Lyttia McManus- Cancinos August 2018

Summary of Capstone Data Wrangling

This project studies the effect of weather and property factors on crimes committed in and around Phoenix, Arizona. Data sources include open data from the City of Phoenix Crime Data website, Realtor.com, and the rwunderground API. Nearly all of the crime data remains intact, joined to columns containing the temperature, weather conditions, median listing price (property value), and number of surrounding listings (a measure of transiency) at the time and location of the crime. A summary of the data wrangling process follows.

To make programming more user friendly, all of the variables in the Phoenix crime data frame (except "INC NUMBER")[1] are subtly renamed. Many of the original variables contain capital letters and blank spaces which can cause problems when writing code in R. The "OCCURRED ON" column is separated into multiple columns ("month", "day", "year", "hour", "minute") to make joining data frames, as well as more specific analysis, possible. The "OCCURRED TO" column is removed as it is irrelevant to the specific conditions leading to the development of each crime type. If any dates are not reported, or NA, the dates are filled by the order (incident number) in which the crime is reported. This is true for 367 observations out of 169,818 total crimes reported from November 1st, 2015 through June 25th, 2018. Dummy binary variables are added for crime categories to use in further analysis. There are originally 95 unique premise types, so smaller premise categories are added with dummy binary variables for a more comprehensible and complete analysis.

Many inconsequential variables are included in the weather and Realtor.com data frames as it pertains to this project. The rwunderground API "history_range()" function calls for observations as requested, but includes many excess variables. Unnecessary variables are removed using dplyr. With API calls made, and tidy data frames, the property data joins the crimes data by "zip code", "month", and "year" to produce the "crimes3" data frame. This matches the crime to the median property value, accurate to the zip code, month, and year of the crime.[2] The weather data joins separately by "zip code", "year", "month", "day", and "hour" of the crime to create "crimes4".

While exploring data for missing values, 6 distinct zip codes missing property values emerge. The 3 NA zip codes are removed since an address or clues to a location are non-existent. The zip codes 85290 (currently does not exist), 85337 (remote location), and 85363, are excluded as each of these zip codes only contain one crime. The 85034 and 85307 zip codes each contain 3,511 and 269 crimes respectively, so it is worthwhile to use historic property values from Zillow to fill these values.

Weather conditions and property values provide a very small picture of the setting in which these crimes occur and offer a limited view on the overall development of crime in Phoenix. Adding more variables such as the education status and quality of education offered to a community can increase the efficacy of training a machine learning model to predict crime.

[1] I chose to leave the incident number column as is to indicate the integrity of these numbers. These numbers may be useful in identifying an observation's original location in a data frame if needed or give additional information to police.

[2] It is not possible to get up-to-date property value estimates on the addresses from the Phoenix Open Crime Data because addresses are displayed to the hundred block for confidentiality.