Final Project

Title: Final Project Prospectus
Notice: Dr. Bryan Runck
Author: Emily Cavazos
Date: 12/22/2021

Project Repository: https://github.com/cavaz020/GIS5571.git
Google Drive Link: Detailed Regression Analysis Reports
Storymap Link: COVID-19 Storymap
Time Spent: 60 hrs

Abstract
The purpose of this report is to visualize, analyze and assess incidence rates between new COVID-19 cases and deaths and the number of arrived flights to each state in the United States, including those outside of the contiguous US such as Alaska, Hawaii, and Puerto Rico. I used flight arrival data from the Bureau of Transportation Statistics, Census and American Community Survey demographics of race and of poverty level, and CDC COVID-19 data to create the visualizations and analyze the data. I also used shapefiles of airport points and of state boundaries. The steps I took to do so are thoroughly detailed below with accompanying screenshots and data flow charts. For results, I visualized the number of flights to each state and then created bivariate choropleth maps of the explanatory variable of interest and the dependent variable. I then ran the explanatory regression tool to assess the prediction ability of different models using all of the explanatory variables I compiled. Within the discussion, I detailed my experience working with the exploratory regression tool, what I learned about the significance of the adjusted R-squared values, potential shortcomings of the approach I took and potential future studies. My sources are cited and I reported a self score at the end of the report.

Problem Statement
The problem I am looking to visualize in my project is the incidence rates between COVID-19 cases and deaths and the number of flights to each state in the United States, including Alaska, Hawaii and Puerto Rico, during spring and summer during the second year of the COVID-19 pandemic, specifically March through August of 2021. I also intend to visualize demographics by county during this time period to attempt to make a commentary on and bring awareness to the ongoing state of settler colonialism in which Hawaii and Puerto Rico reside.

Table 1. Data Categories

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset	Preparation
1	CDC Covid-19 Data	Input Data from CDC - Cases and Deaths by state	Have to be joined to boundary shp	Cases and Deaths by state	CDC COVID Data	Aggregate up to month by state
2	Demographics	Poverty demographics from the American Community Survey	Have to be joined to boundary shp	Race and Poverty status	US Census Bureau - 2020 Census Race ACS 2019 1-year Estimates - Poverty	Clean (some of extra data deleted, looking through metadata to find fields of interest)
3	Flight Counts	Flight counts found within delay causes data	Have to be joined to Airport points	Number of Flights to Airport	Bureau of Transportation Statistics	Clean and aggregate up to state

4	Airport Points	Shapefile of points of all airports in the United States	Ready to be used in visualization	Points with airport, county, and state info	US Department of Transportation	Summarize in ArcPro by state (used for normalizing data)
5	State Boundary Shapefiles	Base shapefiles to easily make tabular data spatial	Shapefile		US Census Bureau	Field formats may need to be edited (Names uppercase to match other shapefiles)

Input Data

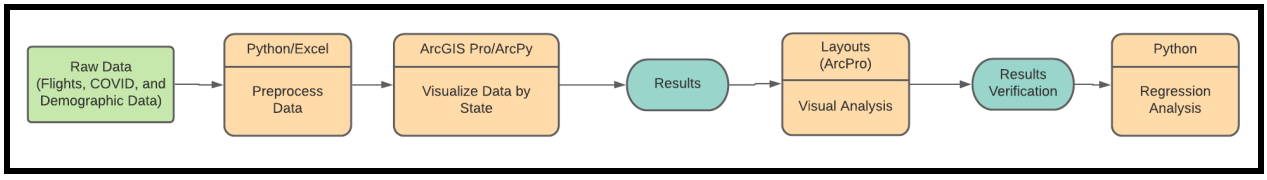
The data used in this project is varied in order to attempt to model a real-world situation. Data in this project is used for visualizations as well as regression analysis to verify the findings. The data is described in detail below.

Table 2. Data used.

#	Title	Purpose in Analysis	Link to Source
1	United States State Boundaries	Dataset to visualize and perform analysis at various scales	US Census Bureau - State Boundaries
2	ACS 2019 1-year estimate	Dataset to visualize and demonstrate the communities that are vulnerable to COVID-19 (because of compounding effects of economic inequity and systemic racism)	ACS 2019 1-year Estimates - Poverty
3	Census 2020 Race by County	Dataset to visualize and demonstrate the communities that are vulnerable to COVID-19 (because of compounding effects of economic inequity and systemic racism)	US Census Bureau - 2020 Census Race
4	Airport Locations	Dataset used to visualize number of airports per state, totals used to normalize flight data, and used to make flight arrival data spatial	US Department of Transportation - Airport Locations
5	Number of Flights	Dataset used to extract total number of flights that arrived at each airport (US)	Bureau of Transportation Statistics - Airline On-Time Statistics and Delay Causes
6	COVID-19 Data	Dataset used to visualize trends in COVID-19 cases and deaths by state	CDC COVID Data

Methods

Figure 1. Data Flow Chart of Project Progression.



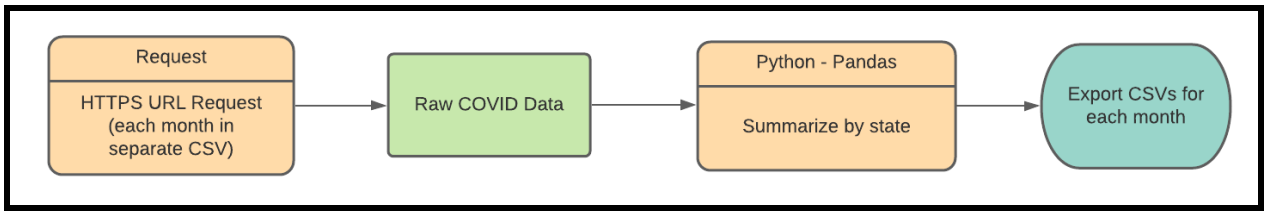
The COVID-19 data for this project was acquired via ETL processes with requests being sent to the CDC's API for each month of interest in this study. The first part of this project consisted of preprocessing the various data tables to ensure they are pared down to the necessary data that will be manipulated and visualized within ArcGIS Pro. Next, visualizations are made of COVID-19 cases by state as well as visualizations of the number of flights to each area by month normalized by the number of airports in each state. Finally, regression analysis was done to find the best fit model for each month and to document the correlation between COVID-19 and the number of flight arrivals for each month.

Data Preprocessing

This project has required a lot of data pre-processing because I ultimately joined many different datasets from a variety of sources.

COVID Data

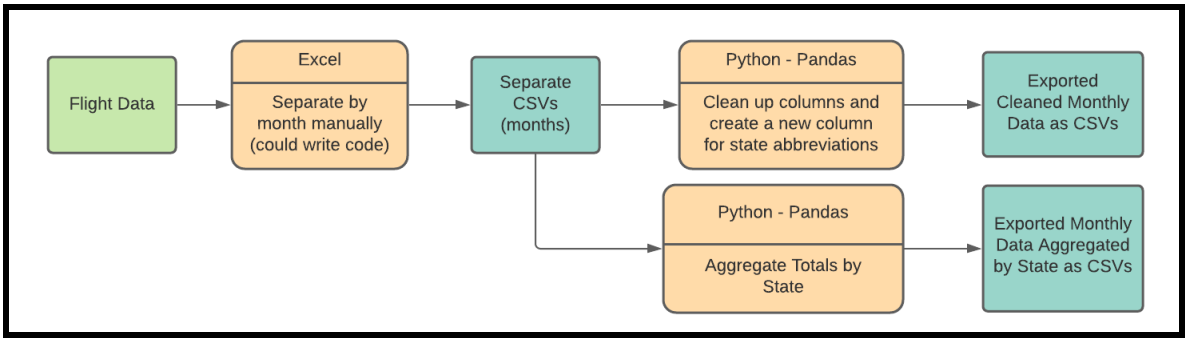
Figure 2. Data Flow Diagram of preprocessing COVID datasets.



I utilized an ETL to send a request to the CDC API to download the CDC cases and deaths data listed by state. After I had each CSV, I used Pandas to summarize the data by state as the original was listed daily and I wanted to aggregate each month. I then used the CSVs to attach to the state shapefile to do further analysis.

Flight Data

Figure 3. Data flow diagram of processing Flight Data.



It was surprisingly hard for me to find flight arrival data that was historical. The raw flight data that I was able to find was actually a csv that was focused on discussing the causes of delays in flights. However, this table also included the total number of flights that arrived by airport. The data I am focused on for this project is from spring and summer of 2021 and I defined that as months March through September of that year.

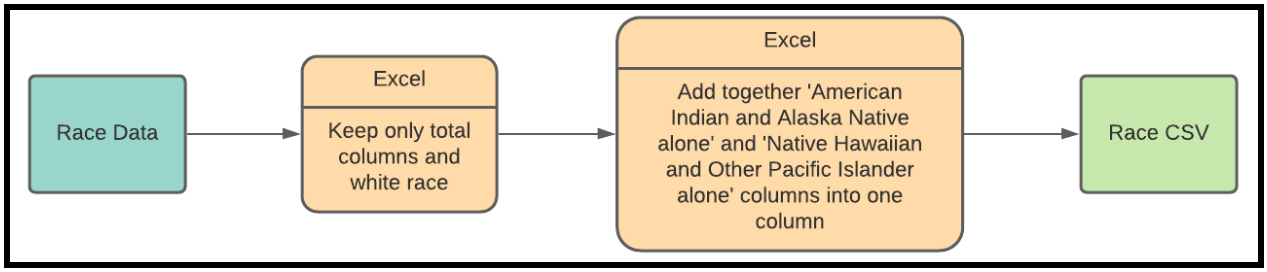
For pre-processing, I wanted to delete the erroneous data about delay causes and make it ready to join with state or county boundaries. To do so, I first separated the data into separate CSVs by month. Next, I input the CSVs into the IDE I use for Python coding and used pandas to delete all of the columns of erroneous data leaving only the columns of airport codes, airport names, and the count of flights that arrived in each one. I used pandas to create a column for state abbreviations. Finally, I created a function to take in the dataframes of the CSVs and aggregate the number of flights by state by using the column name of interest. In the end, I exported the various dataframes as CSVs for use in ArcGIS Pro.

Image 1. Function for aggregating flight data by state.

```
def agg_by_state(data_frame):
    # Returns a dataframe of the data grouped and summed by state
    grouped_state_df = data_frame.groupby(['state_abbrev']).sum()
    return grouped_state_df
```

Census Data - Race

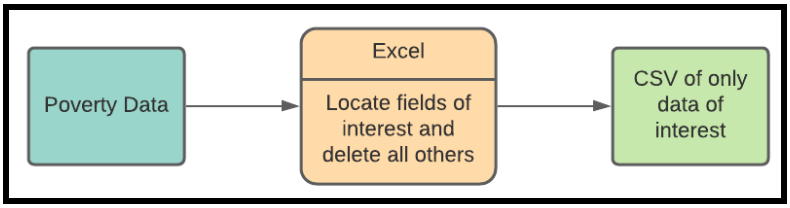
Figure 4. Data flow diagram of processing Demographic data - Race.



To pre-process the race data, I used Excel. I simply kept the state column and the GEOID column, the total population column as well as the 'total one race' column in case I needed to use it for normalization. I also kept the 'Population one race - white alone' column. I then added the 'American Indian and Alaska Native alone' column to the 'Native Hawaiian and Other Pacific Islander alone' column to create a new column I called 'total_native'. For this indicator I only looked at populations who marked one race total (meaning not including mixed race individuals).

ACS Data - Poverty

Figure 5. Data flow diagram of processing Demographic data - Poverty.

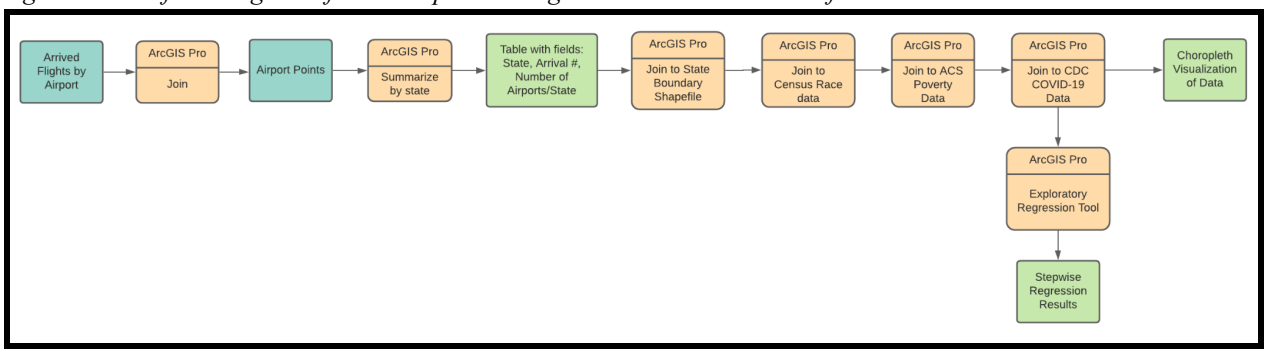


For this indicator, I used Excel again and simply located the state column, the GEOID column, the total column and the column of population below 125% poverty level and saved this as a new CSV.

Results - Visualization

For visualization and results verification using the exploratory regression tool, I joined all of the data to the shapefile of the state boundaries to create shapefiles for each month with all of the data by state. I was then able to create the following visualizations and, as discussed in the next section, the results verification in the form of the stepwise regression of various models.

Figure 6. Data flow diagram of detailed process to get results and results verification.

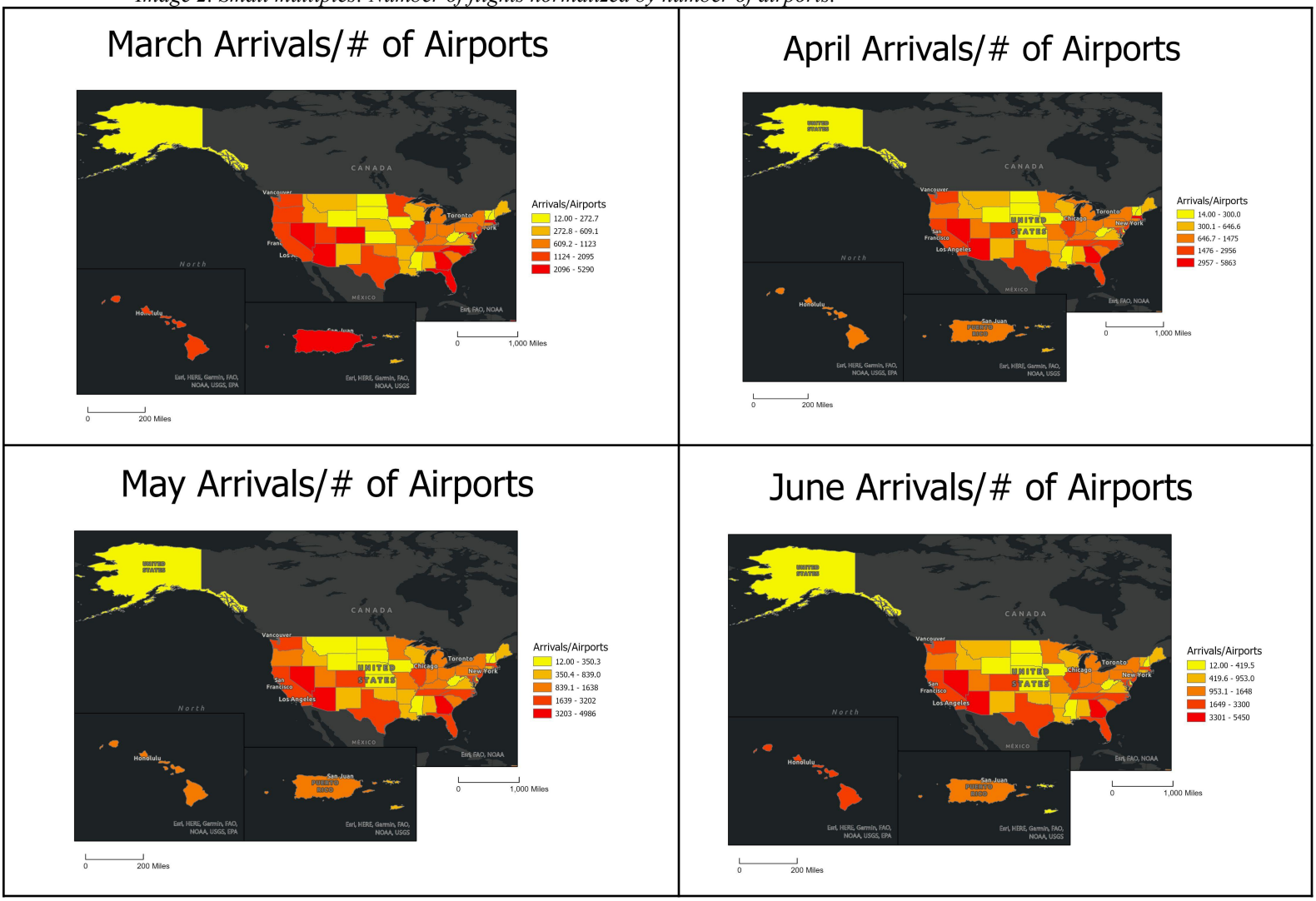


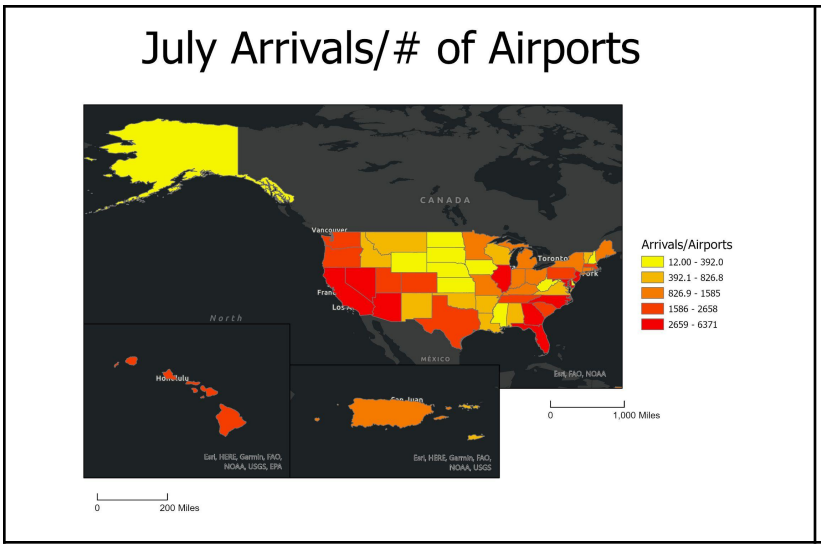
Number of flights by state

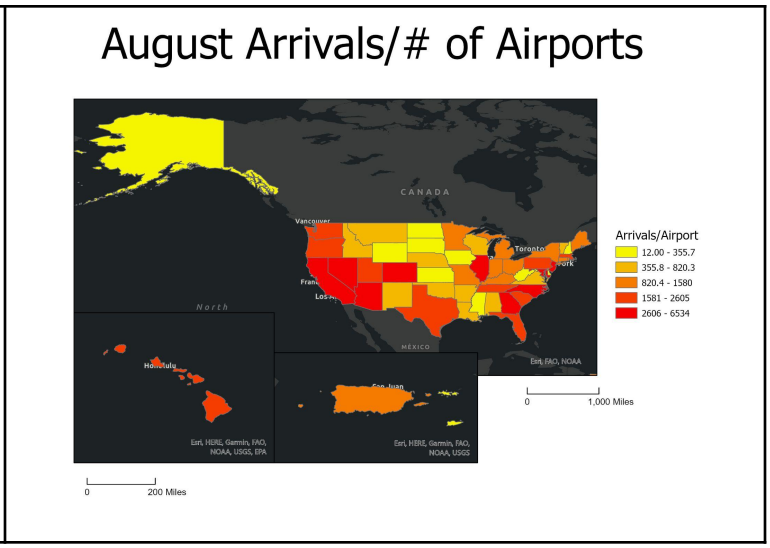
I wanted to map the number of flights by state to see if my theory of why I wanted to include Hawaii and Puerto Rico in my study intentionally would hold true. With this project I was hypothesizing that Hawaii and Puerto Rico

were more visited during the pandemic because of their tourist economies and beaches and that the flights would be tied to increasing COVID-19 cases. Thus, I first had to test and see if they were more frequently visited than other states. I mapped the number of flights by number of airports in each state and I used Quantile breaks with five classes. **As seen below, both Hawaii and Puerto Rico were always in the mid to high range of number of arrivals by number of airports.**

Image 2. Small multiples: Number of flights normalized by number of airports.



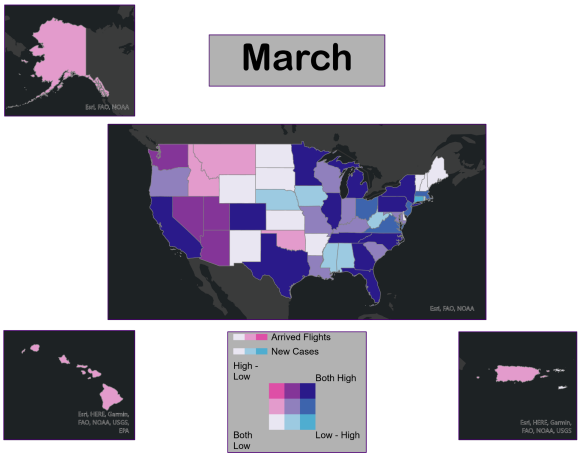




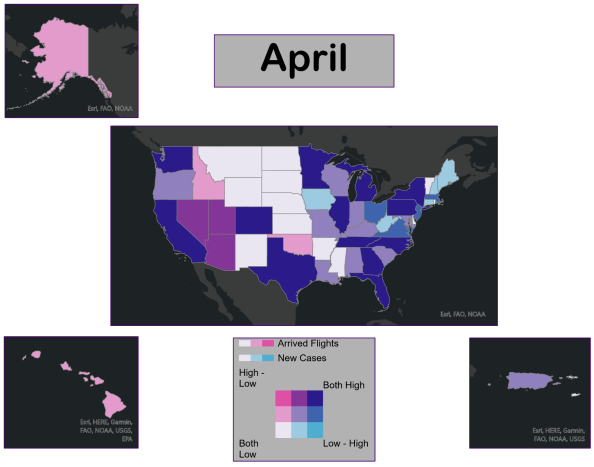
Multivariate Mapping: Number of flights by State

I used the CSVs of the monthly flight data grouped by state to join to the shapefile of the state boundaries from the Census Bureau then joined the case data to the shapefiles as well. This allowed me to create visualizations of each month. Below I have inserted bivariate choropleth maps of this information and added some counts below of each type of correlation. The majority of states had correlation (both high, both low, both mid) between the two variables in each month.

Image 3. Small multiples of bivariate choropleth maps: Number of Flights Arrived by COVID Cases.

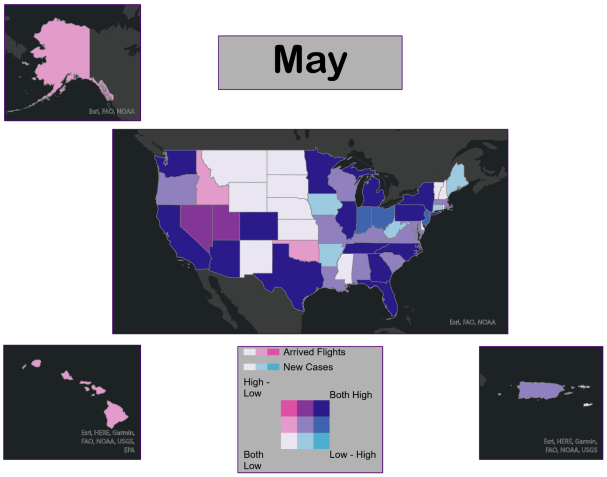


Both High: 12
Both Mid: 7
High Flights - Mid Cases: 4
High Cases - Mid Flights: 4
High Flights - Low Cases: 0
Mid Flights - Low Cases: 6
Both Low: 10



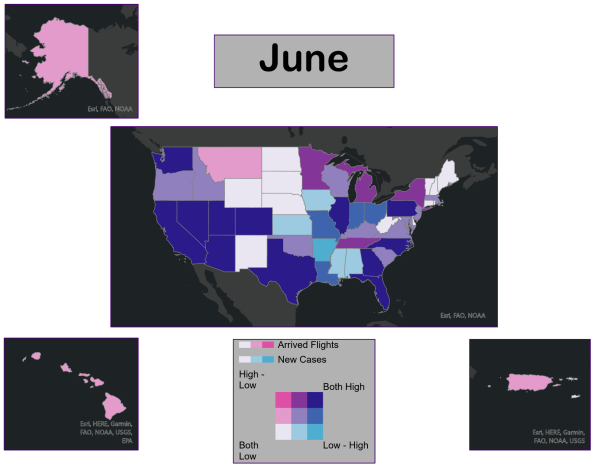
Both High: 13
Both Mid: 9
High Flights - Mid Cases: 3
High Cases - Mid Flights: 4
High Flights - Low Cases: 0
Mid Flights - Low Cases: 4
Both Low: 12

May



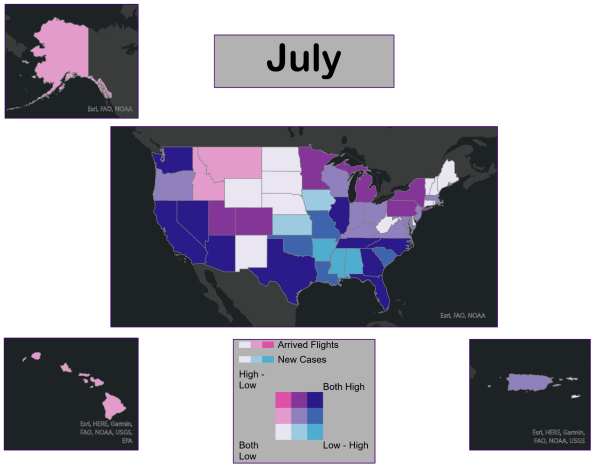
Both High: 14
Both Mid: 10
High Flights - Mid Cases: 2
High Cases - Mid Flights: 3
High Flights - Low Cases: 0
Mid Flights - Low Cases: 4
Both Low: 12

June



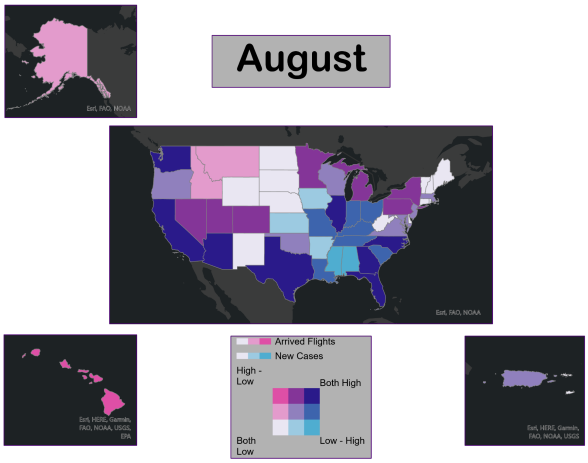
Both High: 12
Both Mid: 8
High Flights - Mid Cases: 4
High Cases - Mid Flights: 4
High Flights - Low Cases: 0
Mid Flights - Low Cases: 4
Both Low: 11

July



Both High: 10
Both Mid: 7
High Flights - Mid Cases: 6
High Cases - Mid Flights: 4
High Flights - Low Cases: 0
Mid Flights - Low Cases: 4
Both Low: 12

August



Both High: 8
Both Mid: 7
High Flights - Mid Cases: 7
High Cases - Mid Flights: 7
High Flights - Low Cases: 1
Mid Flights - Low Cases: 3
Both Low: 12

Results Verification - Stepwise Regression

I used the 'Exploratory Regression Tool' to evaluate the significance of the explanatory variables on the dependent variables. This tool evaluates all possible combinations of the input candidate explanatory variables, looking for Ordinary Least Square (OLS) models that best explain the dependent variable. I ran this for each month for both 'New Cases' and 'New Deaths' as dependent variables. Below are the results for each dependent variable. In each of these models, the stars next to the explanatory variables correlate to differing levels of significance (* = 0.10; ** = 0.05; *** = 0.01) where a lower value, or more stars, means a higher significance.

figure 7. Tables of models for each month - Dependent Variable: New Deaths.

March		**April**		**May**		**June**		**July**		**August**	
Model	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**
+Arrived Flights***	0.48	+Arrived Flights***	0.54	+Arrived Flights***	0.59	+Arrived Flights***	0.68	+Arrived Flights***	0.50	+Arrived Flights***	0.34
-Arrived Flights +Native +Poverty***	0.79	+Poverty* +Arrived Flights +White	0.74	+White*** -Native +Arrived Flights	0.75	+Arrived Flights +White*** +Native	0.79	+Arrived Flights -Native +Poverty	0.49	+Arrived Flights -Native +Poverty	0.37
-Arrived Flights +Poverty*** -White +Native	0.79	+Poverty* +Arrived Flights +White +Native	0.73	+White*** -Native +Arrived Flights -Poverty	0.74	+Arrived Flights +White*** +Native +Poverty	0.78	+Arrived Flights -White -Native +Poverty	0.48	+Arrived Flights -White -Native +Poverty	0.36

figure 8. Tables of models for each month - Dependent Variable: New Cases.

March		**April**		**May**		**June**		**July**		**August**	
Model	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**	**Model**	**AdjR² value**
+Arrived Flights***	0.54	+Arrived Flights***	0.44	+Arrived Flights***	0.72	+Arrived Flights***	0.73	+Arrived Flights***	0.57	+Arrived Flights***	0.68
+Arrived Flights +White*** -Native**	0.78	-Poverty** +Arrived Flights +White***	0.68	+White*** +Arrived Flights*** -Poverty**	0.87	+Arrived Flights*** -Native +Poverty	0.74	+Arrived Flights -Native +Poverty	0.58	+Arrived Flights -White +Poverty**	0.73
+Arrived Flights* -Poverty +White*** -Native*	0.78	-Poverty* +Arrived Flights +White*** -Native*	0.69	+White*** -Native* +Arrived Flights*** -Poverty**	0.88	+Arrived Flights*** +White -Native +Poverty	0.73	+Arrived Flights* -White -Native +Poverty	0.58	+Arrived Flights* -White*** -Native* +Poverty***	0.74

Discussion and Conclusion

The results that I got from my analysis were promising. I appreciated that I was able to view the adjusted R-squared value for the various models because it is more accurate when adding in more and more variables in a stepwise manner. The adjusted R-squared value only increases if adding more variables enhances the model above what would be obtained by probability and it will decrease when a predictor improves the model less than what is predicted by chance. Generally, a higher adjusted R-squared indicates a better fit for the model. Comparing the R-squared between stepwise regression models, if the adjusted R-squared increases when adding in a new variable, this is evidence that the variable contributes to explaining the dependent variable.

As you can see above, the number of arrived flights had larger adjusted R-squared values when modeled against 'New Cases' rather than 'New Deaths'. This initially puzzled me. However, I did some research and learned that it is commonly understood, especially with COVID-19, that deaths are a lagging indicator. Basically this means that it takes several weeks after a diagnosis for a patient to die. Then it takes more time for them to be added to the official state tally. This makes sense why cases would be more linked to the number of arrived flights in a month as they are both variables that are updated almost in real time.

There are some issues with using the Exploratory Regression tool. Because it is a data mining tool, a strong proponent of the scientific method might object to the use of this tool because, from their perspective, you should formalize your hypotheses before exploring your data to avoid creating models that fit only your data, but don't reflect broader processes. For future studies, I would like to run more verification analyses such as bootstrapping and then move on to running models based on the results that seemed to be most successful in my initial use of the exploratory regression tool.

References

1. *Malfer, Lindsay. 2019. "The Devastating Effects of Colonization on Hawai'i." ArcGIS StoryMaps. July 27, 2019. https://storymaps.arcgis.com/stories/83474c5d6077492d990b961bab0bcd74.*
2. *Nast, Condé. 2020. "As Puerto Rico Prepares to Reopen, Residents Are Concerned." Condé Nast Traveler. July 13, 2020. https://www.cntraveler.com/story/as-puerto-rico-prepares-to-reopen-residents-are-concerned.*
3. *"The History of Hawaii Tells the Story of a Violent Colonization." 2021. Study Breaks. May 20, 2021. https://studybreaks.com/thoughts/hawaii/.*
4. *"Traveling to Puerto Rico? Here's What Locals Want You to Know - The Washington Post." n.d. Accessed September 29, 2021. https://www.washingtonpost.com/travel/tips/puerto-rico-covid-local/.*
5. *Mollalo, Abolfazl, Behzad Vahedi, and Kiara Rivera. 2020. "GIS-Based Spatial Modeling of COVID-19 Incidence Rate in the Continental United States." Science of The Total Environment 728. https://www.sciencedirect.com/science/article/pii/S0048969720324013.*
6. *"Interpreting Exploratory Regression Results—Help | ArcGIS Desktop." n.d. Accessed December 22, 2021. https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/interpreting-exploratory-regression-results.htm.*
7. *"Virus Fatality Picture Is Obscured by Ultimate Lagging Indicator (Correct)." n.d. Accessed December 22, 2021.*

https://news.bloomberglaw.com/coronavirus/virus-fatality-picture-is-obscured-by-ultimate-lagging-indicator.

Self-score

Fill out this rubric for yourself and include it in your lab report. The same rubric will be used to generate a grade in proportion to the points assigned in the syllabus to the assignment.

Category	Description	Points Possible	Score
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	24
Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	20
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	26
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	18
		100	88