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 decided to mainly focus on the weather data at ten selected cities; the primary cleaning was done by selecting the weather station with the most data at each city, the time range of interest, and the parameters that we want to discuss. Then we separated some columns into variables that contain more specific kinds of information and converted them to appropriate data types accordingly; at this step, we get useful new variables (e.g. state, exact location of the station, year, month). See the documentation file for the full list of variables in cleaned datasets. Apart from this, we impute the missing values using logic and our knowledge of the real world, setting missing snow data of cities in hot areas to be 0 (this decision is also backed by our observation on the raw dataset; see the comment documentation in code script for more detail); after this step, all 4 columns associated with snow data are fixed with no NA values.

Joining: Our raw data contains several csv files for each city. To make “city” a factor, we join all the data into a unified dataset, making it convenient for us to compare different cities with encoding elements like faceting (datasets of a different resolution are created after this step).

Tidying: The dataset we first cleaned has data entry of each month; using this, we then derived a dataset of a different resolution, year; that is, it only has data entry of each year. Some of the variables in the derived dataset directly root from the variables in the old one (e.g. the new DSNW equal to the sum of a station’s monthly DSNW data of that year, since it represents the number of days with snowfall >= 1in); others are decided to be useful for visualization and derived using old variables. Apart from this major move, we also created an additional dataset ordered by time, making observing patterns across time directly from a portion of the dataset easier (potentially helpful for making preliminary choices on types of visualizations to make).

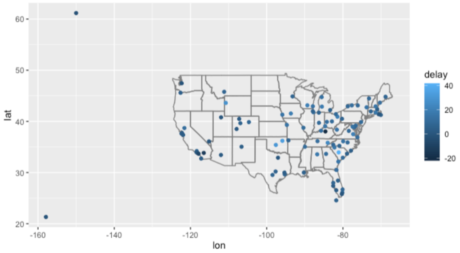
Deriving: Like discussed above, we derived new variables that might be useful to have when doing visualization. For example, with an interest in the scale of temperature variation, we made a variable, MXTRM, to store the maximum temperature range in a month (see documentation file for full details on each derived variable). In the new dataset of resolution "year", standard deviations of some variables in the "monthly" dataset are included, storing fluctuation of a certain element of weather in a year (e.g. standard deviation of monthly average wind speed).

Design Details

Interesting visualization ideas:

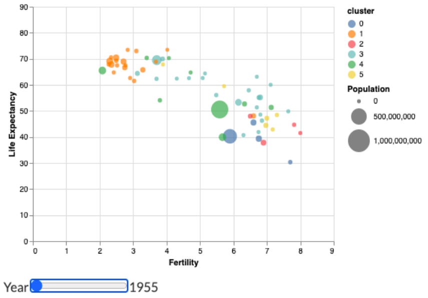
1. US map (selection slider over time range, color coding each (alpha=0.5) circle area centering each city and by temperature)
2. Potential cross-sectional analysis with homeless data (haven't decided, and data yet to be found)
3. Box and whisker plot
4. (simple line data plot: plotting different values over time and making lines grouped by "city")

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

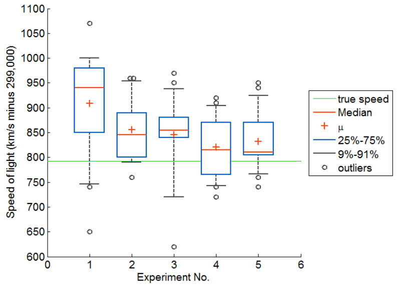


design is that you can see the US map and find out lots of information with little to no effort. Another strength with the design is that you can add variables with ease and there's plenty of space to do so. A weakness is that there's not much you can do with each axis except for put latitude and longitude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


Speed of light (km/s minus 299,000)
Experiment No.
true speed
Median
μ
outliers