Nathan Timmerman and Micah Thompkins
CS 396 Data Science
Submission 1
1/15/19

Q1:1223094
Q2:512
Q3:973
Q4:McCarran International Airport
Q5:9

First, we printed out all the unique state names checking to make sure that AZ wasn't misspelled and then data be under that spelling. We found no misspelled AZ abbreviations, the closest was AB and that is for Albuquerque. We then used regex to check for variations of AZ (i.e. Az; aZ; Az, etc.) and compile that into a dataframe. We found a non-exhaustive list of city names in Arizona to cross reference against. We then printed out all the unique city names from that data frame and examined them to find dirty data. Below is a list of dirty data and how we fixed it.

1) Missing Cities from our internal list
    a) If the unique city name was actually a city in Arizona and not inside our internally managed list of city names we added the city to our internal list to check against.
2) Business name instead of the city
    a) Additionally, some of the locations had something that was not a city name as the name of the city and we fixed that by finding the location of the city using latitude and longitude and replacing these false "cities" with those correct city names.
3) Cities spelled wrong with Levenshtein
    a) Using the Levenshtein distance, a metric for checking how different strings are, we checked city names against each other from the unique list to find the closest city to the misspelled one.
4) Stripping space at end and capitalization
    a) We got rid of both of these issues by grouping separate words together and then only capitalizing the first letter
5) Unicode in front of city
    a) We stripped the Unicode with a method that removes Unicode characters
6) End with Az and two names for same location/extra word in city name
    a) We had a check to see if a city from our internal list of cities was contained within the string of the city name i.e. there was a location written as "Westworld Scottsdale" and this check would catch that and replace that city name with Scottsdale, the correct city we want.
7) Nothing inside
    a) There was one location that had no city name so we found out what city that was then replaced the empty string with a string with Mesa, the where the business was located, inside.