Final Project: First Draft

Title: Final Project Prospectus
Notice: Dr. Bryan Runck
Author: Emily Cavazos
Date: 09/29/2021

Project Repository: https://github.com/cavaz020/GIS5571.git
Google Drive Link:
Time Spent: 30 hrs

Abstract
[I FEEL THIS SHOULD BE COMPLETED AT THE END - AFTER RESULTS ARE FINALIZED]

Problem Statement
The problem I am looking to visualize in my project is the incidence rates between COVID-19 case and the number of flights to 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 and county (2 datasets)	Ready to be joined to boundary shp	Cases and Deaths by county and state	CDC COVID Data	Will need to be aggregated up to month (raw data is daily)
2	ACS or Census Demographics	Demographics from the American Community Survey - speak to COVID-19 vulnerability	Ready to be joined to boundary shp	Race, Age and Sex, Poverty status	Ex: ACS Age and Sex Hawaii	Cleaned (some of extra data deleted, looking through metadata to find field of interest)
3	Flight Counts for time period of interest	Flight counts found within delay causes data	Ready to be joined to Airport points	Number of Flights to Airport	Bureau of Transportation Statistics	Cleaned and aggregated up to county and state
4	Airport Points	Shapefile of points of all airports in the United States (including territories such as Puerto Rico and Guam)	Ready to be used in visualization/ summarization of counts	Points with county and state info	US Department of Transportation	Summarized in ArcPro by county and by state (used for normalizing data)
5	County Boundary and State Boundary Shapefiles	Base shapefiles to easily make tabular data spatial	Shapefiles		US Census Bureau	Field formats may need to be edited (Names uppercase to match other shapefiles)

Input Data

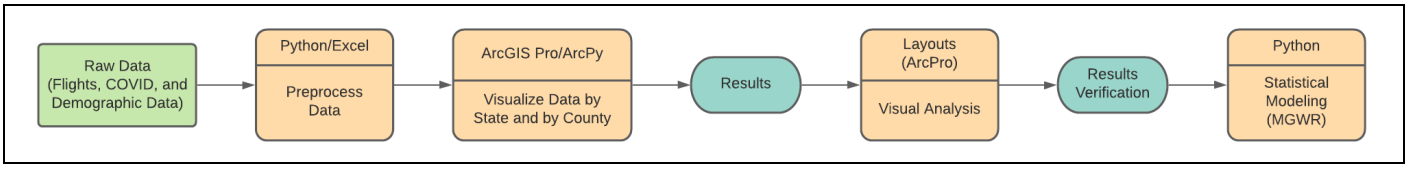
The data used in this project is varied in order to attempt to model a real-world situation. Data in this project is used for visualizations as well as statistical modeling to verify the veracity of the findings. The data is described in detail below.

Table 2. Data used.

#	Title	Purpose in Analysis	Link to Source
1	United States County Boundaries	Dataset to visualize and perform analysis at various scales	US Census Bureau - County Boundaries
2	United States State Boundaries	Dataset to visualize and perform analysis at various scales	US Census Bureau - State Boundaries
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	Census 2020 Demographics	Potential other demographic information to aid in analysis *(Not yet downloaded)*	US Census Bureau
5	Airport Locations	Dataset used to visualize number of airports per state or county and used to make flight arrival data spatial	US Department of Transportation - Airport Locations
6	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
7	COVID-19 Data - by state	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.



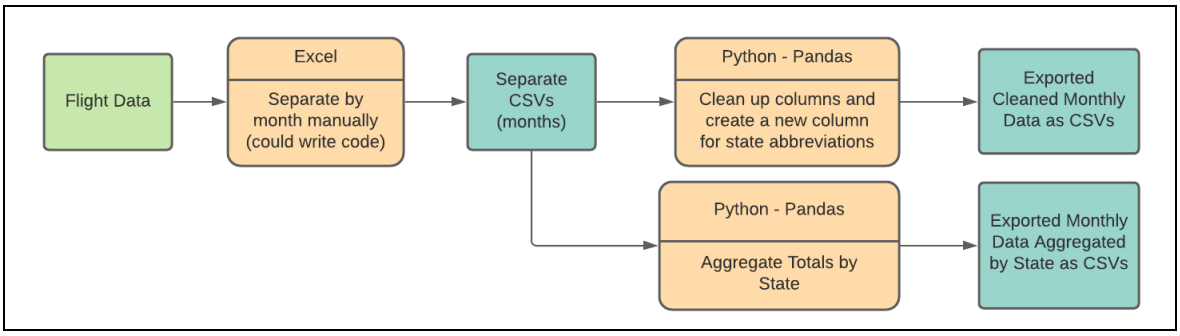
The data for this project was acquired via ETL processes with requests being sent to various APIs. The first part of this project will consist of preprocessing the various data tables to ensure they are pared down to the necessary data that will be manipulated and visualized within ArcGIS Pro. Next, visualizations are made of COVID-19 cases by state and by county as well as visualizations of the number of flights to each area by month. Finally, multiscale geographically weighted regression (MGWR) modeling will be done to test the incidence rates and the correlation between the variables.

Data Preprocessing

This project has required a lot of data pre-processing because I ultimately hoped to join many different datasets from a variety of sources. The python code is in a file titled 'data_clean.py'.

Flight Data

Figure 2. 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 of 2021 and I defined that as months March through June 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 functions to take in the dataframes of the CSVs and aggregate the number of flights by state and by county.

In a later step, I will be creating a table in ArcGIS pro of the flight data joined to the corresponding counties, will export this table, and will use python to group the data by county.

In the end, I exported the various dataframes as CSVs for use in ArcGIS Pro.

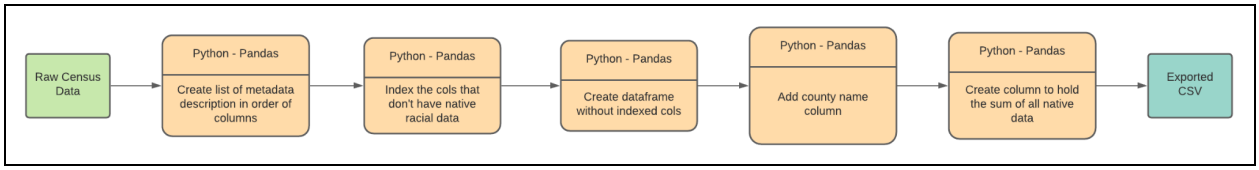
Image 1. Functions for aggregating flight data by state or county.

```
# ------------Functions for aggregating data by state and county------


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


def agg_by_county(data_frame):
    # Returns a dataframe of the data grouped and summed by county
    grouped_county_df = data_frame.groupby(['County']).sum()
    return grouped_county_df
```

Census Data - Demographics
Figure 3. Data flow diagram of processing Census Data (Race).



The census data perhaps took the most pre-processing time. In the census CSVs, the first row is actually metadata descriptions of each column. So this needed to be removed but it was also useful for preprocessing the data and ensuring that I kept columns of interest as the raw table had 73 columns. I used python to loop through the metadata for each column, index the locations of the data I wanted to remove, and create a dataframe without that data. I also added a column with county names properly formatted so that this information could be easily joined to spatial census data. Finally, I created a column to hold the sum of all data on the number of people identifying in some way as Native, the function for which is pictured below. I exported the data as a CSV after it was cleaned.

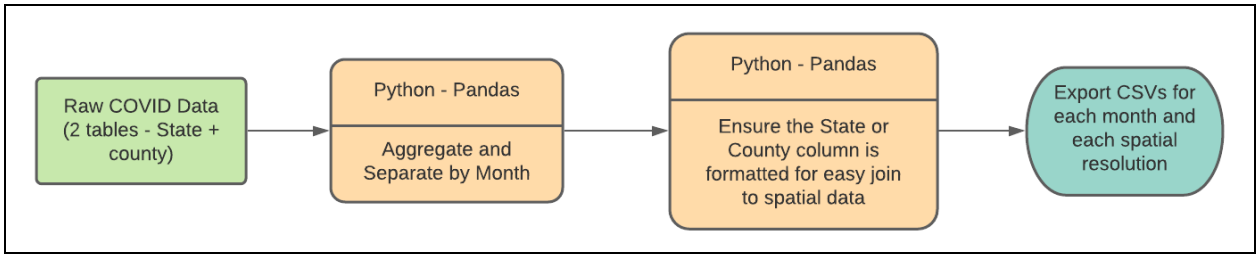
[Could Aggregate by state as well or download data by state from Census and clean that too for visualization]

Image 2. Function for aggregating all of the columns of data on Native populations.

```
def create_sum_col(data_frame):
    # Function to create a column holding the sum of all of the columns
    # of people who have identified in any way as 'American Indian and Alaska Native'
    # or 'Native Hawaiian and other Pacific Islander'
    # And naming this column 'TOTAL_NATIVE'
    data_frame = data_frame.rename(columns={'P1_001N': 'TOTAL_POP'})
    col_list = data_frame.columns.tolist()
    data_frame['TOTAL_POP'] = data_frame['TOTAL_POP'].apply(pd.to_numeric)
    keep_list = ['GEO_ID', 'NAME', 'TOTAL_POP', 'P1_005N', 'COUNTY']
    for i in keep_list:
        col_list.remove(i)
    data_frame['P1_005N'] = data_frame['P1_005N'].apply(pd.to_numeric)
    sum_column = data_frame['P1_005N']
    for col in col_list:
        data_frame[col] = data_frame[col].apply(pd.to_numeric)
        sum_column = sum_column + data_frame[col]
    data_frame["TOTAL_NATIVE"] = sum_column
    return data_frame
```

[NOT DONE - going to add more demographic data]

COVID Data
Figure 4. Data Flow Diagram of preprocessing COVID datasets.



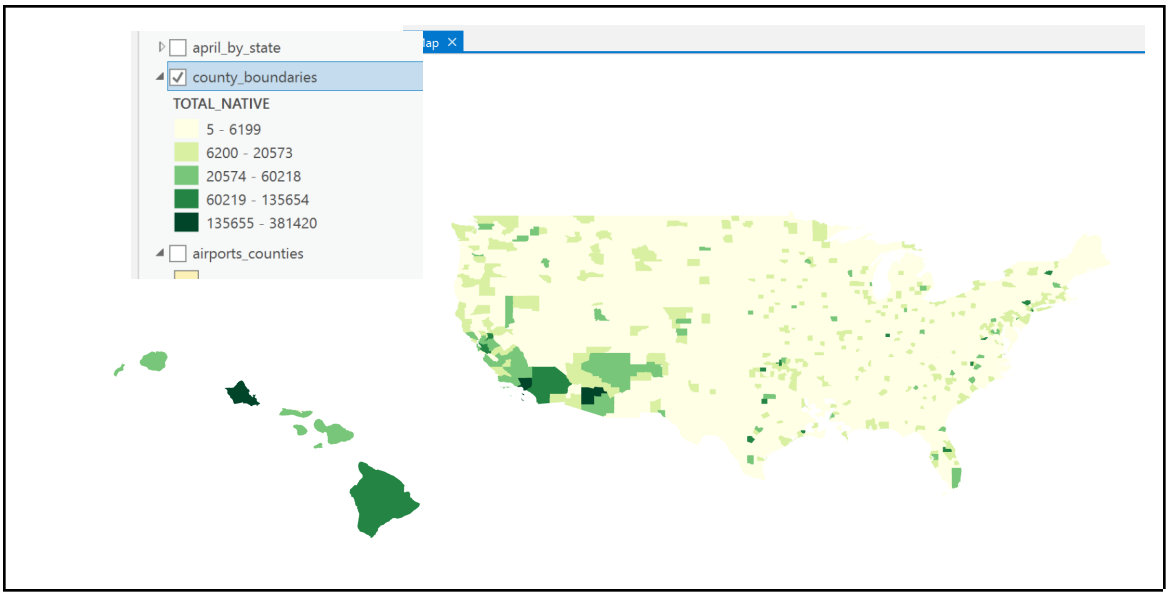
[NOT DONE]

Results - Visualization

Native Population by County
I used the cleaned CSV of the 'Race' demographic from the Census focused on Native populations by county to join to the shapefile of the county boundaries from the Census Bureau. This allowed me to create a visualization of the racial data by county.

I am also planning to join this data to the spatial flight data to normalize the flight data by population of people identifying racially as Native.

Image 3. Visualization of the Native Population by County of the US with Hawaii enlarged.



Number of flights by County
[NOT DONE]

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. This allowed me to create visualizations of each month.

I also joined this data to the table of the summary statistics of the number of airports per state and used that column to normalize the data to see how it affected the visualization.

Image 4. Number of Flights Arrived by State in March, 2021 (not normalized).

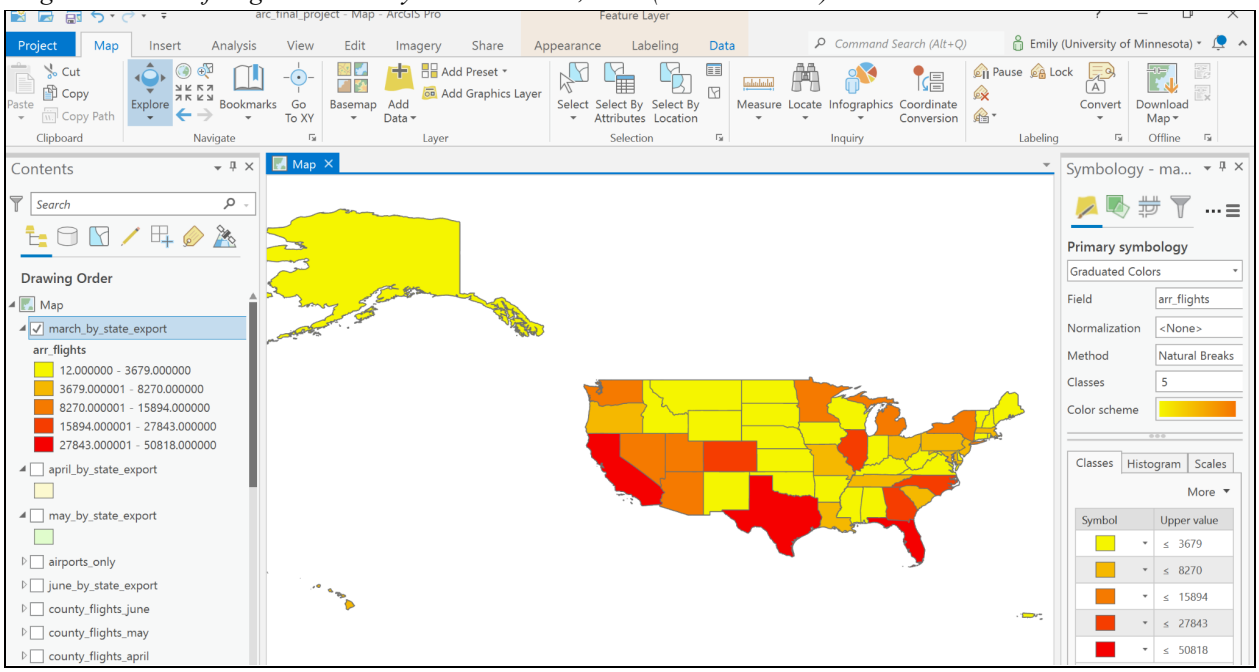
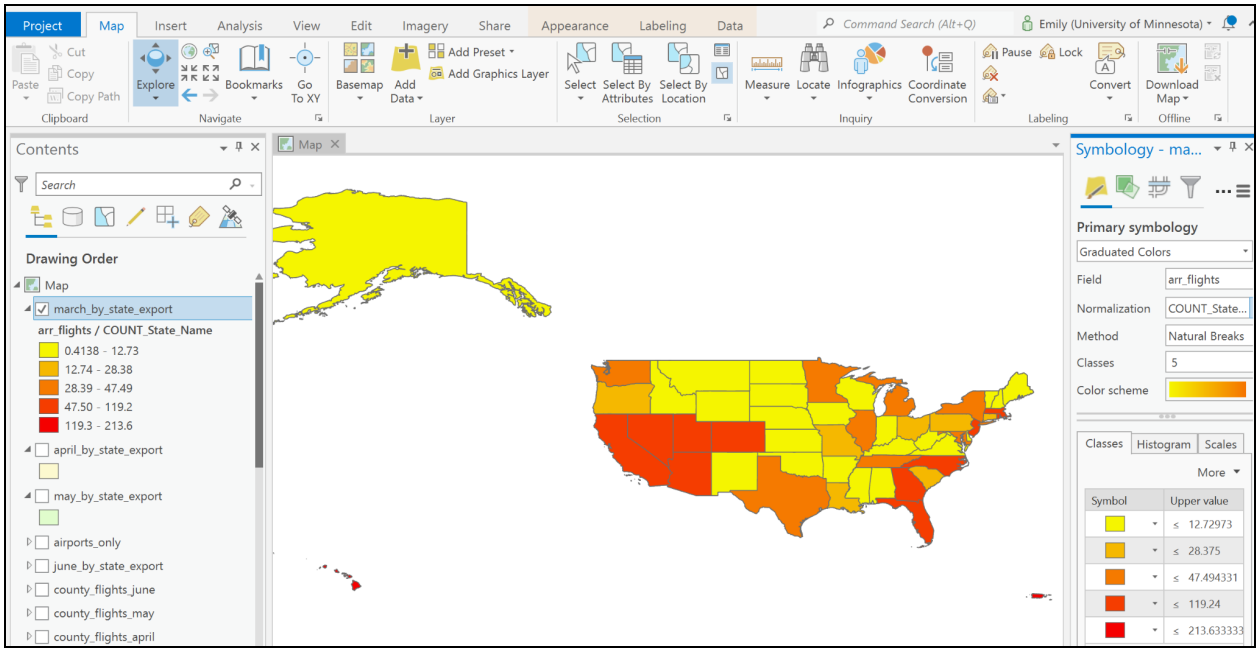


Image 5. Number of Flights Arrived by State in March, 2021 (normalized by count of airports).



Results Verification - Modeling

Multiscale geographically weighted regression (MGWR) modeling is done to test the incidence rates between the variables.

figure 10. Data flow diagram of MGWR modeling.

[NOT DONE].

Discussion and Conclusion
[NOT DONE]

References
[IN PROGRESS]

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.*

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