-datasets#healthcare)
EHDP Large Health Datasets (https://www.ehdp.com/vitalnet/datasets.htm)
CDC Wonder - Public health information system (US federal) (https://wonder.cdc.gov)
Natality Information (Births) (https://wonder.cdc.gov/natality.html)
Natality for 2007 - 2015 (https://wonder.cdc.gov/controller/datarequest/D66)

2- Description of types and semantics

The dataset "Natality" represents the birth counts of residents and non-residents based on the maternal risk factors of Diabetes and Tobacco Use happening within the United States regions. The number of given births are calculated from state, census regions, mother's age, and mother's risk factors such as diabetes and tobacco use. All the data were collected from 2007 to 2015.

Source of dataset: Centers for Disease Control and Prevention (CDC)
United States Department of Health and Human Services (US DHHS)
National Center and Health Statistics (NCHS)
Natality public-use data on CDC WONDER Online Database.

The dataset in the website CDC WONDER provides table, and data extracts according to the chosen Group results. When requesting the data, I can limit and index the data by any and all of the data variables. Follow is the two datasets:

Natality_1.csv

Header	Definition
`State`	The states of the United States
`State Code`	The states of the United States code (1,2,3,...5,6)

`Year`	The year of data
`Year Code`	The year code of data
`Diabetes`	Indicate whether Diabetes is reported as a maternal risk factor
`Diabetes Code`	The code of Diabetes
`Tobacco Use`	Indicate whether tobacco use was reported during the pregnancy
`Tobacco Use Code`	The code of Tobacco Use
`Births`	The total number of births are given

Natality_2.csv

Header	Definition
`Census Region`	Four regions of the United States (Northeast, Midwest, South, and West)
`Census Region Code`	The code of the census regions
`Year`	The year of data
`Year Code`	The year code of data
`Age of Mother Year`	Indicate the age group of the mother at the time of birth
`Age of Mother Year Code`	Mother's age code
`Births`	The total number of births are given

Here is a look at the data:

State	State Code	Year	Year Code	Diabetes	Diabetes Code	Tobacco Use	Tobacco Use Code	Births
Alabama	1	2014	2014	Yes	1	Yes	1	291
Alabama	1	2014	2014	Yes	1	No	2	2619
Alabama	1	2014	2014	No	2	Yes	1	6075
Alabama	1	2014	2014	No	2	No	2	50145

3- Detailed list of the tasks envision project:

- What area of the US see the highest density of birth when mothers have both diabetes and tobacco use between 2007 and 2015?
- Which of the maternal risk factor (Diabetes or Tobacco Use) will affect the most number of birth?
- What different between each state based on the number of births?
- What different between with or without using tobacco and/or having diabetes?
- Does tobacco usage affect the number of given births? By how many times?
- Does diabetes affect the number of given births? By how many times?
- How many baby increase/decrease over period of 8 years?
- Is mother age group affect the number of given birth?
- What mother age group has the highest number of given birth?
- How many percentage does the mother age group from 41 to 49 account for?
- Does the census regions affect the number of given birth?

4- Sketch

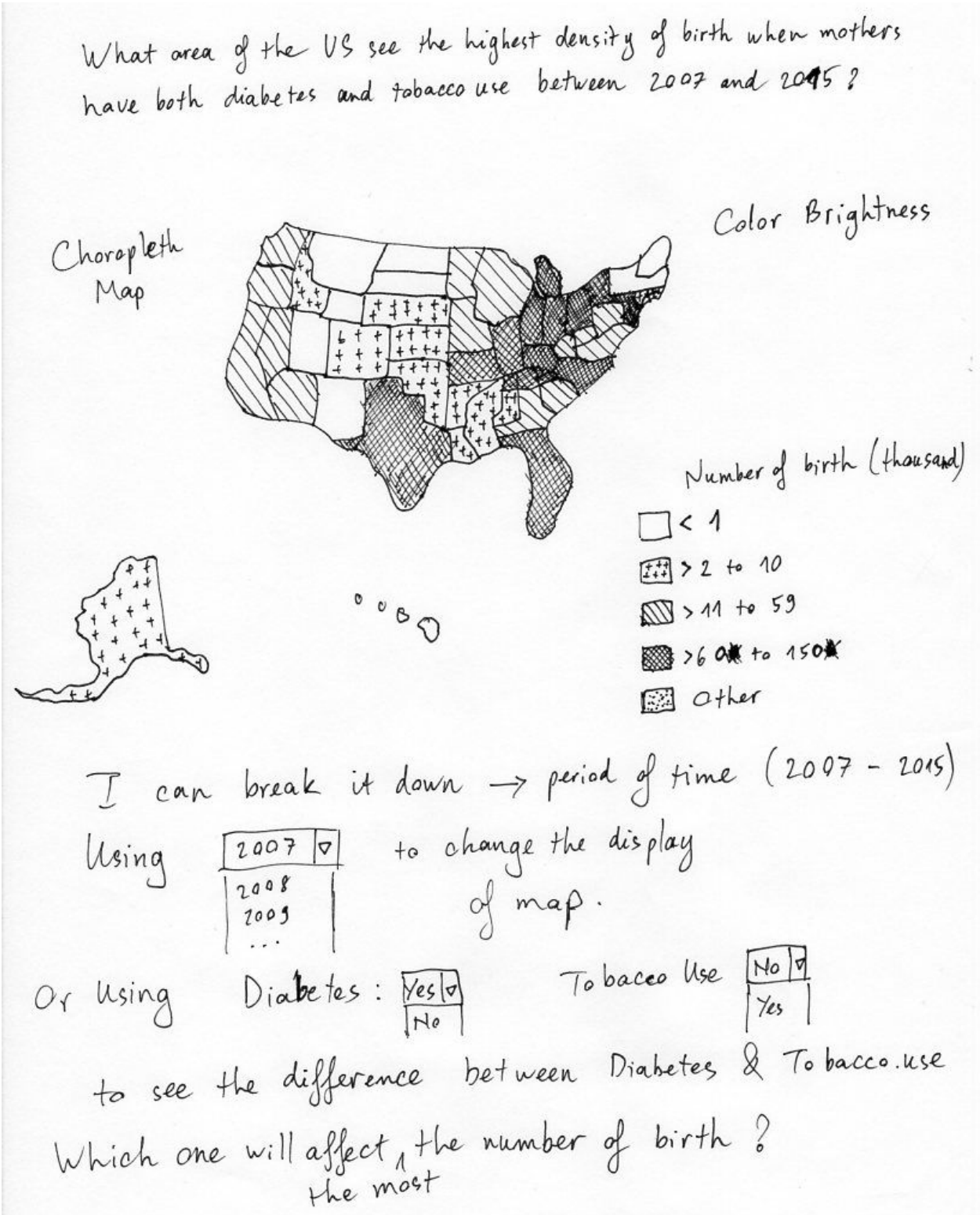

Figure 1

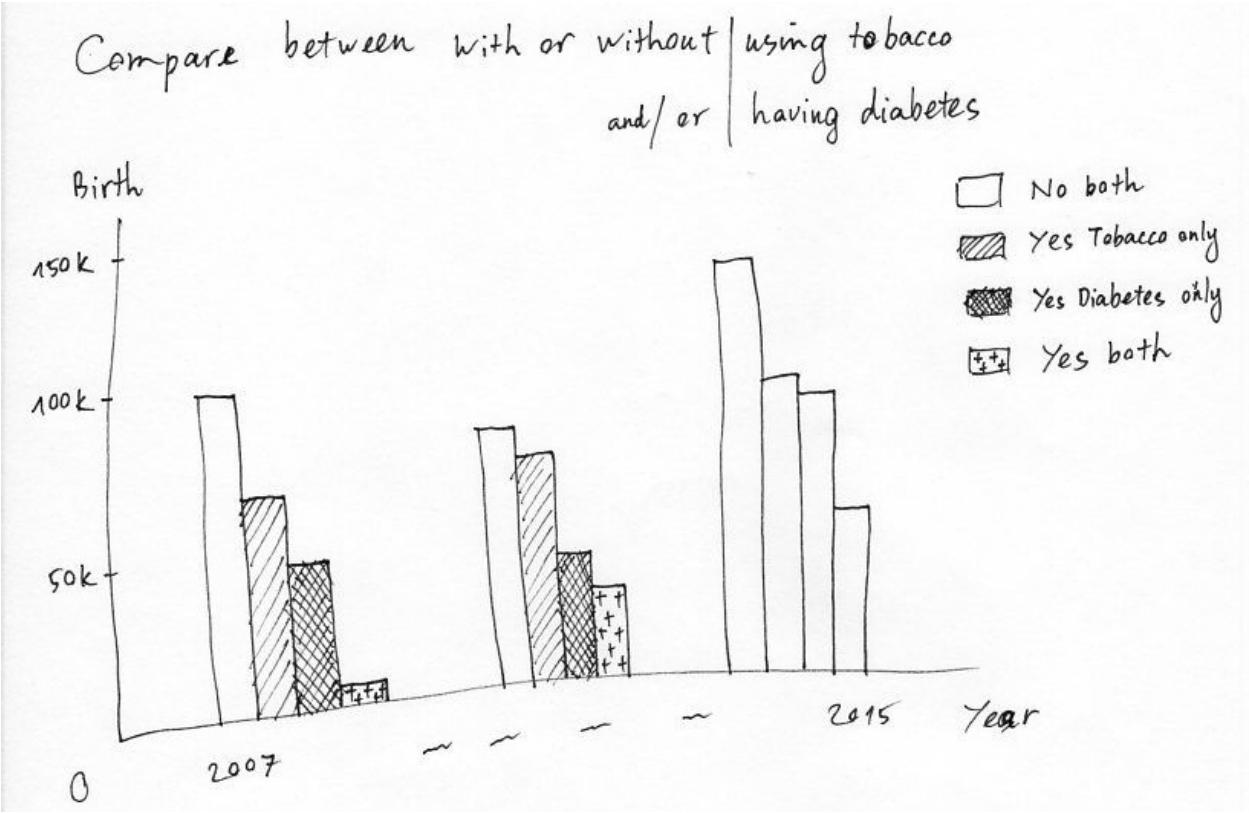
Compare between with or without | using tobacco
and/or | having diabetes
Birth
150k
100k
50k
0
2007
2015
Year
No both
Yes Tobacco only
Yes Diabetes only
Yes both

Figure 2

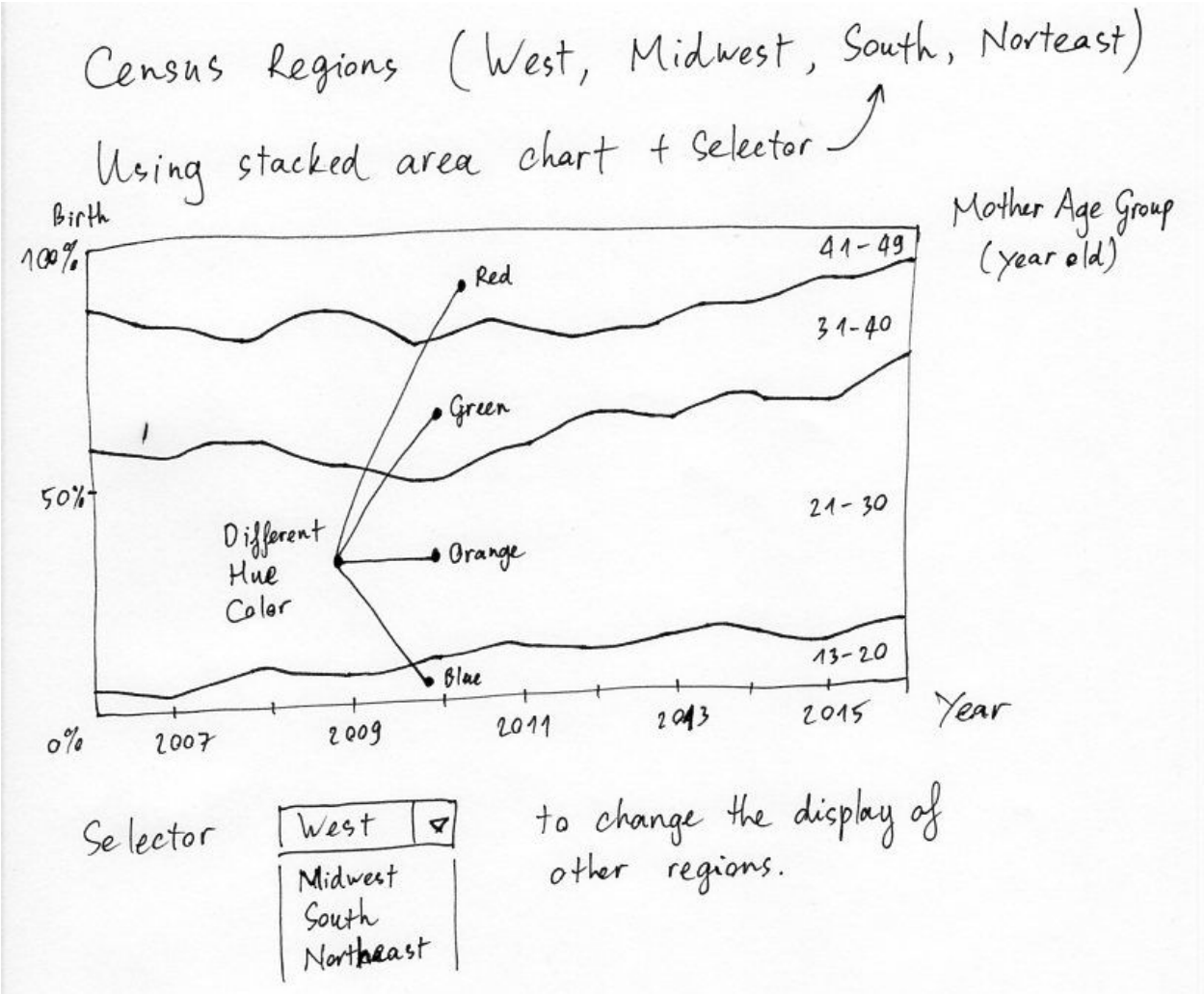

Figure 3

4- Ideas and Requirements

The Natality dataset contains large volume of data, so that I have limited to smaller datasets. D3.js is a javascript library used to bring data to visualization. It is used to perform this project visualization. Additionally, I may use other library such as CARTO and Leaflet to improve the visualization and interaction to the users. Both are open-source Javascript library, and a powerful, intuitive platform for interactive maps (https://carto.com), (http://leafletjs.com).

Healthcare is my interesting section. I have filtered some data from Natality datasets to use in my visualization. There are 3 visualization which will be combined together, I have mention above.

Figure 1:

- Using choropleth map to demonstrate the density of birth for each state in the US.

- Using color brightness to hightlight and separate each state.
- Changing the display of map by using dropdown menu (year), Diabetes (Yes/No), Tobacco Use (Yes/No). So we can see the difference for each selection.
- Reference link (not this project): https://bl.ocks.org/mbostock/4060606.

Figure 2:

- To compare birth effection between with or without using tobacco and/or having diabetes.
- Using grouped bar chart is the best way to compare those values.
- Reference link (not this project): https://bl.ocks.org/mbostock/882152.

Figure 3:

- Using stacked area chart to compare the difference of mother age group.
- Mother age group is divided into group of four: 13-20, 21-30, 31-40, 41-49 years old.
- Using different hue color to classify the age group.
- Using dropdown menu for those census regions to change the display.
- Transition between each selection to make it more interactive.
- Reference link (not this project): https://bl.ocks.org/mbostock/3885211.