DePaul University - Winter, 2021

DSC 465 Data Visualization - Final Group Report

IMPACT OF COVID-19 ON FLIGHTS FROM CHICAGO'S O'HARE INTERNATIONAL AIRPORT (ORD)

Professor: Dr. Eli T Brown

Group Name: Team Sky

Group Members:

Ashish Gare
Jared Young
Hiwot Gebreyohannes
Serena Yang
Shuo Wang

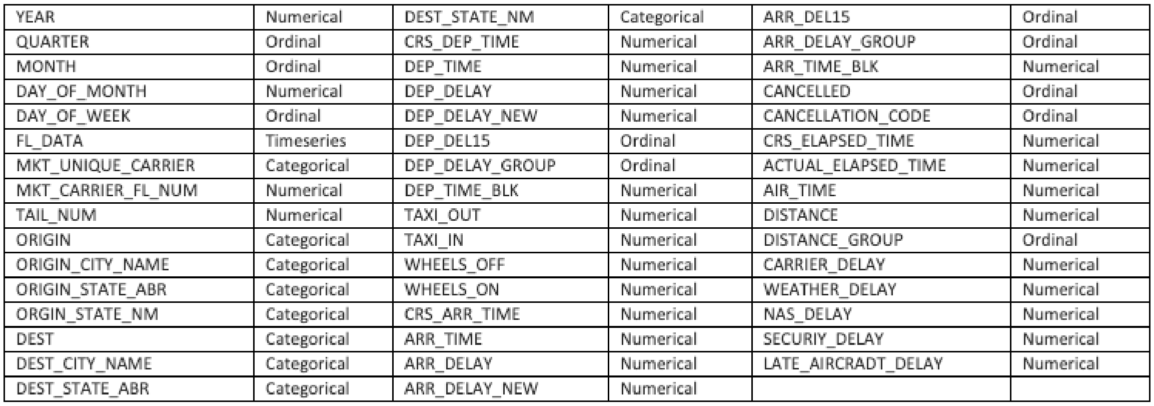
Table of Contents

Introduction

COVID-19 is extremely easy to contract. It is essential to reduce contact time and increase social distance; therefore, many people choose to cancel their travel plans to minimize the chance of being infected. For this project, our group will visualize the dataset from different perspectives to gain insights into the impact of COVID-19 caused on the United States aviation services in the several domestic flight carriers from throughout January to June 2020, which is the beginning stage of the pandemic.

COVID-19 has crippled the global airline industry with air service reductions widespread throughout 2020. The United States Department of Transportation's (DOT) Bureau of Transportation Statistics has tracked the on-time performance of domestic flights operated by large air carriers. The data was collected is from January to June 2020 and contained relevant flight information (on-time, delayed, canceled, diverted flights) from the Top 10 United States flight carriers. We downloaded the data file from the Kaggle website.

The original dataset contained 2.7 million rows. Since our group wanted to analyze mainly all the flights originating from Chicago O'Hare International Airport, we filtered the data through the SQL tool by "ORIGIN='ORD'." After filtered on the dataset, the database contains about 127k rows (flights), which will be used in this project to visualize the topic. The dataset has 47 columns (variables), including 9 categorical variables, 27 numerical variables, 10 ordinal variables, and 1 time-series variables. The variable table is shown below:



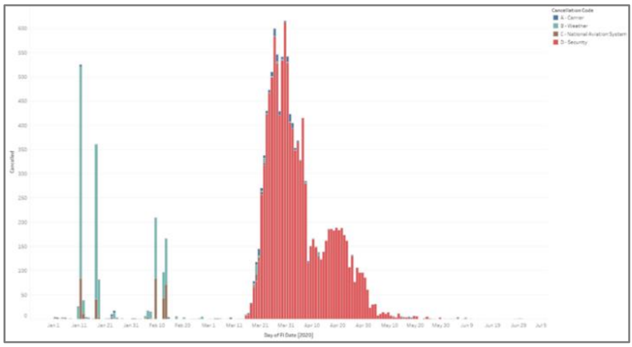
YEAR	Numerical	DEST_STATE_NM	Categorical	ARR_DEL15	Ordinal
QUARTER	Ordinal	CRS_DEP_TIME	Numerical	ARR_DELAY_GROUP	Ordinal
MONTH	Ordinal	DEP_TIME	Numerical	ARR_TIME_BLK	Numerical
DAY_OF_MONTH	Numerical	DEP_DELAY	Numerical	CANCELLED	Ordinal
DAY_OF_WEEK	Ordinal	DEP_DELAY_NEW	Numerical	CANCELLATION_CODE	Ordinal
FL_DATA	Timeseries	DEP_DEL15	Ordinal	CRS_ELAPSED_TIME	Numerical
MKT_UNIQUE_CARRIER	Categorical	DEP_DELAY_GROUP	Ordinal	ACTUAL_ELAPSED_TIME	Numerical
MKT_CARRIER_FL_NUM	Numerical	DEP_TIME_BLK	Numerical	AIR_TIME	Numerical
TAIL_NUM	Numerical	TAXI_OUT	Numerical	DISTANCE	Numerical
ORIGIN	Categorical	TAXI_IN	Numerical	DISTANCE_GROUP	Ordinal
ORIGIN_CITY_NAME	Categorical	WHEELS_OFF	Numerical	CARRIER_DELAY	Numerical
ORIGIN_STATE_ABR	Categorical	WHEELS_ON	Numerical	WEATHER_DELAY	Numerical
ORGIN_STATE_NM	Categorical	CRS_ARR_TIME	Numerical	NAS_DELAY	Numerical
DEST	Categorical	ARR_TIME	Numerical	SECURIY_DELAY	Numerical
DEST_CITY_NAME	Categorical	ARR_DELAY	Numerical	LATE_AIRCRADT_DELAY	Numerical
DEST_STATE_ABR	Categorical	ARR_DELAY_NEW	Numerical		

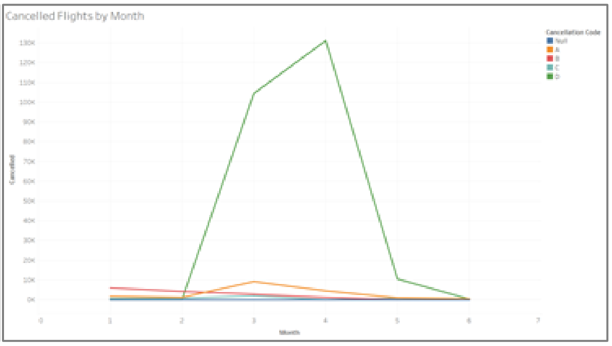
In this project, we used software R and Tableau to achieve these goals. During the process, we first search the Internet for excellent charts made by other analyzers for the same type of dataset, trying to get some inspiration from them. Then we got some topics of several aspects through group discussion and assigned them to each member to make the charts. After each team member made the first draft of individual works, we improved those plots by getting comments from the professor and the team members. Finally, after all the plots were finalized, our group found insights from the visualizations together. To visualize flight information geographically, we appended coordinate information from OpenFlights for all origin and destination airports. OpenFlights is a publicly available dataset updated as of January 2017.

Exploratory Analysis

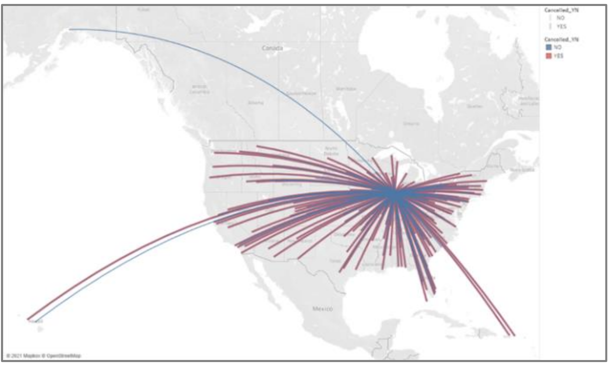
As a first step, we focused on visualizing flight cancellations by month. We started off by looking for any signs of obvious disruptions in flight activity in early March when state of emergency was declared in Illinois. From the quick exploratory visualization below, we were able confirm our assumption and make few more observations. Geographically, flights were cancelled all over U.S. and 'Security' was used as the main cancelation code between March and May. Though the total cancellation was the highest in March, the rate of cancellation relative to available flights was higher in April. About 40% of the available flights were cancelled in April but only about 20% of the available flights were cancelled in March.

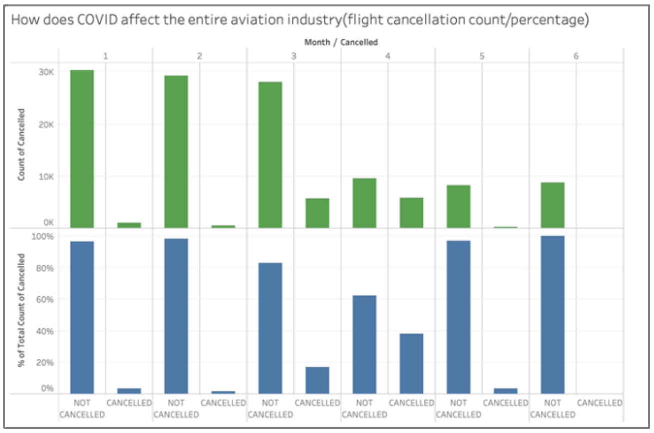
Using these initial visualizations as a starting point, we further analyzed flight cancellations and investigated additional topics and variables. The key set of variables utilized described flight in availability, delay, cancellation and variability by time and region.

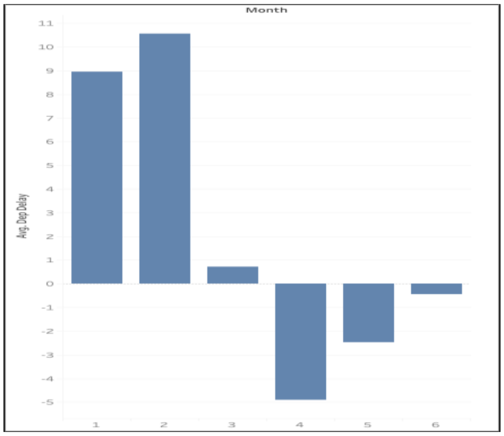




Cancelled Flights by Month (3)
Cancelled
Mkt Unique Carrier



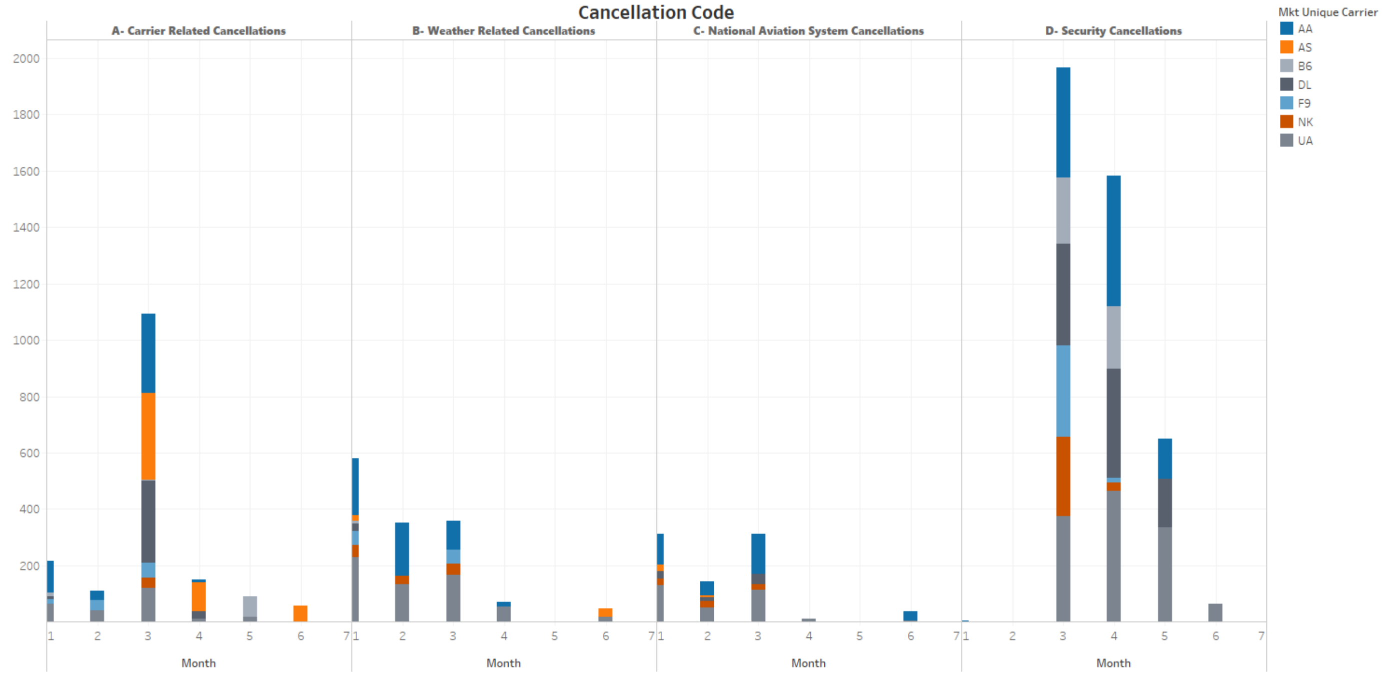




Visualizations

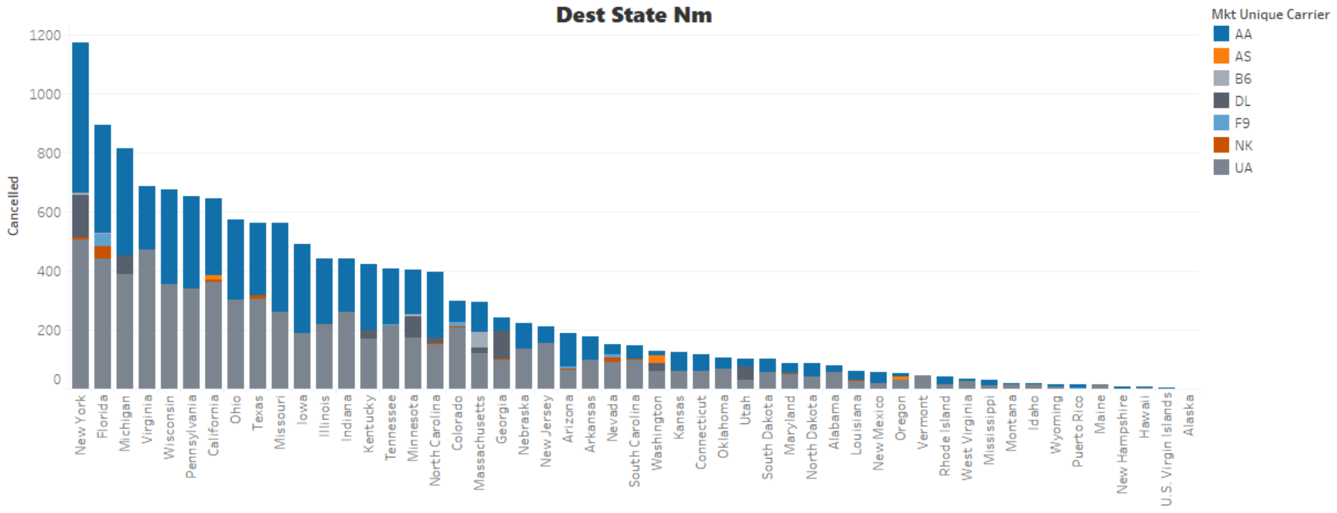
Comparison of Live/Cancelled Ratio by Airline

Reason for Cancellations



Month vs. Day Of Month. Color shows details about Cancellation Code. Size shows details about Mkt Unique Carrier. The view is filtered on Cancellation Code, which keeps A, B, C and D.
CANCELLATION_CODE - Reason for Cancellation - if Cancelled, Letter Present (A: Carrier, B: Weather, C: National Aviation System, D: Security)

Cancelled Airlines By State

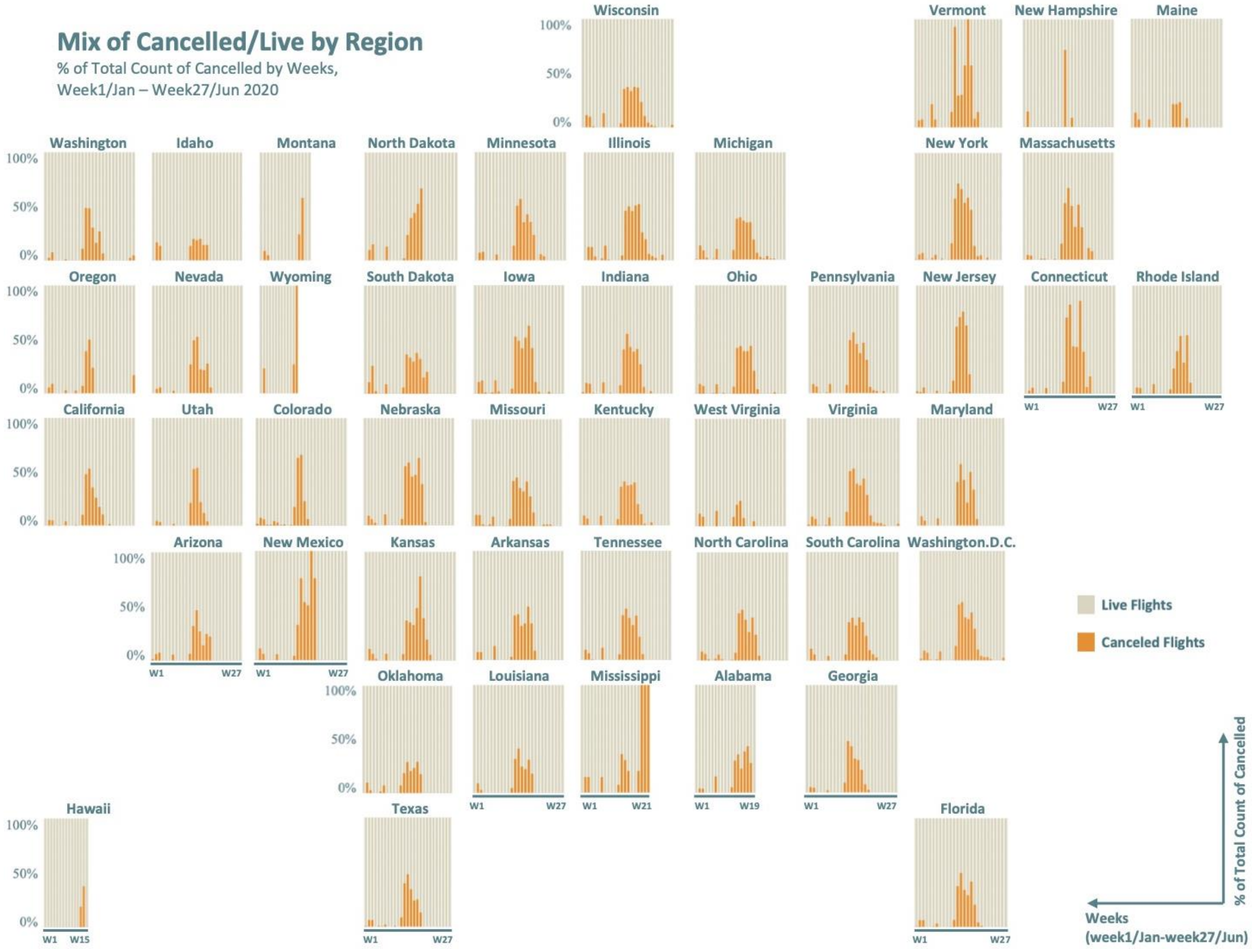


Sum of Cancelled for each Dest State Nm. Color shows details about Mkt Unique Carrier.

We further looked at the reasons for cancellations occurring in various months. At first there seem to be a small number of cancellations due to weather in the month of January and some more in the month of March. It was seen that the greatest number of cancellations occurred in the month of March and April which and it was more due to Security Cancellations (D) which can be the cause of 'mis-reporting' and not a clear way to report the cancellations and security personnel would stop passengers with higher temperatures from travelling.

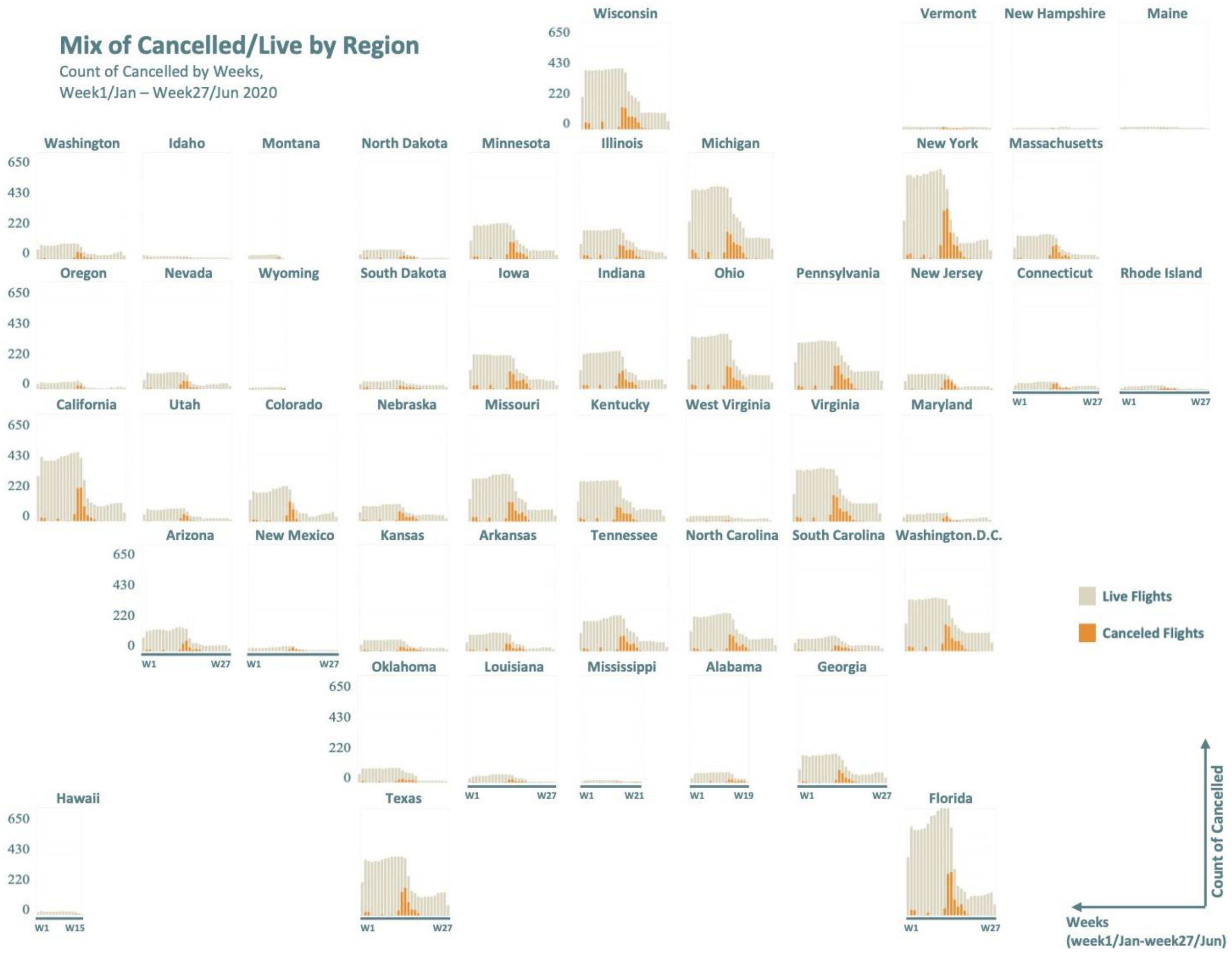
Along with this information it can also be seen that New York had one of the greatest number of cancellations in the country throughout with airline carriers United Airlines and Delta Airlines. This can be because New York had one of the first outbreaks and these airline cancellations can be the cause of that. Florida and Michigan were second and third on the list of cancellations respectively.

Mix of Cancelled/Live by Region



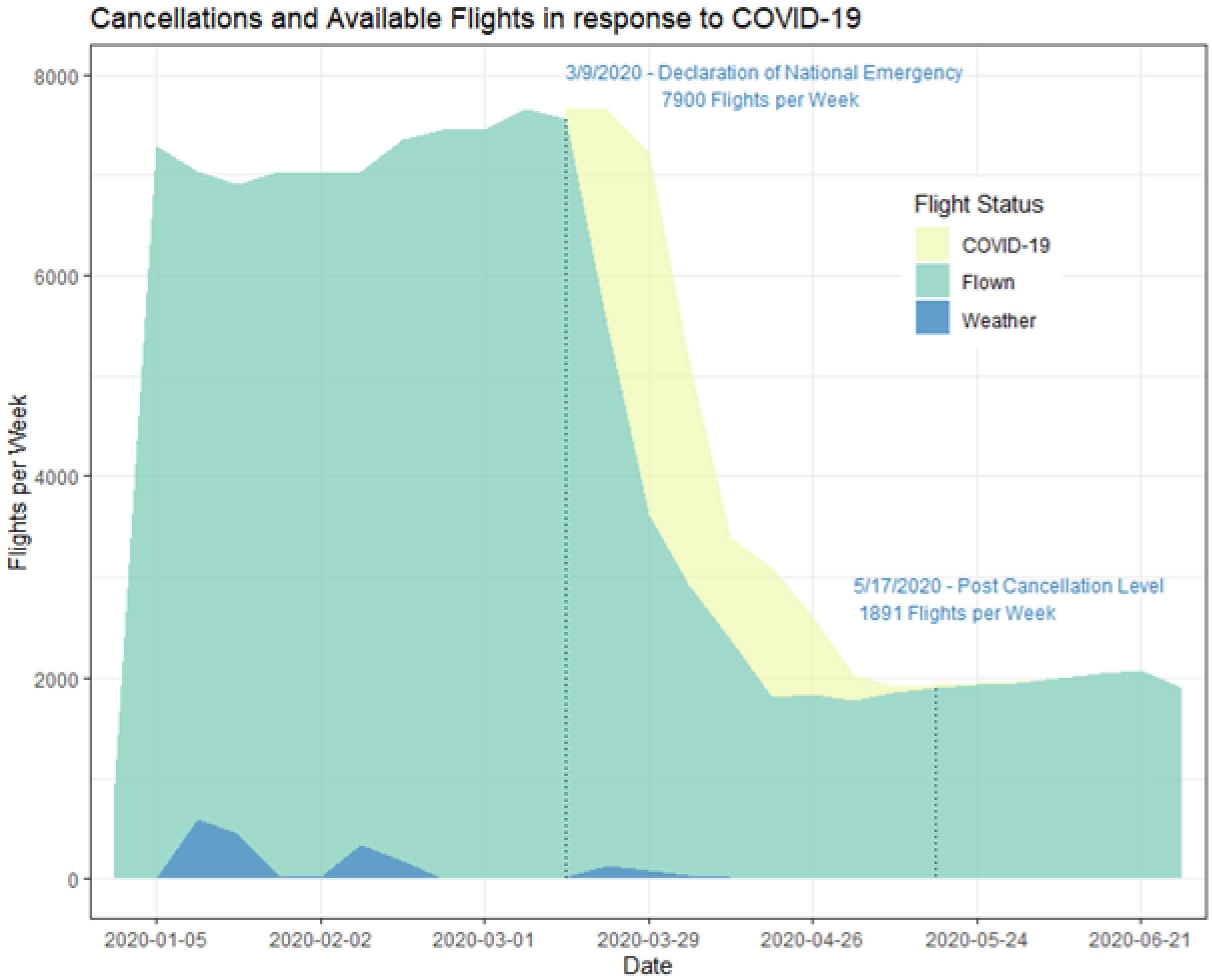
The first graph talks about the percentage of canceled flights of the total count, the X-axis is the timeline from week 1, the start of January, to week 27, the end of the June; and the Y-axis presents the percentage of total count of canceled flights. We can see that the overall trend looks very similar for all the states, and this trend can prove that the impact of COVID on flights. From week 2 to week 3, some flights got canceled due to some weather situation. Starting from around week 10, which was the beginning of the pandemic, flights began to take big hits as the cancellations rate rose by a significant percentage. The large cancellation stayed around for almost 2 months.

From this graph, the audience can only see the trend of canceled flights since the COVID began. There is no way to know the tendency of total flights decreases in the march from this percentage plot; therefore, I also created the second plot to show the total count of canceled flights to see the impact of COVID on the airline industry.



The second graph shows the count of canceled and live flights by region. The x-axis is the timeline from week 1 to week 27, which is from January 1st to the end of June in 2020, and the y-axis is the count of canceled flights. As you can see from this cartogram, although the total number of flights for each state is not the same, the overall trend is very similar. There was a significant decline in the total flights from Chicago to each destination state after week 15. After the big increase in canceled flights, those flight lines got closed for some states like Montana, Hawaii, and Wyoming. Therefore, in the first graph for those states' data were not enough.

Cancellations and Available Flights



The Plot is a Stacked Area plot using the Count of distinct Flights for each Week as the Y axis, and the weeks of 2020 from January through June as the X Axis.

The Fill for the Stacked Area plot is based on a created field 'Flight Status' for each flight in the data. The original data carries four distinct assigned values for CANCELLATION_CODE, and the field is blank for flights that were not cancelled – i.e. 'Flown'.

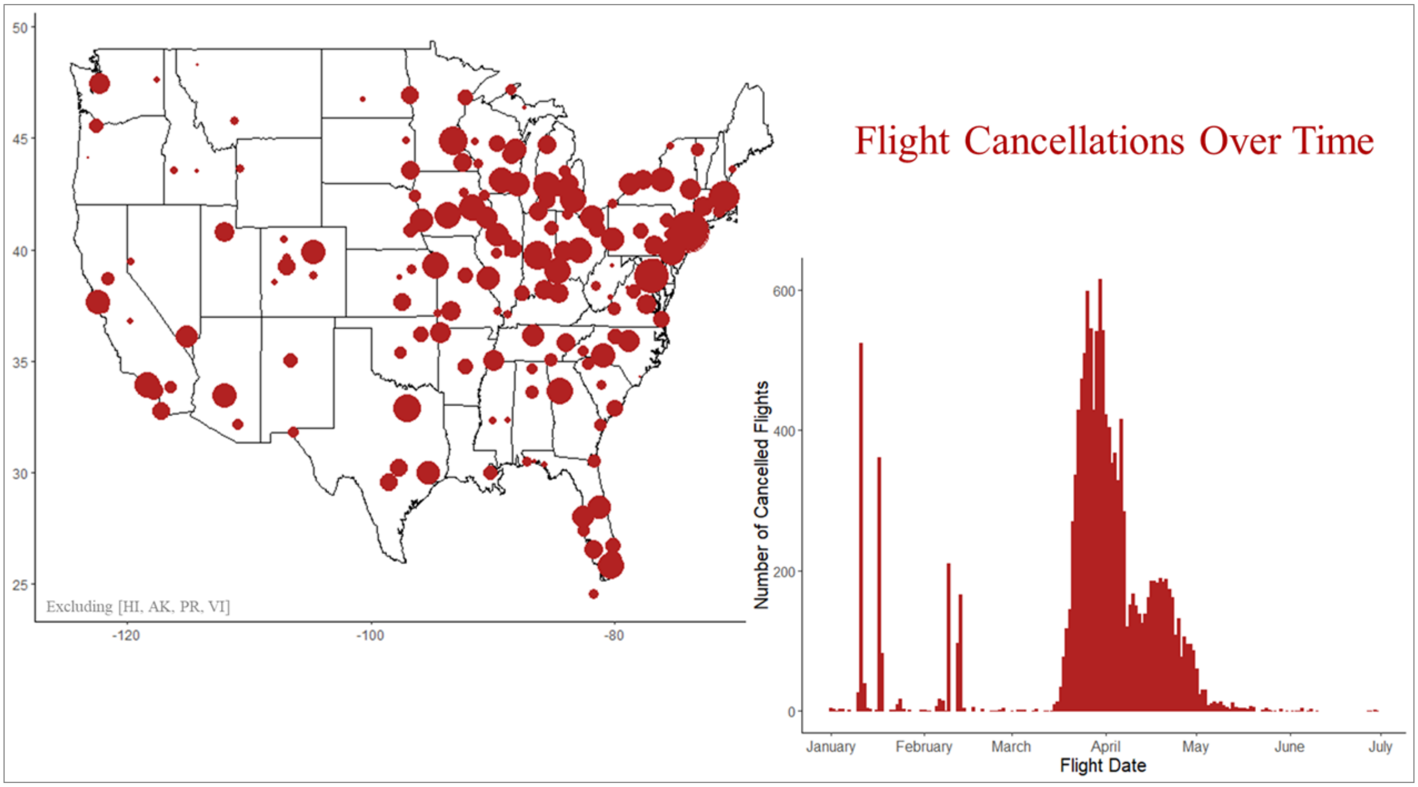
Earlier research into the data indicated that the value of 'D' for CANCELLATION_CODE was unique to the COVID-19 period. All other values were grouped into the general headers of 'Weather' for flights that were cancelled, and 'Flown' for flights that had no populated CANCELLATION_CODE. This decision served the goal of separating normal rates of Cancellation, as seen in Blue, from the COVID-19 event, using Yellow for emphasis. The Color scheme was selected to be Color Blind friendly and still offer significant distinction in hue to make the separation of values clear.

Annotations of the initial declaration of COVID-19 as a state of emergency are included to further tie the wave of cancellations and reduced flight volume to COVID-19. There is no distinct event to tie to the relatively steady state reached by May of 2020 with reduced flight volume, but an annotation to include the specific new average value to compare with the pre COVID-19 value was added to give an easy comparison between the Pre/Post numbers.

Early versions of this display were at the daily volume level, with the intent of isolating specific actions taken by individual states and associating them with state specific shifts in the flight volume for those specific times. The overall patterns did not vary by state to the expected degree, as while state specific actions could affect specific businesses and mask policy, national air travel would be more subject to federal mandates which would then appear identically for all regions. Having learned that, the decision to use a smoother, weekly view was made in order to focus on representing the overall impact on air travel due to COVID-19, and the relative volumes of available flights in the Pre COVID-19 period versus the current steady state.

In this form, this visualization shows how the effects laid out in the other visualizations culminate in overall results for the airlines and for travelers in terms of opportunities to book flights.

Flight Cancellations Over Time



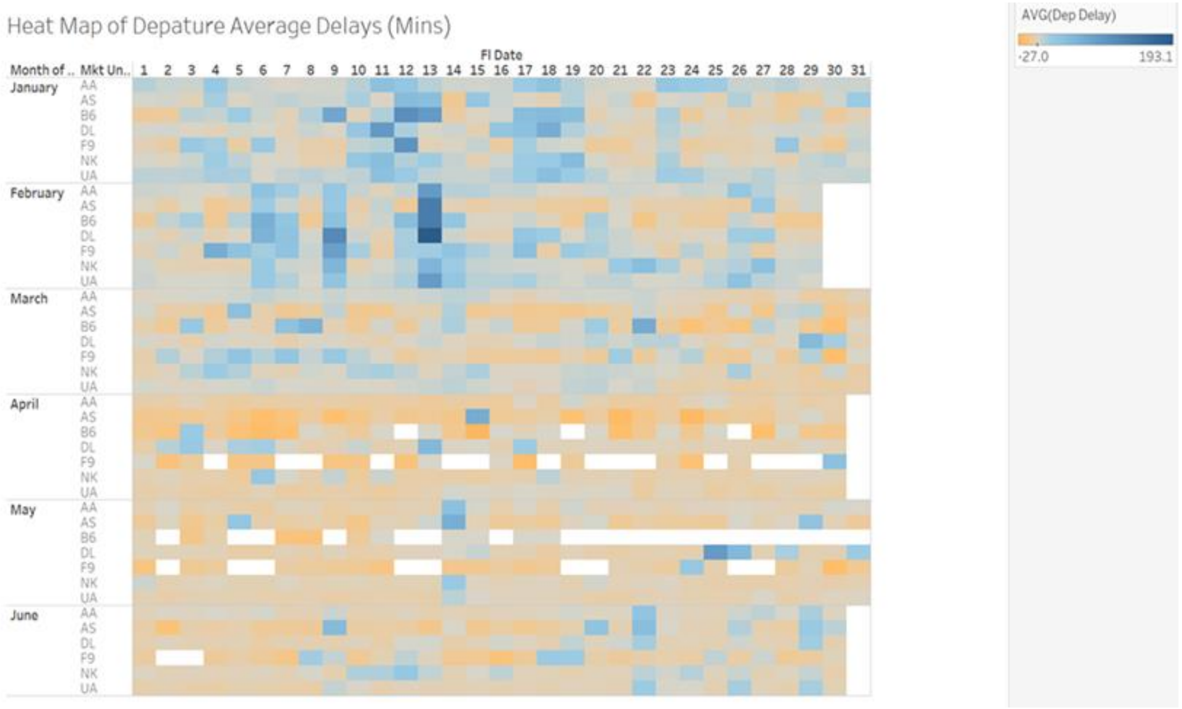
The geographical and time series plots show flight cancellations over time for flights originating out of Chicago’s O'Hare International Airport, from January to June of 2020. The plots as a pair demonstrate the sudden reaction to travel restrictions in early March and how long the impact lasted.

Both plots make use of animation to show the time view of cancellations. It’s easy to notice from these plots the drastic increase in number of flight cancellations in early-mid March. Flights to Hawaii, Puerto Rico, U.S. Virgin Islands and Alaska were excluded from the geographical plots for an easier view of the cancellation. The impact from this exclusion was minimal as most domestic flights out of O’Hare were within contiguous U.S. It’s worth noting that the dates provided in the dataset represent the original scheduled flight date as opposed to the date of cancellation.

From the geographical plot we can see that initially there were only few flight cancellations but suddenly cancellations pop-up all-over U.S. after few frames. Similarly, the timeseries bar plot shows that during the months of March to April there was a long stretch of cancellations compared to prior months. The few cancellations before March were mainly due to weather. State of emergency was declared in Illinois on March 9 and the travel ban, that saw long waits and overcrowded facilities at ORD, took effect on March 14. The travel restrictions timeline aligns with the initial turn of events clearly seen on the plot.

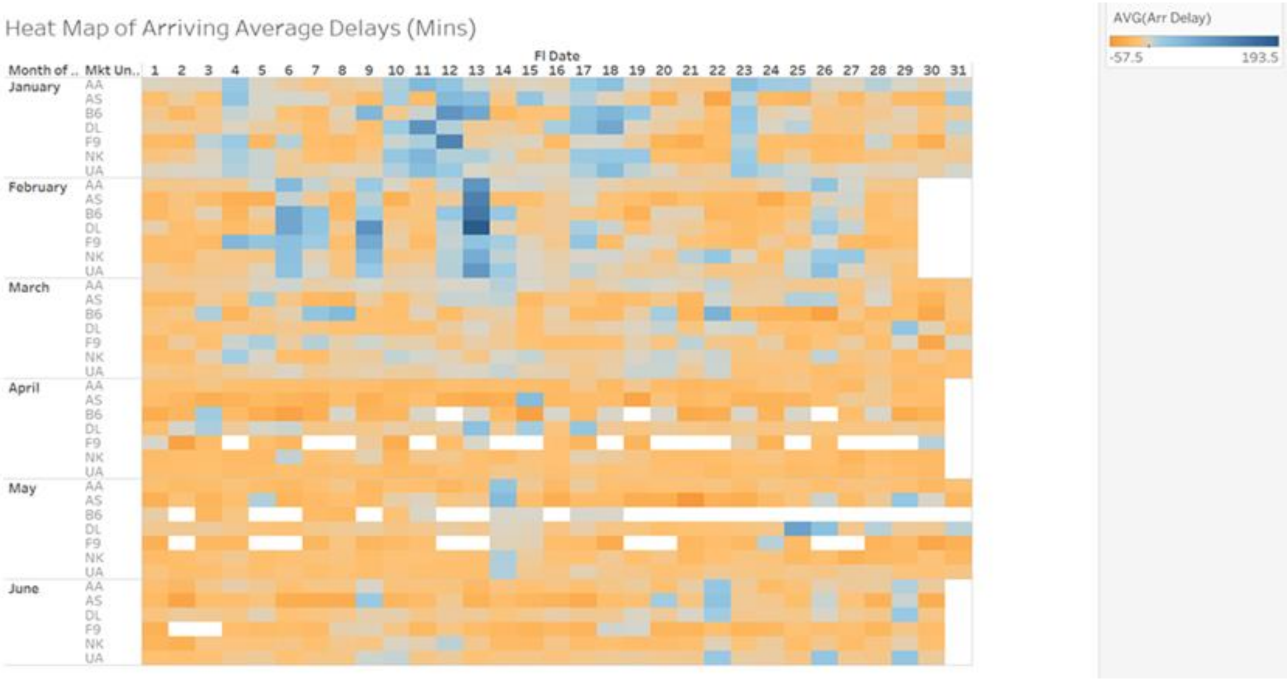
Animated visualization is appropriate for this analysis to create the effect of sudden pop both on the map and the time series plot. Both plots remind the state of shock we were all at the time and how quickly we had to react.

Map View Over Time, Average Delay



My part is to show the departure and arriving average delays on ORD. The reason why I chose to use a heat map is that it can clearly express the variation trend of the data. The color variation will give audiences a first impression on how the delay time changes. And, the audiences can check the scale on a single day within these 6 months. You can see the delay time by month, exact date, and flight carriers. For example, you can check on how one flight carrier's delay times changes in a month. So, it gives a consideration on both outlines and details.

For the data changes in this map, I will use the departure average delays as an example. As you can see, February has the darkest blue, which means it has the most serious delay, and April has more orange, which means it is the best for early departure. I think you can see all the details on the map with the comments I have now. One thing I would like to mention is the comparation between the departure average delays and arriving average delays. Well, obviously, as you can see, the delays on arriving to ORD becomes much better. One reason could the arriving time is also affected by the departure airport from other cities not just ORD. And the other reason could be the time schedule arranged by the ORD.



Based on the heat map I have now. I think there are two aspects that we can do more research in the future. One is that why the arriving delays become better. The previous two reasons were assumptions. We need more data to support them, so we could use more data from other airports to do more analysis. The other is relationship between the average delays and the COVID-19 time's axis to see how the pandemic changes affect the departure and arriving times.

Analysis and Discussion

The flight data for 2020 showed several connected events, all tied to the spread of, and response to, COVID-19 in the United States.

As the ***Comparison of Live/Cancelled Ratio by Airline*** exhibit shows, the volume of cancellations affected all carriers and states, virtually simultaneously. No particular airline appeared to be affected less than others, and through review of the differing Cancellation Codes the COVID-19 related cancellations could clearly be identified. However, United Airlines and Delta Airlines had most of the cancellations affecting them with the impact. The 'Security' category did not appear prior to the March declaration of emergency and aligns with anecdotal knowledge of it being TSA performing the security screenings of passengers.

The ***Map View Over Time, Average Delay*** exhibits for both Departure and Arrival displays show the way that the reduced volume in flights as more and more were cancelled translated into shorter/fewer average flight delays for those flights that were still allowed.

The ***Flight Cancellations Over Time*** exhibit illustrates the sudden increase in cancellations for flights out of O'Hare. This is done using side-by-side timeseries animation for geographical and bar plot. It clearly demonstrates the initial reactions to Illinois' state of emergency orders enacted in early March and how long the impact lasted.

The ***Mix of Cancelled/Live by region*** cartograms displays demonstrate that all states were affected in comparable ways – showing that despite state/regional differences in policies regarding the response to COVID-19, in terms of Air Travel the effects were relatively consistent once the virus had been fully established within the nation. Total flight volume to some states did, as a result of the reduced overall flight volume, drop so low that it could no longer be compared to other states, but all states saw comparable drops within the same time frame. From March to April of 2020 the percentage of flights cancelled vs. flown increased dramatically.

The ***Cancellations and Available Flights*** exhibit then aggregates this National trend data into a single view of how Airlines first responded to COVID-19 restrictions via Cancelled flights, and ultimately the magnitude of reduced flight volume caused by the pandemic.

Individual Report

Ashish Gare

Starting with the exploration of the data, the most important was to first find the trend of the cancellations of the airlines during the 6-month time period. This was first done by the month with the most cancellations and by seeing the greatest number of cancellations occurring in each state. Variables used for both these visualizations were similar and can be found that it was well defined by usage of Airline Market Carriers, Day of the Month, and Cancellation reasons. It was seen every state had a different number of carriers cancelled during certain month with the reason of ‘Security Cancellations’ having the most in the month of March.

Visualization techniques incorporated were stacked bar charts to show the difference between each quantity of flight carriers cancelled. This shows the proportion of flights that were cancelled during the month and the reason for cancellations. This way it was easy to see the cancellation reasons during the different months, see the number of flight cancellations for each month, and separately visualize the reason for these cancellations.

The color choices for the airlines were chosen from the color-blind pallet specifically to avoid color confusion. The data was animated to create a good sense of change as per daily basis of the months. This helped to see the spikes in the data to see the major changes during different months and the cancellation reasons.

Lessons accomplished from this project:

I really loved using Tableau and exploring the various features that were available. I have a better understanding of using the tool. I learnt how to create animations, interactive dashboards and use the data as desired. I was able to utilize the inbuilt tools to create, edit and reform the visualizations as per which conveyed the message best to the audience.

Following my learnings in Tableau and visualizing the various outcomes of the data, it was important to choose the best one that told the story clearly and conveyed its message to the audience in a non-technical form. By keeping the “data, audience, message” it was important to choose only those visualizations that were performing the best in conveying the story I wanted to convey. I also used a little bit of Google Data Studio which had the same ability of Tableau and did some analysis on it. I want to use these tools furthermore to be able to get better and more insightful information and visualizations.

Jared Young

Within the group I participated in the overall research into what data would be a good candidate for visualizations. The group had many ideas to pursue, so most of what I added was simply keeping track of ideas as people noted them and trying to group them into distinct questions that could be answered with a visual. Overall we had lots of willingness to contribute amongst the group members, my assistance beyond pursuing my own assignments was primarily just tracking what needed to be completed and noting who would be completing which part. The group worked together very well, so nothing beyond that was needed in terms of keeping the team on track, which was great. As with everyone on the team, I contributed feedback on early versions of other team member's visualizations as we worked to refine the narrative.

My visualization was the stacked bar showing Cancellations and Total flight volume over time. Originally we noted that we had multiple visualizations using the same general encoding (bar charts) so to add variety and I think better communicate the desired message I adjusted my code to Stacked Area (which was more involved than just changing the geometry, it turned out). I also generated a Mosaic Plot of the same data, which encoded the total flight volume as the width of columns which broke out Cancelled vs. Flown flights by percentage – but vertical height of the area appeared to be a better encoding for comparing volume (of flights) than horizontal widths of columns, so I stuck with the simpler presentation.

I learned two key lessons through the work on this project:

The first being the need to balance exploratory analysis (to start to create specific questions which final visualizations can then answer) with work on refining how that 'answer' is then best stated. Significant time had to be devoted to sifting through the data and deciding what possibilities might be shown before any work could be done on identifying the best type of encoding, etc, but it was difficult to resist the urge to start polishing final materials before it was even clear what the message was.

The second lesson was that there are many ways to accomplish tasks within R code, which is good because the requirements for different geometries can require many adjustments in order to meet the requirements. I'm confident that my existing code has steps that could be consolidated into far fewer statements – but ultimately even taking the long way around there is generally a way to accomplish what was needed, given enough research and rewriting of code.

Hiwot Gebreyohannes

For my visualization, I used two animated timeseries plots side-by-side to show flight cancellations out of O'Hare International Airport from January to June 2020. The dates provided in the dataset represent the original scheduled flight date as opposed to the date of cancellation.

The plots highlight the sudden uptick in cancellations likely in response to state orders.

- For the map view, the increasing count of flight cancellations was plotted with a circle at the destination airport using coordinates pulled from secondary data source. As cancellations increase, the size of the circle on the map increases. This plot emphasizes the sudden increase impacting flights to all states.
- For the time series view, bar plots were used to show total cancellations per day over time. This plot emphasizes the sudden increase in cancellations in early-mid March.

For both plots 10 frames per second were used in sync. The sudden increase in cancellations pop up at the same time for both plots in early March. The same red color scheme was used for both plots so it's easier to relay the main message.

The time element of the cancellation was the focus of these plot. As to not to distract from this message, the stacked color-coded bar plots used during exploratory analysis were further refined to use only one color and exclude cancellation code information since it did not provide additional meaningful insight. Flights to Hawaii, Puerto Rico, U.S. Virgin Islands and Alaska were excluded so it's easier to view on the map. The impact from this exclusion was minimal as most domestic flights out of O'Hare were within contiguous U.S.

The plots also show cancellations leading up to state orders and the aftermath until June 30, 2020. The few cancellation peaks prior to March were due to weather, which were confirmed from the cancellation code provided in dataset and weather reports released around those times.

R/RStudio was utilized to generated both animated plots. The geographical animated plot required data manipulation to create the growing effect on the circles representing accumulated cancelled flights at destination airport. The accumulated flight cancellations were recalculated for each day so that the size of the circle for a specific day would include count of prior cancellations.

If I had more time, I would have made use of the network visualizations taught in lecture 9 to investigate how cancellations at O'Hare airport affected air travel for the rest of the country as it is a hub for many flights, as suggested in the feedback received.

I filtered the original data with 2.7 million flights to 127 thousand flights with origin ORD origin using SQL. I was also able to format the city information provided so it's able to be geocoded within Tableau. Since our primary dataset only contained city and state information for origin and destination airports, I also appended coordinates for each airport from a secondary source so that exact airport address can be mapped. The two datasets were joined using SQL.

The final filtered dataset with coordinate information was used for our visualizations and analysis.

In addition to the preparing the data for visualization, I made sure we were on track for deadlines and project requirements were met. Working on this project, I've learned creating informative and appealing visualizations is an iterative process and hence R might be better suited for certain visualizations.

R and SQL code used are enclosed in submission and in the appendix.

Serena Yang

In this project, I researched on the topics for this project and later my topic was selected. In addition to completing the part assigned to me for each milestone, I was also responsible for the aesthetics and formatting of the entire project, including PowerPoint slides and reports.

For my visualization, the goal is to present the mix of canceled and live flights by region. The original variables from the data file include the destination state name and the canceled flight; the variables I created based on the variables that already exist are week count by date and the flights' total count. For week count by date, I used an aggregate function to display the week number, and for total count is self-explanatory as I used "COUNT" function. The variable canceled flight is the key variable.

If only using Tableau to create a histogram for each state, 50 states need to present; it will confuse the audience to see all the results and waste space in a long horizontal graph. Therefore, the best way to tell the story is to use a geographical plot to show each state's results. Due to the limited technical knowledge, I first used Tableau to create each state's products, cut them into Adobe Photoshop to make the cartogram. Even though there are some blurs for the cartogram after shrinking each state's result, the overall trend from Jan to June is apparent to find when 50 states' results show simultaneously.

After the technical problem is solved, first, I created the percentage of the canceled flights of all. The audience can only see the trend of rate of canceled flights decrease from the graph since the COVID began. There is no way to know the tendency of total flight count decreases in the march from this percentage plot; therefore, I also created the second plot to show the total count of canceled flights.

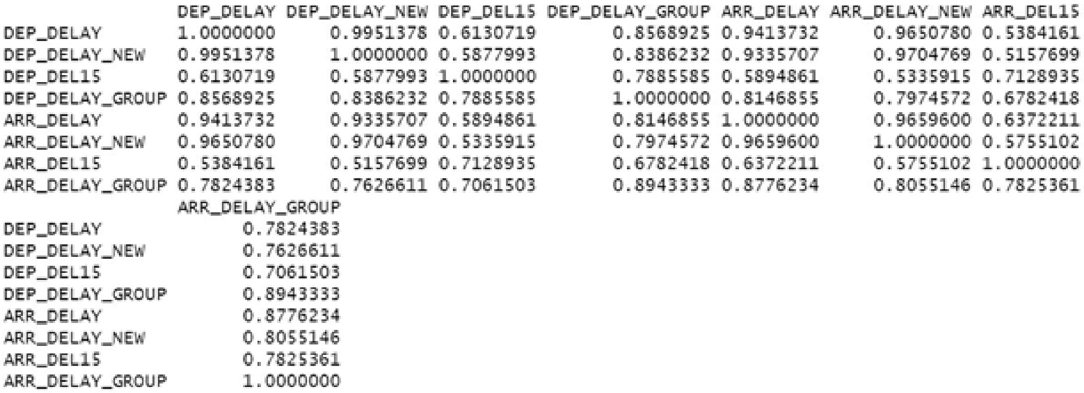
The color I chose to use is the analogous colors since the cartogram will show 50 histograms. If using contrastive colors, the map will be massive, difficult to distinguish the canceled flights part, and challenging to see the overall trend.

From this project, besides, I learned how to use the Tableau tool and used it to generate different type of visualizations for corresponding data insights. I also learned as a data scientist to tell the story from a graph, I should think about "data, audience, message" before creating the chart. I learned that we should try to convert the long horizontal graph into an easily readable graph when dealing with a long horizontal chart. If I have more time, I will try to figure out how to use R or Java to create the cartogram; in other words, trying to improve my technical coding skill. Moreover, I want to learn about dashboard which can be very useful for future projects or jobs as I can form multiple graphs into a single Tableau sheet.

Shuo Wang

Data Review

In this part, I will do the analysis on the average delays of ORD. The variables include flight date, airline carriers, departure delay, departure delay without early departures, departure delay greater than 15 mins, departure delay in number of 15-mins level, arrival delay, arrival delay without early arrivals, arrival delay greater than 15 mins, and arrival delay in number of 15-mins level. Here are the correlations between variables.

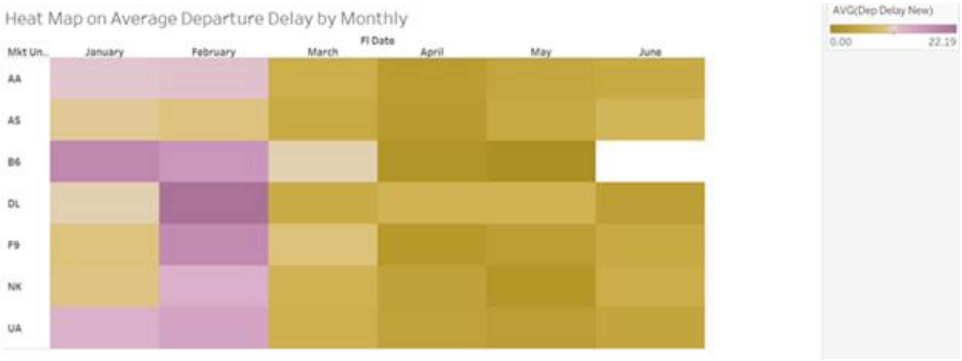


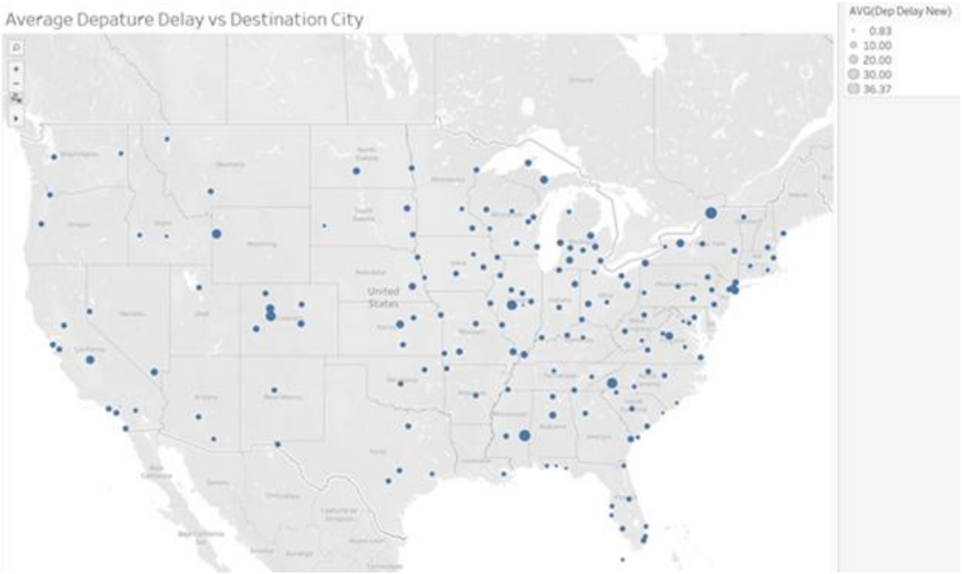
```
                DEP_DELAY DEP_DELAY_NEW DEP_DEL15 DEP_DELAY_GROUP ARR_DELAY ARR_DELAY_NEW ARR_DEL15
DEP_DELAY       1.0000000     0.9951378 0.6130719       0.8568925 0.9413732     0.9650780 0.5384161
DEP_DELAY_NEW   0.9951378     1.0000000 0.5877993       0.8386232 0.9335707     0.9704769 0.5157699
DEP_DEL15       0.6130719     0.5877993 1.0000000       0.7885585 0.5894861     0.5335915 0.7128935
DEP_DELAY_GROUP 0.8568925     0.8386232 0.7885585       1.0000000 0.8146855     0.7974572 0.6782418
ARR_DELAY       0.9413732     0.9335707 0.5894861       0.8146855 1.0000000     0.9659600 0.6372211
ARR_DELAY_NEW   0.9650780     0.9704769 0.5335915       0.7974572 0.9659600     1.0000000 0.5755102
ARR_DEL15       0.5384161     0.5157699 0.7128935       0.6782418 0.6372211     0.5755102 1.0000000
ARR_DELAY_GROUP 0.7824383     0.7626611 0.7061503       0.8943333 0.8776234     0.8055146 0.7825361
                ARR_DELAY_GROUP
DEP_DELAY             0.7824383
DEP_DELAY_NEW         0.7626611
DEP_DEL15             0.7061503
DEP_DELAY_GROUP       0.8943333
ARR_DELAY             0.8776234
ARR_DELAY_NEW         0.8055146
ARR_DEL15             0.7825361
ARR_DELAY_GROUP       1.0000000
```

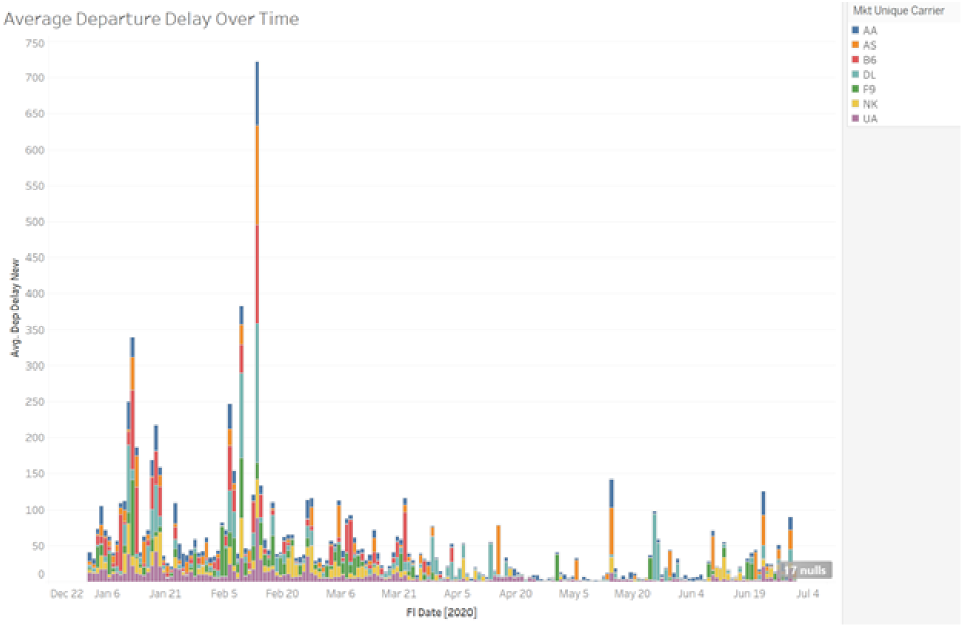
As we can see, the departure delay and the departure delay new has a very high correlation value because the departure delay new is the departure delay ignoring early departures, so they are highly related to each other. And the departure delay also has a high correlation value with arrival delay, which need to be more researched in the future. There could be some reasons for this result. One is that it could be just the number related and no connections between the departure and arrival. The other one is that the departure and arrival has the connections, and that’s the reason why the correlation value is very high, and we need to do more research on that to find out the inner connections. And, the same result of correlation is on the arrival delay and arrival delay new with the same reason.

Visualization analysis

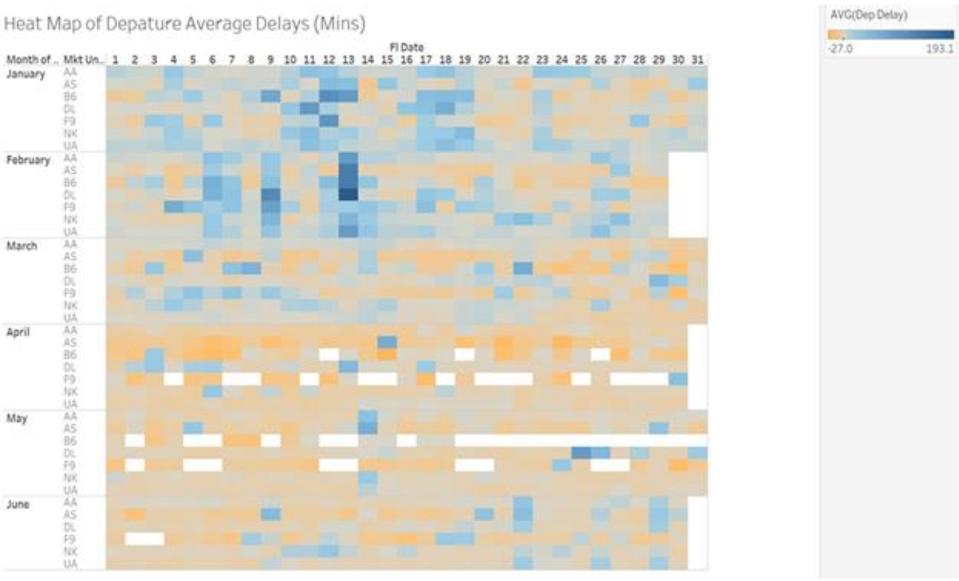
Here are some images that I built on M04 for the first time. The reason why I use the heat map, geography map and bar chart are that I hope to provide a multidimensional visualization to the audiences. I hope to tell them the trend of the delay, how it spread and how it changes on the time series. However, when I double reviewed with the slide about the three most important elements of data visualization (data, audience, and message), I found that my images are too complicated to audiences. So, I was dedicated to add more information one map.


Heat Map on Average Departure Delay by Monthly
Fl Date
Mkt Un..
January
February
March
April
May
June
AA
AS
B6
DL
F9
NK
UA
AVG(Dep Delay New)
0.00
22.19

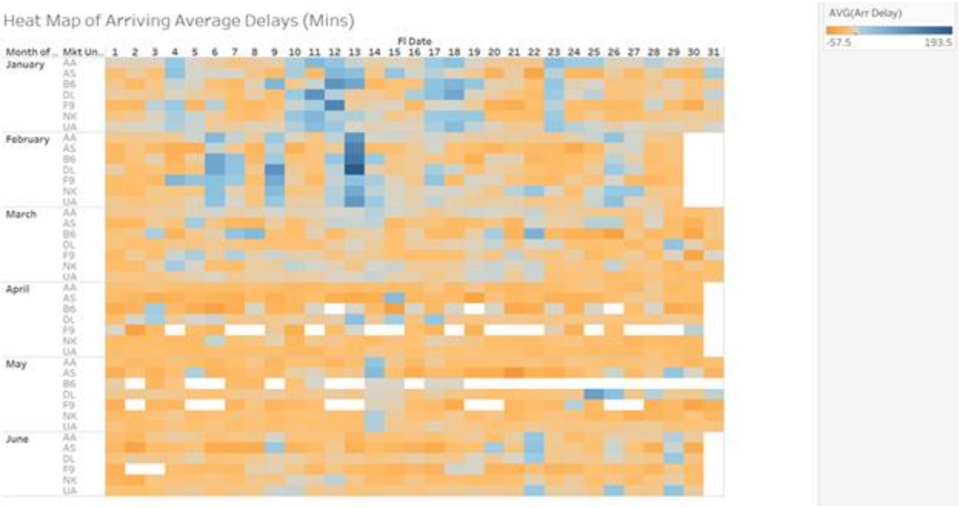

Average Depature Delay vs Destination City
United States
AVG(Dep Delay New)
0.83
10.00
20.00
30.00
36.37


Average Departure Delay Over Time
Avg. Dep Delay New
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
Dec 22
Jan 6
Jan 21
Feb 5
Feb 20
Mar 6
Mar 21
Apr 5
Apr 20
May 5
May 20
Jun 4
Jun 19
Jul 4
Fl Date [2020]
17 nulls
Mkt Unique Carrier
AA
AS
B6
DL
F9
NK
UA

Here are the two heat maps that I used in my final presentation. After searched on the visualizations we have learned in this quarter, I this heat map that will provide more information I hope to give to the audiences.



This is a heat map with average delay time (mins) of departure from Chicago ORD. In this plot, we can clearly see the variation trend of delay time by month, exact date, and flight carriers. The audiences can check with one single day, but also can have a first impression on how the delay time changes. Each color grid means the average delay time (mins) with flight numbers. Dark blue means positive delay time, and the dark orange means early departure. As we can see in this map, February has the most serious delay, and April is the best for early departure.



This is a heat map with average delay time (mins) of arriving to Chicago ORD. As we can see in this map, February has the most serious delay and April to May has the best early departure time. Comparing with the departure average delay map, in obviously, the delays on arriving to ORD becomes much better. One reason is that the arriving time is also affected by the departure airport not just ORD. And the other reason could be the time schedule arranged by the ORD.

Conclusion

From my view, the variation trend of the data can be easily expressed by the visualizations. Still some points need to be improved. One is that there are some empty on the heat map. The reason is because there is no data on that day with one specific airline carrier. But it will be confused to audiences because there is not blank in the color scale. I think it will be better if I change them to another color and add more comments to explain it. The other one is that some label one the graphs are not very clear, because I used some abbreviations on it. This will confuse audiences with the meaning of the graph, so I will definitely change them on the next time.

Further Research

In the future, I think there are two aspects need to be done by more research. During my work, I have analyzed the delay time with date and airline carriers, but I hope to do more analysis on the delay time with different flight to provide the audiences with more accurate information. Also, based on the correlations result I have found, I hope to try to find the relations between departure delays and arrival delays. I may need to use more data analysis from my group members from other aspects.

Reflection Summary.

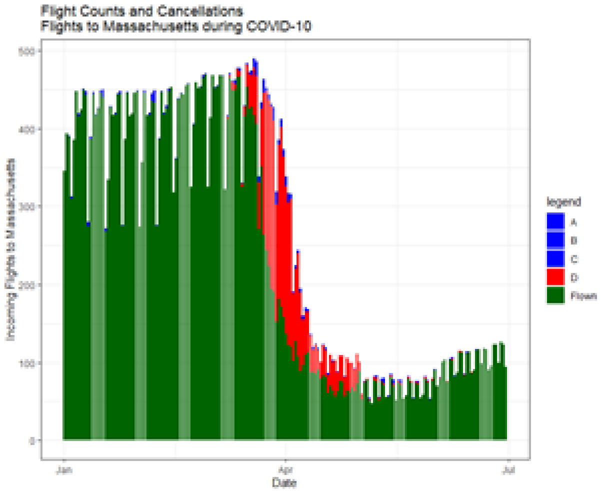
During the whole project, I attended all meetings, and I was able to provide my own ideas to the group conversations based on what I have learned on this course. I wrote and organized the introduction and data set part at the beginning milestones. I did researches on different visualizations online to provide more ideas on how to build the graphs to my group. I picked the delay time analysis to finish the visualization's part and shared my result to my group. What I did learn from this project is that I should consider more building graph method with composite factors and try to provide a simpler graph with more information on it. Also, I should take care more about the details of the visualization’s process such as colors, labels, and outliers. I think I will do better in the future works.

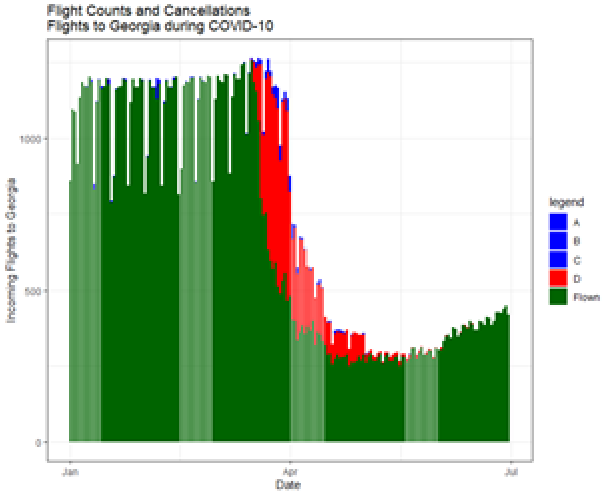
Appendix

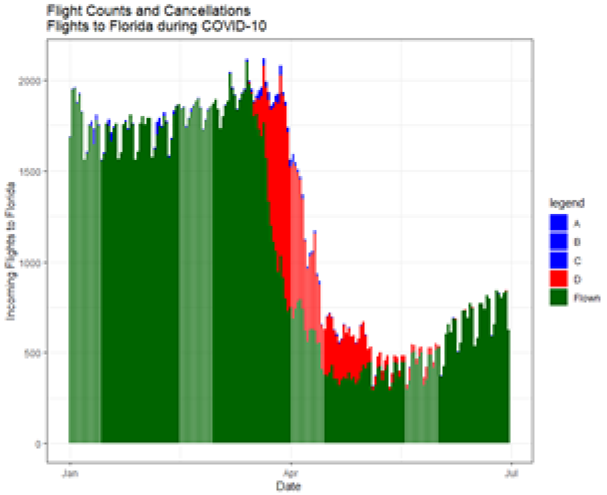
Appendix – Jared Young - Code in R Studio

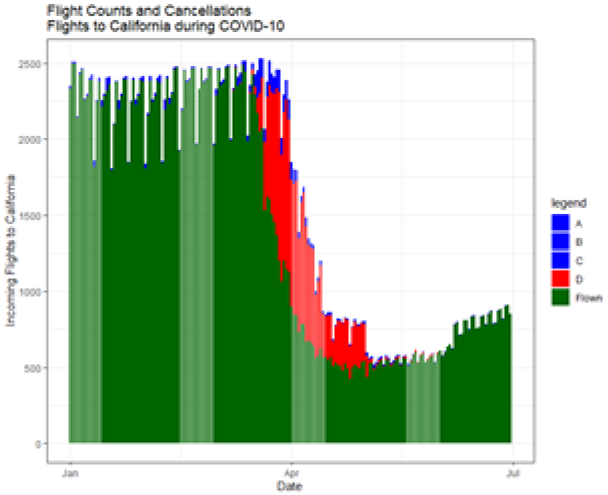
```
jantojun2020 <- read.csv("C:/Users/jster/Desktop/Projects/DePaul/DSC_465/Project_Work/jantojun2020.csv")
ORD_Dep <- jantojun2020 %>%
  filter(ORIGIN=="ORD")
ORD_Dep$CANCELLED <- as.factor(ORD_Dep$CANCELLED)
ORD_Dep$Date<- anytime(as.factor(ORD_Dep$FL_DATE))
ORD_Dep$Date <-as.Date(ORD_Dep$Date)
ORD_Dep$CANCELLATION_CODE <- sub("^$","Flown",ORD_Dep$CANCELLATION_CODE)
ORD_Dep$Date<- anytime(as.factor(ORD_Dep$FL_DATE))
ORD_Dep$Date_Week <- round_date(ORD_Dep$Date,"week")
ORD_Dep$DWK_Char <- as.factor(ORD_Dep$Date_Week)
Ord_Dep_2 <- ORD_Dep %>% count (DWK_Char, CANCELLATION_CODE)
Ord_Dep_3 <- Ord_Dep_2 %>% group_by(CANCELLATION_CODE,DWK_Char) %>% summarize(flights=sum(n))
Ord_Dep_3$CANCELLATION_CODE <- sub("A","Weather",Ord_Dep_3$CANCELLATION_CODE)
Ord_Dep_3$CANCELLATION_CODE <- sub("B","Weather",Ord_Dep_3$CANCELLATION_CODE)
Ord_Dep_3$CANCELLATION_CODE <- sub("C","Weather",Ord_Dep_3$CANCELLATION_CODE)
Ord_Dep_3$CANCELLATION_CODE <- sub("D","COVID-19",Ord_Dep_3$CANCELLATION_CODE)
Final_Out <- Ord_Dep_3 %>% group_by(CANCELLATION_CODE,DWK_Char) %>% summarize(flights=sum(flights))
Final_Out <- Final_Out[-c(1),]  # Outlier record in January 2020 with CANCELLATION_CODE ‘D’
ggplot(Final_Out, aes(x=DWK_Char, y=flights, group = CANCELLATION_CODE, fill=CANCELLATION_CODE)) +
  geom_area(alpha=3/4) +
  scale_fill_manual("Flight Status",
            values = c("Flown" = "#7fcdbb", "Weather" = "#2c7fb8", "COVID-19" = "#edf8b1",
            limits = c("Flown","Weather","COVID-19"))) +
  annotate("text",x="2020-03-15", y=7900,
       label = "3/9/2020 - Declaration of National Emergency\n          7900 Flights per Week",
       colour = "#2c7fb8",
       hjust = 0,
       size = 3.1) +
  annotate("text",x="2020-05-03", y=2801,
       label = "5/17/2020 - Post Cancellation Level\n 1891 Flights per Week",
       colour = "#2c7fb8",
        hjust = 0,
        size = 3.1) +
  theme_bw() +
  ylab("Flights per Week") + xlab("Date")  + ggtitle(("Cancellations and Available Flights in response to COVID-19")) +
  scale_x_discrete(
    breaks = c("2020-01-05",  "2020-02-02","2020-03-01","2020-03-29","2020-04-26","2020-05-24","2020-06-21") ) +
  theme(legend.position = c(0.8,0.75)) +
  #geom_vline(xintercept="2020-03-15", linetype="dotted", color="red", size = .5) +
  #geom_vline(xintercept="2020-05-17", linetype="dotted", color="red", size = .5) +
  geom_segment(aes(x="2020-05-17", y=1891, xend = "2020-05-17", yend = 0), linetype = "dotted") +
  geom_segment(aes(x="2020-03-15", y=7540, xend = "2020-03-15", yend = 0), linetype = "dotted")
```

Exploratory Analysis:









The initial intent of this visualization was to highlight the effects of different changes made at the State level regarding stay at home orders and other restrictions related to COVID-19. Initial versions were made as Bar Plots, at the Day level for maximum granularity and included only total flight counts.

Separating out the Cancelled volume through Color was added as a second layer, and an initial version of small-multiples was created in order to find states with clear differences in the pattern of Cancellations and reduced flight volume. However, no real separation appeared in terms of flight volumes and cancellations by state, even with states known to have had very different levels of response to the pandemic.

This consistency is likely due to the lack of impact on Air traffic that statewide orders and policies would have in comparison to the restrictions placed on Airlines in general regarding screening.

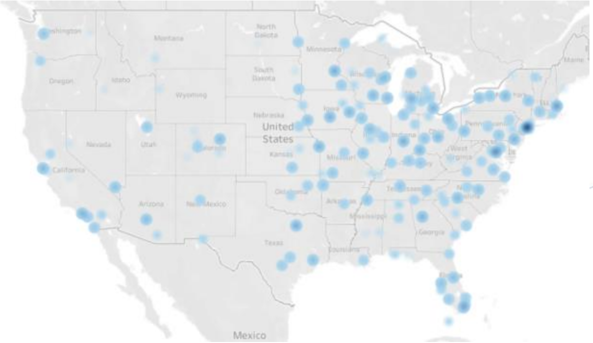
That discovery then led to the choice to show the National level impact, and focus more on the impact of COVID-19 on Total flight availability and volumes than on trying to isolate individual state level differences. Because national air traffic is subject to number of forces beyond what individual states could influence, Flight data does not present itself as a sound basis for isolating the effects of state policies. This finding was further supported by the Cartogram generated by S. Yang, which broke out the patterns of flight counts across all states, which made further analysis at that level unnecessary.

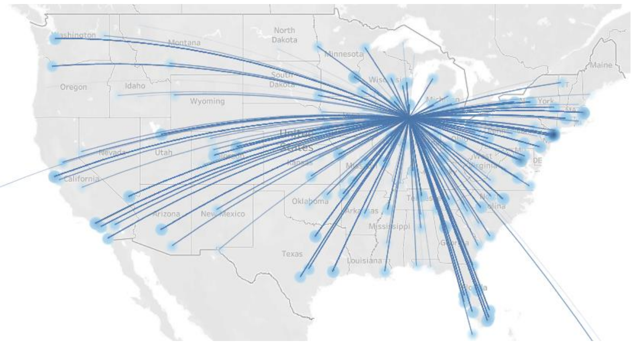
The decision to move to showing the overall impact of COVID-19 on Flight counts and to illustrate the unique event of COVID-19 specific cancellations was the driver for shifting from Stacked Bars to Stacked Area. This took the focus off of individual volatility between the weeks and smoothed the overall presentation, simplifying the presentation of the three key features. The unique wave of COVID-19 cancellations, the High number of average Pre-COVID-19 Flights, and the significantly reduced number of flights during the Pandemic.

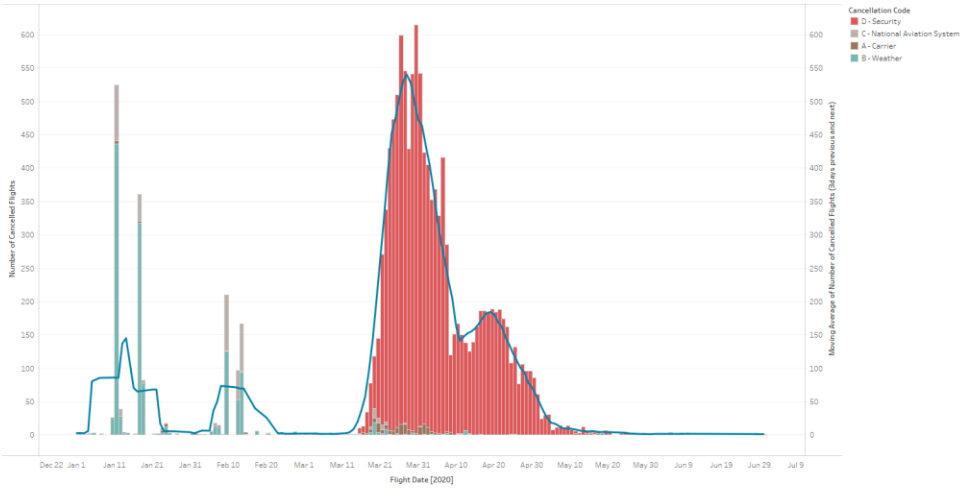
Appendix – Cancellations Overtime - Hiwot Gebreyohannes

Exploratory Visualizations – Hiwot Gebreyohannes

For my initial visualizations, I used Tableau to see if there is any pattern that tells a good story. Geographically, it seemed like cancellations happened across US. To show the effect of time on cancellations on a map, I decide to switch to R as Tableau did not allow for specific customizations. Since both the geographical view and the timeseries bar plot emphasized the story well, I paired them up as side-by-side views.







R Code – Hiwot Gebreyohannes

```
library(tidyverse)
library(ggplot2)
library(gapminder)
library(gganimate)

getwd()
setwd("C:/Users/Hiwot/Desktop/Depaul/OneDrive - DePaul University/2021
Winter/DSC465/Project")
Data <- read_csv("Data/jantojune_2020_ORD_FormatCity_wAirportCoord.csv")

# Remove 1 destination city since coordinates were not available
Data <- Data %>% filter(Dest_City_Name_v2!='Kearney')

##### Data Prep

# Filter for Contiguous US and select desired columns
Data2 <- Data %>% filter(!(DEST_STATE_ABR %in% c('AK','HI','PR','VI'))) %>%
  select(FL_DATE,MONTH,DEST_STATE_NM,Dest_Lon,Dest_Lat,CANCELLED)

# Format Date Column
Data2$FL_DATE <- as.Date(Data2$FL_DATE,format='%m/%d/%Y')

# FIlter for cancelled flights
Data3 <- Data2 %>% filter(CANCELLED==1)

################## Visualization 1 ##############################

# Create a day column and accumulate prior cancellations
Data4 = data.frame()

days =
seq(as.Date('1/1/2020',format='%m/%d/%Y'),as.Date('6/30/2020',format='%m/%d/%Y
'),1)

for (day in days) {
  currentdata = cbind(as.Date(day, origin="1970-01-
01"),Data3[Data3$FL_DATE<=as.Date(day, origin="1970-01-01"),])
  Data4 = rbind(Data4,currentdata)

}

names(Data4)[1] <- 'day'

Data5 <-
Data4[c("day","MONTH","DEST_STATE_NM","Dest_Lon","Dest_Lat","CANCELLED","FL_DA
TE")]

# Base map
states_map <- map_data("state")

# Plot map animation
p1 <- ggplot() +
  geom_polygon(data=states_map,
               aes(x=long, y=lat, group=group),
               colour='black',
               fill=NA) +

geom_count(data=Data5,aes(as.numeric(Dest_Lon),as.numeric(Dest_Lat)),color="fi
rebrick", show.legend = FALSE) +
  scale_size(range = c(0, 13)) +
  labs(x=NULL,y=NULL) +
  theme_classic() +
  transition_time(as.Date(day)) +
  shadow_mark() +
  enter_grow()

# Animate and Save to file
animate(p1,fps=10,height = 550, width =700)
anim_save("DSC465_Project_Animation1_HiwotG.gif")

################# Visualization 2 ##############################

# Plot timeseries
p2 <- ggplot(data=Data3, aes(x=FL_DATE, y=CANCELLED)) +
  geom_bar(stat="identity",color="firebrick")+
  labs(x='Flight Date',y='Number of Cancelled Flights') +
  scale_x_date(date_breaks = "1 month", date_labels =("%B")) +
  theme_classic() +
  theme(axis.text.x = element_text(size=12)) +
  theme(axis.title.x = element_text(size=15)) +
  theme(axis.text.y = element_text(size=11)) +
  theme(axis.title.y= element_text(size=15)) +
  transition_time(as.Date(FL_DATE)) +
  shadow_mark() +
  enter_grow()

# Animate and Save to file
animate(p2,fps=10,height = 550, width =700)
anim_save("DSC465_Project_Animation2_HiwotG.gif")
```

SQL Code – Hiwot Gebreyohannes

```
---------------------
DROP TABLE #TEMP2

SELECT A.*
  ,LEFT([ORIGIN_CITY_NAME],LEN([ORIGIN_CITY_NAME])-4) Origin_City_Name_v2
  ,B.[Latitude] Origin_Lat
  ,B.[Longitude] Origin_Lon
  ,LEFT([DEST_CITY_NAME],LEN([DEST_CITY_NAME])-4) Dest_City_Name_v2
  ,C.[Latitude] Dest_Lat
  ,C.[Longitude] Dest_Lon
INTO #TEMP2
FROM [jantojun2020_v2] A
LEFT JOIN [AIR2] B ON A.ORIGIN=B.[IATA]
LEFT JOIN [AIR2] C ON A.DEST=C.[IATA]
WHERE ORIGIN='ORD'

SELECT COUNT(*) FROM #TEMP2
SELECT COUNT(*) FROM [jantojun2020_v2] WHERE ORIGIN='ORD'

SELECT DISTINCT ORIGIN, Origin_Lat,Origin_Lon
FROM #TEMP2

SELECT DEST,Dest_City_Name_v2, Dest_Lat,Dest_Lon,COUNT(*)
FROM #TEMP2
GROUP BY DEST,Dest_City_Name_v2, Dest_Lat,Dest_Lon
ORDER BY COUNT(*) DESC


SELECT *
FROM #TEMP2


SELECT *
FROM #TEMP2
WHERE Cancelled=1 AND FL_Date='1/11/2020'
```