**Data Cleanup & Exploration**

​

* <Pauline> creating and cleaning up the combined census csv file
* <Pauline> Discuss insights you had while exploring the data that you didn't anticipate
* On the combined census csv file, removed all rows with blanks/ NA and extreme values
* The lat/lng dataframe and the census dataframe were from different websites, and differed slightly in length and in the zipcode format. So faced a challenge when trying to merge both DFs on zipcode.
* We resolved this by using the lambda function to change the zipcode format (<https://stackoverflow.com/questions/13250046/how-to-keep-leading-zeros-in-a-column-when-reading-csv-with-pandas>) and using a left join when merging

​

**Data Analysis & Findings**

​

* In order to come up with our target market, analyzed three key datasets – age, income, and population.
* Based on annual research from Motion Picture Association of America (MPAA), the demographics of a typical movie watcher are
* Used the describe function to see how the data is distributed:
  + Age appears to be close to normal distribution, with majority of people between 40-50 years
  + Population is heavily skewed to the left, as majority of data is below 20,000 people
  + We see a similar trend with Income – majority of the data is between $25000-$75000
* Combining both insights, we came up with our typical target market – which has a median age of 30 (i.e. zip codes with the highest number of 30yr olds), population of 10,500 (mean of the dataset) and household income of $57,600 (mean of the dataset)
* We then calculated z-scores for age, population and income across each zip code, as this would be used as the weights for plotting the heatmap
  + Why manually calculate instead of using the z-score function?
  + For age, created modified z-scores: Instead of taking the difference of each observation from the mean, we took the difference of each observation from 30 (which was our target age group). This was done bcoz majority of data in our dataset was in the 40-50 years age group, which was above our target and may have skewed our analysis
* Combined the three z-scores using a weighted average method based on the assumption that age, population and income have equal weights in deciding the target market. This will also be our feasibility score

​

**Conclusion**

​Based on our calculations, the top 3 cities with the highest combined z-scores (closest to one) were: *Sioux Falls, SD; Page, AZ; and Gonzales, CA*. This was a surprising conclusion as we were expecting populous metros (like NY, LA, Chicago) to come up with the highest scores. We attribute this to using an absolute number for age, as opposed to the mean.

**Key Challenges/ Limitations**

* Gmaps does not plot negative values, hence we were unable to plot the negative z-scores on the heatmap. So only the zipcodes with z-scores above the mean have been plotted
* One of the key challenges was data availability
  + Wanted to run a correlation between our variables (age, income, population) vs box office receipts/ ticket sales. However, we could not do this as box office receipts/ ticket sales data by region/state/zip was not available
  + Performed an estimation using census and MPAA data as indicative. However, we could not run a correlation on this as we had one of the variables used was the population itself