Covid-19 Infection Analysis

Tonirose Babasoro, Elsa Bustos, Utsav Chaudhary, India Scott

Program: Data Analytics
Group Name: Endless Knot
Presentation Date: April 7, 2022

INTRODUCTION

TOPIC

What is the likelihood of being infected by Covid-19? How is infection affected by factors such as vaccination rates, gender, and ethnicity?

PURPOSE

It is important to analyze future trends following the Covid-19 pandemic to understand the prevalence of infection within the American population.

Questions Answered

- Are certain populations more likely to be infected than others?
- How do these factors affect the other?
- What other factors should be considered in identifying risks of infection?

Data Sources

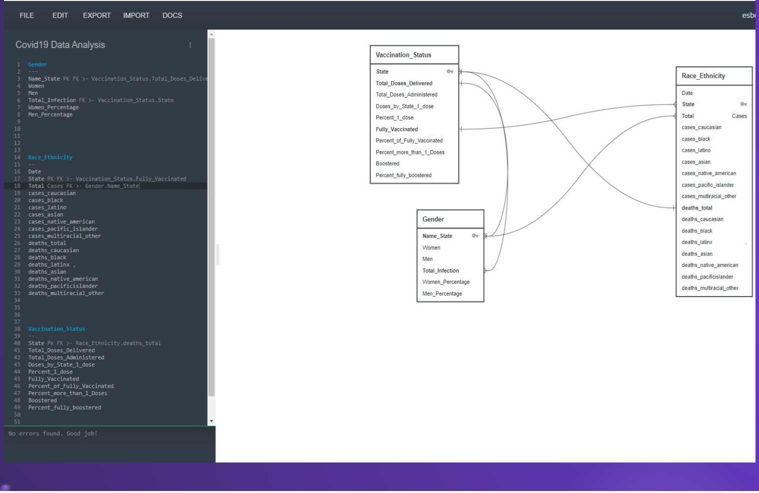
We gathered data from reliable organizations such as New York Times and the Center for Disease Control (CDC) which provide csv files on their findings. Our data identifies:

- vaccination rates
- gender ratios
- ethnicity statistics

Data Exploration

SQL

Entity Relationship Diagram



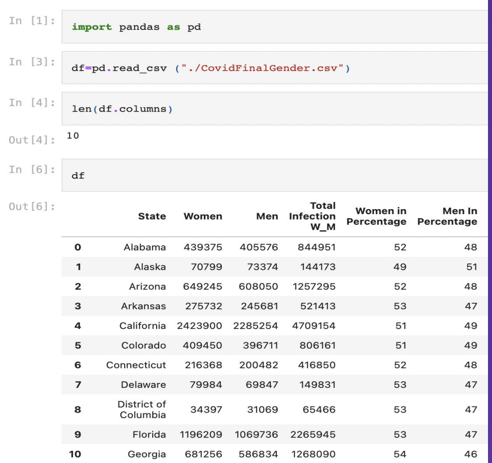
```
Gender
---
Name_State PK FK >- Vaccination_Status.Total_Doses_Delive
Women
Men
Total_Infection FK >- Vaccination_Status.State
Women_Percentage
Men_Percentage




Race_Ethnicity
--
Date
State PK FK >- Vaccination_Status.Fully_Vaccinated
Total Cases FK >- Gender.Name_State
cases_caucasian
cases_black
cases_latino
cases_asian
cases_native_american
cases_pacific_islander
cases_multiracial_other
deaths_total
deaths_caucasian
deaths_black
deaths_latinx ,
deaths_asian
deaths_native_american
deaths_pacificislander
deaths_multiracial_other




Vaccination_Status
--
State Pk FK >- Race_Ethnicity.deaths_total
Total_Doses_Delivered
Total_Doses_Administered
Doses_by_State_1_dose
Percent_1_dose
Fully_Vaccinated
Percent_of_Fully_Vaccinated
Percent_more_than_1_Doses
Boostered
Percent_fully_boostered
```

No errors found. Good job!

Vaccination_Status: State, Total_Doses_Delivered, Total_Doses_Administered, Doses_by_State_1_dose, Percent_1_dose, Fully_Vaccinated, Percent_of_Fully_Vaccinated, Percent_more_than_1_Doses, Boostered, Percent_fully_boostered

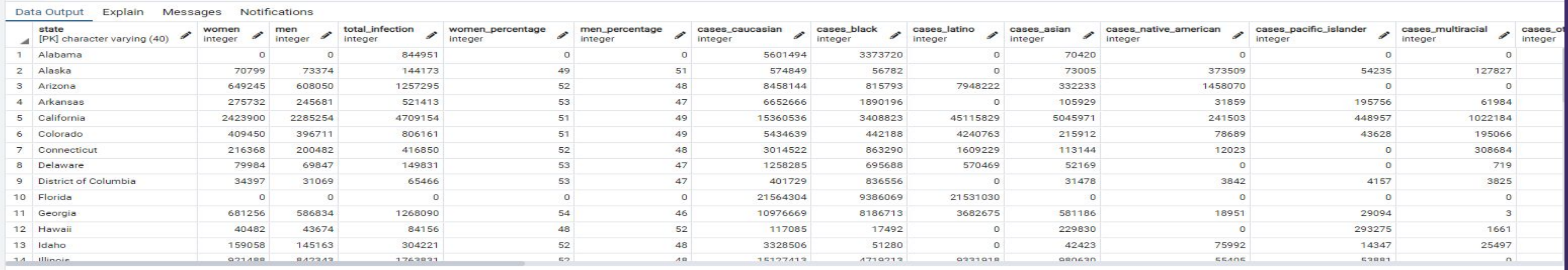
Gender: Name_State, Women, Men, Total_Infection, Women_Percentage, Men_Percentage

Race_Ethnicity: Date, State, Total Cases, cases_caucasian, cases_black, cases_latino, cases_asian, cases_native_american, cases_pacific_islander, cases_multiracial_other, deaths_total, deaths_caucasian, deaths_black, deaths_latinx ,, deaths_asian, deaths_native_american, deaths_pacificislander, deaths_multiracial_other



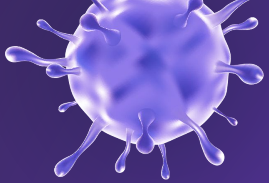
```
In [1]:  import pandas as pd

In [3]:  df=pd.read_csv ("./CovidFinalGender.csv")

In [4]:  len(df.columns)

Out[4]:  10

In [6]:  df

Out[6]:
                                     Total      Women in      Men In
          State      Women      Men  Infection  Percentage  Percentage
                                     W_M
 0      Alabama     439375   405576     844951          52          48
 1       Alaska      70799    73374     144173          49          51
 2      Arizona     649245   608050    1257295          52          48
 3     Arkansas     275732   245681     521413          53          47
 4   California    2423900  2285254    4709154          51          49
 5     Colorado     409450   396711     806161          51          49
 6  Connecticut     216368   200482     416850          52          48
 7     Delaware      79984    69847     149831          53          47
 8  District of      34397    31069      65466          53          47
       Columbia
 9      Florida    1196209  1069736    2265945          53          47
10      Georgia     681256   586834    1268090          54          46
```

- After manual review of multiple csvs, we were able to identify a primary key: “States”
- Data is filtered and aggregated via SQL server and filter null values with pandas (Python)



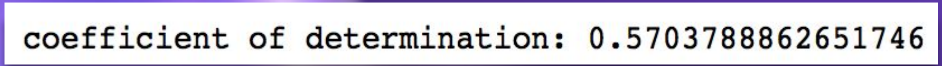
	state [PK] character varying (40)	women integer	men integer	total_infection integer	women_percentage integer	men_percentage integer	cases_caucasian integer	cases_black integer	cases_latino integer	cases_asian integer	cases_native_american integer	cases_pacific_islander integer	cases_multiracial integer	cases_o integer
1	Alabama	0	0	844951	0	0	5601494	3373720	0	70420	0	0	0	
2	Alaska	70799	73374	144173	49	51	574849	56782	0	73005	373509	54235	127827	
3	Arizona	649245	608050	1257295	52	48	8458144	815793	7948222	332233	1458070	0	0	
4	Arkansas	275732	245681	521413	53	47	6652666	1890196	0	105929	31859	195756	61984	
5	California	2423900	2285254	4709154	51	49	15360536	3408823	45115829	5045971	241503	448957	1022184	
6	Colorado	409450	396711	806161	51	49	5434639	442188	4240763	215912	78689	43628	195066	
7	Connecticut	216368	200482	416850	52	48	3014522	863290	1609229	113144	12023	0	308684	
8	Delaware	79984	69847	149831	53	47	1258285	695688	570469	52169	0	0	719	
9	District of Columbia	34397	31069	65466	53	47	401729	836556	0	31478	3842	4157	3825	
10	Florida	0	0	0	0	0	21564304	9386069	21531030	0	0	0	0	
11	Georgia	681256	586834	1268090	54	46	10976669	8186713	3682675	581186	18951	29094	3	
12	Hawaii	40482	43674	84156	48	52	117085	17492	0	229830	0	293275	1661	
13	Idaho	159058	145163	304221	52	48	3328506	51280	0	42423	75992	14347	25497	
14	Illinois	921488	842343	1763831	52	48	15127413	4719213	9331918	980630	55405	53881	0	

SQL

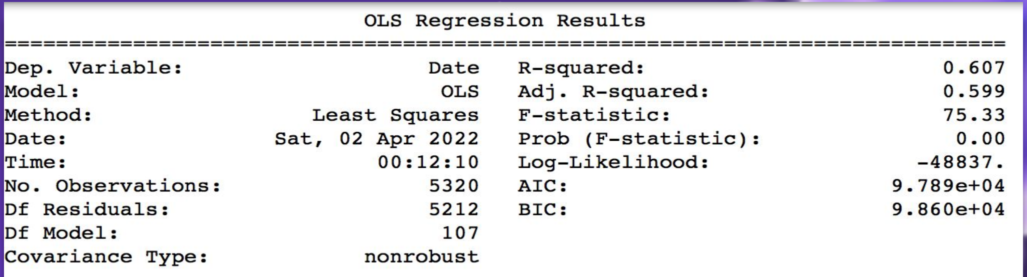


Data Analysis

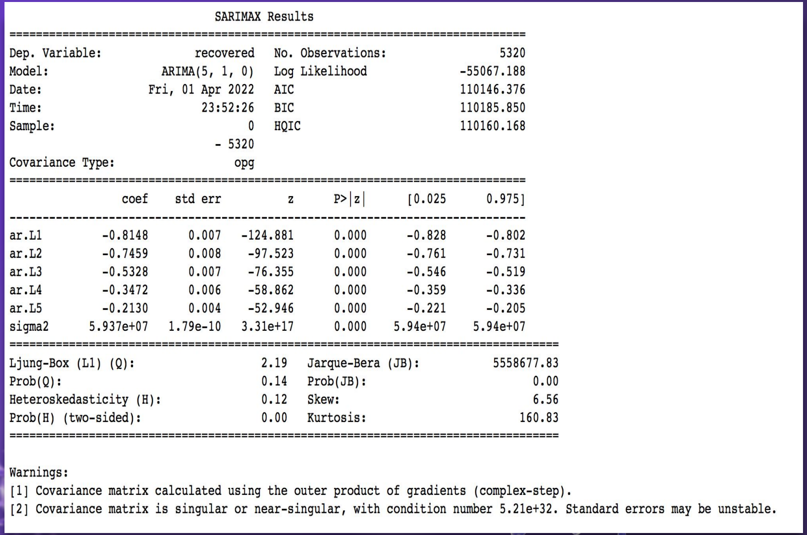
- Ordinary Least Squares (OLS)
- Linear regression
- SVM support vector machine
- Autoregressive Integrated Moving Average (ARIMA) for Time series model



```
coefficient of determination: 0.5703788862651746
```

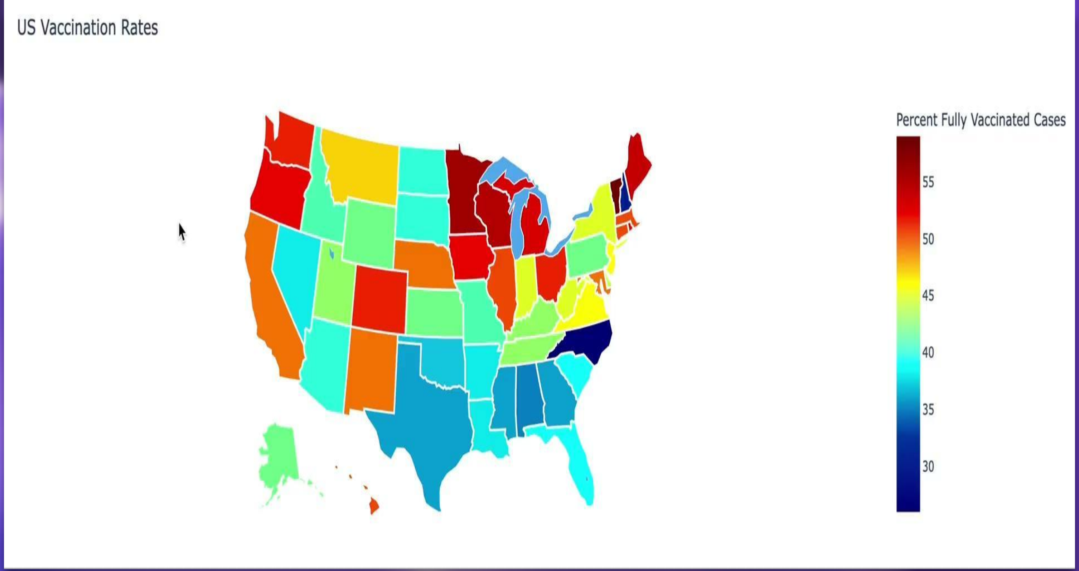


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Date   R-squared:                       0.607
Model:                            OLS   Adj. R-squared:                  0.599
Method:                 Least Squares   F-statistic:                     75.33
Date:                Sat, 02 Apr 2022   Prob (F-statistic):               0.00
Time:                        00:12:10   Log-Likelihood:                -48837.
No. Observations:                5320   AIC:                         9.789e+04
Df Residuals:                    5212   BIC:                         9.860e+04
Df Model:                         107
Covariance Type:            nonrobust
```



```
                               SARIMAX Results
==============================================================================
Dep. Variable:              recovered   No. Observations:                 5320
Model:                 ARIMA(5, 1, 0)   Log Likelihood              -55067.188
Date:                Fri, 01 Apr 2022   AIC                         110146.376
Time:                        23:52:26   BIC                         110185.850
Sample:                             0   HQIC                        110160.168
                               - 5320
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8148      0.007   -124.881      0.000      -0.828      -0.802
ar.L2         -0.7459      0.008    -97.523      0.000      -0.761      -0.731
ar.L3         -0.5328      0.007    -76.355      0.000      -0.546      -0.519
ar.L4         -0.3472      0.006    -58.862      0.000      -0.359      -0.336
ar.L5         -0.2130      0.004    -52.946      0.000      -0.221      -0.205
sigma2      5.937e+07   1.79e-10    3.31e+17      0.000    5.94e+07    5.94e+07
===================================================================================
Ljung-Box (L1) (Q):                   2.19   Jarque-Bera (JB):            5558677.83
Prob(Q):                              0.14   Prob(JB):                          0.00
Heteroskedasticity (H):               0.12   Skew:                              6.56
Prob(H) (two-sided):                  0.00   Kurtosis:                        160.83
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 5.21e+32. Standard errors may be unstable.
```

Interactive Element



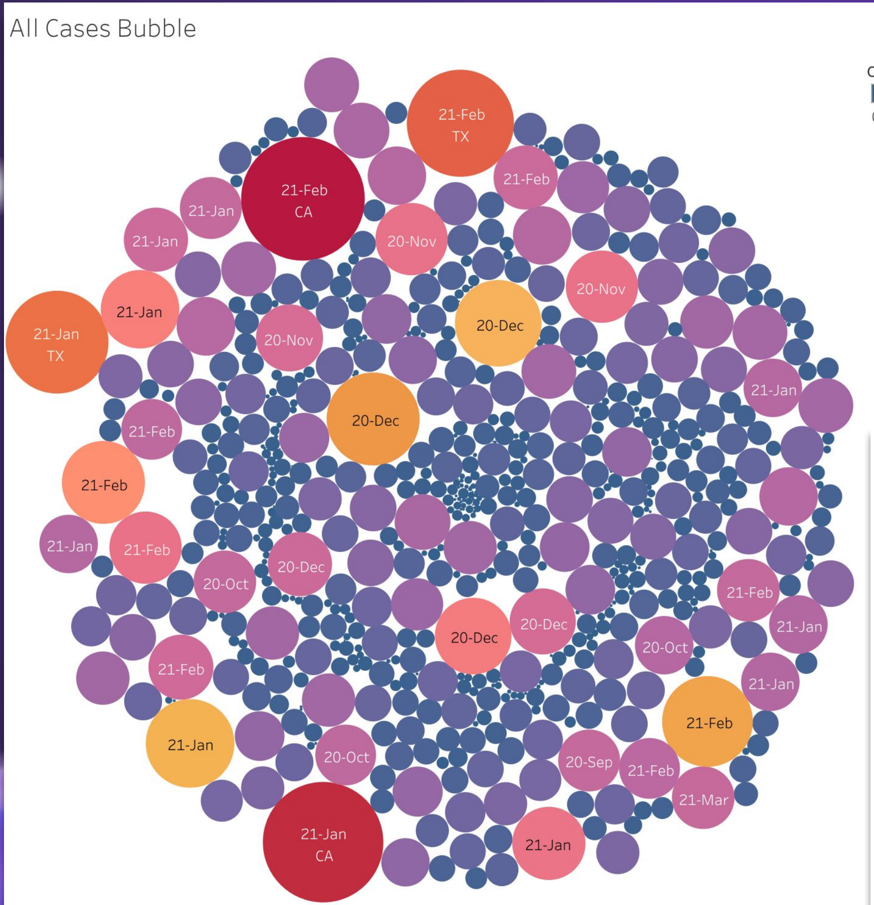
Vaccination Rates by State

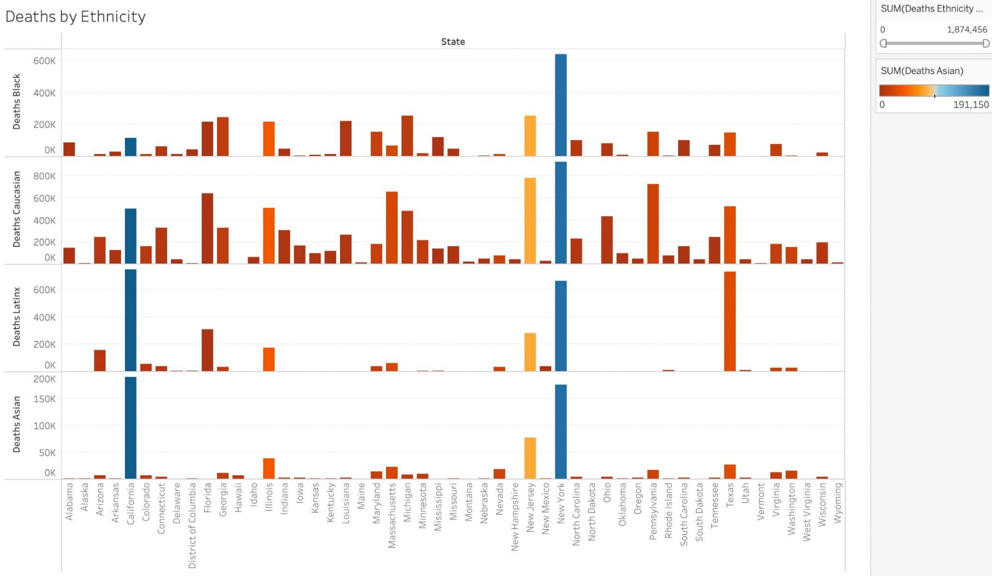
Plotly

- Map with options to toggle different factors of each state such as:
  - Vaccination rates
  - Infections
  - Gender data

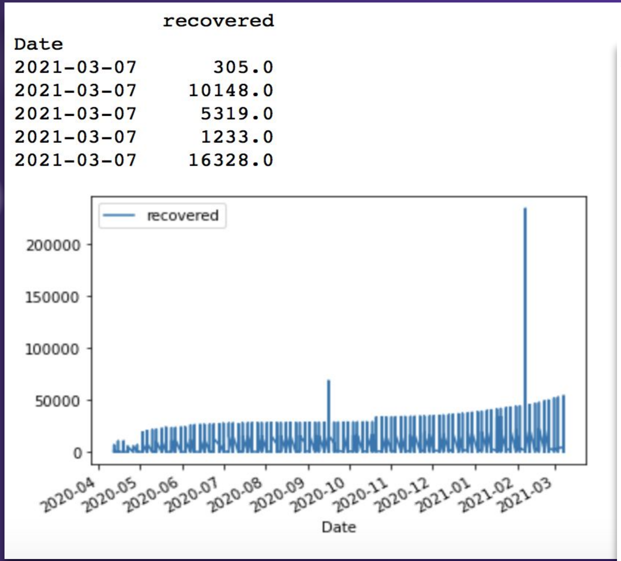
Visualization

- Tableau
    - Correlation graphs

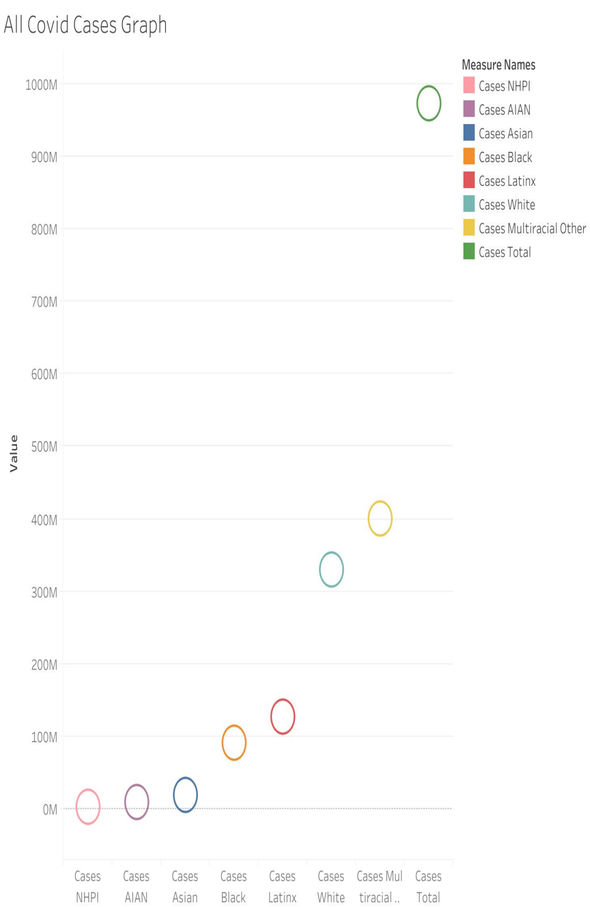

All Cases Bubble

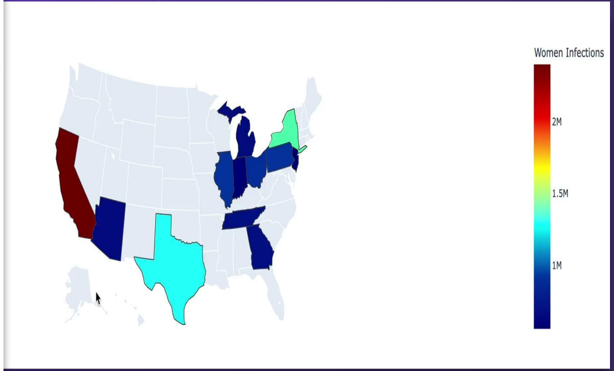

Deaths by Ethnicity

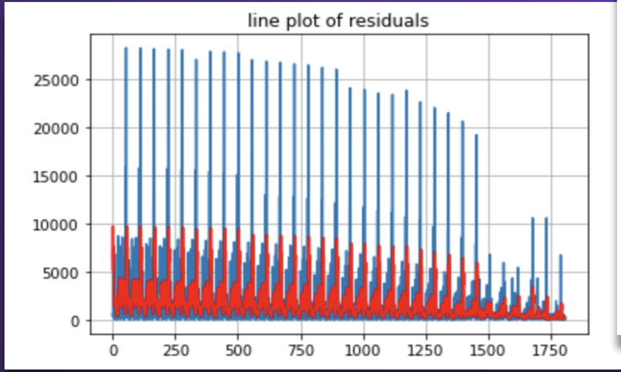
Dashboard

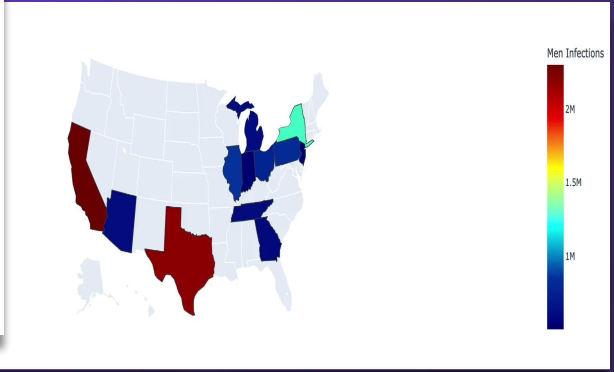


```
            recovered
Date
2021-03-07      305.0
2021-03-07    10148.0
2021-03-07     5319.0
2021-03-07     1233.0
2021-03-07    16328.0
```









Results of Analysis

- No significant differences between gender infection rates
- More vaccinations → lower cases and death rates
- People who were vaccinated and boosted had lower amounts of cases and fatalities
- On the time series analysis, the line plot is created, showing the expected values (blue) compared to the rolling forecast predictions (red). Values show some trends and are in the correct scale.

What We Would Do Differently

- Before deciding on a topic, look into data resources
- Compare more factors from different data sets
    - Identify more primary keys to connect data
- Added other factors such as economic impact

Recommendations for Future Analysis

- Statistical analysis can be used with bigger data sets
- Explore other factors that may have a correlation with the data we already have
- Expand to global research
    - countries that are not as ethnically homogenous
- Project future peak infections

Gender

https://www.genderscilab.org/gender-and-sex-in-covid19/#CaseDeathRatebySex

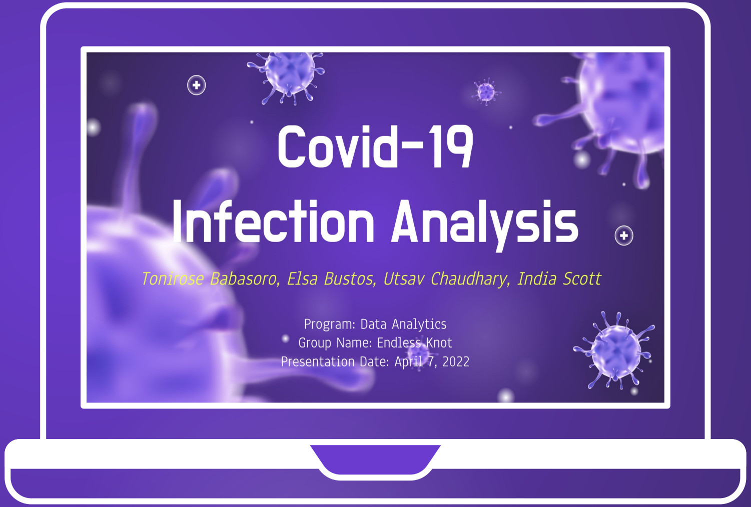
Ethnicity

https://covidtracking.com/race/dashboard

Vaccination

d.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total

Check out our project at:
https://github.com/antirose/CovidInfectionAnalysis



Covid-19 Infection Analysis

Tonirose Babasoro, Elsa Bustos, Utsav Chaudhary, India Scott

Program: Data Analytics
Group Name: Endless Knot
Presentation Date: April 7, 2022