# IMPACT OF COVID-19 ON AIPORT TRAVEL IN THE UNITED STATES AND CANADA

Contreras, Wally
Le, Cody
Sandoval, Iliana
Tian, Chloe
Team: 4Travel

DSC 465 | Winter 2022

March 15, 2022

## Table of Contents

## Introduction

During the first year of the COVID-19 pandemic, countries enacted various measures to mitigate the spread of the virus. In the United States, federal and state mandates including travel restrictions, border closures, lockdowns, and mask requirements were enforced as part of public health mitigation efforts. With lockdowns in place and multiple travel bans locally and abroad, what happened to the travel and aviation industry? With various states experiencing different levels of community spread and adopting different social distancing measures, which airports were most affected by the pandemic? In this analysis, we visualize the impact of the COVID-19 pandemic on airport travel and compare how the pandemic has affected travel levels between the United States and Canada. Through geospatial and time-series visualization techniques, we compare how travel changed throughout 2020 by state/province, by city, and by airport.

The dataset for this analysis comes from Kaggle and was curated by Terrance Shin, a Toronto based data scientist, with data from Geotab Inc., a commercial telematics provider. The dataset shows airport traffic by day, over 8 months, from March 16th through December 2nd of 2020, at various international airports in the United States, Canada, Chile, and Australia. The dataset contains 11 variables and 7,247 observations, where each observation represents a day for one of the 28 different airports. Two variables are identifiers and were not used in the analysis. There are 6 categorical variables, 1 ordinal variable, and 3 numeric variables. The categorical variables represent names of airports, cities, states, or countries. The ordinal variable represents the date. Out of the 3 numeric variables, two are geography points or coordinates. The variable *Geography* representing polygon coordinates that represent a mapping of each airport was also not used in the analysis. The remaining numeric variable, *Percent of Baseline,* is the primary response variable. *Percent of Baseline* is a metric that measures the relative traffic on a given day compared to the same day of the baseline period, this number is expressed as a percentage. The baseline period used for this dataset is February 1st to March 15th of 2020.

*Original Attributes from the Dataset:*

| Attribute | Type | Description |
|---|---|---|
| Aggregation Method | Identifier / Categorical | Aggregation period used to calculate this metric. |
| Version | Identifier / Numeric | Version number for the data. |
| Airport Name | Categorical | Name of airport (28 names total) |
| City | Categorical | The city within which the airport is located. |
| State | Categorical | The state or province within which the airport is located. |
| ISO_3166_2 | Categorical | Two-character code that represents the location of the airport. The United States includes the state abbreviation in the code and Canada includes the province abbreviation in the code. |
| Country | Categorical | The country within the airport is located. |
| Date | Ordinal | Date formatted as YYYY-MM-DD. |
| Centroid | Numeric / Geography | Geography point representing the centroid of the airport polygon. |
| Geography | Numeric / Geography | Polygon coordinates that represent a mapping of each airport. |
| Percent of Baseline | Numeric | Proportion of trips on the record date compared to the average number of trips on the same day of the week in the baseline |

*Percent of Baseline* is a key variable for creating the visualizations and drawing conclusions since this variable is the only numeric variable to determine airport traffic levels. This variable is compared against categorical variables such as country, state, city, and airport to visualize impacts of the COVID-19 on airport travel. The date variable is the second key variable for drawing conclusions, as different trends and patterns occur in the *Percent of Baseline* across time.

For the final analysis, the dataset was filtered to observations of airports in the United States and Canada. With the filtered data, there are 2,311 instances for Canada and 4,441 instances for the United States. There are 6 provinces, 9 cities, and 9 airports represented in Canada. There are 15 states, 16 cities, and 17 airports represented in the United States. The dataset contains missing values for some entries for some airports. Missing values were handled by imputation, creating weekly summaries for the *Percent of Baseline,* or displaying a new variable to show the first day of each week, starting with March 16th. This variable was then used for aggregation and became the average percentage of airport travel by week relative to the baseline period.

Several preprocessing techniques were implemented to clean and transform the data and prepare the data for visualization. First, the *Date* variable was transformed to a date type then mutated into numeric variables *Year, Month*, and *Day*. The *Month* variable was converted to factors with levels in chronological order. Second, the latitude and longitude characteristics were extracted from the *Centroid* variable which originally read as a string variable. Third, the abbreviation for states and provinces were extracted from the *ISO_3166_2* variable and added to names of states and provinces. Lastly, country names were edited and modified for readability in visualizations.

*Additional Attributes Created from the Dataset:*

| **Attribute** | **Type** | **Description** |
|---|---|---|
| Year | Numeric | Numeric number representing Year. |
| Month | Numeric | Numeric number representing Month. |
| Day | Numeric | Numeric number representing Day. |
| Week | Numeric | First day of the week. |
| POB | Numeric | Average percentage of baseline by week. |
| long | Numeric | Longitude coordinate for participating states and provinces. |
| lat | Numeric | Latitude coordinate for participating states and provinces. |
| MonthX | Categorical | Month represented as levels in chronological order by month name. |

There are several limitations to the dataset. First, the main response variable, *Percent of Baseline*, is pre-aggregated and raw data including data on flights and trips are not available. Instead, the measure to determine airport travel is a relative comparison calculated by the number of trips recorded at the airport each day, compared to the average number of trips for the same day of the week during the pre-covid baseline period. Second, data is unavailable for the period before March 16, 2020, which means that for this analysis, we are unable to make comparisons between distribution of airport travel prior to March 16, 2020. Third, the dataset only contains data for 8 months of the COVID-19 pandemic in 2020, which means that this analysis is limited in scope and the results only show a snapshot of the impact of COVID-19 at the specified time. These limitations were kept in mind when determining the audience and message for the visualizations.

## Message and Audience

By focusing on the numeric variable, *Percent of Baseline*, the fluctuations in airport traffic across the dataset's timeframe could be seen from different perspectives, from large scope to local impacts, by country, state/province, city, or airport. Overall, the visualizations show that COVID-19 had a greater impact on air travel in the United States than Canada during the first year of the pandemic. Most importantly, using the *Percent of Baseline* as a metric showed that airport travel levels improved and returned to pre-covid levels in Canada by the end of the year, but not in the United States. By contrasting the airport traffic levels between the two countries, our story is developed and focused on the aviation industry in the United States.

Our audience is the Federal Aviation Association (FAA), the largest transportation agency in the United States, and their affiliated government officials. We are representing the Airport Council International (ACI), the largest organization representing local, regional, and state governing bodies that operate airports in the United States and Canada. The federal government is considering granting additional federal support to the aviation industry due to the COVID-19 pandemic. The FAA needs data to determine why additional support should be granted and if granted, how to allocate the additional support. This analysis will show that additional support should be granted to the aviation industry in the United States since by the end of 2020, airport travel levels were unable to recuperate to pre-covid levels. As the data shows that the airport travel levels did recover to pre-covid levels in Canada, this comparison provides evidence of why additional support should be granted. The additional support may be allocated based on states, cities, or airports that were most impacted.

## Exploratory Analysis

In the exploratory analysis, the key variables of interest were *Percent of Baseline* and *Date*. Since the dataset contains only one numeric variable, a time variable, several categorical variables, and geography variables, geographical and time-series visualization techniques were most appropriate for the dataset. In researching the dataset, we discovered three types of analyses that had already been created by others: geographical, time-series analysis, and summary statistics. Geographical visualizations included choropleths and glyph maps showing the *Percent of Baseline* by country and state. Time-series line graphs and heatmaps have been performed showing the *Percent of Baseline* by airport by time. Lastly, violin plots and density plots have been performed to show the differences in distributions of *Percent of Baseline* by country.

Using the two key variables, the distribution of the *Percent of Baseline* was visualized with a ridgeline plot as this would allow us to see changes in the distribution over time. In addition, the distribution was further explored with violin plots but instead of creating violin plots for each country which others have performed, violin plots were created for each state to explore the relationship between *Percent of Baseline* and *State*. Counts of observations did not reveal any meaningful trends, instead time and *Percent of Baseline* became the focus.  Further exploration with time was explored through a heatmap to examine the relationship between *Percent of Baseline* and *City*. These techniques were beneficial as they showed trends across specific times and can handle variables with many categories, and in this case, variables such as *Airport Name* and *City* have a high number of categories. Lastly, choropleths were created for each country to visualize the *Percent of Baseline* by state or province and compare the differences between countries. This technique was beneficial since the data contained variables for areas of geographic regions, and choropleths are best at showing differences in trends by region. This exploration led to the decision to focus the analysis on two countries, the United States and Canada, since these two countries showed the most variation in trends by region.

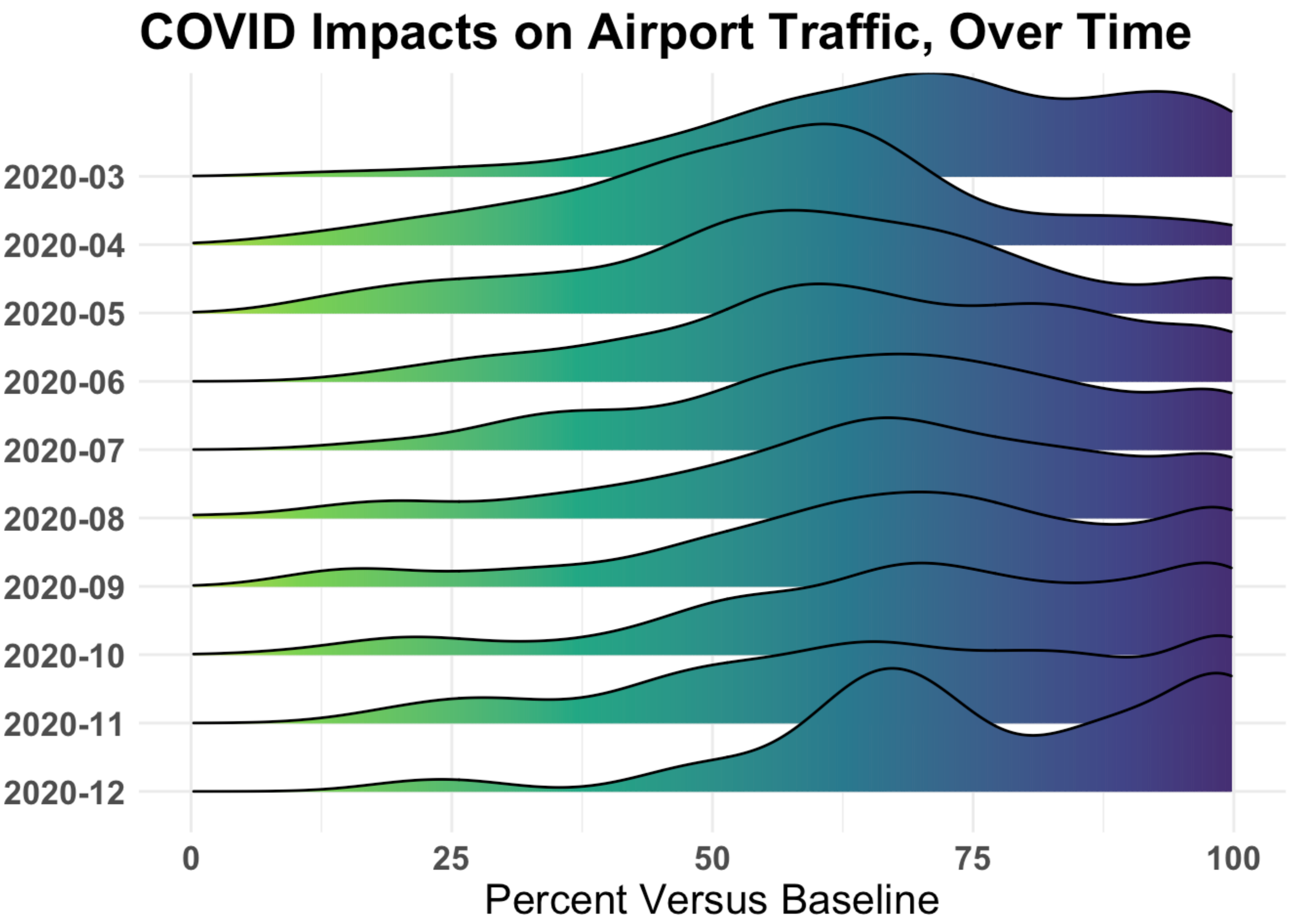


*Figure E1: Initial Exploratory Ridgeline Plot of Distribution of Percent of Baseline*

In this exploratory ridgeline plot, we see that the distribution of *Percent of Baseline* started as a bimodal distribution in March, changed to a near-normal distribution in April as more lock-downs took effect, and then slowly reverted to a bimodal distribution over the remaining months. The full dataset was used for this visualization, meaning all countries were included. Looking into the underlying cause for the changes in distribution led us to uncover that the peaks were driven by differences between American and Canadian airports. *Figure E1*, was part of the reason for focusing our analysis on these two countries.

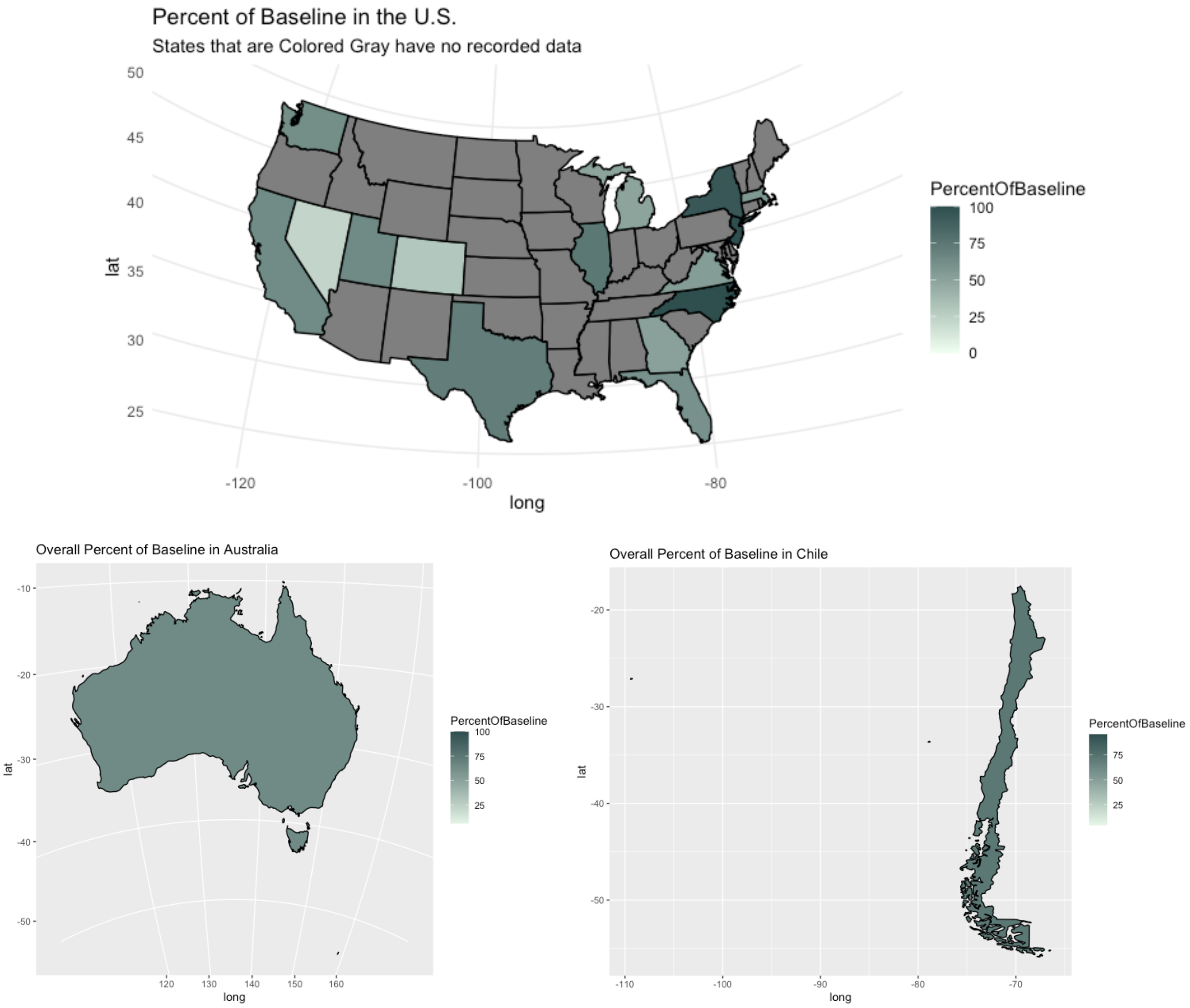


*Figure E2: Initial Exploratory Choropleths by Country by Percent of Baseline*

These exploratory choropleths shown in *Figure E2* were created from the *mapdata* library in R. Displaying all countries at once was a challenge while initializing the choropleths. While it was possible to have a graph that displayed all four countries, the story seemed compromised. At the time, the only solution was to show all participating countries through a world map. If a world map was used, the visualization would be difficult to understand, and it would also have a lot of grey areas since the dataset does not have information for the remaining +190 countries. The dataset had more observations representing states and Canadian provinces, and only one observation each for Australia and Chile. Upon initial research, it was discovered that the province in Chile is one of the six provinces of the Santiago Metropolitan Region. In terms of graphing, it would have been a small region to highlight and not representative of the entire country. This ultimately led to the decision to focus the analysis with only the observations in United States and Canada. It was decided that the next iterations of visualizations should be animated to show how airport traffic changed for states and provinces over time.

### Final Visualization #1: Ridgeline Plot

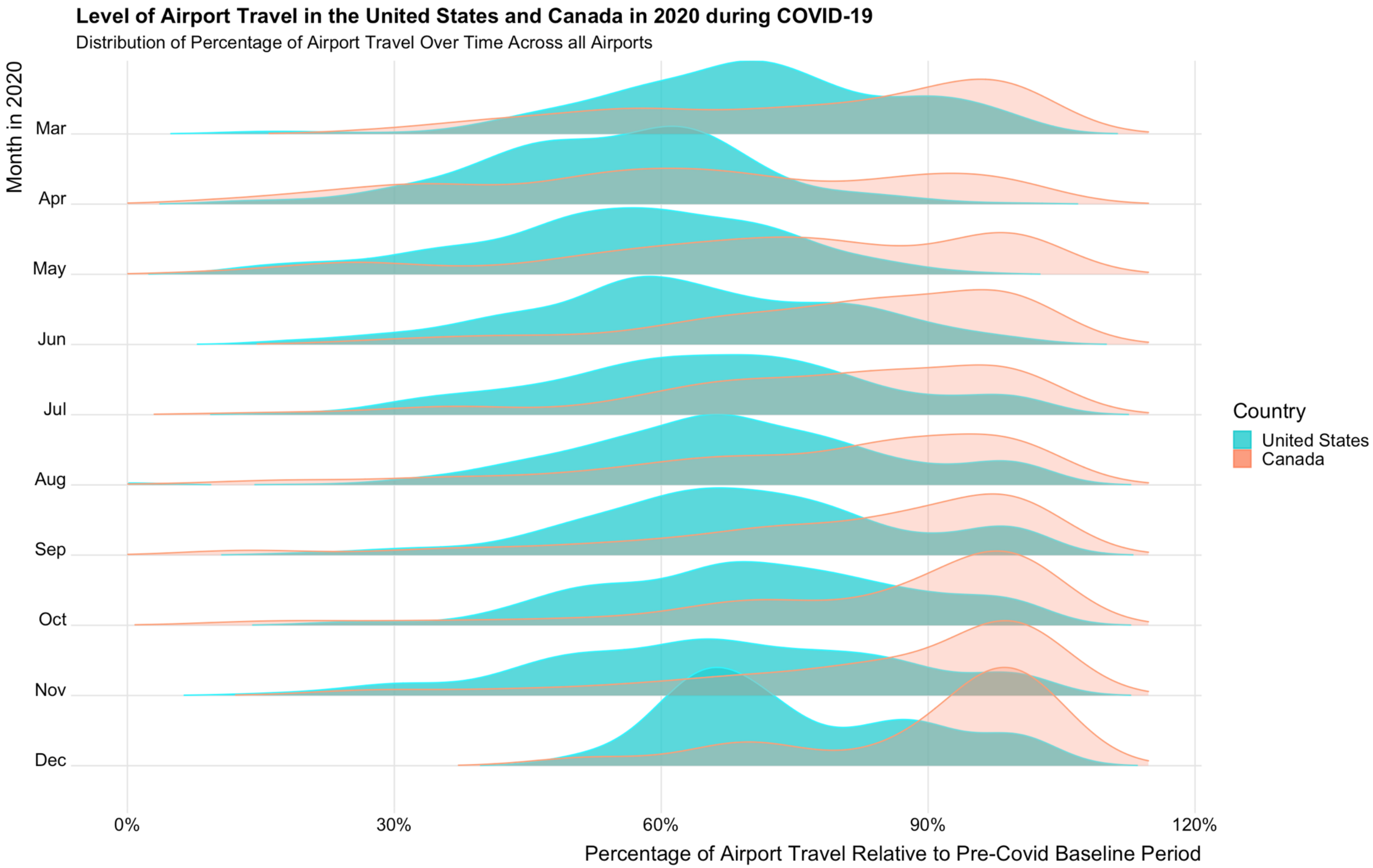


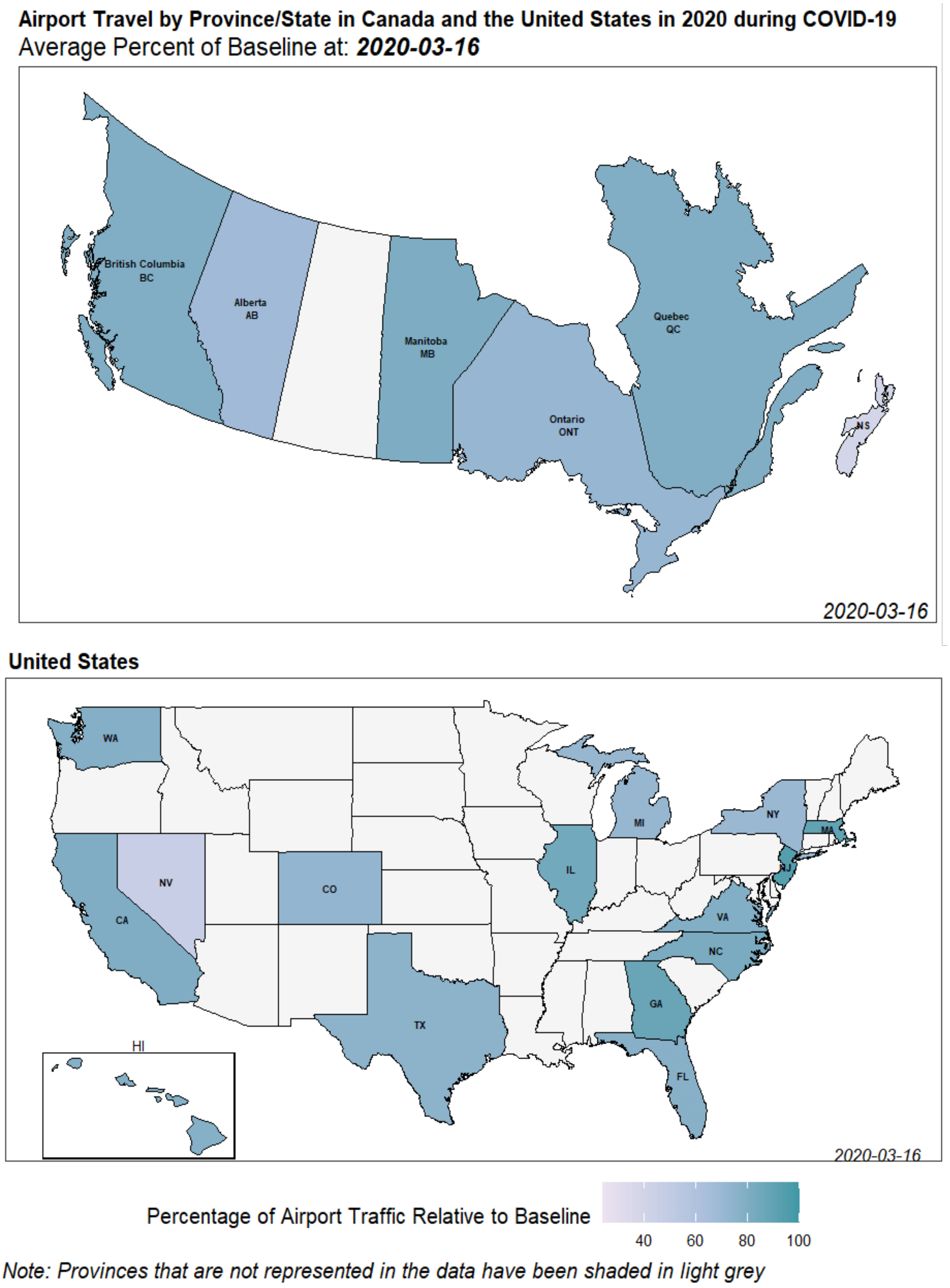
*Figure V1: Ridgeline Plot for Distribution of Airport Travel Levels Between United States and Canada*

*Figure V1* shows the distribution for the *Percent of Baseline* over time through a ridgeline plot, which is an optimal visualization type to show summary statistics and trends that occur along the year. The ridgeline plot is created in R using the *ggplot* library and the *ggridges* library. In the first version of the ridgeline plot, the distribution for the *Percent of Baseline* was mapped for all countries at once. After narrowing the analysis to two countries, two separate distributions were mapped to the plot, one for the United States and the other for Canada, each having their own layer using *geom_density_ridges*. The color and fill were added to each layer to differentiate the two countries using a complementary color palette with turquoise for the United States and salmon for Canada based on the color wheel. For aesthetic purposes, a lighter shade of each fill was used for the color, the height and transparency were modified so that overlap between the two countries display well visually.

The x-axis, a continuous scale, represents the *Percent of Baseline*, and the y-axis, a discrete scale, represents the *Date* in months. The *MonthX* variable used for the y-axis was engineered and transformed from a numeric variable into a categorical variable representing each month in chronological order. The y-axis was scaled to reverse the order on the axes starting with March on the top and December on the bottom and the names of the months abbreviated to three letter abbreviations for consistency with all other visualizations in the analysis. The *scales* library was used to scale the x-axis with a modified limit to display the scale in percentage format. In the final version of the plot, the legend was created to match the fill and color of the two countries by overriding the legend guide and manually scaling the breaks and labels. The x-axis was relabeled to provide a clearer understanding of the meaning of *Percent of Baseline*. The title was enhanced, and a subtitle was added to better represent the distributions and the message shown by the ridgeline plot. Lastly, *theme_ridges* was used as the theme since this theme in *ggplot* is native to ridgeline plots and visually enhances the plot.

The ridgeline plot compares the percentage of airport travel relative to the pre-covid baseline period between the two countries. The means of the distributions show that the United States operated at around 60% of the baseline period from April through June and close to 70% of the baseline period between July through December. The range of distribution varies widely month by month. Canada also operated with varied distribution during the first month of the pandemic, but in contrast starting in May, the mean distribution is closer to 90%. With each subsequent month until the end of the year, Canada's airport travel levels were relatively the same level compared to the pre-covid baseline period and almost at full capacity by December. Overall, COVID-19 had a greater impact on airport travel in the United States than Canada. By December, the peak of the distribution for the United States remained closer to 60% which shows that the airport travel levels were unable to recover to the pre-covid levels by the end of the year. By contrasting the distribution of the airport travel levels between the two countries, the visualization provides support to our story and indicates that there may be a need for additional federal support to the aviation industry in the United States, since the United States was unable to recuperate to pre-covid levels by the end of the year as shown in the ridgeline plot.

### Final Visualization #2: Animated Choropleths



Airport Travel by Province/State in Canada and the United States in 2020 during COVID-19

Average Percent of Baseline at: ***2020-03-16***

*Note: Provinces that are not represented in the data have been shaded in light grey*

*Figure V2: Animated Choropleths by State/Province Over Time*

*Figure V2*, shows animated choropleths to visualize airport travel by state or province for all 37 weeks. Both animated choropleths show airport traffic on a sequential color palette from Color Brewer, 'PuBuGn'. As the color shifts from a light purple to a darker teal, the audience is made aware of the rise and fall in airport traffic from March throughout the rest of 2020. Variables are mapped by province/state for the week. Because the airport traffic is represented as a daily interval in the dataset, the weekly average is calculated from Monday to Sunday to create the weekly average airport traffic based on *Percent of Baseline*.

From the first version of the choropleth that was created, several updates were made to enhance the visualization. Beginning with the animation portion, the visualizations were animated using the *animate* function in *gganimate* library. The transition time used is the weekly average of the airport traffic. The legend has a light purple color to signify a lower *Percent of Baseline*. As airport traffic increases, we see the color scale change to a darker teal. The color scale used helps identify noticeable changes in an easier manner through the animation. Next, an individual map of Hawaii was created and animated then added to the United States' mainland graphic using the *draw_plot()* in the *cowplot* libary. This was done so that the audience was able to view the United States and Hawaii together without affecting the size of the graph.

Another enhancement made to the graph is the theme used. The choropleths are using the black and white theme, *theme_bw()*. After implementing the theme, gridlines were removed. Because the two countries geographical data comes from two different sources, their longitude and latitude coordinates were not the same. To avoid confusion, the axes and their titles were removed from both graphs. For the final version of the visualization, labels were added for each state to better identify the states. For Canada, the full name as well as the abbreviation of each province were added. These labels were created using a function that calculates the centroid for all participating states/provinces. From there, adjustments were made to Florida and Michigan's coordinates so that the labels were centralized. Lastly, the size of the legend was increased using the *unit()* function. Also, the font of the time frame was bolded and italicized to make it easier to identify in the visualization. A caption was included to let the audience know that states/provinces shaded in a light grey color signified that there was no data for that region. Finally, both graphs were stacked using the *image_append()* function given by the *magick* library. A loop function was then created to animate a stacked image. From there, the function *image_write_gif()* was used to save the final visualization which can be viewed as a moving image in file name *canada_us_vis.gif.*

The animated choropleths reveal that the airport travel levels that provinces in Canada overall have higher averages compared to states in the United States. For Canada, all provinces experience a lower volume in airport traffic in March and April, this may suggest that Canadians took the initial lockdown in earnest. Thereafter, provinces like Ontario, Manitoba, Quebec, and Alberta experienced higher averages through the rest of the year. British Columbia is the only province that seems to have a lower airport traffic. here seems to be a pattern among all provinces where they go through a continuous trend of a sudden lower-than-their-own normal in airport traffic after seeing a peak just weeks before.
By the end of the year, all provinces except Novia Scotia display averages closer to darker teal.

In the United States, we see a rapid decrease in airport traffic from March to April. This light purple and muted-blue color remain that way for the most part up until July when it experiences a small surge. The choropleth shows that the states that remained relatively low averages were Nevada, Michigan, and Virginia. One observation worth mentioning is that Hawaii and New Jersey's airport traffic maintained a higher average throughout the 37 weeks. As the animation continues throughout the year, the airport traffic never really returns to the low percentage of baseline it once had in the beginning of the pandemic. Instead, by the end of the year, all states have varying colors, showing that all states were affected differently by the pandemic, with most of the states displaying averages closer to lighter-purple and muted-blue.

### Final Visualization #3: Violin Plots

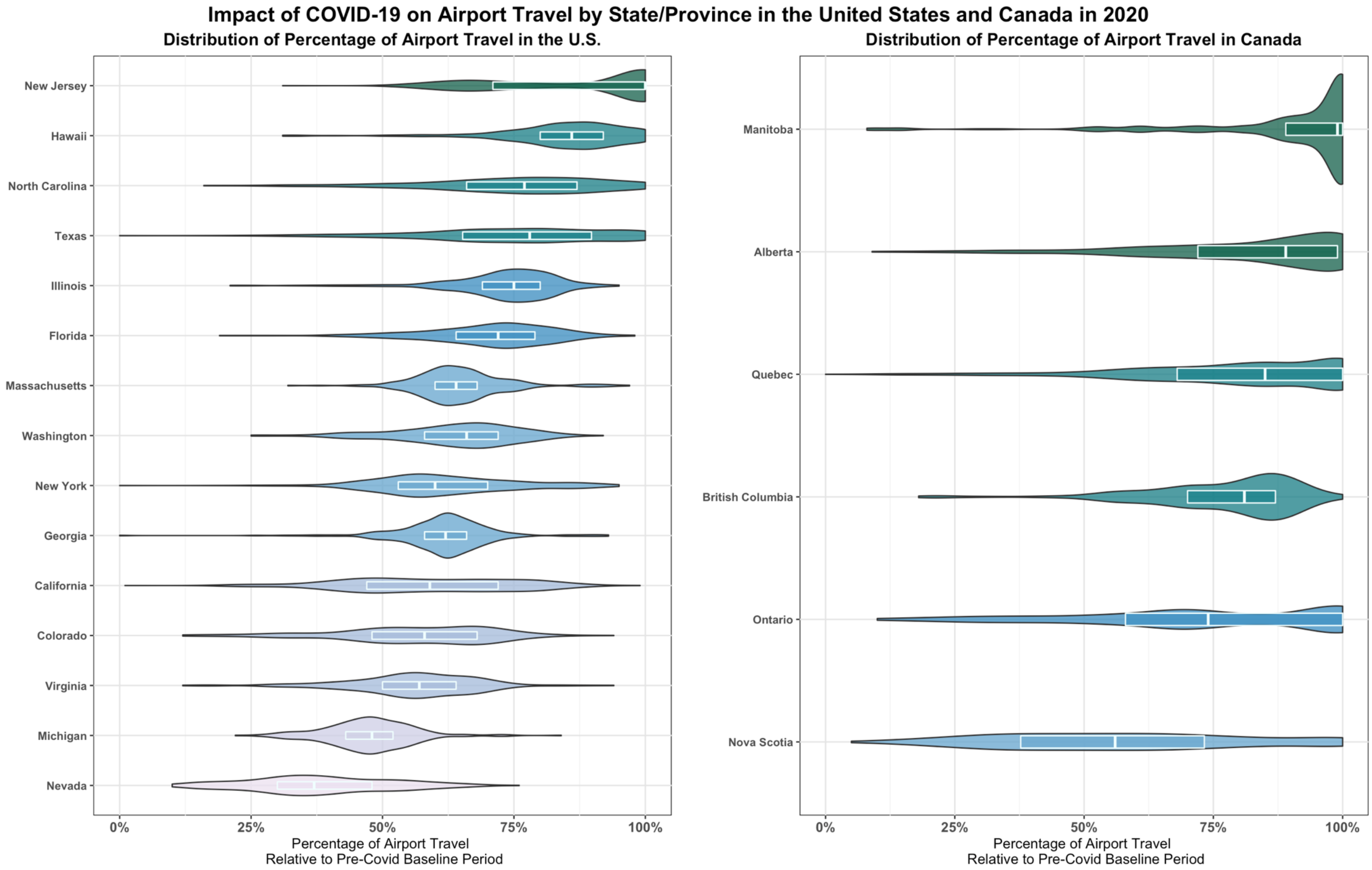


*Figure V3: Violin Plots for Distribution of Airport Travel by State/Province*

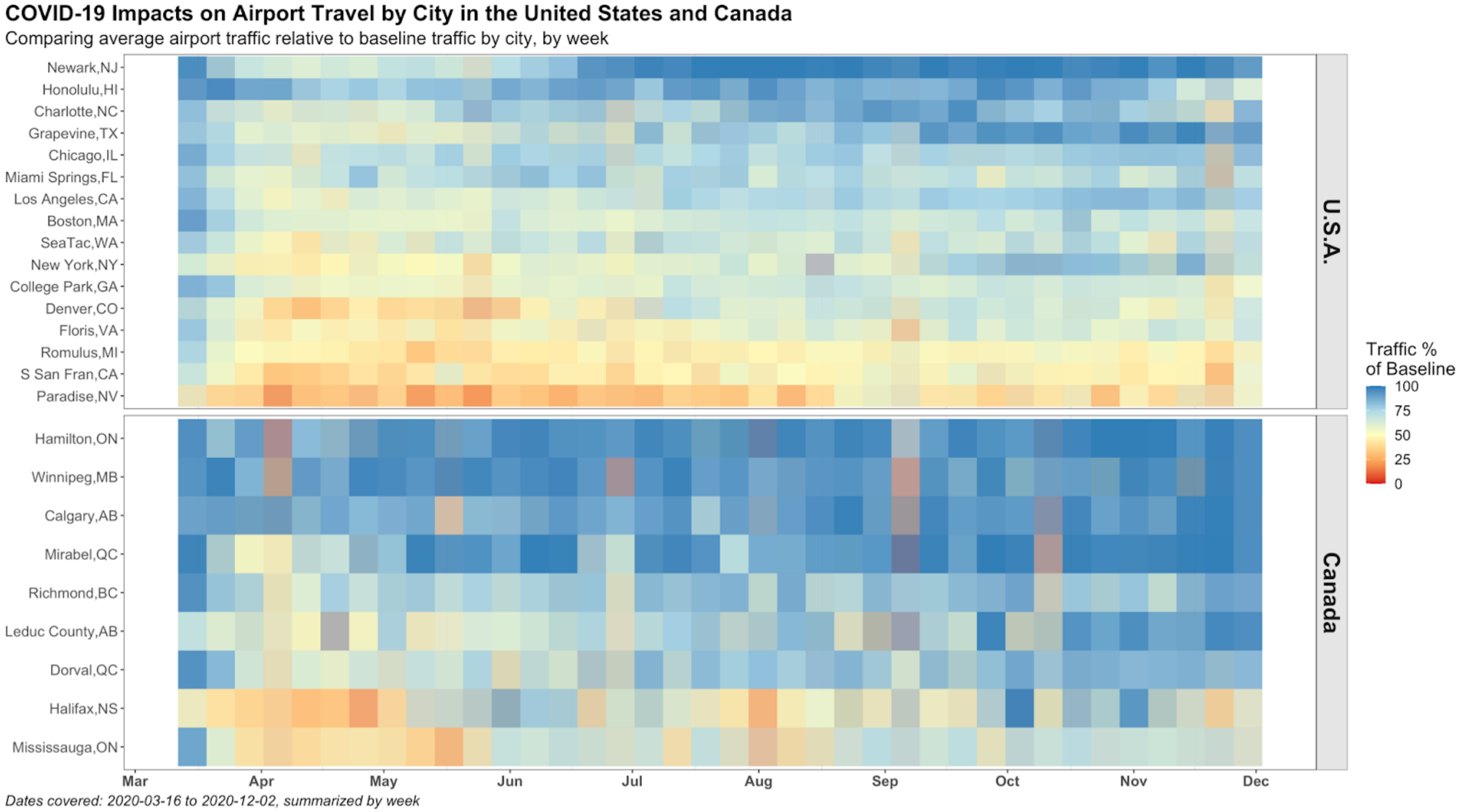
*Figure V3*, shows violin plots to highlight in detail the comparison by state/province since violin plots are useful for comparisons across variables containing many categories. Violin plots for each country are created separately in R using the *ggplot* library. In the first version of the violin plot, the distribution of the *Percent of Baseline* was mapped for all states for all countries. The second version focuses on two separate plots, one for each country, making it clearer to see a pattern by region. The violin plots contain boxplots without whiskers inside each violin to easily see the distribution and the median for each violin. The *gridExtra* library was then used to arrange the two violin plots side-by-side for comparison.

The x-axis, a continuous scale, represents the *Percent of Baseline*, and the y-axis, a discrete scale, represents the *State,* which includes both states and provinces. The variable *State* is reordered by *Percent of Baseline* on the y-axis so that each state or province is shown in descending order with the state or province on top having the highest percentage of airport travel relative to pre-covid baseline period and the lowest percentage at the bottom. This reordering of the states and provinces allows the visual to easily show the highs and lows in airport traffic levels between each state or province. The *scales* library was used to scale the x-axis to display the scale in percentage format.

In the final version of the plot, the fill color of each violin was manually scaled. Since this visualization is meant to be a complementary to the animated choropleths as both focus on visualizing the relationship between *Percent of Baseline* and *State*, the fill color is manually scaled to match the colors used in the choropleths. The fill color is a sequential palette from Color Brewer, ‘PuBuGn’. The palette is manually scaled based on median distributions of *Percent of Baseline*. The legend is removed from both plots since it did not add to the overall visual since the state and province names are displayed on the y-axis. The panel background, panel grids, and axis texts were all manually scaled for aesthetic purposes. Lastly, the *ggpubr* library was used to add text and a main title for the arranged violin plots.

The violin plots show that almost all provinces in Canada had median distributions above 75% with only Nova Scotia at around 50%. In contrast, the states in the United States varied widely in the distribution of percentage of airport travel with Nevada and Michigan having median distributions below 50% and New Jersey and Hawaii with well above 75%. Moreover, the impact of COVID-19 on airport travel varied widely in the United States by state. States such as New Jersey, Hawaii, North Carolina, and Texas were able to maintain a relatively high percentage of airport travel and had similar levels of airport travel compared to pre-covid levels. States such as Colorado, Virginia, Michigan, and Nevada had much lower percentages of airport travel with the remaining states somewhere in between. By comparing the distribution of airport traffic by state, the violin plot makes it clear which states in the United States had the most impact from COVID-19. This visualization provides support to our story as it shows that on a regional level certain states had much lower levels of airport traffic compared to other states and the difference between airport traffic between states could be considered in allocating additional federal support.

### Final Visualization #4: Heat Map



**COVID-19 Impacts on Airport Travel by City in the United States and Canada**

Comparing average airport traffic relative to baseline traffic by city, by week

*Dates covered: 2020-03-16 to 2020-12-02, summarized by week*

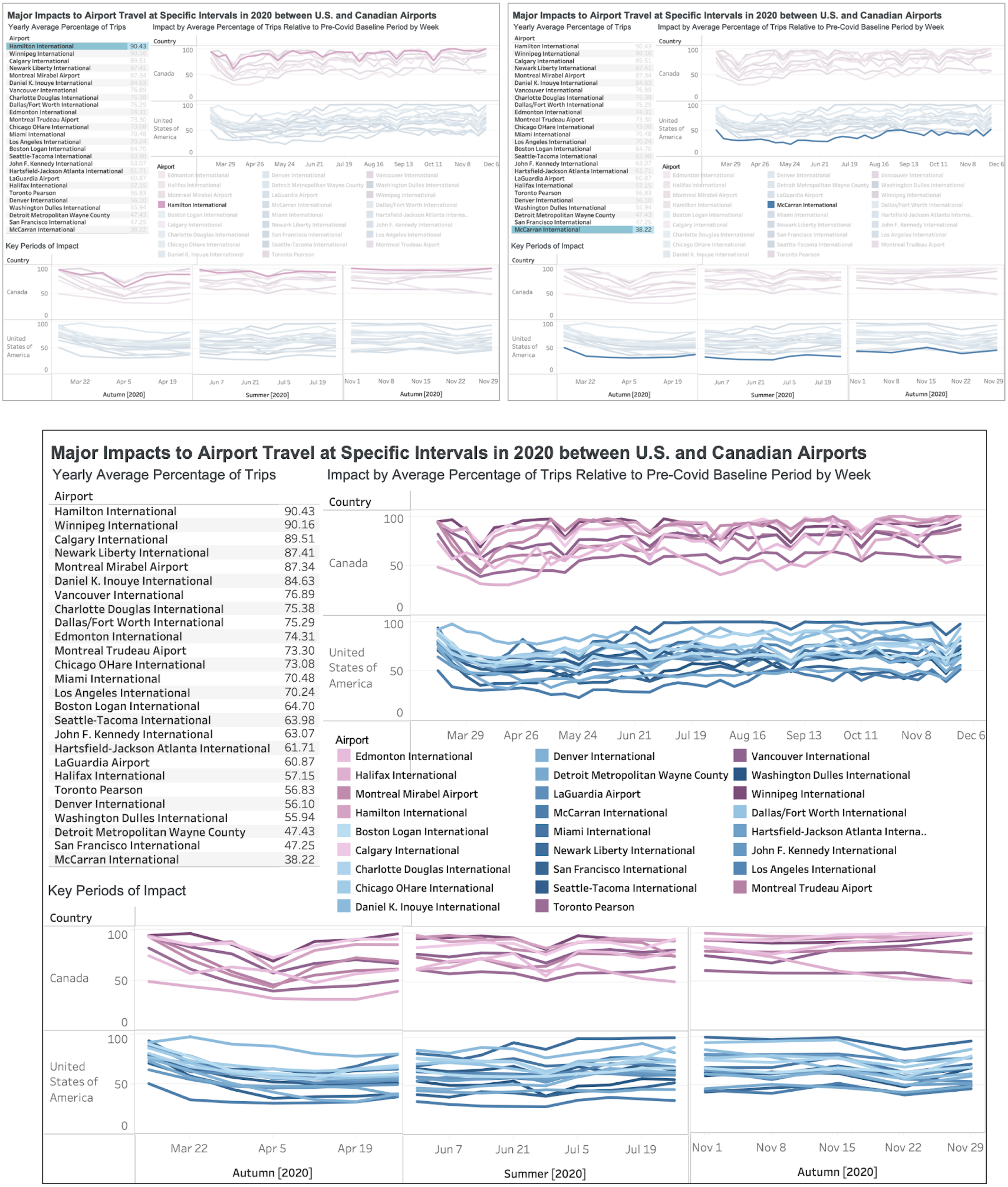
*Figure V4: Heat Map for Airport Travel by City in the United States and Canada*

The heat map in *Figure V4*, shows the average airport traffic as a *percentage of the baseline* across time. Each block on the chart represents a week for the airport on the y-axis, color is used to encode the average airport traffic as a *percentage of the baseline*: the darkest red color shown on the legend indicates that the airport operated at zero percent for the week, the darkest blue color indicates that the airport operated at 100 percent for the week, most airports operated somewhere in between. The cities are arranged in descending order by their average airport traffic as a *percentage of baseline* so that the least impacted cities are at the top and the most impacted cities are at the bottom.

This chart was created using the *ggplot* library in R and customized with the “RdYlBu” divergent color palette obtained from Color Brewer. The exploratory version of this chart was unfiltered, unaggregated, and sorted in ascending order so that the most impacted cities were listed on top; this showed us daily observations and revealed that data were missing for a few airports. For the final visualization, data were aggregated into weekly averages to keep us from having to impute data or drop airports with missing data from the dataset. Facets were added to group cities by country and highlight the overall difference. To aid the audience with visual decoding, the cities were ordered in a descending order so that the most impacted cities are now shown at the bottom; the color palette was also updated from a continuous palette to a divergent one. Labels were also reorganized and resized, a state abbreviation appended to the state name, and a subtitle and caption added.

From *Figure V4***,** we see that the U.S.A facet is mostly between yellow and red, and that the Canadian facet is mostly blue, implying airports in the United States were more affected. We can see that airports in the two countries were impacted similarly towards the beginning of the pandemic from March to late May but then started diverging. This chart also allows us to dig a little deeper to point out the most affected cities. We can observe that the most impacted airports were in Paradise (Nevada), South San Francisco (California), and Romulus (Michigan). Newark (New Jersey) had the least impacted airport in the United States and Hamilton (Ontario) had the least impacted airport in Canada. Overall, this heatmap reinforces the observations made in the previous figures: COVID-19 had a larger impact on airport traffic in the United States than in Canada.

## Final Visualization #5: Interactive Dashboard



### Major Impacts to Airport Travel at Specific Intervals in 2020 between U.S. and Canadian Airports

Yearly Average Percentage of Trips

| Airport | |
|---|---|
| Hamilton International | 90.43 |
| Winnipeg International | 90.16 |
| Calgary International | 89.51 |
| Newark Liberty International | 87.41 |
| Montreal Mirabel Airport | 87.34 |
| Daniel K. Inouye International | 84.63 |
| Vancouver International | 76.89 |
| Charlotte Douglas International | 75.38 |
| Dallas/Fort Worth International | 75.29 |
| Edmonton International | 74.31 |
| Montreal Trudeau Aiport | 73.30 |
| Chicago OHare International | 73.08 |
| Miami International | 70.48 |
| Los Angeles International | 70.24 |
| Boston Logan International | 64.70 |
| Seattle-Tacoma International | 63.98 |
| John F. Kennedy International | 63.07 |
| Hartsfield-Jackson Atlanta International | 61.71 |
| LaGuardia Airport | 60.87 |
| Halifax International | 57.15 |
| Toronto Pearson | 56.83 |
| Denver International | 56.10 |
| Washington Dulles International | 55.94 |
| Detroit Metropolitan Wayne County | 47.43 |
| San Francisco International | 47.25 |
| McCarran International | 38.22 |

Impact by Average Percentage of Trips Relative to Pre-Covid Baseline Period by Week

Country
Canada
United States of America
100
50
0
Mar 29
Apr 26
May 24
Jun 21
Jul 19
Aug 16
Sep 13
Oct 11
Nov 8
Dec 6
Airport
Edmonton International
Halifax International
Montreal Mirabel Airport
Hamilton International
Boston Logan International
Calgary International
Charlotte Douglas International
Chicago OHare International
Daniel K. Inouye International
Denver International
Detroit Metropolitan Wayne County
LaGuardia Airport
McCarran International
Miami International
Newark Liberty International
San Francisco International
Seattle-Tacoma International
Toronto Pearson
Vancouver International
Washington Dulles International
Winnipeg International
Dallas/Fort Worth International
Hartsfield-Jackson Atlanta Interna..
John F. Kennedy International
Los Angeles International
Montreal Trudeau Aiport

Key Periods of Impact

*Figure V5: Interactive Dashboard for Comparison of Airport Traffic by Airport Over Time*

*Figure V5*, shows an interactive dashboard created in Tableau. The dashboard includes the yearly average percentage of airport travel sorted in descending order and the average percentage of airport travel by week for all airports in the United States and Canada as a time series line graph. This visualization started out as line graphs of airports for all countries, then line graphs for airports for the United States and Canada, and lastly line graphs for the two countries separated at different time periods. To see clearer trends between the two countries, time-series panel plots were created for the *Percent of Baseline* versus the airports for March to April, June to July, and November. The time-series panel plots were enhanced through color selection to provide better comparison between the two countries. The color palette selected is native to Tableau, with blue gradient for the United States and purple gradient for Canada.

In the final visualization, using the interactive dashboard to lay out the time-series panels made the visualization easier for the audience to understand. Because of the interactivity in the dashboard, it is easy to select the airport, timeline, or average percentage of baseline and the connecting information will pop-up making it useful to see trends at various airports. In this way, the gradient color scheme shows the different shades of blues representing the United States airports and the different shades of purple representing the Canadian airports. The Key Periods of Impact shows Spring (March and April), Summer (July and August), and Autumn (November and December). This shows a clearer compassion of the average percentage of airport travel between the United States and Canadian airports at key impacted periods. When looking at the yearly average percentage of trips at all airports, the Canadian airports have the highest yearly average with three of the airports averaging above 89%. In contrast, the U.S. airports have the lowest yearly average with Detroit, San Francisco, and McCarran (in New York), averaging below 50%.

Using the Dashboard, when selecting Hamilton International, all information about this airport is displayed showing an average parentage of the baseline is 90.43%, compared to McCarran International, showing the lowest average percentage of the baseline at 38.22%. The dashboard, not only analyze the airport traffic information directly, but shows a variety of displays, calculation, filtering, and grouping of data through the controls in Tableau. The Dashboard presents critical information to the user and highlights key important aspects for the user through various levels of filters.

## Analysis and Discussion

COVID-19 had a greater impact on airport travel in the United States than Canada. By comparing the distributions of airport travel between the two countries in a ridgeline plot, the percentage of airport traffic was lower in the United States throughout the months in 2020. The ridgeline plot clearly showed with each subsequent month from the start of the pandemic, that the airport traffic levels increase and return to pre-covid levels in Canada, but not in the United States. When visualizing the percentage of airport travel by state and province, the comparison of choropleths between the two countries showed that Canada was able to maintain a relatively higher percentage of airport travel compared to the United States. At the state and province level, the violin plots show that the airport traffic levels are vastly different by state, especially in the United States.

Impact on airport travel at the city level was also greater in the United States than Canada. As shown through the heat map, cities in the United States and Canada had similar levels of airport traffic at the beginning of the pandemic, but then by May cities in the two countries started to diverge. By the end of the year cities in Canada maintained a relatively higher percentage of airport travel compared to the United States. Major impacts to airport travel are highlighted through the Dashboard and likely follow the waves of COVID-19, during the onset in March and April, during the summer in June and July, and towards the end of 2020. Moreover, the data confirms that the airport travel levels improved and returned to pre-covid levels in Canada by the end of the year, but not in the United States. As these impacts are ongoing due to the pandemic, the visualizations support the story and confirm that additional support could be granted to the aviation industry in the United States and can be allocated by looking at the impact levels by state, city, and/or airport.

This analysis was limited in scope as it focuses on the percentage of airport travel based on a specific baseline through a pre-aggregated metric. In the future, the analysis can be expanded by focusing on flight data including flight cancellations, delayed flights, or airport closures. In addition, the dataset only contained data for selected airports. The analysis may have different results if the dataset contained data for all airports in the United States and Canada. There is variation between airport traffic levels within the United States, in a future analysis, the comparison between West Coast and East Coast which experienced the COVID-19 pandemic differently may provide further insights. If data for 2021 was available, it could also provide a better understanding of how COVID-19 impacted airport travel long term and how the data changes considering variants and mutations from the pandemic.

## Appendix Part 1: Individual Report

### Individual Report for Cody Le

Cody's role in the project was group leader and researcher. Cody organized weekly meetings through Zoom, sent recap of the meeting minutes, and created and organized the shared G-drive and Google Docs for each milestone. Cody organized each of the Google Docs in an informative way so that each team member was aware of their expectations and goals for each milestone. After each milestone, Cody formatted each milestone submission and submitted the milestones on behalf of the team. Cody provided feedback and constructive comments to each team member regarding each iteration of their visualization and their presentation. As researcher, in the exploratory analysis phase, Cody researched visualizations that have been performed by others on the dataset and reported the findings to the team, serving as inspiration for the teams drafts and final explanatory analysis. Cody created the idea for the story, message and audience after the explanatory analysis phase and did research on the aviation industry to make the story realistic and believable. Cody maintained active communication with the team by email and through Google Docs and provided guidance to team members as needed.

Cody contributed to the team's visualizations by creating the ridgeline plot and the violin plots. These two plots represent summary statistics and were explored during the exploratory analysis by other team members. After the exploratory analysis phase, the team decided to focus on two countries which led the team to decide that including summary statistics with the ridgeline plot and the violin plots would be an important comparison for the analysis. Cody used R and various libraries to create the ridgeline plot and violin plots. The two visualizations were then enhanced with various iterations after feedback from the team including updates to color palette, axes labels, font, font size, title, subtitle, and theme.

Cody wrote the introduction for Milestone 2 and Milestone 3. Cody also wrote the introduction, message and audience, and analysis and discussion for the final report. For the final report, Cody also wrote the introductory portion of the exploratory analysis and the analysis for two of the main visualizations: Ridgeline Plot and Violin Plots. For the presentation, Cody wrote the parts on limitations, audience and message, and key takeaways, discussion, and further analysis, which are all incorporated into the final report. Cody provided feedback to team members regarding their visualizations and assisted with their presentation scripts.

From this project, I learned that data visualization is an iterative process. Like design, data visualization goes through an initial drafting stage which is the exploratory phase, then a more detailed stage, which is the explanatory phase, and lastly a final stage where the visualization is finalized with finishing touches made to enhance the message and story. The key takeaway to data visualization is to consider the audience and message in designing the visualization. The data may be applicable in various aspects, but the main goal is to extract the data to map the data into appropriate types of visualization. From this project, I had to look very closely at the visualization, think in terms of the audience, and have a keen eye for details. The coding for the visualizations in R was challenging because of the little details when it came to the axes, labels, color, and various other factors regarding aesthetics. It was a rewarding experience to understand design principles and learn about color theory as it related to data visualization. This project was different from other data science projects because in other projects, the focus is on algorithms and often the visualization aspect is not emphasized in detail as in this project. I have a greater appreciation for design now after this project and realize the importance of audience and story in designing visuals which I had not thought of prior to this project. In addition, this course emphasized critique and feedback, which was not heavily emphasized in other data science courses. Through this project, I learned to provide constructive feedback and positive critique especially when it came to design and visualization, which is important for the iterative nature of the process. Ultimately, this project was a great experience to further coding skills in R, understand the data visualization process, and understand how the design process works within a team.

## Individual Report for Iliana Sandoval

For this project, Iliana had given herself the responsibility of creating a choropleth with the data to visualize the airport traffic. Because we were about five or six weeks into the semester, she was familiar with geographical visualizations and was looking forward to the opportunity to create these maps. One roadblock that I faced was the borders for all non-U.S. countries. Being introduced to choropleths in this class, I was familiar with the U.S. map in *map_data()* that was already pre-set with state borders. I was unsure on how to include borders into Canada, Australia, or even how to represent the province of a province, like in Chile's situation.

Because the dataset did not have many variables, making an interactive map seemed like an obvious *must-do* for this project. In the beginning weeks, there were multiple issues with getting the animation going. After reading online and seeing how others fixed their errors, Iliana found herself still unable to make this animation work. It did not really make sense why this data was not animating, Iliana had taken extra steps to make sure the date was in the correct format and condensed the data to a weekly average, thinking that the data may have been too large to animate. After discussing with the group, we considered the possibility of just making an interactive plot with *plotly* and seeing if other group members may be of help. Iliana communicated with the team the issues, potential solutions, and other alternatives. After reading out the specific errors in the data to them, Iliana mentioned that she never tried to remove the states with no airport traffic data. As it turns out, this ended up being the reason the code was unable to animate.

With more time, one thing I would want to work on is merging the U.S. and Canada map together. Their neighboring countries, so it would only make sense to have the two connected. I spent a good amount of time looking for a map of Canada and the U.S. that also included borders, but to no avail. What I discovered about data visualization throughout the project is the importance of perspective. When creating my visualizations, the things that had made sense to myself and my own *pretend audience*. The biggest example with this would be the choice of phrasing for *Percent of Baseline*. Being comfortable with the dataset by now, I had no issue with leaving all titles, legends, and descriptions as relatively base level. It wasn't until one group member had mentioned that when they described their visualization to their housemate, they [the housemate] did not understand what percent of baseline was. From there, extra steps were made so that visualizations were titled with a more descriptive title that leaves room for little no confusion. I also learned just how customizable *ggplot* really is and how useful other packages are (especially cowplot for being able to have Hawaii visualized from a closer distance).

## Individual Report for Wally Contreras

Wally's primary role in the group was providing visualizations and initial analysis to help develop and support our story. In the explanatory phase of the analysis, Wally created various visualizations to share with the group to help draw out the story from the Data. Visualizations he created during the exploratory phase include a ridgeline plot with unfiltered data, a violin plot by State, a heatmap by city, and an interactive plotly map for airport centroids. Through discussion of these (and other) visualizations our story started to take form. While working on milestone 4 the group shared additional visualizations, Wally shared a parallel sets plot and 2-D density chart which further supported our conclusion. For the final report, Wally polished up the heatmap with feedback from the other group members and provided the final analysis. Wally also compiled the final R file - transforming the dataframe to suit everyone's analysis. During our weekly meetings, he provided feedback and recommendations on visualizations, the presentation, and written assignments.

For Milestone 1, Wally explored the dataset and described 4 of the 11 variables to the group (Percent of Baseline, Centroid, City, and State), this fed directly into our Milestone 2 submission. Also, for the milestone 2 submission, Wally provided the charts and initial analysis of the violin plot, the ridgeline plot, the heat map, and Plotly glyph map. For the Milestone 3 submission, Wally refined the heat map with feedback from the other group members and expanded on the initial analysis. In the group presentation, Wally presented the variable transformation section and explained the heat map visualization.

Throughout the course of this project, I've gained invaluable experience customizing visualizations in R and gained an appreciation for all of the work that has gone into the tidyverse, ggplot2, lubridate, and plotly packages. I've learned that the perfect visualization does not happen the first time around, but that persistence pays off. While I spent inordinate amounts of time reading through documentation to learn how to control the most minute details, it was all worth it in the end. I also learned that the right aggregation, color palette, and sort variables can make all the difference in the world. This is an experience I'll be able to apply in different aspects of my life.

## Individual Report for Chloe Tian

In the team, Chloe mainly uses Tableau to do data visualization. At the beginning she created chart shows the month vs. the count of airports names (Australia, Canada, Chile, United States of America). But in this visualization is not successful as a time series, because the count of each airport is relatively the same. So, Chloe used percent of Baseline vs. all airports. But the visualization displaying all the airports results in a very busy graph and difficult to compare trends between the various locations. The clutter of the data line makes Chloe unable to read the valid information, so Chloe filtered the airports for U.S. and Canada. After discussion with the group, Chloe used Tableau to create a panel plot (multiple time-series graphs stacked on top of each other). We choose March and April, June and July, and November and December for comparison, because these months are particularly affected. To draw a clearer visualization, Chloe also learned online how to filter out the months wanted for comparison. Also thought about whether x-axis used days or weeks. If Chloe used days, there would be a lot of detail, so Chloe couldn't read important information properly, so Chloe ended up choosing week as the x-axis.

To see the comparison more intuitively between the United States and Canada for each period, Chloe also learned to use Tableau to make panel plot, to compare the differences and similarities between the airports in Canada and the United States due to the epidemic in a specific period. Chloe creates different worksheets to create different views of the data, so flexible analysis can ensure continuity. Lastly, Chloe created a dashboard, in order have more intuitive perspective to compare the United States and Canada. The group discussion decided to use blue to represent the United States and purple to represent Canada. Because of the interactivity in the dashboard can see the data’s information of various airports. If we want to know which airport has the highest average percentage of the baselines? What are the trends over time? Just click on the airport name, timeline, or percentage of the baselines, all the information will pop out.

Through standardized and structured processing of original data, they are sorted into data tables. Convert these numbers into visual structures and represent them visually. Combine visual structures to better understand the problems and patterns behind the data. Make sure the data is the same size and font. A lot of data is intended to highlight important data, so give it a different identity by distinguishing it with colors, sizes, and notes to make the communication more direct. The use of charts also has its own principles, classification, norms, charts are not just a list of data, the use of charts is to make people understand, but this is not enough, to make people understand quickly and efficiently, then to draw a good chart also need to have a certain ability, not graffiti. It's not what data we have that determines the format of our reports, it's what topic we want to express. The same data, want to express different topics, chart form is different. Use a special way to highlight key data that reflects the problem. Use different colors or text boxes to assist and describe the information behind the data.

For Tableau, I learned data connection and management, basic and higher-order graphic analysis, map analysis and advanced data operation from different videos and practices. Secondly, data is given roles, modified, filtered, grouped, and split. This part is the key to data cleaning, which is oriented to data exceptions and errors. According to the needs of analysis, we need to carry out in-depth processing of data, such as merging and connecting multiple data sources, adjusting the granularity of data (level of detail), and even doing data transpose when necessary. The last step is to export the data and share it locally or as a data source.

### Appendix Part 2: Additional Exploratory Analysis

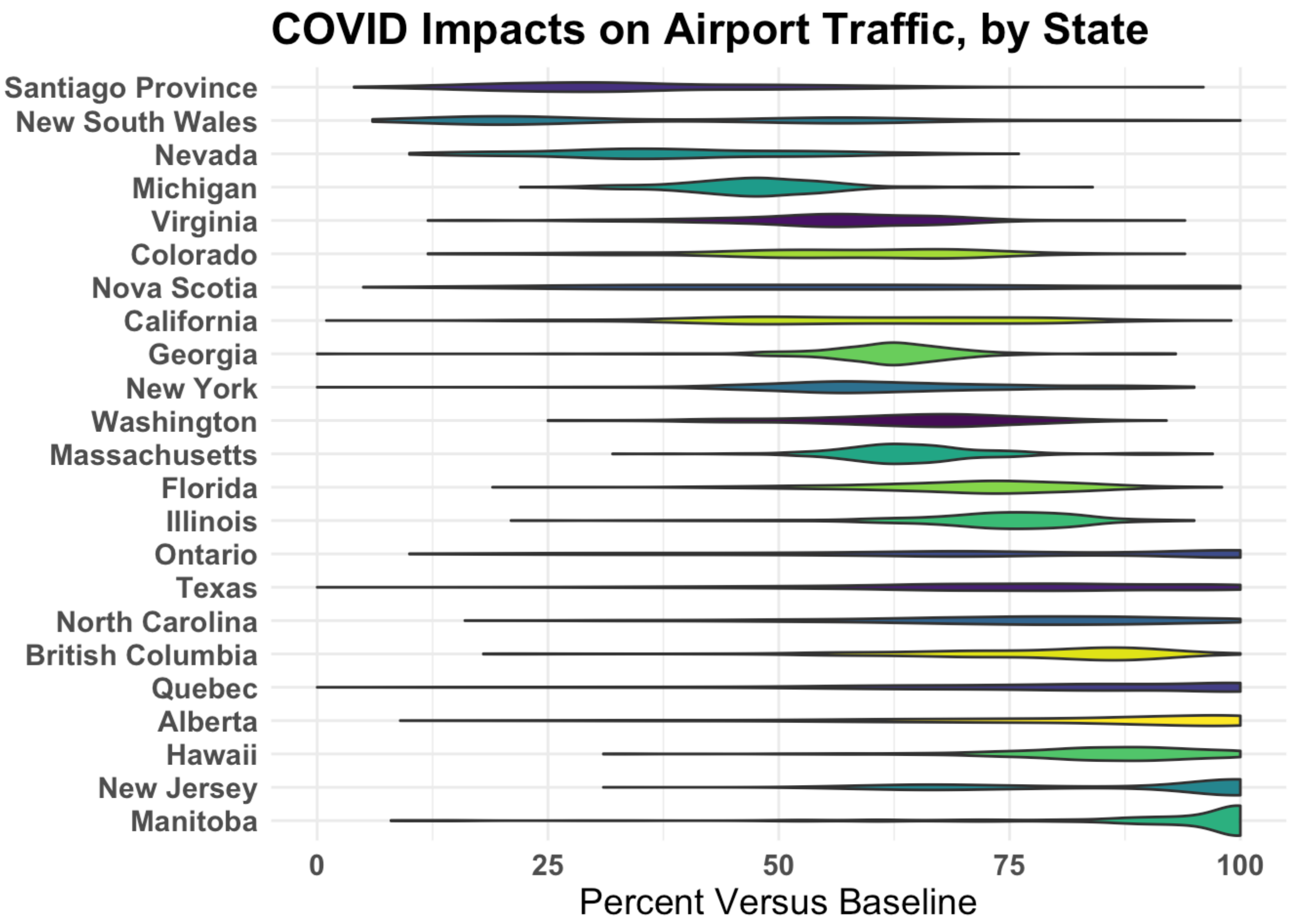


*Figure E3: Initial Exploratory Violin Plot of Distribution of Percent of Baseline*

This exploratory violin plot in *Figure E3*, shows the median percent of airport traffic versus baseline. Each datapoint in the violin is a different day. The “states” are sorted in ascending order by the median *percent of baseline* for the year so that the most impacted states are at the top of the list and the least-impacted states are at the bottom. We see that Santiago Province was the most affected and Manitoba was the least impacted in this timeframe. Because this visualization was based on the full dataset and the country is not labeled, it is easy to overlook the differences between the American and Canadian states/provinces. The color here is just used to differentiate the states, it is not an encoded variable. At the very least we could see that American states tended to group at the top of this chart.

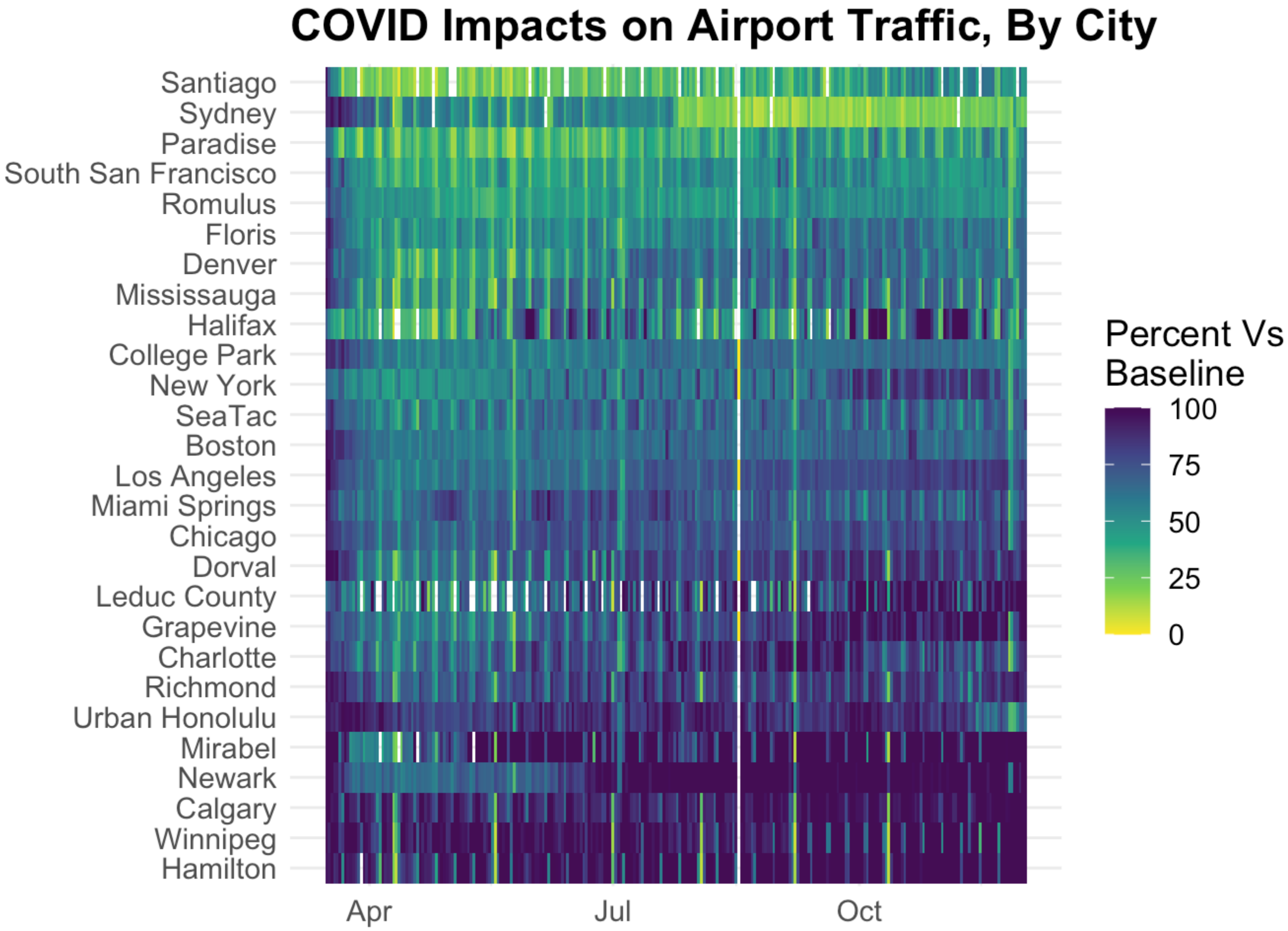


*Figure E4: Initial Exploratory Heatmap of Distribution of Percent of Baseline*

In this exploratory heat map, *Figure E4* had no aggregation or filtering, and we can see that there are missing data points. The y-axis labels display the names of the city where the airport is located, each rectangle represents a single day (one observation). The color encodes percent of traffic relative to the baseline with yellow indicating that the airport operated at 0% and dark blue indicating the airport operated at 100%. The cities are sorted in ascending order by median percent of baseline for the timeframe so that the most impact cities are at the top of the list. While it may not be the most beautiful chart in the world, we see that there is a wide range of effects; moreover, we can pinpoint the effect on a few cities. We can tell that Santiago was the most affected city while Hamilton was the least affected, for example. We can see many of the American cities at the top of the chart and most Canadian cities towards the bottom of the chart.

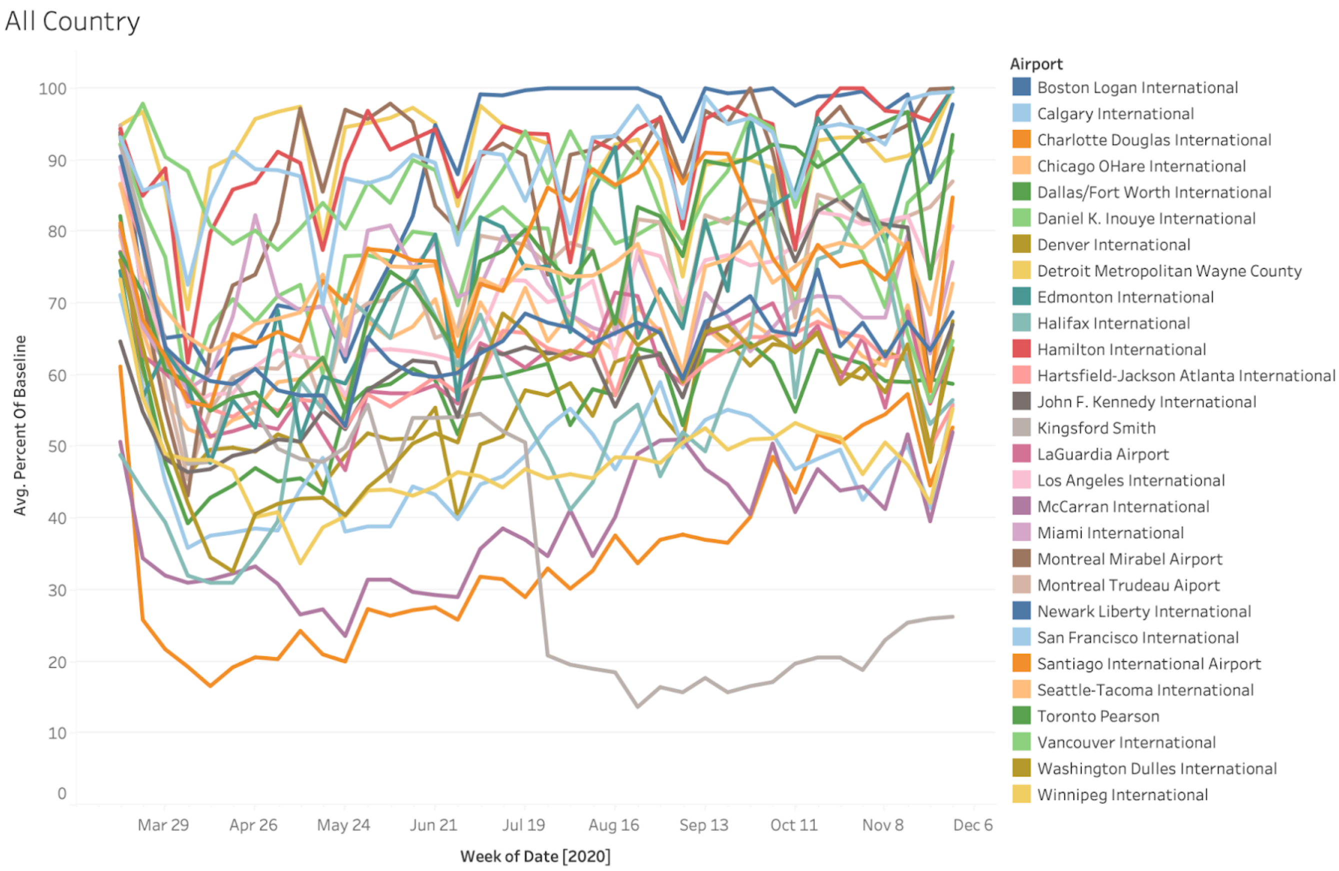


*Figure E5: Initial Exploratory Time-Series Line Graph of All Airports*

This exploratory line graph displays the time series using the average percent of baseline by week for all countries: Australia, Canada, Chile, United States. Because this visualization displays all airports, the graph appears very busy and makes comparisons between airports difficult. The biggest impact on the airports appears in March. Most airports have significant decreases in average percentage of baseline around June 21 and the middle of November. The line graph in *Figure E5* was created using Tableau. Through this line graph, it was determined that more trends appear from airports in the United States and Canada. Since Australia and Chile only has one airport, this ultimately led to the decision to focus the analysis on two countries. In addition, from this line graph, it was determined that the aggregation for percentage of baseline would be an average by week instead of by day, since missing values occurred in some of the airports, and this would allow for consistency in metric across all visualizations.

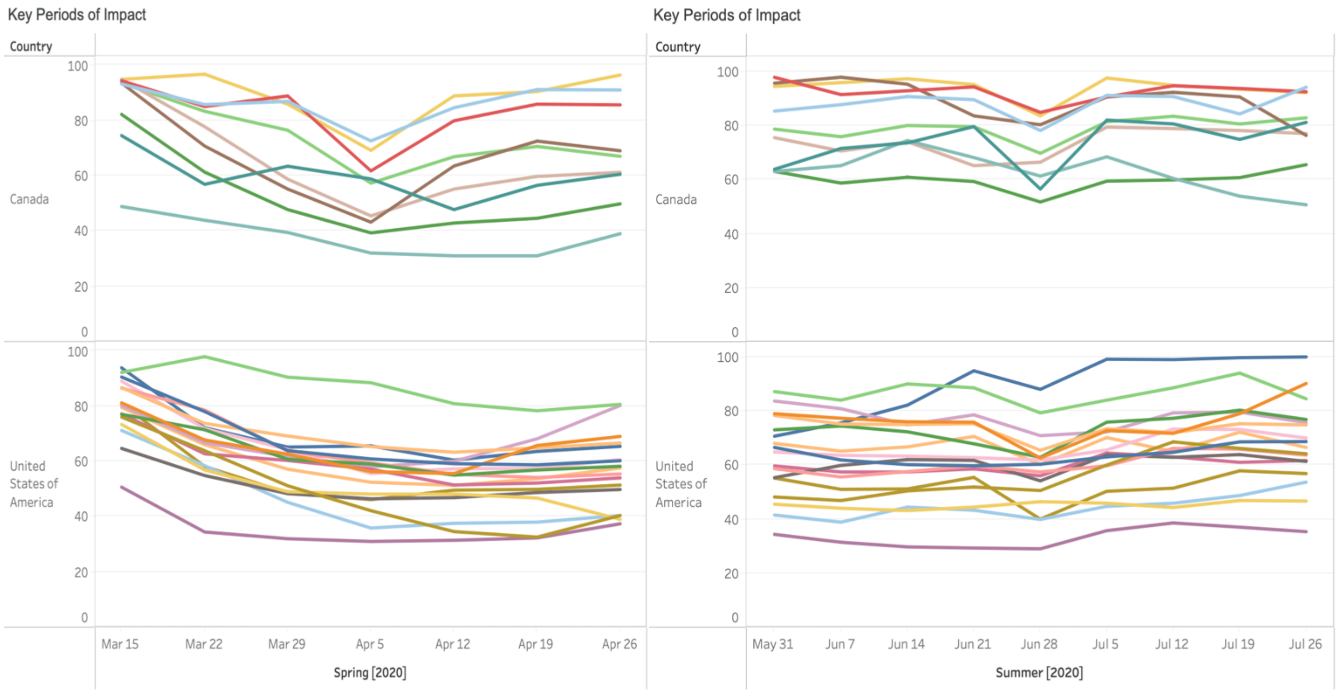


*Figure E6: Exploratory Time-Series Panel Plot for United States and Canada at Impact Periods*

Additional line graphs displayed in a panel for specific months were explored to better compare the trends around specific weeks as shown in *Figure E6.* From this timeline, every airport in U.S. has had some decline from March 15 until March 22 except Daniel K. Inouye International in Hawaii. San Francisco International airport has been most affected from March 15 to April 5, the average baseline decreased from 71.17 to 35.86. And Canadian airport has had some decline except Winnipeg International. Vancouver International, Toronto Pearson, Montreal Trudeau, Montreal Mirabel has big decreases between March 15 until April 5. Halifax international does not seem to have much effect. Every airport has had some decline from June 21 until June 28 expected Newark Liberty International. But in Canada, the average percentage of baseline for all airports continued to decrease through April 5 and then rapidly increased by April 12. All airports in the United States and Canada were affected on June 21, in which a sharp decrease in the average percentage of the baseline occurred. Then on June 28 the average percentage of the baselines rose again for both the United States and Canada.

To explore the underlying differences for the observations made in the other exploratory visualizations, we found it useful to visualize these line graphs at various levels with different filters. For the first level we wanted to see the overall effect by country, and then for individual states, cities, and airports. Using's tableau's interactivity was perfect for this; the dashboard layout made it easy to highlight the same airport across multiple timeframes of interest. This final implementation of this was the dashboard chosen as one of our main visualizations.

## Appendix Part 3: R-Code

```
### DSC 465 --- Group Project ---
## Final Code

#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("scales")
#install.packages("gridExtra")
#install.packages("ggridges")
#install.packages("ggpubr")
#install.packages("gapminder")
#install.packages("gganimate")
#install.packages("mapproj")
#install.packages("mapdata")
#install.packages("maps")
#install.packages("magrittr")
#install.packages("usdata")
#install.packages("reshape2")
#install.packages("ggforce")
#install.packages("cowplot")
#install.packages("ggtext")
#install.packages("geosphere")
#install.packages("sp")
#install.packages("gifski")
#install.packages("mapcan")
#install.packages("dplyr")

# General use libraries
library(tidyverse)
library(lubridate)

# Ridgeline/Violin plots
library(scales)
library(gridExtra)
library(ggridges)
library(ggpubr)

# Animated Choropleth
library(gapminder)
library(gganimate)
library(mapproj)
library(mapdata)
library(maps)
library(magrittr)
library(usdata)
library(reshape2)
library(ggforce)
library(cowplot)
library(ggtext)
library(geosphere)
library(sp)
library(gifski)
library(mapcan)
library(dplyr)

### --- Dataframe setup --- ###

# import the dataset
filename ='covid_impact_on_airport_traffic.csv'
```

```
df <- read_csv(filename,col_names=TRUE, show_col_types = FALSE)

# see the structure
str(df)
# view the dataframe
View(df)

# Extract Dates/Date Parts
df <- df %>%
  mutate(DateX = as.POSIXct(Date, format = '%Y-%m-%d'),
         Week = floor_date(Date,unit='week'),
         Month = factor(month(Date,TRUE)),
         MonthX = factor(month(Date,label=TRUE,abbr=FALSE),
                         levels = c("March", "April", "May", "June",
                                    "July", "August","September",
                                    "October", "November", "December"))
  )

# Add State/Province abbreviations
df <- df %>%
  mutate(stateAbb = str_sub(ISO_3166_2,-2),
         cityState = paste(City,stateAbb,sep=","))


### --- Ridgeline Plot --- ###

# Create Ridgeline Plot for U.S. and Canada:
ggplot(df, aes(x = PercentOfBaseline, y = MonthX)) +
  geom_density_ridges(data=filter(df, Country=='United States of America (the)'), aes(fill='darkturquoise'), color = 'turquoise1',
alpha=0.8, rel_min_height = 0.003, scale = 1.4) +
  geom_density_ridges(data=filter(df, Country=='Canada'), aes(fill='salmon1'), color = 'lightsalmon1', alpha=0.3, rel_min_height
= 0.005, scale = 1.4) +
  scale_x_continuous(limits = c(0, 115), labels = scales::percent_format(scale = 1)) +
  scale_y_discrete(limits=rev(levels(df$MonthX)),
                   labels=c('Dec', 'Nov', 'Oct', 'Sep', 'Aug', 'Jul', 'Jun', 'May', 'Apr', 'Mar')) +
  labs (x ='Percentage of Airport Travel Relative to Pre-Covid Baseline Period',
        y = 'Month in 2020') +
  scale_fill_identity(name = 'Country',
                      breaks = c('darkturquoise', 'salmon1'),
                      labels = c('United States', 'Canada'),
                      guide = guide_legend(override.aes = list(color=c('darkturquoise', 'salmon1')))) +
  ggtitle(label='Level of Airport Travel in the United States and Canada in 2020 during COVID-19',
          subtitle='Distribution of Percentage of Airport Travel Over Time Across all Airports ') +
  theme(plot.title = element_text(face='bold'),
        axis.text = element_text(family='Arial', size = rel(1.25))) +
  theme_ridges(font_size = 16, font_family = 'Arial')


### --- Heat Map --- ###

# custom palette - diverging red-yellow-blue, colorblind safe
cpal = c('#d7191c','#fdae61','#ffffbf','#abd9e9','#2c7bb6')

# filter -> re-code -> ggplot
pCity <- df %>%
  filter(Country %in% c('United States of America (the)','Canada')) %>%
  mutate(Country = recode_factor(Country,
                                 'United States of America (the)' = 'U.S.A.'
  ),
  cityState= recode_factor(cityState,
                           'South San Francisco,CA' = 'S San Fran,CA',
                           'Urban Honolulu,HI'= 'Honolulu,HI')
  ) %>%
  ggplot(aes(x=Week,
             y=reorder(cityState,PercentOfBaseline,fun=average),
             fill=PercentOfBaseline))

# customize ggplot
plot_CityHeatMap <-pCity + geom_tile(alpha=0.4) +
  # scale
  scale_x_date(breaks='1 month',date_labels = '%b') +
  # labels
  labs(title = 'COVID-19 Impacts on Airport Travel by City in the United States and Canada',
       subtitle = 'Comparing average airport traffic relative to baseline traffic by city, by week',
       caption = 'Dates covered: 2020-03-16 to 2020-12-02, summarized by week',
       x = NULL,
       y = NULL,
       fill='Traffic % \nof Baseline') +
  # theme
  theme_bw() +
  scale_fill_gradientn(
    colors=cpal,
    values = NULL,
    space = "Lab",
    na.value = "grey50",
    guide = "colourbar",
    aesthetics = "fill"
  ) +
  theme(axis.title.y = element_text(face='bold',size = rel(1.1)),
        axis.text.x= element_text(face='bold',size = rel(1.1)),
        plot.title = element_text(face='bold',lineheight=0.9),
        plot.caption = element_text(hjust=0, face = 'italic'),
        plot.title.position = 'plot',
        plot.caption.position = 'plot',
        strip.background = element_rect(fill = 'grey90',
                                        color = 'black',
                                        size=0.25),
        strip.text = element_text(face='bold', size = rel(1.2)),
        panel.grid.major = element_blank(),
        panel.grid.minor = NULL
  ) +
  # facet
  facet_grid(Country ~ .,
             scales = 'free_y')

# show the visualization
plot_CityHeatMap

# save the visualization
#ggsave('COVIDCityHeatmap.png',plot_CityHeatMap,width=9,height=5, scale=2,dpi=300)


### --- Violin Plots --- ###

# Create DF for U.S. Airports:
us_airports <- df %>%
  filter(Country == 'United States of America (the)')

# Create DF for Canada Airports:
can_airports <- df %>%
  filter(Country == 'Canada')


# Manual Scale of Colors for States:
```

```
us_pal = c('#ece2f0', '#d0d1e6', '#a6bddb', '#a6bddb', '#a6bddb', '#67a9cf', '#67a9cf', '#67a9cf', '#67a9cf', '#3690c0', '#3690c0',
'#02818a', '#02818a', '#02818a', '#016450')
can_pal = c('#67a9cf', '#3690c0', '#02818a', '#02818a', '#016450', '#016450')

# Create Violin Plot Ordered by Percent of Baseline (US):
us_dist = us_airports %>%
  mutate(State = fct_reorder(State, PercentOfBaseline)) %>%
  ggplot(aes(y=reorder(State, PercentOfBaseline), x=PercentOfBaseline, fill=State)) +
  geom_violin(alpha = 0.8) +
  geom_boxplot(color = 'azure1', alpha=0.7, width=0.15, coef = 0, outlier.color = NA) +
  scale_x_continuous(limits = c(0, 100), labels = scales::percent_format(scale = 1)) +
  scale_fill_manual(values = us_pal, guide = 'none')+
  labs (x ="Percentage of Airport Travel \n Relative to Pre-Covid Baseline Period",
        y = '') +
  ggtitle('Distribution of Percentage of Airport Travel in the U.S.') +
  theme(plot.title = element_text(face='bold', family = 'Arial', hjust=0.5),
        panel.background = element_rect(fill = 'white', color = 'black'),
        panel.grid.major = element_line(color = 'gray90'),
        panel.grid.minor = element_line(color = 'gray95'),
        axis.text.y = element_text(face="bold", family='Arial'),
        axis.text.x = element_text(face='bold', family='Arial', size=rel(1.2)))


# Create Violin Plot Ordered by Percent of Baseline (Canada):
can_dist = can_airports %>%
  mutate(State = fct_reorder(State, PercentOfBaseline)) %>%
  ggplot(aes(y=reorder(State, PercentOfBaseline), x=PercentOfBaseline, fill=State)) +
  geom_violin(alpha = 0.8) +
  geom_boxplot(color = 'azure1', alpha=0.7, width=0.1, coef = 0, outlier.color = NA) +
  scale_x_continuous(limits = c(0, 100), labels = scales::percent_format(scale = 1)) +
  scale_fill_manual(name = 'Province', values = can_pal, guide = 'none') +
  labs (x ="Percentage of Airport Travel \n Relative to Pre-Covid Baseline Period \n",
        y = '') +
  ggtitle('Distribution of Percentage of Airport Travel in Canada') +
  theme(plot.title = element_text(face='bold', family = 'Arial', hjust=0.5),
        panel.background = element_rect(fill = 'white', color = 'black'),
        panel.grid.major = element_line(color = 'gray90'),
        panel.grid.minor = element_line(color = 'gray95'),
        axis.text.y =  element_text(face="bold", family='Arial'),
        axis.text.x = element_text(face='bold', family='Arial', size=rel(1.2)))


# Display U.S. and Canada Distribution Side-By-Side:
grid.arrange(us_dist, can_dist, ncol = 2,
             top=text_grob('Impact of COVID-19 on Airport Travel by State/Province in the United States and Canada in 2020',
                           face = 'bold', size = '16'))


### --- Animated Choropleth --- ###
covidbyweek <- df %>%
  group_by(State,Country,Week) %>%
  summarise(POB = mean(PercentOfBaseline)) %>%
  ungroup() %>%
  mutate(WN = format.Date(Week, format='%m.%d.%y'),
         WN_2 = format.Date(Week, format='%m.%d.%y')
  )

names(covidbyweek)[3] <- "WeekNumber"

#US mainland map data
mainland <- map_data("state") %>%
  mutate(subregion = toupper(state2abbr(region)))
```

```
#covid_by_week is the dataset that will be used for the chloropleths
#abbreviating state data in covid, this is only for the US
covid_by_week <- covidbyweek %>%
  mutate(State_revised = state2abbr(State))

#map data for Hawaii
maps <- map_data("world")
hawaii <- subset(maps, subregion=="Hawaii") %>%
  mutate(region=subregion) %>%
  mutate(region=state2abbr(region))

#HAWAII COVID
covidHI = hawaii %>%
  left_join(covid_by_week ,by=c("region" = "State_revised"), all=TRUE)
covidHI <- na.omit(covidHI)

#MAINLAND COVID
covidUS = mainland %>%
  left_join(covid_by_week ,by=c("subregion" = "State_revised"), all=TRUE)
covidUS <- covidUS %>%
  mutate(WN = format.Date(WeekNumber, format='%m.%d.%y')) %>%
  filter(subregion != "UT")
covidUS <- covidUS %>%
  mutate(WN_2 = format.Date(WeekNumber, format='%m.%d.%y'))
covidUS <- na.omit(covidUS)

#Creating a seperate Chloropleth for Hawaii, and then using the draw_plot() feature to have it more visible in the mainland
graphic

hawaii_map <- ggplot(aes(long,lat,group=group, fill = POB), data = covidHI) +
  geom_polygon(colour = "black") + theme_void() +
  panel_border(color = "black") +
  scale_fill_gradient2(low =  "#ece2f0", mid = "#a6bddb", high = "#1c9099", midpoint = mean(covidHI$POB)) +
  ggtitle("HI", ) +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5))

#Lines 87 - 103 related to adding labels to the U.S. Map
#Read different articles/forums on how to approach this, original idea was to use the centroid points given in the dataset,
however the points were pretty random
#Eventually saw an example on how to find the centroid for multiple factors

#Here, a function is created to find the centroid for all US states using the Polygon() function
statenames <- function(State)
{
  Polygon(State[c('long','lat')])@labpt
}

centroids <- by(covidUS, covidUS$State, statenames)
centroids2 <- do.call("data.frame", centroids)
centroids3 <- t(centroids2) %>%
  as.data.frame()

centroids_final <- rownames_to_column(centroids3, "State")
names(centroids_final)[2] <- "long"
names(centroids_final)[3] <- "lat"

centroids_final$short_state <- state2abbr(centroids_final$State)


#chloropleth of the U.S. mainland
us_map <- covidUS %>%
  ggplot(aes(long,lat,group=group, fill = POB)) +
  borders("state", colour = "black", fill = "whitesmoke") +
  geom_polygon(colour = "black") + theme_bw() +
  scale_fill_gradient2(low =  "#ece2f0", mid = "#a6bddb", high = "#1c9099", midpoint = mean(covidUS$POB)) +
  with(centroids_final, annotate(geom="text", x = long, y= lat, label = short_state, size = 3.25, fontface = "bold" ))

us_map

#USANIMATION is the mainland map with the Hawaii plot on the lower left hand corner, transition time is the variable WN_2
#from the cowplot library, using draw_plot so Hawaii is included with the states while also not compromising size
USANIMATION <- us_map + draw_plot(hawaii_map, width = 13, height = 7, x = -125, y = 24) +
  #transition_time(WN_2)+
  labs(title = "Airport Travel by State in the United States in 2020 during COVID-19",
       subtitle = "Average Percent of Baseline at: {frame_time}",
       fill = "Percentage of Airport Traffic Relative to Baseline",
       caption = "Note: States that are not represented in the data have been shaded in light grey") +
  theme(axis.text.x = element_blank(), #axes changes begins
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        text = element_text("Arial"), #adjusting font type
        plot.title = element_text(face = "bold", size = 21), #title is made bold
        plot.caption = element_markdown(face = "italic", vjust = 1,size = 12),
        plot.title.position = "plot",
        legend.position = "bottom", #legend changes
        legend.key.size = unit(1,'cm'),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 10),
        plot.subtitle = element_text(size = 19.5),
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank())
#??valign
USANIMATION
typeof(USANIMATION)
#animating USANIMATION
USANIMATION_FINAL <- gganimate::animate(USANIMATION, nframes = 450, height = 600, width = 900,
                                        renderer=file_renderer())
warnings()
#saving animation
anim_save(filename = "US_Animation.gif", animation = USANIMATION_FINAL)


#CANADA ANIMATION PROCESS
#Discovered a library in R, mapcan that provides Canada's map with the provinces outlined
library(mapcan)
library(dplyr)
Canada_States <- mapcan(boundaries = province, type = standard)

CAN <- Canada_States %>%
  mutate(region = pr_english) %>%
  mutate(subregion = pr_alpha) %>%
  as.data.frame() %>%
  mutate(as.numeric(Canada_States$group))

CAN1 <- subset(CAN, select = c("long","lat","group","order","region","subregion"))
CANADA_BORDER <- CAN1 %>%
  filter(region == "Saskatchewan")
ggplot(aes(x=long, y=lat, group = group), data = CANADA_BORDER) + geom_polygon(colour = "black")
CANcovid <- CAN1 %>%
  left_join(covidbyweek, by=c("region" = "State"), all=TRUE)
CANcovid <- na.omit(CANcovid)
```

```
canada_centroids <- by(CANcovid, CANcovid$region, statenames)
canada_centroids2 <- do.call("data.frame", canada_centroids)
canada_centroids3 <- t(canada_centroids2) %>%
  as.data.frame()
canada_centroids4 <- rownames_to_column(canada_centroids3)
names(canada_centroids4) <- c("Province", "long","lat")
#adjusting british columbia and nova scotia

canada_centroids4$Province [canada_centroids4$Province == "British.Columbia"] = "British Columbia \nBC"
canada_centroids4$Province [canada_centroids4$Province == "Alberta"] = "Alberta \nAB"
canada_centroids4$Province [canada_centroids4$Province == "Manitoba"] = "Manitoba \nMB"
canada_centroids4$Province [canada_centroids4$Province == "Ontario"] = "Ontario \nONT"
canada_centroids4$Province [canada_centroids4$Province == "Quebec"] = "Quebec \nQC"
canada_centroids4$Province [canada_centroids4$Province == "Nova.Scotia"] = "NS"

#Chloropleth for Canada
CANmap <- ggplot(CANcovid, aes(x = long, y=lat, group=group, fill = POB)) +
  geom_polygon(colour = "black") + #transition_time(WN_2) +
  labs(title = "Airport Travel by Province in Canada in 2020 during COVID-19",
       subtitle = "Average Percent of Baseline at: {frame_time}",
       fill = "Percentage of Airport Traffic Relative to Baseline",
       x = "",
       y = "",
       caption = "Note: States that are not represented in the data have been shaded in light grey") +
  theme_bw() +
  scale_fill_gradient2(low =  "#ece2f0", mid = "#a6bddb", high = "#1c9099", midpoint = mean(CANcovid$POB)) +
  theme(text = element_text("Arial"),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        plot.title = element_text(face = "bold", size = 19.5),
        plot.subtitle = element_text(size = 19),
        legend.position = "bottom", #legend changes
        legend.key.size = unit(1,'cm'),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 10),
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        plot.caption = element_text(face = "italic", hjust = 1, vjust = 0.5, size = 12)) +
  with(canada_centroids4, annotate(geom = "text", x = long, y = lat, label = Province, size = 3, fontface = "bold")) +
  with(CANADA_BORDER, annotate(geom = "polygon", x = long, y = lat, group = group, fill = "whitesmoke", colour = "black"))

CAN_ANIMATION <- gganimate::animate(CANmap, nframes = 450, height = 600, width = 900)

#saving animation
anim_save(filename = "CAN_Amination.gif", CAN_ANIMATION)
```