ETL-Project Report

Extract:

Erin and I are interested in how weather influences behavior. We first found a weather dataset from The National Center for Environmental Information. This climatic data contained information on the amount of cloud cover for U.S. Cities collected from the different weather stations. The weather database came in a .dat file from The National Centers for Environmental Information ([www.ncdc.noaa.gov/ghcn/comparative-climatic-data](http://www.ncdc.noaa.gov/ghcn/comparative-climatic-data)). We opened the file in excel and converted it to a csv file. We were then able to use pandas to read the csv file and convert it into a data frame.

We then found a crime database which contained crime rates for U.S. Cities also from 2015. The crime database came in an xls file from DataWorld. We uploaded into pandas and converted it to a data frame using pd.read\_excel function.

Transform:

Since Erin and I were interested in how weather influences crime rates, we needed both the weather data and the crime rate data for U.S. cities. This meant that we needed to merge our two datasets by city. For the crime dataset, the location included not only a major city, but also the cities in the metropolitan area. A visual inspection at the weather dataset showed that the location of the weather stations was individual cities. In order to merge the two datasets, we needed to join the datasets based on city. As we worked on our project, the steps we needed to take evolved and end up becoming a little more complicated than we realized.

For the crime data set, the first step was to simplify the names of the variables and create a data frame based on the variables that we were interested in using (Location, Aggravated Assault, Violent Crime, Murder and Robbery). Since the location include a major city and then other cities and/or states that were boarding the metropolitan area, multiple steps were taken to separate the major city from nearby cities and states. We were able to break apart the original column into multiple columns and eventually had the major city in a column by itself. Once we had the major city in a column by itself, we realized that due to cities having same name and different states, we had to eventually put the major city and state abbreviation in a cell by itself. For cities that were on the border and had two states, we dropped the second state and kept the first one. We then switched the city names to uppercase letters and sort alphabetically.

For the weather dataset, the first step was breaking apart the location column into two columns since we originally thought we were joining on city names alone. The location column was broken into a “City” column and “State” column. The data set had the number of clear days, partially cloudy days and cloudy for each month and then an annual total for the 3 types of cloud cover. We decided that the annual number of clear days for a city would be the best way to measure weather for a city. We then looked for empty cells and other non-numerical data. We found that there were multiple cities with missing data. However, instead of an empty cell, there was an asterisk for missing data. We searched for the rows of data that contained asterisks and then deleted those rows. Next, we added up the number of clear days each month and compared it with the annual total listed in the dataset. The two columns were not identical, so we used the annual total of clear days listed in the database. We then made a data frame that only contained the columns that we needed.

Once we started joining the datasets, there were additional complications that needed to be addressed. To solve these problems, we took the “City” column and “State” column in each dataset and merged them back together. When the join did not work, we realized that one dataset has a space after the comma and before the state abbreviation, while the other dataset did not contain the space. Once the issue was addressed, the combined data sets were put into a data frame with 130 rows (or 130 matches between the 2 datasets).

Load:

The last step was to load the database into SQL Workbench. Since we had joined the two data frames in pandas, we uploaded our final data frame into SQL as a database. We loaded our final database into a relational database. We then created one table that contained the city & state, annual number of clear days, aggravated assault rates, violent crime rates, robbery rates and murder rates.