Laura Wiseman
Clarissa Qian

Submission 1 - Data Management and Cleaning

Part I

Q1: 1223094
Q2: 512
Q3: 8159
Q4: McCarran International Airport
Q5: 9

Part II

We found formatting issues by printing the unique city names in Arizona and manually looking to see common issues. The issues we found and how we handled them are:

- Upper vs. lowercase letters: most entries followed the convention that only the first letter of each word in a city is capitalized. To unify the entries, we capitalized all city names using the pandas string method str.upper().
- Whitespace: To clean up any leading or trailing whitespace, we used the string method str.strip().
- Typos: we manually wrote a list of valid cities in Arizona and did a string compare when cities values weren't in the list. Using the built-in python library, "difflib", if there was a close enough match to a city in the list, we assumed that was a typo and replaced the entry.
- Missing entries: To clean our data, we deleted any rows with missing entries using the pandas dropna(axis=0) method.
- Naming conventions eg Laveen vs. Laveen Village: handled in the same way as typos
- Extraneous information such as including direction eg East Mesa vs. Mesa: To account for directional extraneous information, we can delete every instance where "North", "East", "South", or "West" appears. To do this, we used the str.replace() method to replace every instance of these words with whitespace.
- Including state eg Arizona or AZ in city name: there are no valid cities with "Arizona" (or AZ) in their city name, so we went through and dropped any rows where the city field was invalid.
- False locations eg Las Vegas. When checking for typos, if a city name wasn't in the list of valid cities and the "difflib" library couldn't find a close enough match, we changed the value of this city to "" (space) and dropped these rows later

The list of unique cities found in Arizona is: ['PHOENIX' 'GOODYEAR' 'GLENDALE' 'SCOTTSDALE' 'MESA' 'GILBERT' 'LITCHFIELD PARK' 'TEMPE' 'PEORIA' 'CHANDLER' 'SURPRISE' 'BUCKEYE' 'QUEEN CREEK' 'AVONDALE' 'HIGLEY' 'CAVE CREEK' 'SUN CITY' 'CAREFREE' 'EL MIRAGE' 'PARADISE VALLEY' 'LITCHFIELD' 'FOUNTAIN HILLS' 'TOLLESON' 'SUN LAKES' 'FORT MCDOWELL' 'APACHE JUNCTION' 'LAVEEN VILLAGE' 'MARICOPA' 'YOUNGTOWN' 'ANTHEM' 'SOMERTON' 'GUADALUPE' 'VALLEYWIDE' 'RIO

VERDE' 'WADDELL' 'PASADENA' 'AHWATUKEE' 'STETSON VALLEY' 'SEDONA' 'APACHE TRAIL' 'RAINBOW VALLEY' 'RED ROCK' 'DESERT RIDGE' 'ESTRELLA VILLAGE' 'SUNNYSLOPE' 'SAN TAN' 'CENTRAL CITY VILLAGE' 'ARROWHEAD' 'TUCSON' 'CENTRAL' 'RED MOUNTAIN' 'GREENWAY']
The length of this list is 52 cities.