Introduction

Our group started out with a interest in COVID Data and how different factors effected this. We hoped to determine which socioeconomic factors effected COVID outcomes and general information on COVID vaccines, testing and cases worldwide and per country. We knew this was too large and would have to decrease the scope of the project but we would proceed and get as much as we could done.

Project Scope

Our group was interested in COVID-19 data and understanding COVID trends on an international level. As such, we had a few precursory questions that drove our data selection:

The relationship between vaccination and covid case growth

The relationship between testing and covid case growth

What factors, if any, are leading indicators of higher COVID rates?

Which countries are able to ramp up their vaccination efforts the fastest? What factors contribute to a country's vaccine implementation?

Are certain vaccines more effective in curbing COVID spread? On a more general level, does vaccination rate impact COVID infection rate? What is the respective time lag?

Scrum Roles

Project Manager - Matthew Sachs

Software Architect - Thomas Butler

Quality Assurance Manager - Karan Manwani

User Interface Designer - Matthew Sachs

A Project Manager makes sure every aspect of the design will fit into our primary goal. Matthew's job was to keep us on track and make sure we weren't straying from our project scope.

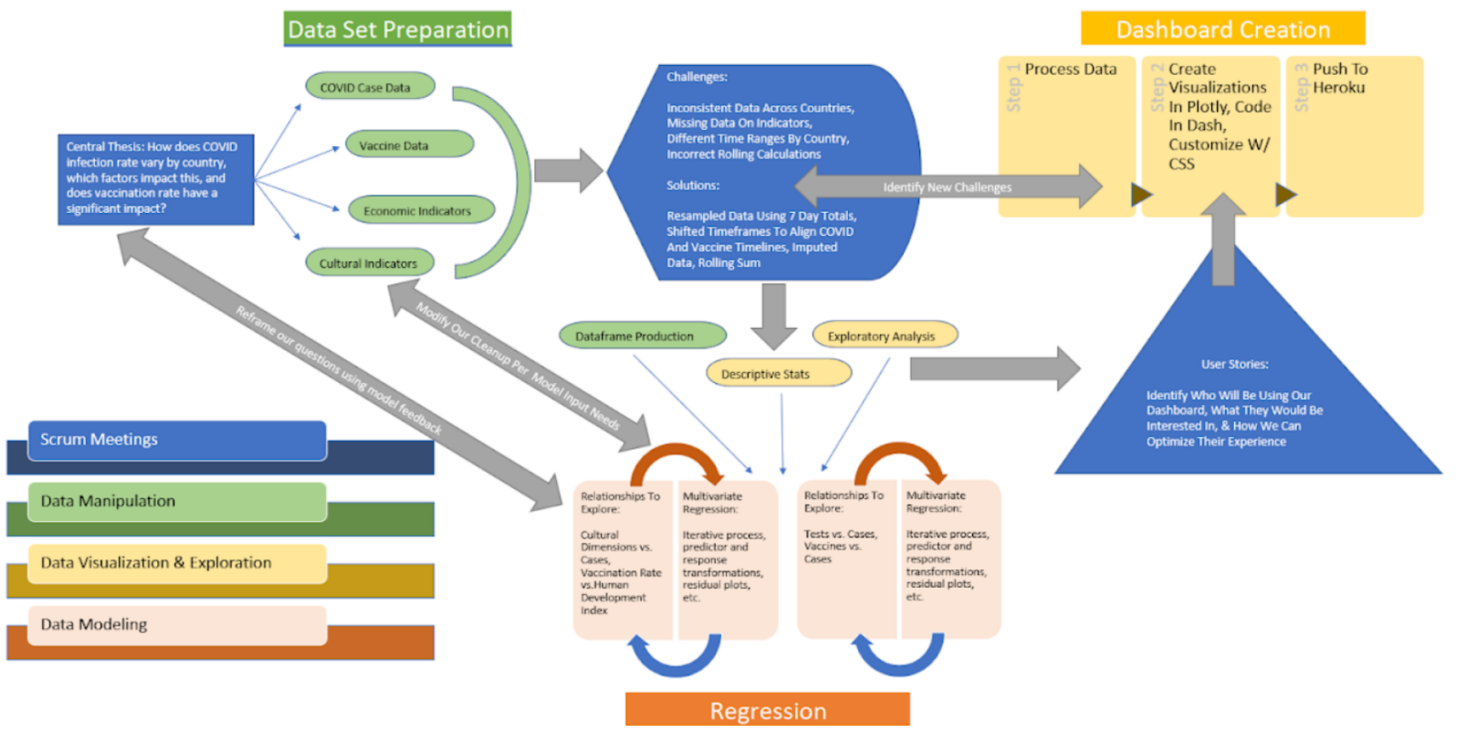
A Software Architect makes sure the idea we think of is actually possible to code. Thomas's job was to make sure everything was possible to code and help those who were having trouble or come up with the code framework for the problem if they were off course.

A Quality Assurance Manager makes sure the code is robust, there job is to make sure all potential aspects of the code will not cause errors for our intended use case. Karan ensured we had unit testing and helped us understand if a particular part of our code wasn't coded in a proper manner.

A User Interface Designer makes sure the results we come to is shown in a easy to understand way for our client. This role makes it possible for other people to understand our work without looking through our code and allows those without the required background knowledge to understand what we did. Matthew's job was to make sure all our results were presented in a understandable way for our client.

Even though each of the group members were assigned or fell into these roles, every scrum role was filled by every group member at some time during the project.

Project Workflow



Our project workflow was very consistent. We broke everything up into pieces. Our first meeting was determining what we had to do to finish the project and dividing the work as evenly as possible for the rest of our project. Every meeting we used our plan from the first meeting and determined more details who would work on what so we could show what we did for that week. An example of this is the first meeting we decided on our main topic and our job was to find datasets and go think about what questions we wanted answered on those datasets. We should post what we find on slack and then we would discuss which datasets we wanted to use and which questions we wanted to answer the next week.

Project Management

Our project management was all over the place. We started out with a google document detailing our ideas and the datasets we were using. Eventually we exclusively moved over to a group chat in Slack where we would post our code various ideas and talk about our meetings. Right at the end of our project we set up a Github project. The way Github tracks changes would have been very useful to share code and see various changes in our code. If I was going to redo this project right now I would set up a Github project for all of our work to go on and use Slack for communication through the team. This would make us much more efficient.

The Data

Describe your data set and its significance. Where did you obtain this data set from? Why did you choose the data set that you did? Indicate if you carried out any preprocessing/data cleaning/outlier removal, and so on to sanitize your data.

Vaccine Data

The crux of our project resides with accurate COVID-19 vaccine data. While it is difficult to believe in the veracity of data reported by countries with low transparency scores (see https://www.transparency.org/en/cpi/2020/index/nzl), we arrived at the conclusion that we could only do so much to clean up the data; the majority of reported data needs to be taken at face value. For this reason, further analysis likely warrants grouping customers by transparency score (ie: high trapnsparency, average transparency, and low transparency) and evaluating COVID-related numbers through this lens.

We gathered our vaccine data from Kaggle, https://www.kaggle.com/gpreda/covid-world-vaccination-progress. The dataset is updated daily with the latest reported numbers and contains 15 fields, largely representing one of four categories: country level, vaccine level, date level, and source-related information. The country level data encompasses the country name and country code, which are of important note because these fields serve as the connective tissue between our datasets. Vaccine data includes the raw number of vaccines given on a particular day, the raw number of vaccinated people on a particular day, smoothed vaccination numbers, adjusted vaccination numbers, and vaccine types. There are two glaring omissions from these fields: partial vaccinations given and vaccination rate. Both of these metrics can be calculated after bringing in population data, but we still need to account for missing data and incorrectly calculated rolling sum columns.

161 countries are represented in the data, and the date ranges (from when we downloaded the data) are from 12/13/2020, which is the earliest vaccination implementation for any country, to 3/30/2020. This gives us a best-case scenario of 3.5 months of vaccination data. The daily numbers fluctuate wildly from country to country, and the standard deviation of daily vaccinations across all the countries is 244,102. This is a strong indicator of a) the inherent reporting inconsistencies and b) the disparity in vaccination implementations between countries.

COVID Case Data

To supplement our vaccine data, we needed the actual COVID-19 case data as well; this is updated daily and can be downloaded from https://ourworldindata.org/covid-cases?country=IND~USA~GBR~CAN~DEU~FRA. Spanning 59 fields and 215 countries, a couple take-aways are immediately apparent: there are quite a few countries which have reported COVID 19 cases and are yet to implement a vaccine program, and the breadth of data here indicates the likelihood of non-COVID fields. Sure enough, there are a number of potential infection predictors, including but not limited to: population size, median age, and average life expectancy. Predictors that are directly related to COVID's spread and potency, like cardiovascular deaths, handwashing facilities, and prevalence of diabetes, complement our COVID-specific metrics like daily cases reported, daily deaths reported, and daily tests conducted. Unfortunantely, a number of fields, like patients in ICU, number of smokers, etc, have a paucity of countries represented and, as a result, can only be effectively modeled for a subset of countries.

The data has observations from as early as 1/1/2020, which is nearly a full year's worth of data more than our vaccine data. Assuming the dataset's reported numbers are roughly as accurate as our vaccine numbers, we would also assume that the increased number of observations would help us to train our models more effectively. As far as the reporting inconsistencies we saw in the vaccine dataset, the COVID case data has less missing data in the COVID-related numbers and the daily observation should prove more reliable.

Cultural Dimension Data

Geert Hofstede is a social scientist who rates countries on 0-100 scale on 6 cultural dimensions: power distance, individualism, masculinity, uncertainty avoidance, long-term orientation, and indulgence. Each of these dimensions warrants explanation and can offer insight into any potential vaccination program.

Per their website, https://hi.hofstede-insights.com/national-culture, power distance "expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally". A country with a high power distance score is representative of a prevalent culture where hierarchical order is accepted and social inequalities are not questioned. When we consider this in a COVID context, countries with high power distance scores might roll out vaccine programs early to a select few, prior to ramping up vaccination for the rest of the country. Similary, a high power distance country might only hospitalize certain individuals and provide more promising medical treatments to a select few; a higher mortality rate may, in fact, be expected for these countries and should not be internally considered as a cause for alarm.

Individualism is defined as the "preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families". High-scoring countries on this dimension follow an "I" paradigm, whereas low-scoring countries follow a "we" paradigm and expect to be taken care of by relevant in-groups. Interpreting this in a COVID context, there are two low-hanging hypotheses with very different consequences: 1) individualist countries are likely to be more self-centered and less likely to follow policies designed to mitigate COVID's spread and 2) individualist countries are more likely to self-isolate and forego collective activities, enabling a given country to slow COVID's spread. Obviously, the implications of these two hypotheses are polar opposites and they require further exploration.

Masculinity is best interpreted for the lack thereof; low-scoring countries (ie: countries with high feminity) stand "for a preference for cooperation, modesty, caring for the weak and quality of life". As such, low-scoring countries on this dimension should be expected to have high hospitalization rates and public policies designed to limit COVID exposure for more susceptible demographic groups.

Uncertainty avoidance "expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity". Countries with higher uncertainty avoidance scores stringently follow codes of behavior and belief, and are less likely to deviate from these pre-conceived notions. In a COVID context, we could expect countries with higher scores to cling to established medical protocols and be less likely to adopt innovate mitigation policies, try new vaccinations, and ultimately treat the pandemic as seriously as they should. Subsequently, we would not be surprised to see higher infection rates in these countries.

Countries high in long-term orientation "encourage thrift and efforts in modern education as a way to prepare for the future", whereas low-scoring countries are more adherent to established social norms. We would expect long-term oriented countries to have more robust responses to the COVID pandemic, and low-scoring ones to be slow to adopt new vaccines as they come out.

Indulgence "stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun", and the COVID implications here are fairly self-evident. A country comprised of highly indulgent individuals would have a hard time convincing individuals to forego their immediate self-gratification and act in the better interest of society. We might see relatively low infection rates for a while, but as COVID continues to drag along, the self-restraint of individuals in these countries is likely to decline and we should expect to see jumps in infection rate after a certain time. The question becomes: how can we translate an indulgent score to a tangible time threshold where indulgence overcomes public policy?

We downloaded this data from https://geerthofstede.com/research-and-vsm/dimension-data-matrix/. There are some noteworthy characteristics of the data:

1. The sociological nature of the analysis means this data was collected over many years. Culture is not static, and immigration patterns and globalism can significantly alter these cultural scores over time. Unfortunately, we are only given one data point per country and we must come to terms with the fact that are current scores aren't very accurate, and that we lose the dynanism aspect inherent to culture.
2. There are countries where Hofstede has not scored all of the dimensions, so there are gaps with missing data. Non-country entities, like regions and intra-country cultural divides, are also included. Countries and country codes are included as well, which in theory will allow us to merge this to our other datasets.

Economic Indicator Data

To understand a country's access to medical supplies, existing infrastructure for vaccine distribution, and (generally speaking) financial heft, we recognized that economic indicators would be an effective proxy. We went to the World Bank, https://data.worldbank.org/topic/economy-and-growth, and found a comprehensive dataset. There are quite a few economic indicators (245) included here, complemented with country and country code designations (264 represented in total). The sheer breadth of economic indicators was quite intimidating, especially since we didn't yet have exposure to shrinkage techniques, factor analysis, or partial components analysis.

The indicators have observations in the year range of 1960-2020, but countries were poorly represented across some metrics, and there were gaps as to which years had reported values. The metric frequency was a deciding factor when delineating highly similar metrics (for example, we have GDP in US dollars, GDP per capita, GDP adjusted for purchasing power parity).

The data itself was in an odd wide-long mesh, meaning it did not follow a typical wide or long format. Every represented year had its own column, but all indicators were lumped into one column. Each metric value was included in the year column. This format required significant reshaping if we wanted to include it on our analysis.

Experimental Design

Describe briefly your process, starting from where you obtained your data all the way to means of obtaining results/output.

Vaccine Clean-Up

With the vaccine data, using the head() function gave us a precursory look into what the data looked like, while the info() function highlighted the data type and non-null counts. We decided to immediately drop the source-related data from our dataset and noticed that, beyond the general missing data, there was substantially more missing data when it came to fields relating to full vaccinations. This can be due to the inherent time lag between initial vaccination or the data not being recorded in some countries; to verify this, we took a vaccine that we knew was one shot, Johnson & Johnson, and counted how many instances there were of the "people_fully_vaccinated" being populated and "total_vaccinations" was not. Since the returned count was 0, we could conclude that any variance in null counts was due to the time lag.

We continued the clean-up by field. All observations with a null value in the country field were dropped. The date field was converted to a datetime object and, along with country, was used to re-sort the dataframe. Since we were new to pandas at the time of this script, we created our own for loop to generate lagged values. There are a number of reprecussions handling the data this way, pending the field it is applied to, but it works well for our applications. We used the newly configured "total_vaccinations" field to calculate a "Daily Change" field with the diff() function.

Most importantly, we decided on a resampling period. By using the cumsum() function, we were able to create a new "Day Count" field by country. Then we divided the "Day Count" field by 7 (rounding up), to generate a vaccination program week number for each observation point. We needed to resample the data due to reporting inconsistencies (there were instances where data would not be reported for days at a time), but did not have sufficient data for a 15-day or 30-day resampling window. We had vaccination programs for 161 programs, and with the resampling, the average country had a program in place for 7.6 weeks; the longest program had been in place for 15 weeks at the time of analysis, a distinction shared by a handful of countries including Canada, the UK, and Russia.

Lastly, we aggregated our fields of interest by country and our newly generated week field. Our end result was a dataset with country, vaccine information, vaccination week number, total vaccinations given over the period, people vaccinated over the period, and full vaccinations given over the period.

COVID Case Clean-Up

For our case data, we first converted the date field to a date time object. Then, we grouped our data by country and counted the non-null values present by field. Looking at the aggregated data, we could see which countries had a scarcity of reported observations and which COVID predictors were not well represented. We selected predictors and demographic variables that had wide representation and dropped the rest; leaving us with 'continent', 'location', 'date', 'total_cases', 'new_cases', 'total_deaths', 'new_deaths', 'icu_patients', 'hosp_patients', 'new_tests', 'total_tests', 'population', 'population_density', 'median_age', 'aged_65_older', 'cardiovasc_death_rate', 'diabetes_prevalence', 'handwashing_facilities', 'life_expectancy', and 'human_development_index'.

The next step was making sure that our case data aligned with our vaccine data from a date standpoint. We calculated the last reported vaccination date by country, and dropped any reported case dates greater than this. Our time series data was now parallel, even if there were different begin dates for each country. To calculate the COVID case week number, we used the same methodology we used in our vaccine dataset: we created a day counter by country, divided by 7, and rounded up. When trying to merge this with our vaccination data by date, and then attempting to aggregate our data by either vaccination program or COVID case week, testing revealed that our weeks were getting duplicated; further inspection revealed that although our dates aligned, a COVID case week could include two different vaccine weeks, and vice-a-versa, due to non-uniform starting dates.

To resolve this, we re-sequenced the vaccine weeks so their start date wasn't the beginning of the vaccination program, but rather the closest date to the vaccination program start that coincided with the start of a COVID case week. This erased duplicate weeks upon aggregation and put our COVID datasets on parallel weekly timeframe. We were then able to replace our missing data with 0's, where applicable.

Cultural Dimension Clean-Up

The cultural dimension data required us to identify which country names and codes didn’t align with our other data. To do so, we merged the cultural data with a subset of our economic data, "Country Lookup", using a left join. Any country that did not join would need to be renamed in our cultural dataset, of which there were 15 total countries. Our dataset had 111 countries, some of which weren't actual countries, and of these only 65 were complete observations--meaning we had numerous observations with missing data.

In order to keep as many countries as possible, we decided to deal with our missing data with imputation. We grouped our countries by region, for each cultural dimension, and calculated the median. We then backfilled our missing data with the relevant region's median dimensional value. By doing so, we kept all 111 countries. The downside of this imputation is that we are assuming cultural similarities across region, which, given the prior discussion on cultural dynanism, is overly simplistic; this could impact the strength of future models.

Economic Indicator Clean-Up

The melt() function was pivotal in reorganizing this dataset--our dataframe was now in long format and we could proceed to inspect indicator coverage by country as well as an indicator's relative recency. The aggregated dataframe contained the following: 1) a given indicator, 2) the number of countries with some value for given indicator, 3) the average year for the indicator's most recent data by country, and 4) the most out to date observation for a given indicator. We were effectively able to eliminate indicators with poor country coverage and outdated data, and we were then able to qualitatively identify proxies for a country's access to medical supplies, existing infrastructure for vaccine distribution, and financial heft. We eventually settled on GDP per capita, foreign direct investment, education expenses, manufacturing as a percentage of GDP, and services as a percentage of GDP.

Merging Datasets

With the countries renamed in our cultural dimension dataset, it was fairly straightforward to merge our economic indicators to cultural dimensions by country; the fact that there was only one observation per country made this much easier. We were then able to merge this to our COVID data, which joined vaccine and COVID case data by country and case week. From a modeling standpoint, this means the stagnant nature of our socioeconomic and cultural standpoint is useless for time series (it doesn't change); these predictors can only be helpful in attempting to predict infection, mortality, or vaccination rate for a given country at a given week in time (relative to that country's own timeframe).

Testing

Our testing at this stage was fairly limited in scope, considering the extensive clean-up we performed, but was supplemented by testing in the modeling stage--this feedback helped implement changes as needed. The unit testing here is as follows:

1. Are there duplicate case weeks for a given country? We did not use an actual unit test for this, but used the following in console. It should return 0 for our final dataset: new_cases_df.duplicated(subset=['location','Case Week'])
2. Are there no null values in columns that should have been back-filled? We created a function to test this as needed: def na_counter(df, col): return df[df[col].isna()].count()
3. Do the number of countries represented match our expectations? We used unit testing for this, with the expectation of failure. We then had to assess which countries dropped and why.

After multiple rounds of testing, the datasets we produced were able to meet our expectations.

Beyond the original specifications

Highlight clearly what things you did that went beyond the original specifications. That is, discuss what you implemented that would count toward the extra-credit portion of this project (see section below).

User Stories

Prior to any "product development", we identified user stories that would guide our choices and understanding of the task at hand. We came up with four user stories:

1. As a “concerned citizen”, I want to understand how my country’s COVID-19 efforts compare to other countries’ so that I can see how different policy decisions affect COVID spread.
2. As a “government official”, I want to see how country COVID metrics differ by vaccine type, because this might impact the vaccine manufacturer I pursue business with.
3. As a “concerned citizen”, I want to know when there are spikes in COVID cases, because if I can understand when there might be dangerous exposure to the virus, I can opt to self-quarantine.
4. As a “vaccine manufacturer”, I want to know which countries have low vaccination rates because there might be an opportunity for me to step in and increase market share.

Requirements

The requirements for our end product can be divided between functional and non-functional. We used these to design our product

Functional requirements:

1. Provide the dynamic capability to make 4 queries against our dataset.
2. Provide overview of COVID metrics by country.
3. Provide ability to explore COVID efforts over time.
4. Provide method to group countries and compare COVID metrics to similar countries.

5. Provide capability to visualize relationships between different predictors and COVID metrics.
6. Provide access to users who do not have Python.

Non-functional requirements:

1. House functional requirements in an interactive dashboard.
2. Host dashboard on a URL so multiple users can engage and access from any device.
3. Make sure the data can be drilled into and has clear, understandable hover labels.
4. Verify that chart controls and charts are grouped together.

COVID Dashboard

To meet our product requirements, we built a COVID-19 dashboard using the "dash" and "plotly" modules. The dasboard pulls from our cleaned COVID datasets, performs some data restructuring as needed, accommodates layout preferences via HTML and CSS, and then provides user interaction via functions in the dash module. In order to have the dashboard hosted on a URL, we needed to push our code and source files to Heroku, which provides a free server to host our dashboard. The entire scope of this endeavor goes beyond the initial project requirements and, we believe, should be considered extra credit.

The dashboard can be visited here https://covid-dashboard-msachs.herokuapp.com/. Per our functional requirements, we have distinct "queries" to explore:

1. The first chart grants the user the option to choose between two map views--"cases" and "vaccines". The first view produces a heat map by infection rate, where countries with higher infection rates appearing more red and countries with lower infection rates appearing more blue. The second view produces a heat map vaccination rate (either full or partial), with a similar color scale. The user can quickly see abnormalities; for example, Mexico's positivity rate is incredibly high, but the overall infection rate is extremely low. Interpreting this in the context of other countries, it seems that tests are only administered to those with the most severe symptoms and there is a strong chance that the reported number of COVID cases in Mexico is lower than the true number.
2. The second chart allows the user to explore time series data for a specific country. Four metrics are presented over time: new tests, new cases, new deaths, and new vaccines. It should be noted that the time series window here is daily, as opposed to an aggregated 7 day total, so the user can see for themselves any reporting inconsistencies. Looking at the United States, you can see a slow decline in daily reported COVID cases once the vaccine program started to ramp up.
3. The third chart is a little more complex. The user first selects a COVID-specific metric they wish to explore, then selects a categorical variable to explore the data by (continent, region, income group, or vaccine type). The user then has the capability to explore the different levels for the categorical variable and see how various countries meeting the chose criteria perform as it relates to COVID.
4. The last chart is a simple correlation heat map.The user can hover over specific colors and readily see which variables have high correlations with specific COVID metrics.

Testing

The testing performed here was not via unit testing, but rather an iterative exploration of different user actions within the dashboard. Below is a list of some of the tests performed, along with the corrective action where appropriate:

1. Did the dashboard work on multiple devices? It works on laptop and mobile, though the mobile experience is admittedly less than ideal. No corrective action was taken at this time, but we should explore a more robust formatting for mobile.
2. Is the formatting consistent and are the chart controls appropriately paired with their respective charts? Our product failed this test multiple times, and we had to become much more familiar with CSS in order to get the formatting correct. Pairing the chart controls and charts was corrected once we understood the ordering of the HTML containers within the dash framework.
3. Do the map views toggle easily between each other? Our product failed this test, as there is considerable lag when switching from the "cases" to "vaccines" view. Even when the re-rendering has completed, there are issues with the hovering abilities and there seems to be a lag between user action and hover text. We believe this has to with how we handled our if-else statement for this chart; ideally, our statement will prevent the graph from re-rendering, it just re-plots the new data. Further investigation is needed.
4. Do the four COVID metrics appear accurately in the second chart? Again, our intital tests failed. Further exploration revealed that this was because our testing rate was not being calculated correctly; a simple modification of the script, outside of the dash application, corrected this.
5. Does the third chart accurately respond to the selected levels? The testing here went well until we looked at segmenting countries by vaccine type. If a country utilized multiple vaccines, they would be listed in string format in their respective data column. We were able to enforce accuracy here by comparing the user's level selection to the respective column with the str.contain() function.
6. Is the last chart readable and easy to navigate? The first iterations had cut-off labels and was very difficult to parse; modifying the format in CSS made it much easier for the user to interpret.

Results

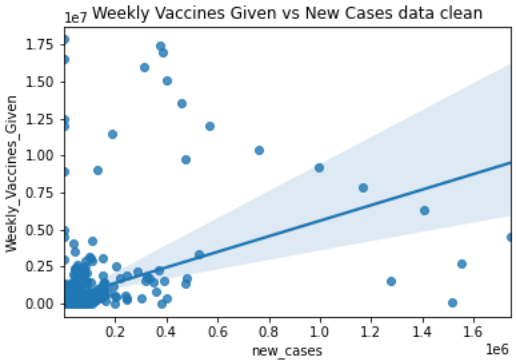
Display and discuss the results. Describe what you have learned and mention the relevance/significance of the results you have obtained.

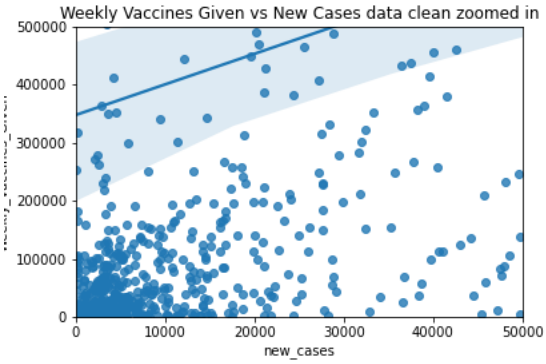
Hypothesis 1: COVID Vaccinations vs. Lagged COVID Cases

This model intents to answer the relationship between COVID Case growth and COVID vaccination worldwide. Code for model can be found https://github.com/Toble007/COVID-19_CS_5010_Project/blob/main/Code/Covid_Vaccines_Tests_Cases_Regression_Code_and_Unit_Testing.py in the 'Vaccine v cases regression cell', testing is found inside the 'testing' cell and the data is loaded in inside the 'data load in and test v cases regression' cell. My code can also be found https://github.com/Toble007/COVID-19_CS_5010_Project/blob/main/Code/Covid_Vaccines_Tests_Cases_Regression_Code_and_Unit_Testing.ipynb and https://colab.research.google.com/drive/1tt usp=sharing

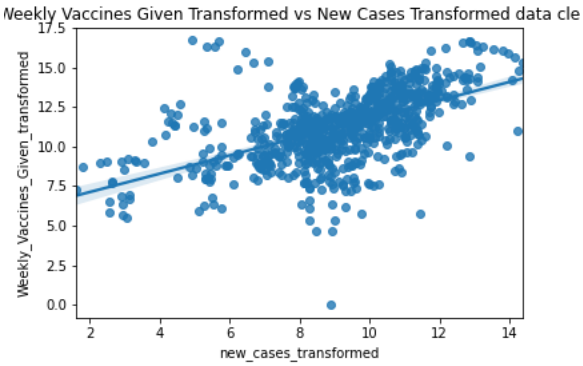
Model Preparation

The COVID Cases and COVID Vaccines datasets were used in this model. All data points that did not have Vaccine happening during that week were removed. All zero value data points for Weekly Vaccines Given and new COVID cases were removed. Any country that was vaccinating but not reporting any covid cases were removed. Any country that didn't have at least 3 weeks of vaccine data was removed. The reason why any country that didn't have at least 3 weeks of vaccine data were removed was because they were not actively vaccinating. Alot of countries had vaccines donated to them from various sources and either can't afford and/or get access to more vaccines.





Above is the graph for the linear relationshp between COVID Vaccines and COVID Cases. The data is not a linear function so a log-log transformation was applied to the data.

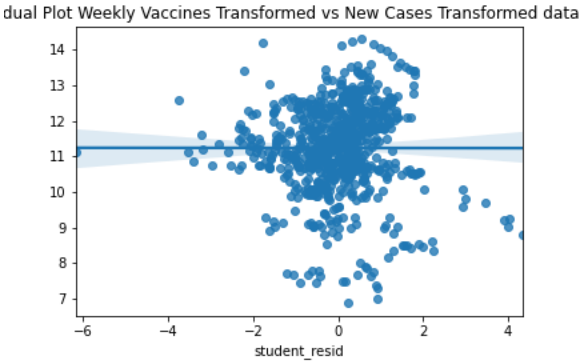


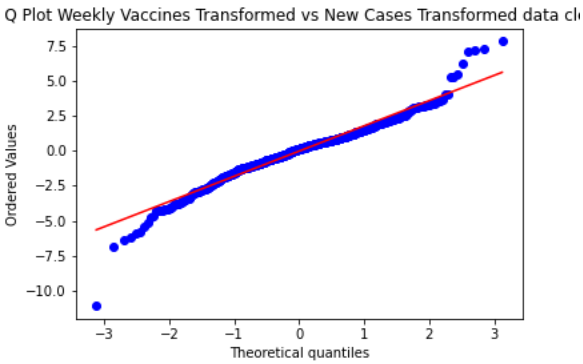
The transformed data looks more linear now so proceeded to model fitting.

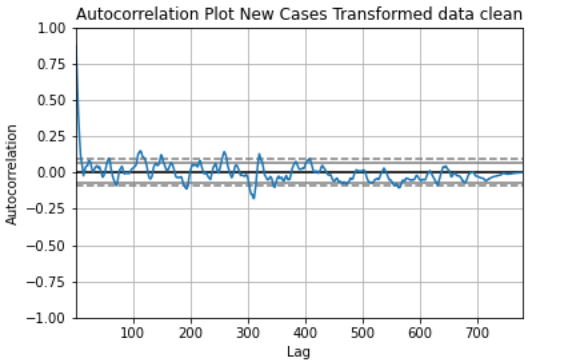
Model Description

Weekly Vaccines vs New COVID Cases were fitted with a simple linear regression model. Output and graphs are below.

```
                                OLS Regression Results
==============================================================================================
Dep. Variable:     Weekly_Vaccines_Given_transformed   R-squared:                       0.308
Model:                                         OLS   Adj. R-squared:                  0.307
Method:                              Least Squares   F-statistic:                     345.3
Date:                             Sat, 08 May 2021   Prob (F-statistic):           4.92e-64
Time:                                     17:44:21   Log-Likelihood:                -1579.9
No. Observations:                              778   AIC:                             3164.
Df Residuals:                                  776   BIC:                             3173.
Df Model:                                        1
Covariance Type:                         nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                          5.9545      0.292     20.404      0.000       5.382       6.527
new_cases_transformed          0.5799      0.031     18.581      0.000       0.519       0.641
==============================================================================
Omnibus:                       77.660   Durbin-Watson:                   0.660
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              360.955
Skew:                          -0.315   Prob(JB):                     4.17e-79
Kurtosis:                       6.277   Cond. No.                         41.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```







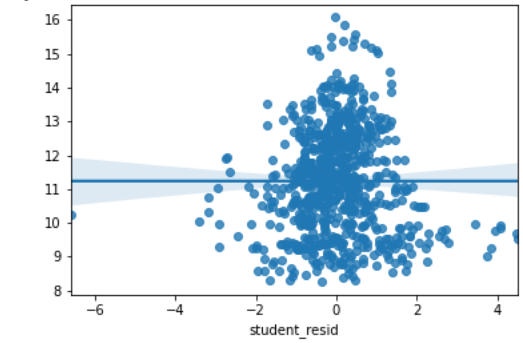
There are some issues with the model. The residual plot shows a nonconstant variance and non-zero mean. The QQ Plot has a light-tailed distribution. The Autocorrelation plot shows that an autocorrelation issue exists. The nonconstant variance and non-zero mean means that there is probably a issue with my model. When I made these models I didn't know how to fix these issues. So I tried fitting more variables to the model in hopes it would solve some of the issues.

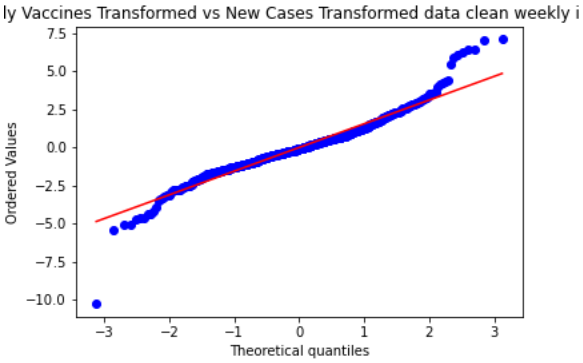
```
                                OLS Regression Results
==============================================================================================
Dep. Variable:     Weekly_Vaccines_Given_transformed   R-squared:                       0.474
Model:                                           OLS   Adj. R-squared:                  0.468
Method:                                Least Squares   F-statistic:                     69.23
Date:                               Sat, 08 May 2021   Prob (F-statistic):          3.47e-100
Time:                                       17:44:25   Log-Likelihood:                -1472.9
No. Observations:                                778   AIC:                             2968.
Df Residuals:                                    767   BIC:                             3019.
Df Model:                                         10
Covariance Type:                           nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                          6.3033      0.258     24.466      0.000       5.798       6.809
new_cases_transformed          0.3268      0.034      9.710      0.000       0.261       0.393
Week_1                         0.1147      0.027      4.179      0.000       0.061       0.169
Week_2                         0.0527      0.027      1.919      0.055      -0.001       0.107
Week_3                         0.0399      0.027      1.452      0.147      -0.014       0.094
Week_4                         0.0273      0.028      0.972      0.331      -0.028       0.082
Week_5                         0.0163      0.029      0.557      0.578      -0.041       0.074
Week_6                         0.0157      0.031      0.514      0.607      -0.044       0.076
Week_7                         0.0035      0.032      0.111      0.912      -0.058       0.065
Week_8                         0.0415      0.033      1.276      0.202      -0.022       0.105
Week_9                         0.0084      0.027      0.309      0.757      -0.045       0.062
==============================================================================
Omnibus:                       88.446   Durbin-Watson:                   0.498
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              644.321
Skew:                           0.166   Prob(JB):                    1.22e-140
Kurtosis:                       7.446   Cond. No.                         89.8
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Ran this and check for significant terms and then interactive terms. All the additional terms are already log transformed.

```
                                OLS Regression Results
==============================================================================================
Dep. Variable:     Weekly_Vaccines_Given_transformed   R-squared:                       0.484
Model:                                           OLS   Adj. R-squared:                  0.481
Method:                                Least Squares   F-statistic:                     144.9
Date:                               Sat, 08 May 2021   Prob (F-statistic):          2.08e-108
Time:                                       17:44:25   Log-Likelihood:                -1465.6
No. Observations:                                778   AIC:                             2943.
Df Residuals:                                    772   BIC:                             2971.
Df Model:                                          5
Covariance Type:                           nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                      8.2904      0.465     17.812      0.000       7.377       9.204
new_cases_transformed          0.1139      0.053      2.133      0.033       0.009       0.219
Week_1                        -0.2062      0.070     -2.928      0.004      -0.344      -0.068
new_cases_transformed:Week_1   0.0355      0.007      5.229      0.000       0.022       0.049
Week_3                         0.0884      0.018      5.014      0.000       0.054       0.123
Week_6                         0.0688      0.015      4.667      0.000       0.040       0.098
==============================================================================
Omnibus:                       90.372   Durbin-Watson:                   0.576
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              641.355
Skew:                           0.211   Prob(JB):                    5.39e-140
Kurtosis:                       7.428   Cond. No.                         749.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

ekly Vaccines Transformed vs New Cases Transformed data clean week





This is the model I ended up with. My issues from before are still there but the residual plot looks better and the autocorrelation issue might have been fixed, I just had no clue how to check. From here I didn't know how to improve the model further.

Model Interpretation

```
                                OLS Regression Results
==============================================================================================
Dep. Variable:     Weekly_Vaccines_Given_transformed   R-squared:                       0.484
Model:                                          OLS   Adj. R-squared:                  0.481
Method:                               Least Squares   F-statistic:                     144.9
Date:                              Sat, 08 May 2021   Prob (F-statistic):          2.08e-108
Time:                                      17:44:25   Log-Likelihood:                -1465.6
No. Observations:                               778   AIC:                             2943.
Df Residuals:                                   772   BIC:                             2971.
Df Model:                                         5
Covariance Type:                          nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                      8.2904      0.465     17.812      0.000       7.377       9.204
new_cases_transformed          0.1139      0.053      2.133      0.033       0.009       0.219
Week_1                        -0.2062      0.070     -2.928      0.004      -0.344      -0.068
new_cases_transformed:Week_1   0.0355      0.007      5.229      0.000       0.022       0.049
Week_3                         0.0884      0.018      5.014      0.000       0.054       0.123
Week_6                         0.0688      0.015      4.667      0.000       0.040       0.098
==============================================================================
Omnibus:                       90.372   Durbin-Watson:                   0.576
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              641.355
Skew:                           0.211   Prob(JB):                    5.39e-140
Kurtosis:                       7.428   Cond. No.                          749.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

A 10% increase in new cases for current week leads to a 1.1% increase in Vaccines for the current week.

The Week 1 coefficient is negative but it has a interactive term and the break even point which causes more vaccines is at least 334 cases on new cases. To come to this number take the absolute value of the coefficient Week_1 / the interactive term (new cases transformed:Week 1) and set that equal to ln(new cases). Solve that equation. Math is shown below.

```
import math

x = 0.2062/0.0355

math.exp(x)
333.10265116201714
```

A 10% increase in cases from 3 weeks ago leads to a 0.85% increase in Vaccines for the current week.

A 10% increase in cases from 6 weeks ago leads to a 0.66% increase in Vaccines for the current week.

Testing

All columns used in this model were tested to ensure there were no NA values in them and that all the values inside the columns were greater than zero. All countries were tested to ensure they held at least 3 weeks of vaccine data. Created columns for Week 1 to Week 9 were tested to insure they held the correct values.

Improvements

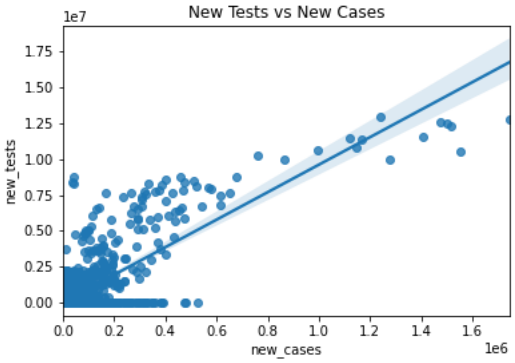
Starting out our group asked this question in a way that we didn't know how to answer. The question was reworded a few times but we never got a perfect answer. If we had more time we would have figured out how to create an ARIMA (auto regressive integrated moving average) model to answer this question. Originally we didn't know what a autoregressive model or partial autocorrelation function (PACF) was and would now use those to sovle the time lag issue present in the model.

Hypothesis 2: COVID Testing Rates vs COVID Cases

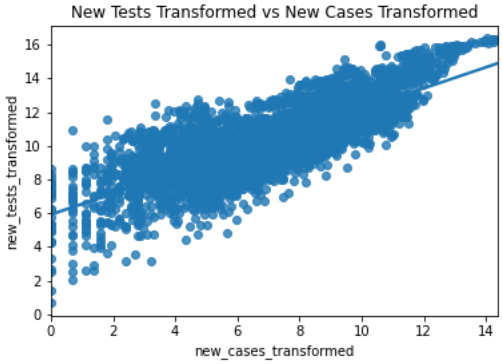
This model intents to answer what the relationship between COVID Case growth and COIVD Testing is worldwide. Code for model and data load in can be found https://github.com/Toble007/COVID-19_CS_5010_Project/blob/main/Code/Covid_Vaccines_Tests_Cases_Regression_Code_and_Unit_Testing.py in the 'data load in and test v cases regression' cell and testing is found inside the 'testing' cell. My code can also be found https://github.com/Toble007/COVID-19_CS_5010_Project/blob/main/Code/Covid_Vaccines_Tests_Cases_Regression_Code_and_Unit_Testing.ipynb and https://colab.research.google.com/drive/1tt usp=sharing

Model Preparation

The COVID Cases dataset was used in this model. All zero value data points for new COVID tests and new COVID cases were removed.



As shown above the data was not a linear function. So a log-log transformation was tried on the data.

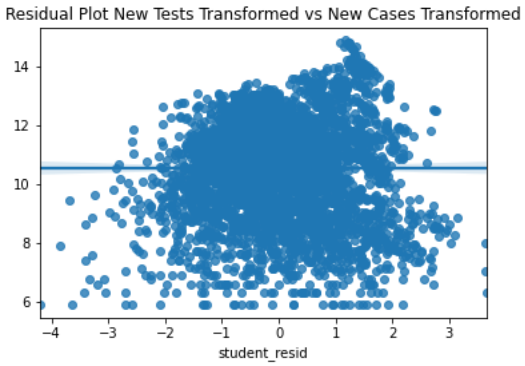


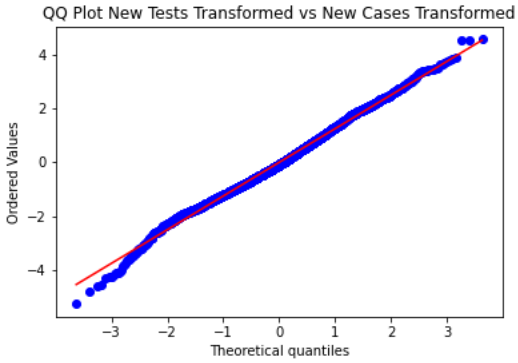
This somewhat solved the linearity problem so proceed to model fitting.

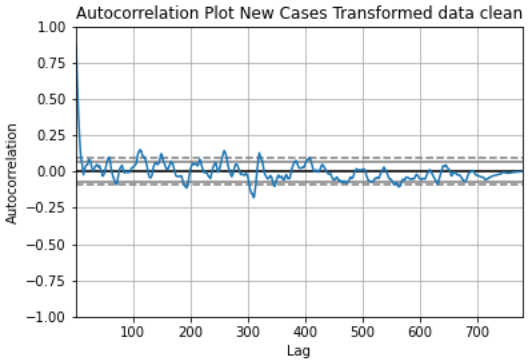
Model Description

New COVID Tests vs New COVID Cases were fitted with a simple linear regression model. Output and graphs are below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     new_tests_transformed   R-squared:                       0.632
Model:                               OLS   Adj. R-squared:                  0.632
Method:                    Least Squares   F-statistic:                     8400.
Date:                   Sat, 08 May 2021   Prob (F-statistic):               0.00
Time:                           17:44:06   Log-Likelihood:                -8022.7
No. Observations:                   4891   AIC:                         1.605e+04
Df Residuals:                       4889   BIC:                         1.606e+04
Df Model:                              1
Covariance Type:               nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                        5.9256      0.054    110.419      0.000       5.820       6.031
new_cases_transformed        0.6240      0.007     91.654      0.000       0.611       0.637
==============================================================================
Omnibus:                       15.703   Durbin-Watson:                   0.222
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.968
Skew:                           0.124   Prob(JB):                     0.000341
Kurtosis:                       3.131   Cond. No.                         24.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```







There are some issues with the model. The residual plot shows a nonconstant variance and non-zero mean. The QQ Plot has a slight downward skew. The Autocorrelation plot shows that an autocorrelation issue exists. The nonconstant variance and non-zero mean means that there is probably a issue with my model. When I made these models I didn't know how to fix these issues.

Model Interpretation

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     new_tests_transformed   R-squared:                       0.632
Model:                               OLS   Adj. R-squared:                  0.632
Method:                    Least Squares   F-statistic:                     8400.
Date:                   Sat, 08 May 2021   Prob (F-statistic):               0.00
Time:                           17:44:06   Log-Likelihood:                -8022.7
No. Observations:                   4891   AIC:                         1.605e+04
Df Residuals:                       4889   BIC:                         1.606e+04
Df Model:                              1
Covariance Type:               nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     5.9256      0.054    110.419      0.000       5.820       6.031
new_cases_transformed     0.6240      0.007     91.654      0.000       0.611       0.637
==============================================================================
Omnibus:                       15.703   Durbin-Watson:                   0.222
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.968
Skew:                           0.124   Prob(JB):                     0.000341
Kurtosis:                       3.131   Cond. No.                         24.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Overall, this model states that a 10% increase in cases leads to a 1.1^Bcase or 1.1^0.6240 = 1.0613 or 6.13% increase in testing.

Testing

All columns used in this model were tested to ensure there were no NA values in them and that all the values inside the columns were greater than zero.

Improvements

Starting out our group asked this question in a way that we didn't know how to answer. The question was reworded a few times but we never got a perfect answer. If we had more time we would have figured out how to create an ARIMA (auto regressive integrated moving average) model to answer this question. Originally we didn't know what a autoregressive model or partial autocorrelation function (PACF) was and would now use those to sovle the time lag issue present in the model.

Hypothesis 3: A Country's Ability To Scale Vaccination Efforts Is Related To HDI

The hypothesis of this model is: The more advanced a country, based on the Human Development Index (HDI), the higher their ability to scale vaccination efforts, as measured by vaccination rates.

Model Preparation

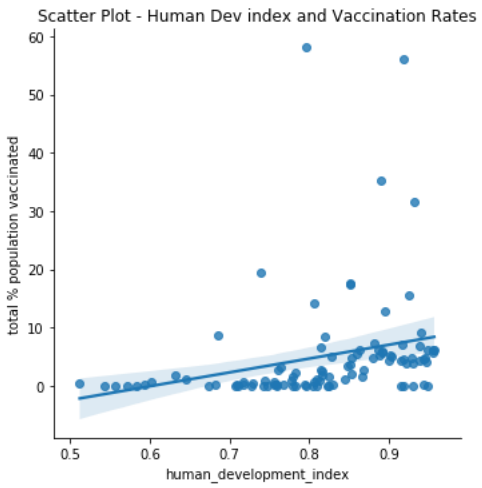
The Vaccination, World bank, and Covid cases datasets were used for this model. Human Development Index is a calculated index using life expectancy, education, and per capita income indicators. It is compiled by the United Nations Development Program (UNDP). Below is the scale and categories of the index:

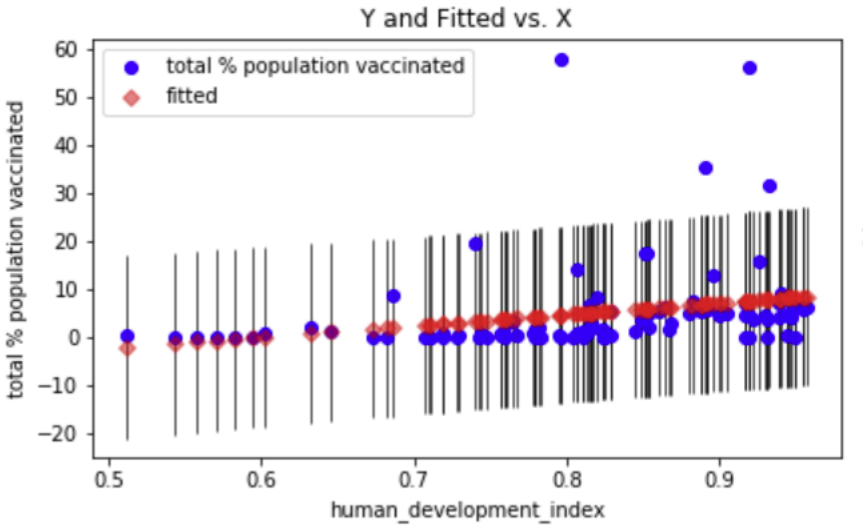
- Score of 0.8 and above considered as very high.
- 0.7 to 0.799 is high.
- 0.550 to 0.699 considered medium.
- Below 0.550 is low

Vaccination rates were calculated for each country using the vaccination data and demographics data i.e. Total People Vaccinated/population. The final percentage of the population vaccinated was taken by finding the highest percentage of vaccination rates per country, which in theory should be the latest day of vaccinations. The dataframe was then grouped by country to produce one datapoint for each country showing their percentage of population vaccinated against their human development index. Countries without a human development index score were dropped from the analysis. This resulted in 15 countries being dropped and the analysis had a total of 100 countries.

Model Description

A scatter plot and fitted plot of the Human Development Index against Vaccination rate was produced to perform initial exploratory data analysis. The HDI is the predictor variable and Vaccination rate is the response variable. Below are the two plots:





As can be seen from the plots, it appears that there could be a linear relationship between HDI and Vaccination rates. There are also some points which are on the top right side of the plots that apprear to be leverage points given their much higher vaccination rates versus most other countries.

Below is the top 10 countries based on vaccination rates along with their corresponding HDI and GDP Per capita numbers (in current $).

	location	total % population vaccinated	human_development_index	GDP per capita (current US$)
306	Seychelles	58.058776	0.796	17448.270293
902	Israel	56.178152	0.919	43592.083582
1857	United Arab Emirates	35.189831	0.890	43103.323058
4072	United Kingdom	31.462767	0.932	42330.117537
1277	Maldives	19.353168	0.740	10626.513402
513	Bahrain	17.523153	0.852	23503.977127
5008	Chile	17.311482	0.851	14896.453867
4676	United States	15.635962	0.926	65297.517508
3690	Serbia	14.270561	0.806	7411.836116
3168	Malta	12.801814	0.895	29820.603247

This does show that most of these countries have an HDI above 0.8. It also shows Seychelles and Israel appear to be the leverage points, given their significantly higher vaccination rates.

The next step was to proceed with building a simple linear regression model.

```
                            OLS Regression Results
==============================================================================================
Dep. Variable:     total % population vaccinated   R-squared:                       0.069
Model:                                     OLS   Adj. R-squared:                  0.060
Method:                          Least Squares   F-statistic:                     7.300
Date:                         Sat, 24 Apr 2021   Prob (F-statistic):           0.00813
Time:                                 12:22:28   Log-Likelihood:                -363.24
No. Observations:                          100   AIC:                             730.5
Df Residuals:                               98   BIC:                             735.7
Df Model:                                    1
Covariance Type:                     nonrobust
===============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                      -14.3231      7.224     -1.983      0.050     -28.659       0.013
human_development_index     23.7212      8.780      2.702      0.008       6.298      41.144
==============================================================================
Omnibus:                      115.947   Durbin-Watson:                   0.159
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1670.263
Skew:                           4.039   Prob(JB):                         0.00
Kurtosis:                      21.320   Cond. No.                         15.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model Interpretation

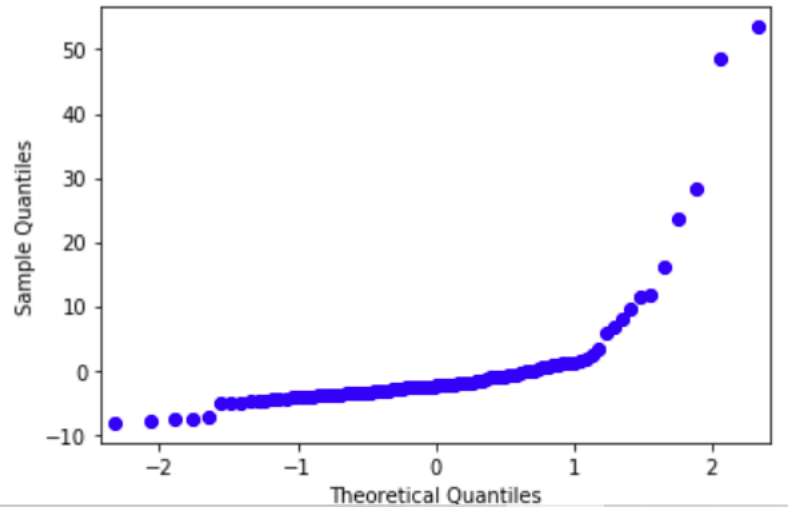
The results show that the HDI (predictor variable) coefficient has a value of 23.7212 and a t-statistic of 2.702. This means it is statistically significant within a 95% confidence interval. This can be interpreted as for a unit change in the Human development index, the vaccination rate of a country increases by 23.7%. However, this model had a low R-Square value of 0.069, which meant that 7% of the variance of Vaccination rates is explained by HDI in this model.

Testing

Unit testing was done to ensure there were no negative values for HDI and Vaccination Rates. Testing was also done to ensure vaccination rates were equal to or lower than 100, given that this was a calculated number in percentage terms.

Improvements

Further improvements would be focused on improving the model to better explain the variance. This could be done by trying models outside of linear regression or finding the appropriate transformation. The data also showed a positive skewness as illustrated by the QQ plot below. So perhaps dealing with the outliers would not only correct for the skewness, but also produce a better R-squared and hence a better explanation of the variance.



Hypothesis 4: Cultural Dimensions Play A Role In COVID Infection Rates

The hypothesis of this model is: Countries with high individualism & uncertainty avoidance had higher covid cases.

Model Preparation

The Covid cases was sourced from Ourworld dataset on Covid cases. Geert Hofstede cultural dimensions data set was our other data source for this model.

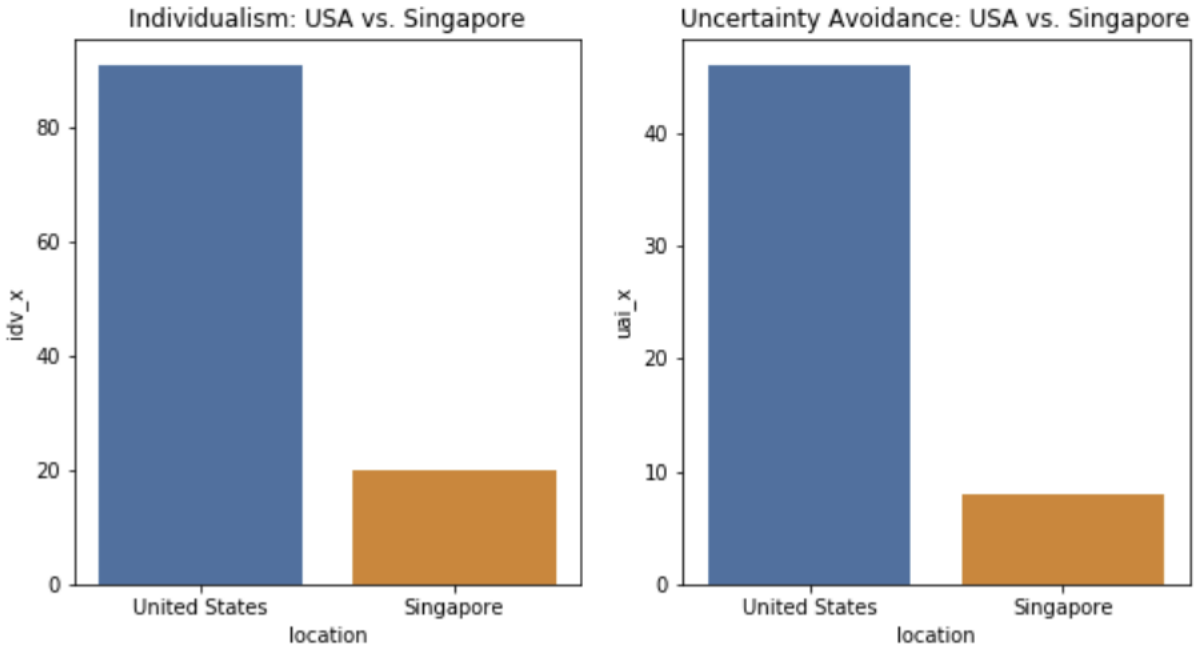
Two cultural dimensions were selected from Hofstede Cultural dimensions : individualism and uncertainty avoidance.

Individualism index measures the ties people have to their society.

- The scores demonstrate loosely knit vs. tightly knit social framework within a country.
- High value indicates weak interpersonal connection to people outside their core family.

Uncertainty avoidance measures the society's tolerance for uncertainty. Countries will high scores tend to be more uncomfortable with uncertainty. The scores range from 1 to 100 for each dimension.

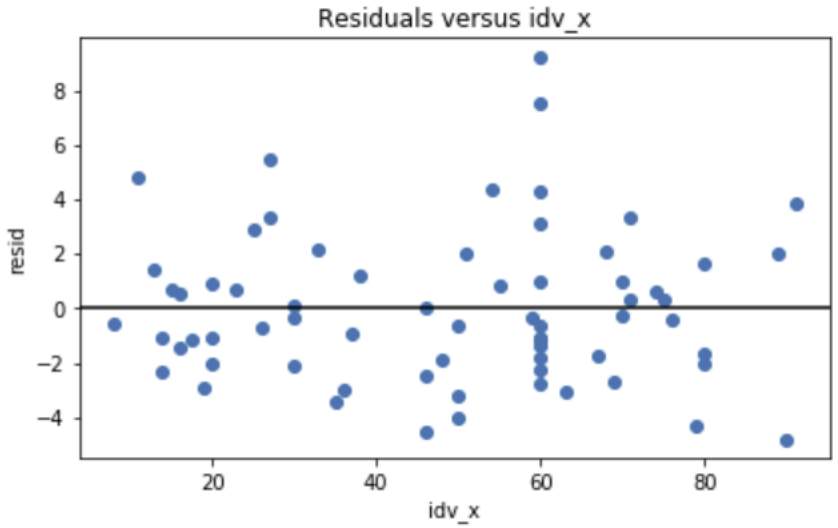
The below bar plot shows an example of two countries with different Individualism and Uncertainty avoidance scores. These are also two countries which had different responses to the Covid 19 pandemic.

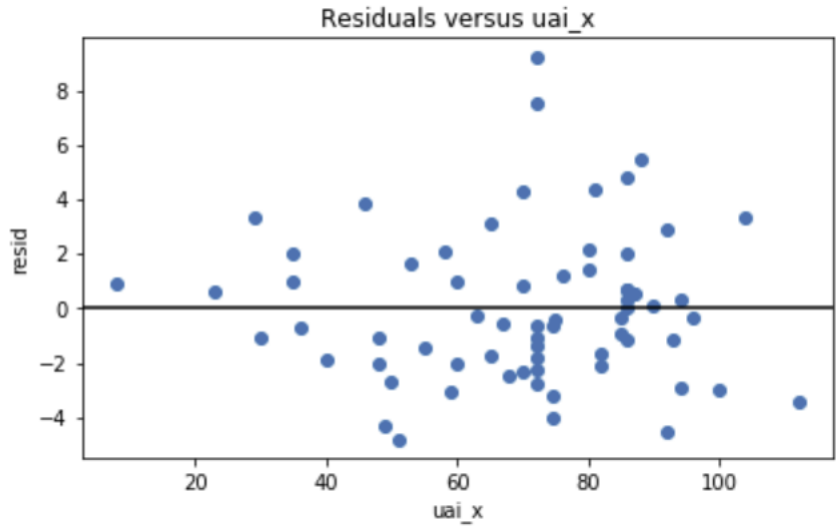


Covid positive rates were calculated for each country using the covid cases data and demographics data i.e. Total population tested positive for Covid/population. The final percentage of covid positive cases was taken by finding the highest percentage of covid rates per country, which in theory should be the latest day of cases. Data was grouped by country to produce one datapoint for each country showing their percentage of covid cases against their individualism and uncertainty avoidance measures. Countries without cultural dimension scores were dropped from the analysis. This resulted in a total of 67 countries for the analysis.

Model Description

A multiple regression model was run using Individualism index and the Uncertainty avoidance index as the predictor variables with percentage of population that tested positive for covid as the response variable. Below is a residual plot for the two predictors.





Visually, it appears that the residuals are appear to be scattered randomly around zero, which is one of the assumptions of a linear regression model.

Here are the results of the regression models:

```
                                OLS Regression Results
==============================================================================================
Dep. Variable:     total % of population with covid   R-squared:                       0.139
Model:                                          OLS   Adj. R-squared:                  0.112
Method:                               Least Squares   F-statistic:                     5.166
Date:                              Fri, 07 May 2021   Prob (F-statistic):            0.00832
Time:                                      01:14:40   Log-Likelihood:                -163.47
No. Observations:                                67   AIC:                             332.9
Df Residuals:                                    64   BIC:                             339.6
Df Model:                                         2
Covariance Type:                          nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0763      1.652     -0.652      0.517      -4.376       2.224
idv_x          0.0423      0.016      2.670      0.010       0.011       0.074
uai_x          0.0432      0.017      2.478      0.016       0.008       0.078
==============================================================================
Omnibus:                       11.293   Durbin-Watson:                   1.596
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               11.510
Skew:                           0.881   Prob(JB):                      0.00317
Kurtosis:                       4.009   Cond. No.                         414.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model Interpretation

The results show that coefficients of the Individualism and Uncertainty avoidance are 0.0423 and 0.0432 respectively. Their T-values are 2.67 and 2.478, which means they are both statistically significant. This means that for a unit change in the Individualism index score that covid positive cases increase by 0.04 while holding the other predictors constant. This also about the same for the uncertainly avoidance index score. However, the model only explains 14% of the variance given the R-square of 0.139.

Testing

Unit testing was done to ensure there were no negative values for the cultural dimensions and covid positive cases. Testing was also done to ensure covid positive rates were equal to or lower than 100, given that this was a calculated number in percentage terms.

Improvements

Further improvements would be focused on improving the model to better explain the variance and improve the R-squared value. Perhaps, trying models outside of linear regression would be more helpful here.

Conclusions

In conclusion, after pulling multiple datasets and performing data cleaning and merging, we were able to run a few regression tests on hypothesis we were interested in exploring. We also performed unit testing on our code using Glass-box/white-box testing methods.

Key Takeaways

Some of the conclusions we arrived at included:

- For a 10% increase in Covid cases, testing efforts increased by 6.1%.
- A 10% increase in new Covid cases from the current week led to a 1.1% increase in Vaccines.
- There is a positive relationship between a country's level of development and vaccination rates.
- There is a positive relationship between individualism, uncertainty avoidance and the number of covid cases. In addition, we also produced a dashboard with visualizations to show the trend by week in vaccination rates and covid cases by multiple factors such as income group and region.

Opportunities For Further Investigation

There were also some areas we were interested in investigating further to produce better and more robust models. Some of the topics for further investigation include:

- Data Pipeline: automate the downloading of our dynamic data
- Data Timeline: explore alternative time windows (3 days, 10 days, etc)
- Data Selection: could use PCA or factor analysis to assess additional economic variables
- Ancillary Data: including public policy as categorical data (i.e.: 1 for lockdown imposed, 0 for no lockdown)
- Data Clean-Up: outlier detection would help with modeling
- Data Modeling: time-series (ARIMA) or maybe even Naive Bayesian Regression
- Data Visualization: annotated graphs