Summary of Capstone Data Wrangling Lyttia McManus- Cancinos

For this project, I chose to analyze the conditions that produce crimes committed in and around Phoenix, Arizona. The variables I am exploring now, include weather (temperature and conditions), property value (median listing price) as well as transiency (number of listings). These variables account for only a small picture of the conditions in which crimes occur and will give me a limited view on the overall development of crime in this city. After completing this course I plan to add more variables to my analysis to increase the efficacy of a machine learning model to predict crime.

I subtly renamed all of the variables in the Phoenix crime data frame (except INC NUMBER)[1] to make writing code more user friendly. There were many variables with capital letters and spaces which can cause problems when programming in R. I separated the OCCURRED ON column into multiple columns (month, day, year, hour, minute) to make joining data frames (as well as more specific analysis) possible. I removed the OCCURRED TO column because I am not analyzing the duration of the crime, but the specific conditions leading to the development of each crime type. If any dates were not reported, or NA, I filled the dates by the order (incident number) in which the crime was reported. At this time, this is true for 367 observations out of 169,818 total crimes reported between June 25th, 2018 through November 1st, 2015. I created dummy variables for crime categories to aid in performing analysis. There are currently 95 unique premise types, so I am currently thinking about how to create new premise categories for a more comprehensible analysis.

As it pertains to my analysis, many inconsequential variables were included in the Rwunderground and Realtor.com dataframes. The "Rwunderground" API's history_range() function calls for all of the observations I need, but many more variables than needed. Once the data frames were all cleaned up, I joined the property data frame to the crimes data frame them by zip code, month, and year in the "crimes3" data frame. This allowed me to connect the crime to an average property value, accurate to the zip code, month, and year of the crime. It is not possible to get up to date real estate value estimates to the address on the Phoenix Open Crime Data because addresses are displayed to the hundred block for confidentiality. I then joined the wunderground data by zip code, year, month, day, and hour of the crime to create "crimes4".

As I explored my data for missing values, I found that there were 6 distinct zipcodes missing property values. I decided to remove the 3 NA zip codes because I could not find an address or clues to a location. I also excluded 85290 (this zip code currently does not exist), 85337(Gila Bend far out location), and 85363 from the data frame, as each of these zip codes only produced one crime. I will be using the historic market values from Zillow for the 85034 and 85307 zip codes as they each contained 3,511 and 269 crimes respectively.

[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 dataframe if needed or give additional information to police.