




Covid-19 EDA and Visualization



Oluwafemi Ajeigbe
Pierce Cappa
Vincent Lombardi

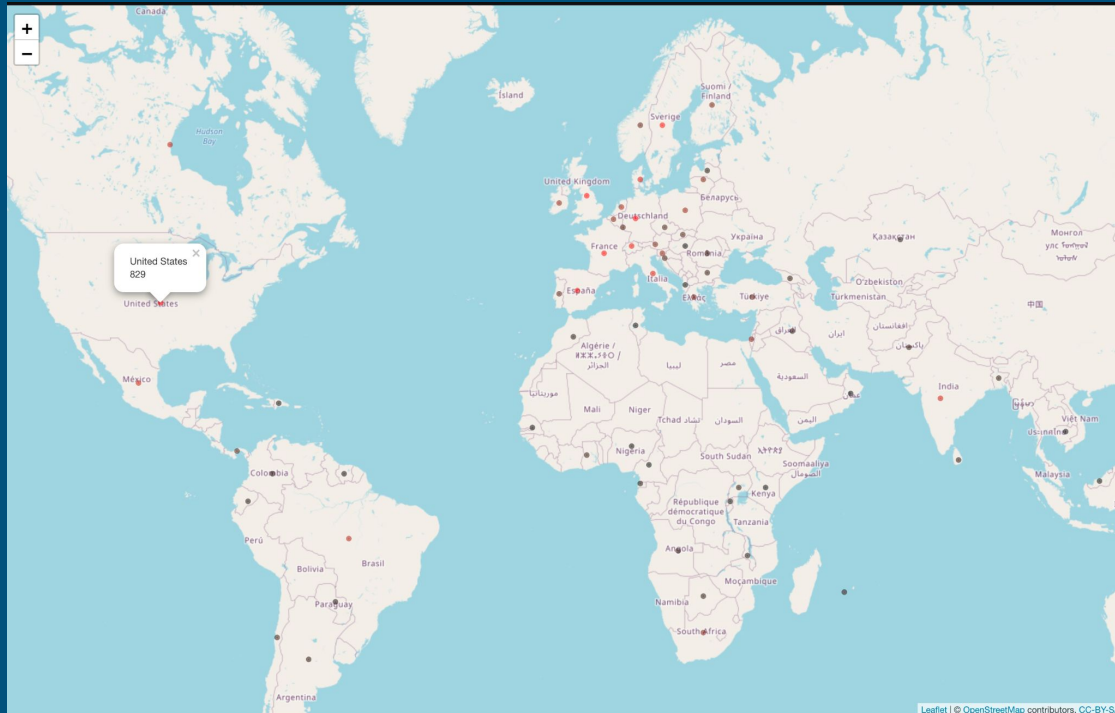


Goal Of The Project

- The spike protein is what the virus uses to penetrate the cell membrane.
- Data collected from ncbi and gisaid.
 - Removed all strains not collected worldwide at least 30 times.
- Visualize the spike proteins mutation over time
 - Checked for patterns or trends over time.

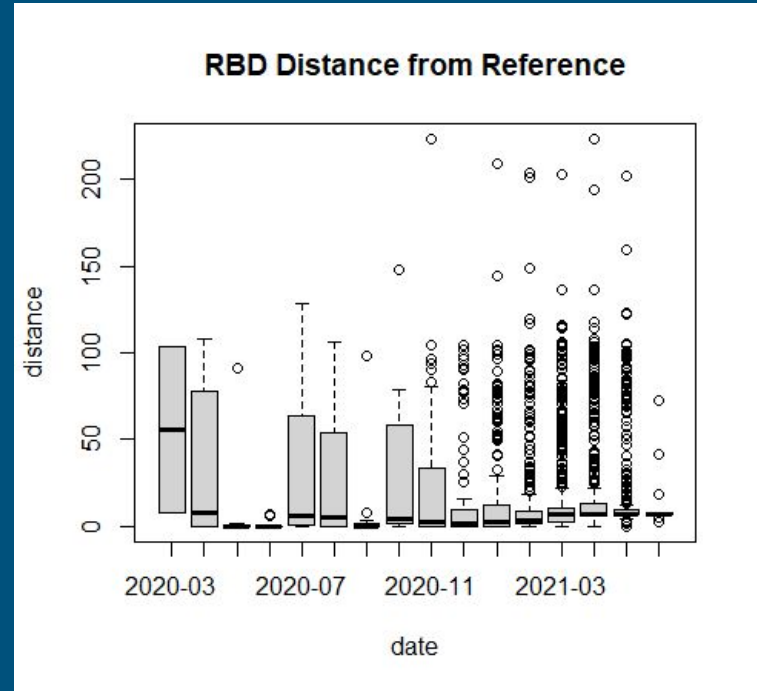
Sequences grouped by Countries

- Grouped the sequences by countries
- Created a visualization using R(leaflet package) and mapped frequencies to country of origin.
- Country with largest number of new sequences: USA,



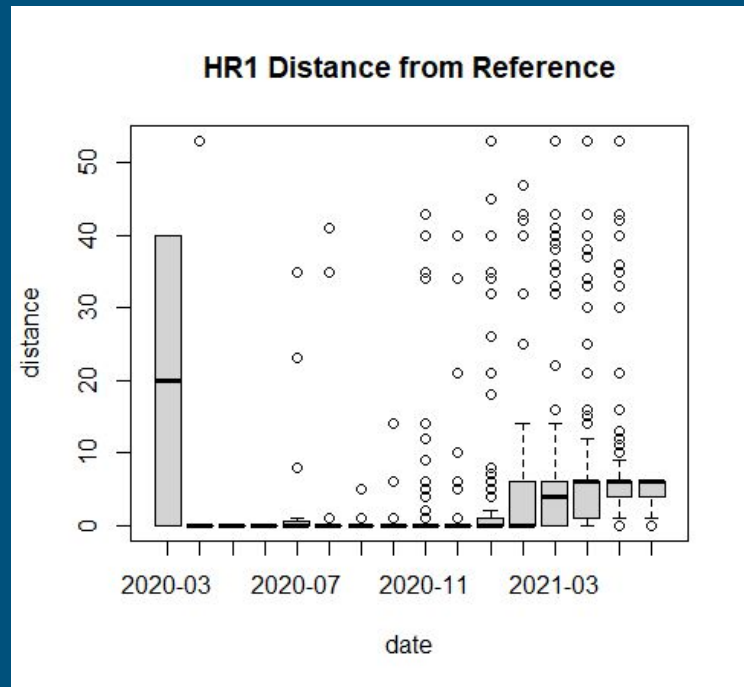
Receptor Binding Domain (RBD)

- Binds with enzymes on the cells surface.
- Is crucial to initiating fusion and infiltrating the cell.
- The delta variant had multiple spike protein mutations.



HR1

- Many of the other subunits saw variance later on.
- HR1 and HR2 specifically mutated more into 2021.
- HR1 and HR2 help regulate RBD.



Tree Maps

- Used python
 - Used the pandas library for data management
 - Plotly.express for tree map creation
- Organized each data point based off of time a strain was first collected
- Used the Damerau–Levenshtein distance algorithm to determine parent strain and strain success
 - Strain success is determined by parent count + all children count