397A Final Project: Museums in the US Data

Corinne Greene & Victoria Okoro

A. Wrangling the Data

a. Thought process

When considering what type of information would be interesting to do analysis on, one topic that one of us came up with was museums, since museums have a lot of data about basically infinite topics. Choosing what kind of data to look for relating to museums was actually a sort of challenge. This led us to find a dataset from the Institute of Museum and Library Services, which provides a CSV of all recorded museums in the US in 2018 and data about those museums. This is a rather extensive list, so when considering how to break it down I noticed that one column labelled what type of museum it is - children's, art, science, ect. This made me curious about the availability of children's museums, and if the average income of a certain area predicts the amount of children's museums in the area. This was our starting point when going into Jupyter Notebook to start inputting and editing the data.

b. Process in Jupyter Notebook

The first step was downloading the CSV from the website with only the columns that we wanted:

```
import pandas as pd
import numpy as np

df = pd.read_csv (r'MuseumFile2018_File1_Nulls.csv')
df = pd.DataFrame(df, columns= ['MID', 'DISCIPL', 'COMMONNAME', 'LEGALNAME',
'ALTNAME', 'PHSTREET', 'PHCITY', 'PHSTATE', 'PHZIP5', 'INCOMECD15', 'LONGITUDE',
'LATITUDE', 'AAMREG', 'LOCALE4'])
df['PHZIP5'].replace(' ', np.nan, inplace=True)
df.dropna(subset=['PHZIP5'], inplace=True)
df.shape
```

The total size of our dataset is over 7000 rows. Because we were initially interested in children's museums, we started by also filtering by that type of museum, but this resulted in a dataset of less than 500 hundred which was surprising. Therefore we ended up changing our hypothesis to investigate the correlation between average income and any museum availability. For our purposes, we also decided to only use data points that have a zip code value, so we ended up removing rows that had those columns empty. This results in 2064 rows and 14 columns.

Next, the Missouri Census Data Center maintains a database of median family income based on zip code for all states. We download this data and change the column names to the same format:

```
zipDf = pd.read_csv (r'ZIP_codes_2018.csv')
zipDf = pd.DataFrame(zipDf, columns= ['ZIP Code', 'Median family income (2018)'])
zipDf = zipDf.rename(columns = {'ZIP Code':'PHZIP5'})
zipDf = zipDf.rename(columns = {'Median family income (2018)':'MEDIAN_FAM_INCOME'})
zipDf
```

Next, we want to merge the two tables. We need to do this in zip code (PHZIP5), which is an object for df, so we must convert the column type first. Since analysis will be difficult with null values for the median we take those out (but can always change this if we want to try different analysis in the future)

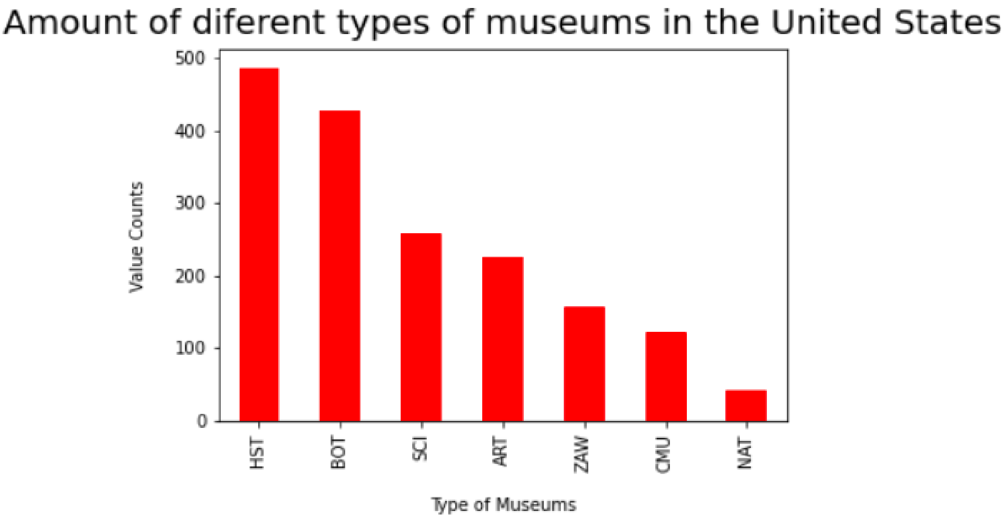
```
df["PHZIP5"] = df["PHZIP5"].astype(str).astype(int)
finalDf = pd.merge(df, zipDf, on='PHZIP5', how='inner')
finalDf.dropna(subset=['MEDIAN_FAM_INCOME'], inplace=True)
finalDf
```

This results in a dataframe with 1715 rows and 15 columns.

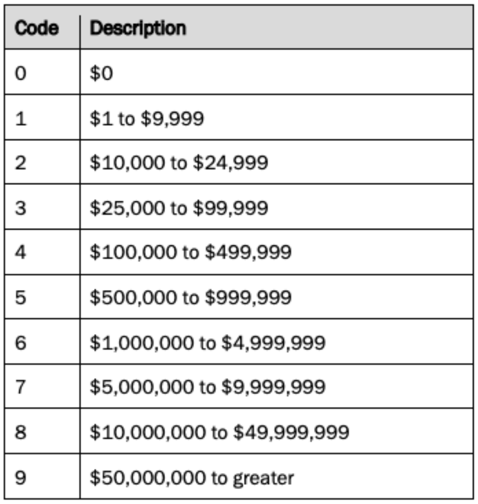
B. Visualizations

For our visualizations, we believed it was beneficial to illustrate the different types of museums in the United States, how the income of the museums differs from the average income of U.S citizens and how the average income varies depending on the State. Our first visual is a bar chart that illustrates the number of different museums in the United States. The majority of the museums in the United States focus on history, while the least amount of museums focuses on historical societies and historic preservation.

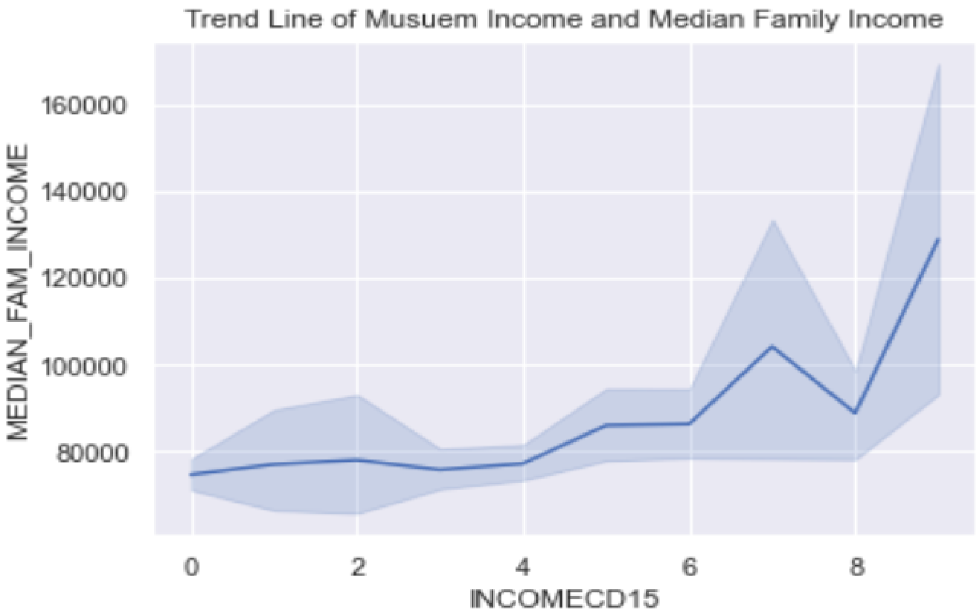
- HST = History Museums
- ART = Art Museums
- BOT Arboretums, Botanical Gardens, & Nature Centers
- SCI = Science & Technology Museums & Planetariums
- CMU = Children Museums
- ZAW = Zoos, Aquariums, & Wildlife Conservation
- NAT= Historical Societies / Historic Preservation

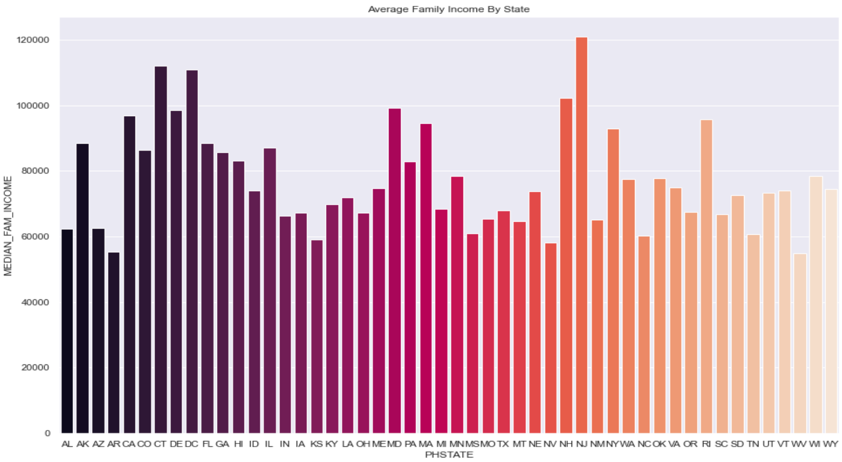


For the second visualization, we created a line plot that illustrates a trend line between the income of museums vs. the median family income in the United States. The visualization allowed us to see how ceratin families' incomes grant them access to more expensive museums. There is an increase at the 8 level code, which shows that families who make $90,000 or more can afford to go to museums worth $10,000,000 to $49,999,999. The majority of families in the United States can afford to attend museums that are code level 1 through 6.



Code	Description
0	$0
1	$1 to $9,999
2	$10,000 to $24,999
3	$25,000 to $99,999
4	$100,000 to $499,999
5	$500,000 to $999,999
6	$1,000,000 to $4,999,999
7	$5,000,000 to $9,999,999
8	$10,000,000 to $49,999,999
9	$50,000,000 to greater





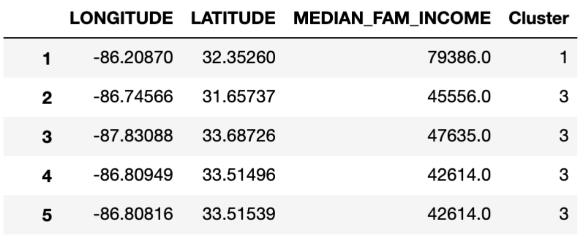
We wanted to see how the median family differs depending on the State for our third visualization. We discovered that New Jersey has the highest median family income, while Wisconsin has the lowest median family income. East and West coast States such as Massachusetts and Connetiut have higher median family incomes than Southern and MidWestern States such as Texas and

WyomingOverall, this visual gave us a deeper understanding on where families make more money in the United States.

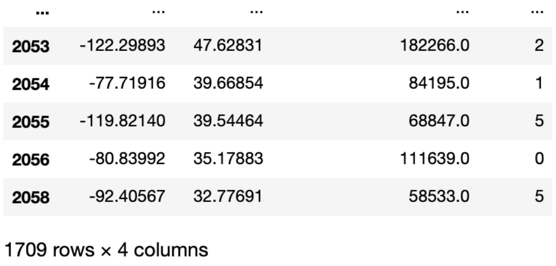
C. Unsupervised Learning

a. K-Means Clustering

Since we have latitude and longitude and neither of us had these values in the datasets we used in the original homeworks, we thought it could be cool to see what K-means clustering would look like for this data and what clusters would result. We start by making a new dataframe with only the columns we want. In this case, we decided to use the median family income as our third column. Setting kmeans to 6 clusters, we use kmeans.fit_predict to assign each datapoint to a cluster. The resulting data frame look like this:



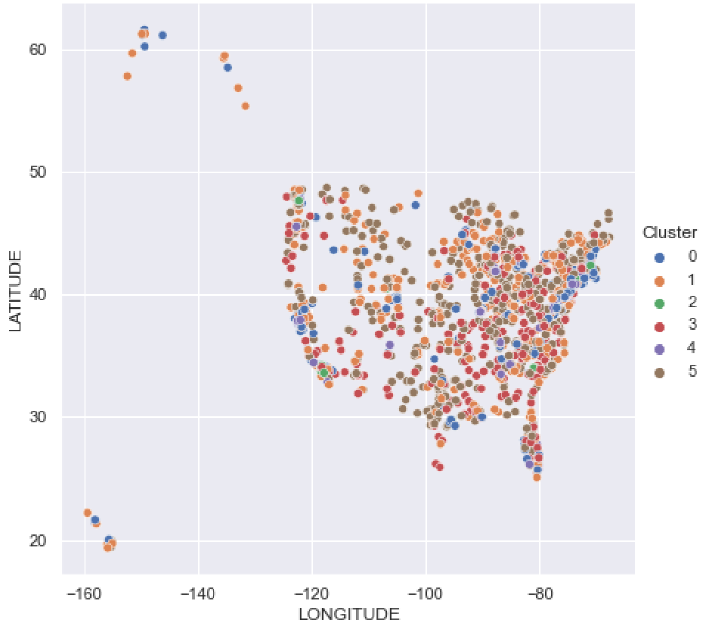
	LONGITUDE	LATITUDE	MEDIAN_FAM_INCOME	Cluster
1	-86.20870	32.35260	79386.0	1
2	-86.74566	31.65737	45556.0	3
3	-87.83088	33.68726	47635.0	3
4	-86.80949	33.51496	42614.0	3
5	-86.80816	33.51539	42614.0	3



...	...	...	...	...
2053	-122.29893	47.62831	182266.0	2
2054	-77.71916	39.66854	84195.0	1
2055	-119.82140	39.54464	68847.0	5
2056	-80.83992	35.17883	111639.0	0
2058	-92.40567	32.77691	58533.0	5

1709 rows × 4 columns

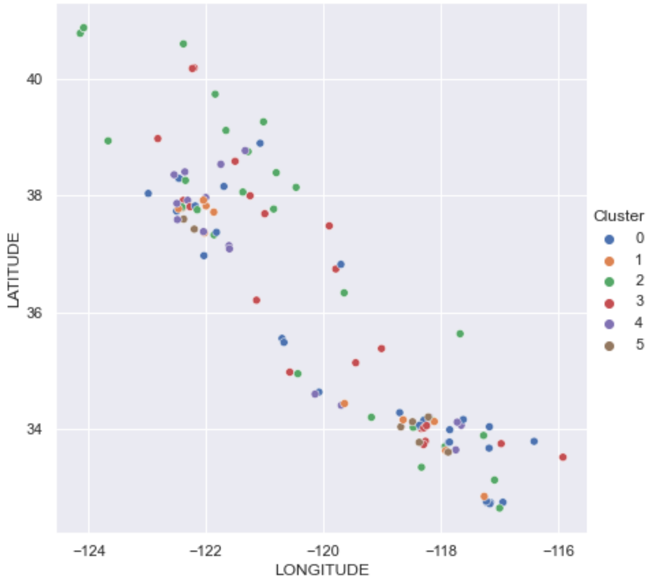
In this case, the number of clusters were chosen arbitrarily, as we just wanted to see what would be produced. Using sns.relplot, we print out the following representation of our clustering:



Now, this representation surprised us on multiple levels. For one, it does kind of look like the US - cool! - but it seems like our US is completely missing any data on the west coast. This led me into the excel to investigate what was going on.

First, I did a ctrl+F search to final examples of data points form CA. Since we took out all data points with no zip code, I wanted to see if this data set for some reason just doesn't have zip codes for west coast states, which would be odd, but also wasn't the case - there were many such data points with zip codes. I checked this in the median income csv as well, but many data points in CA had income recorded. Finally, I realized - how could a data point have no zip code for the museum location but still have a latitude and longitude, as many of them did. Checking the informational document on the descriptions of each column I found a mistake.

We had assumed that 'institution' in the description of latitude and longitude referred to the actual museum, but in fact they were referring to "parent or affiliated academic institution." Therefore, we weren't quite right about what was being plotted, but this did answer my question because we should still be seeing institutions in the west coast. Unfortunately, I did quite a bit of analysis before realizing that I probably should have googled the general longitude of California, which is about 120 degrees. This shows that although I had a different idea of what I thought this map should have looked like, it made me confused in the comprehension of the result. However, it did uncover that the data we were using was not actually the data we were looking for!



So, in terms of what the clusters actually show, it's pretty different from what we were looking for. While we were looking to see if there would be certain groupings of museums based on family income levels of the zip code, this was telling us groupings based on affiliated academic institutions. Unfortunately, it doesn't really seem to show anything - nothing like if you wanted to look at the median incomes of a specific state or region and the breakdown of the incomes of certain areas.

For curiosity's sake, I wanted to see what this would be like on a smaller scale to see if I could understand what was going on in the clustering better. Therefore I plotted only the data points in California, and this is the result, which looks like the right size. Unfortunately this doesn't show much correlation between the clusters, and if anything this shows that there is less correlation than I thought. If I took an attempt at reasoning this, because these are institutions, i.e. schools, there likely isn't much correlation between schools and median income based on zip code. I wish I had the latitude and longitude of the actual museums themselves, then I could run k-means on the correlation between the availability of museums to the average income.