DSCI 510 - Final Project

Kristian Castellanos
December 14, 2022

1. Title: **Uncovering and Defining the Relationship between Voter Turnout rates and Covid-19 Vaccination rates in the San Diego County**
    a. Description: The intention of this project is to uncover distinct relationships between Voter Turnout, as defined by the California Healthy Places Index (HPI) survey, and the Covid-19 Vaccination rate, defined by the number of people administered vaccines in a given census tract over the last two years; both datasets have been primed to the regions of San Diego county based on the available census tract information.

       This project seeks to answer the question: What is the correlation between Voter Turnout rates and Vaccination rates for the census tracts in San Diego county?

       Furthermore, the analyses provided hope to define, or "paint the picture", of the above relationship through statistical analysis methods and visual aids.

2. The code and analyses/visuals are provided in a Jupyter Notebook file (.pynb). The code may be run by selecting the "Run All" button located in the tool bar using Microsoft Visual Studio Code; otherwise, the file can be launched directly in Jupyter Notebook.

   Disclaimer: Encountered an issue executing the file using Jupyter Notebook; run individually from top to bottom if running in Jupyter. When using Visual Studio Code, the "Run All" button perfectly executes the script. (See **"Example Execution" video in folder)

   Each block can be executed individually, starting from the top-most block of code and working down to the bottom.

    a. Dependencies: see "requirements.txt" file located in the zip folder.

```
matplotlib          3.6.2
numpy               1.23.3
pandas              1.5.0
plotly              5.11.0
requests            2.28.1
session_info        1.0.0
-----
Click to view modules imported as dependencies


kaleido             0.2.1
nbformat            5.7.0
urllib3             1.26.12

-----
IPython             8.4.0
jupyter_client      7.3.5
jupyter_core        4.11.1
-----
Python 3.10.6 (tags/v3.10.6:9c7b4bd, Aug  1 2022, 21:53:49) [MSC v.1932 64 bit (AMD64)]
Windows-10-10.0.22621-SP0
-----
Session information updated at 2022-12-14 16:38
```

    i.
    b. To reproduce the same results, perform Cell-> Run All. The file was originally written in Visual Studio Code.
        i. A variety of dataframes will populate, along with the Plotly visuals toward the bottom of the notebook file.

c. Github Repository: kacastel/DSCI-510-Final-Project (github.com)

3. The data was collected from three separate sources using an API and key or by importing a file (dataset) hosted on an external server.
    a. Data Sources:
        i. County of San Diego - Vaccines by Census Tract
            1. **Method of Import: JSON File Import**
            2. This dataset has approximately 60k rows of data, yet was segmented for the purposes of this project. The majority of rows were reiterations from previous entries; only the useful entries were used in this project, which amounted to approximately 1000 entries. This data tracks Covid-19 vaccines administered to census tracts in San Diego County since 2020, incrementing on a daily basis for every census tract in the county (629 census tracts).
        ii. California Healthy Places Index - Voter Turnout Rate
            1. **Method of Import: API Access and Key, JSON Import**
            2. This dataset details indicators listed in the California Healthy Places Index (HPI) for all census tracts in California; the data was segmented to only utilize San Diego county's information, resulting in over 7k rows of data.
        iii. Master HPI data
            1. **Method of Import: CSV Import**
            2. This dataset fully encompasses all of the indicators, and their respected values, that are a part of the Healthy Places Index (HPI) survey. This dataset was received by special request directly from the California HPI agency, which includes over 8k rows of data. The file was received in tabular format (MS Excel).
            3. Importance: This dataset is needed to normalize the census tracts; it serves as an intermediary (foreign key) to both the Voter dataset and Vaccine dataset. It will allow the joining of the Voter data to the foreign data.
        iv. Approach: The method employed was to join all three datasets together on common census tracts (primary and foreign keys). Since all three datasets were either narrowed to or specific to San Diego county, shared census tracts would allow the unification between all three data sources. Below is a screenshot of the unified datasets:

	voting	voting_pctile	NAME	GEO_ID	county	pop	hpi_pctile	hpi_quartile	hpi_least_healthy_25pct	economic_pctile	census_tract	gis_hpi_quartile_sd	vaccination_percentage_all	Vaccine Quartile	geoid	name	population	value	percentile
0	0.924053	0.994865	1	6073000100	San Diego	3093	0.975995	4.0	No	0.971759	1	4	86.387274	Q4	6.073000e+09	1	3093.0	0.924053	0.994865
1	0.910670	0.977664	2.01	6073000201	San Diego	1891	0.884339	4.0	No	0.850578	2.01	4	76.140808	Q4	6.073000e+09	2.01	1891.0	0.910670	0.977664
2	0.910670	0.977664	2.01	6073000201	San Diego	1891	0.884339	4.0	No	0.850578	2.01	4	76.140808	Q4	6.073000e+09	2.01	1891.0	0.910670	0.977664
3	0.895648	0.927599	2.02	6073000202	San Diego	4542	0.783440	4.0	No	0.894737	2.02	4	69.598617	Q3	6.073000e+09	2.02	4542.0	0.895648	0.927599
4	0.895648	0.927599	2.02	6073000202	San Diego	4542	0.783440	4.0	No	0.894737	2.02	4	69.637027	Q3	6.073000e+09	2.02	4542.0	0.895648	0.927599
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
981	0.680180	0.187548	220	6073022000	San Diego	4681	0.230424	1.0	Yes	0.272144	220	1	83.682394	Q4	6.073022e+09	220	4681.0	0.680180	0.187548
982	0.680180	0.187548	220	6073022000	San Diego	4681	0.230424	1.0	Yes	0.272144	220	1	83.843498	Q4	6.073022e+09	220	4681.0	0.680180	0.187548
983	0.895080	0.925160	221	6073022100	San Diego	10005	0.791656	4.0	No	0.716431	221	4	79.802020	Q4	6.073022e+09	221	10005.0	0.895080	0.925160
984	0.895080	0.925160	221	6073022100	San Diego	10005	0.791656	4.0	No	0.716431	221	4	79.752025	Q4	6.073022e+09	221	10005.0	0.895080	0.925160
985	NaN	NaN	9901	6073990100	San Diego	0	NaN	NaN	NaN	NaN	9901	NaN	NaN	Unknown	NaN	NaN	NaN	NaN	NaN

    b. Changes from Original Plan, Challenges:
        i. The most deterring issue I faced was exporting the visuals to a pdf. I employed the use of the kaleido package to export a static visual, however the code would not execute after an initially successful run. I consulted Professor Gleb on the matter, yet I was still unable to correct the issue. I believe the code is correct, as it had previously executed.

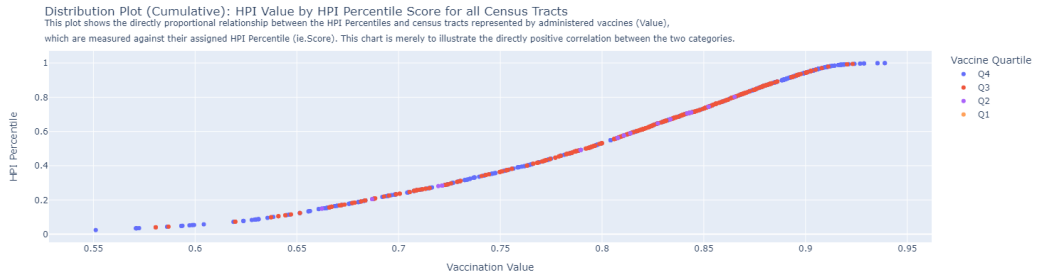
        Although my general plan of importing, joining, and visualizing the datasets went according to plan, my analysis did not reach the depths that I had originally intended. Additionally, I discovered that manipulating the data frame was more difficult than I originally expected, which introduced significant time constraints and methodical issues.

This very likely introduced lack of best practices when cleaning and manipulating data and producing the visuals.

The hope was to produce more statistical measures in my analyses instead of merely relying on visuals to define the relationship between Voter Turnout and Vaccination rates. Introducing a more reliable data science approach would be most beneficial to this project.

4. Visualizations: A variety of visualizations were made using the Plotly package: scatter plots, bar charts, pie chart, and a distribution curve. Each chart is axis-labeled and some have a legend section which is intended to provide further segmentation and context to the graphed data.
    a. Description of Figures - The figures below are only for reference. See the 'result' folder for full sized images (pdf)
        i. Distribution Plot (Cumulative): HPI Value by HPI Percentile Score for all Census Tracts

This plot shows the directly proportional relationship between the HPI Percentiles and census tracts represented by administered vaccines (Value), which are measured against their assigned HPI Percentile (ie.Score). This chart is merely to illustrate the directly positive correlation between the two categories.

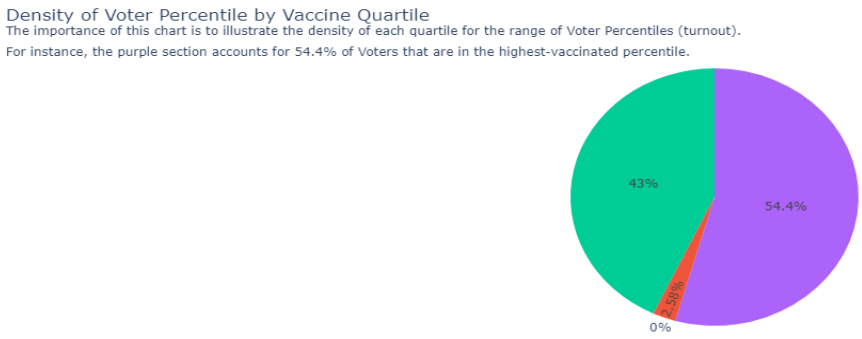


        ii. Count of Census Tracts by HPI Quartile

This chart provides a simple count of the number of census tracts for each HPI Quartile, which is an extension of the above scatter plot.

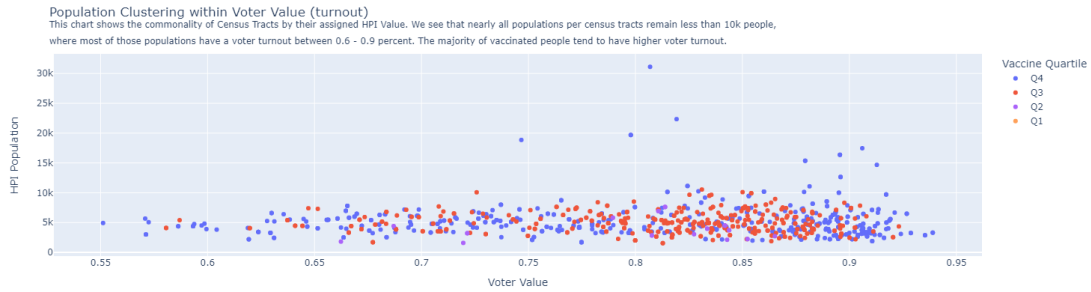
        iii. Density of Voter Percentile by Vaccine Quartile

The importance of this chart is to illustrate the density of each quartile for the range of Voter Percentiles (turnout success). For instance, the purple section attests to 54.4% of Voters that are in the highest-vaccinated percentile.



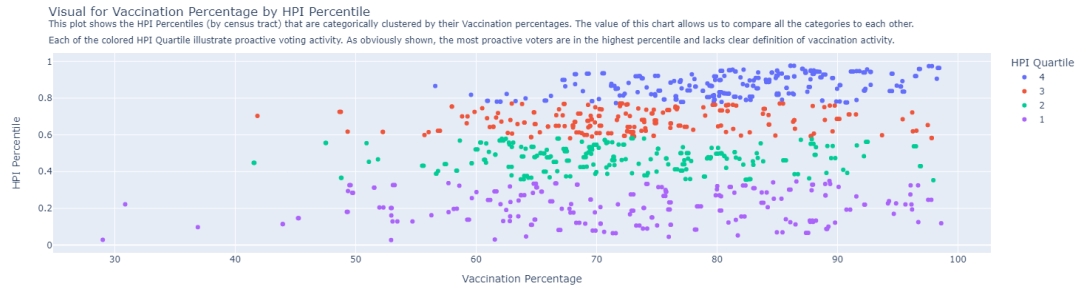
iv. <u>Population Clustering within Voter Value (turnout)</u>

This chart shows the commonality of Census Tracts by their assigned HPI Value. We see that nearly all populations per census tract remain less than 10k people, where most of those populations have a voter turnout between 0.6 - 0.9 percent. The majority of vaccinated people tend to have higher voter turnout.



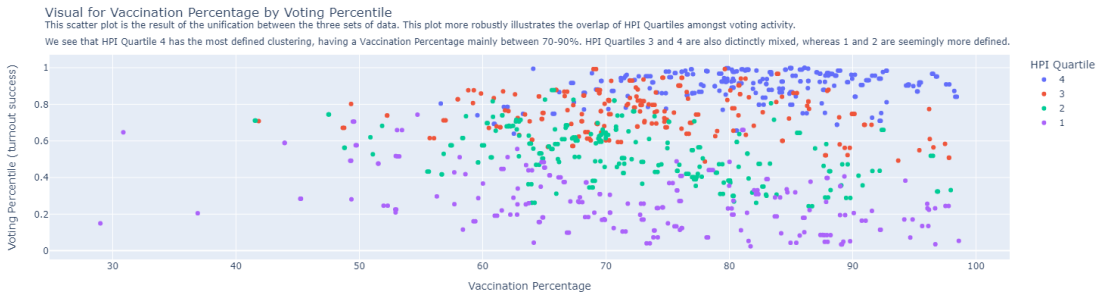
v. <u>Visual for Vaccination Percentage by HPI Percentile</u>

This plot shows the HPI Percentiles (by census tract) that are categorically clustered by their Vaccination percentages. The value of this chart allows us to compare all the categories to each other. Each of the colored HPI Quartiles illustrate proactive voting activity. As obviously shown, the most proactive voters are in the highest percentile and lack a clear definition of vaccination activity.



vi. Visual for Vaccination Percentage by Voting Percentile

This scatter plot is the result of the unification between the three sets of data. This plot more robustly illustrates the overlap of HPI Quartiles amongst voting activity. We see that HPI Quartile 4 has the most defined clustering, having a Vaccination Percentage mainly between 70-90%. HPI Quartiles 3 and 4 are also distinctly mixed, whereas 1 and 2 are seemingly more defined.



b. Observations, Conclusions, and Impact:

i. This research project yielded the following results: Census tracts that had Voter Percentile values above 0.7 coincided with higher Vaccination Percentages. These Vaccination Percentages were almost entirely captured in Vaccine Quartile 4. This indicates a stronger relationship between populations in census tracts with higher Voter turnout and Higher Vaccination rates.

This study also reveals that the number of census tracts in the bottom Vaccination Quartiles account for only 16% of measured tracts with populations that have been vaccinated. Furthermore, the Voting Quartiles 3 and 4 account for nearly 97% of voting activity for voters that have been vaccinated.

The impact of these findings support that there tends to be a positive correlation between populations that have both high voter turnout and higher vaccination rates. However, there are a multitude of factors that can contribute to higher voting activity that may exist completely independently of vaccination rates. This is where more research should be conducted. Although this project has begun the initial investigation into this relationship, more data is absolutely needed for a longer period of time and would better be supported for areas beyond San Diego County.

c. Changes and Challenges

i. Venturing beyond scatter, bar, and pie charts were a challenge. I believe this issue is limited to my abilities in manipulating the dataframes. In addition, presenting concise graphs introduced issues: scale of graphs, thoughtful labeling, and producing interactive images. Unfortunately, my final output was limited to status graphs, whereas they originally existed in an interactive style. The goal was to implement the interactive graphs to a website so that the interactive abilities would be preserved.

ii. Another deviation from the original plan was to incorporate the use of statistical measures, particularly with the scatter plots. This would be the next step in the project if

it were to persist. I believe it is crucial to define correlation based on mathematical models, not only visual models. This is where further work should be performed.

5. Future Work
    a. As mentioned above, future efforts on this project should implement stronger mathematical approaches in the analysis. This is assumingly more of a possibility given the programming and data manipulation ability of the researcher. Additionally, more data would be highly desired; this project only addresses the relationship between Voting and Vaccination Rates in San Diego county. The claim that Voting and Vaccination Rates are positively correlated would best be supported by measurement across the nation. Lastly, a multi-variable analysis would strengthen the findings and possibly provide a more detailed explanation for the influences between Voting and Vaccination Rates. Implementing these steps in the research is the direction I would take to improve the project.