Project Milestone 2: Curation and Design

By Group 6: Zhikai Zheng, Luke Becker, Takahiro Ishikuri, Angelica Sun

Repository URL for code and datasets: https://github.com/zzheng26/Stat-479-Group-6

Data Curation

Cleaning: We are mainly focusing on the weather data at those selected main city airports, so we have to do the cleaning to select the weather stations that are in range, and also the parameters that we want to discuss. To clean the data, we first selected the rows and columns that we are focusing on while dropping irrelevant data. After cleaning, the variables of interest are listed in the documentation doc.

Joining: Our raw data contains several csv files for each city. However, we want the city to be a factor. We have to join all the data into a unified dataset with each city and its data together since this would give us a way of comparing different cities with the faceting function.

Tidying: As we get the cleaned & joint data, it is ordered in date for each city. Although it is easy to make time series comparisons if the data is in chronological order, it is sometimes difficult to make comparisons on different areas. We would derive the dataset to generate one ordered by the station and another one ordered by the time so that we could perform comparison on both between cities and over time.

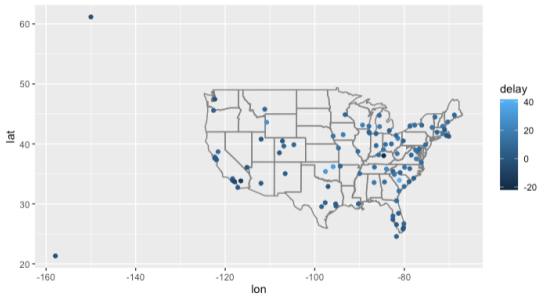
Deriving: We have to derive some data that we are interested in but not provided. For example, we would be interested in the temperature changing each day, thus we would have another variable storing the difference of EMXT (high temperature) and EMNT (low temperature). We also change the resolution from month to year to see a comparison change over time. Standard deviation of temperature is also included to see the fluctuation of the temperature. Also, we should split the date from mon-year into 2 variables containing month and year separately.

Challenging Step: During the examination of the dataset we found several cities with NA data, which would be severe for some years but better for the other. Missing-value visualizations were made to analyze how much missing data was there, and the imputation method was used to impute them according to the value of the other years.

Design Details

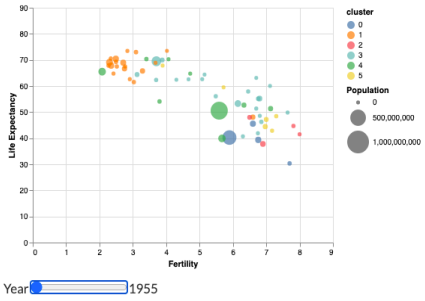
Interesting visualization ideas:

1. US map with “overview+detail”, selection slider over time range, color coding each (alpha=0.5) circle area centering each city, by temperature
2. Cross-sectional with homelessness data
3. Box and whisker plot - x-axis: general regions (e.g. northeast, midwest, etc.)
4. Line data plot: grouped by each city, over x-axis=time

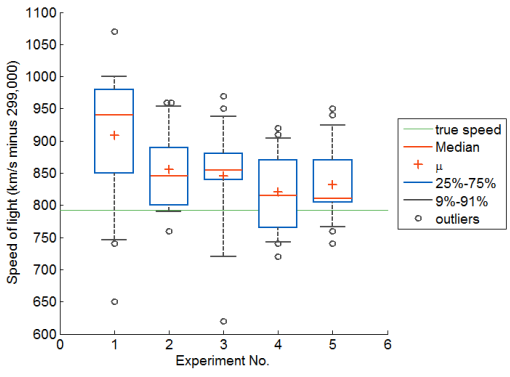


Our first brainstormed visualization design is to use a US map that has points on each of the ten cities we have decided to use for our data. For our encodings, we would have max temperature showing next to each of the points and when you move your mouse to hover over a certain city, we would include the average temperature for each of those cities as well. Another graphical design piece we would include is the color scheme of each data point. Similar to what local and national weather channels produce every day, we would use a color scheme that is darker red for warmer temperatures and darker blue for colder temperatures. Interactive cues will be located at the bottom of the graph with separate sliders for months and years to differentiate between seasons and check for patterns and/or outliers. The strength of this design is that you can see the US map and find out lots of

information with little to no effort. Another strength with the design is that you can add variables with ease and there's plenty of space to do so. A weakness is that there's not much you can do with each axis except for put latitude and longitude as this map is of the US.



A second visualization that we plan to use is a graph very similar to the one shown above. The slider bar at the bottom will allow the user to toggle between years and the second bar to toggle between months if they wish. Instead of fertility on the x-axis, we will have city names. Likewise, on the y-axis, we will have temperatures in degrees Fahrenheit. As far as the population section on this design goes we will utilize a cross-sectional dataset linking homelessness data and the weather data to make the points on the graph bigger or smaller depending on how many people are homeless in that particular city. The strength of this design includes the ability to switch between months and years with a simple drag of the slider. What this means is that the user can switch between many different views and trends with little to no effort. Another strength of this visualization is that it encompasses two different datasets that have rarely been linked together before.



A third visualization that we plan to use is a box and whisker plot that would incorporate general regions of the US on the x-axis (e.g. northeast, midwest, etc.) and temperatures on the y-axis. We would include the median of each in the middle of the graph and would also be able to use a couple more variables as well if we wanted to change the color based on temperature and also maybe put in precipitation or wind data to go along with this.