The Relationship Between Weather and Happiness:

Data Preparation and Visualization

Taylor Callahan

Department of Data Science, Bellevue University

DSC 540: Data Preparation

Professor Catherine Williams

March 2, 2023

Data Preparation

The data retrieved for this analysis was chosen to determine the relationship between the weather and happiness, with a focus on how it relates to suicide rates. To prepare the flat file of city data for use, columns had to be renamed, along with duplicates and NaN's being removed. The preparation of the website data on suicide rates centered around fuzzy matching and renaming to ensure a seamless join across the flat file source. Asterisks and white space also had to be cleaned to allow for seamless matching. The final source obtained contained weather data and was retrieved via API. While API data was received in a relatively clean format, work had to be done to retrieve necessary information from the JSON responses.

Data retrieved from these three distinct sources were joined within a SQL Lite database, where city and weather information were joined on latitude and longitude, which was further joined to suicide rate information on country name. The final dataset contained information on city population, country name, latitude, longitude, suicide rate by country, and various weather statistics.

Data Visualization and What Was Learned

After data preparation was complete, five visualizations were built to help tease out the relationship between the weather and suicide rates. The main points of interest in the creation of these visualizations included differences between genders, an understanding of suicide rates across the world, and which weather patterns, if any, are most likely to result in higher rates of suicide.

When examining the difference between suicide rates across genders, it was found that little to no difference exists. Although the average suicide rate (per 100,000) was 20.84 for females and 21.07 for males, this difference is not distinct enough to be significant. However,

when examining the linear relationship between various weather statistics and suicide rate, the highest positive correlation was found between cloudiness. The visualization created plotting cloudiness (%) versus suicide rate found a slight positive correlation, with locations with an average higher percentage of cloud cover resulting in a slight increase in suicide rates. When comparing the country with the lowest suicide rate to the country with the highest suicide rate, this theme continued, with the country with the highest suicide rate having a significantly higher average percentage of cloud cover throughout the year. Besides cloud cover, temperature appeared to be the next greatest factor, with lower temperatures resulting in slightly higher rates of suicide.

When examining geographical maps of suicide rates across the world, the only correlation that may be found is that coastal countries appear to have lower suicide rates than inland countries. This may directly correlate with higher temperatures and less cloud cover.

Ethical Implications

Most of the ethical implications associated with this data and analysis are related to issues of privacy and sensitivity. Due to the nature of the data, privacy is maintained, but the sensitivity of the topic must be handled with care when presenting information. The analysis also focused on suicide rate as a general indicator happiness, but confounding variables are certainly present within this statistic—such as economic well-being and/or political state.