BE 275: Final Project Proposal
Analysis of COVID-19 cases in Los Angeles county by Demographics

By: KaranDeep Cheema

Why the topic you chose is interesting

COVID-19 cases have disproportionally affected ethnic minorities in US [1]. I aim to study how valid that observation is in Los Angeles county. We need facts and data driven approach to understand health inequities in our system. With better understanding of the COVID-19 cases, we can specifically target the underserved communities.

Demonstrate that your project fits the criteria above

The above project fits the following description

"*Alternatively, you may identify a set of data and corresponding analysis that has not yet been performed. This can be more exploratory in nature. You will implement this analysis with testing and documentation. In your final report, you should discuss whether/how the type of data influenced the analysis that was possible, the findings and limitations of your analysis, and what might be ways to validate your findings.*"

I will acquire data from 3 different sources:

COVID-19 cases by community:
http://dashboard.publichealth.lacounty.gov/covid19_surveillance_dashboard/

Income statistics: http://www.laalmanac.com/employment/em12c.php

Demographics information: https://usc.data.socrata.com/stories/s/pd65-xuak/

The first dataset has 335 rows (each corresponding to a city/ neighborhood in LA county). The income statistics dataset has 285 rows, and the last dataset has 140k rows

What overall approach do you plan to take for the project and why

My plan can be summarized as:

1. Retrieve, check, and organize data (Right now, the 3 datasets do not have matching rows/ observations).
2. Visualize the data (using bar plots or scatter plots or histograms) and select appropriate analysis techniques (OLSR, PCA, PLSR, SVM).
3. Perform the intended analysis and report the results. Verify the results and note down observations.
4. Look for external validation and critically analyze the results to ensure they are correct and make sense.

The above 4 steps are necessary to ensure that the data are accurate, and the results generated make sense.

Demonstrate that your project can be finished within a month

Having identified the datasets, it should take me about a week to have a clean, organized and a coherent dataset that has the following form:

Neighborhood	COVID Cases per 1000 (needed as different neighborhoods have different populations)	Total COVID cases	Median household income	Median household family size
Neighborhood 1	###	####	$$$$	###

(For a total of ~300 rows)

Having done that, the analysis and validation should take a week. The last two weeks will be used for the writing and preparing the presentation and to correct any mistakes. Hence, this project is feasible and should be completed within a month.

Estimate the difficulty of your project

On a scale of 1-10 (with 1 being trivial and 10 being very challenging), I would give this project a 6. The complexity arises from the fact that external data need to be validated and properly organized. Furthermore, identifying assumptions, the correct model and this being an individual project add to the final score of 6.