Analysis of Global Air route Networks

Final review report

Submitted by

Aniket Singh 19BCE2103

Sidharth Bansal 19BCE2080

Ayush Thakur 19BCE0885

For

CSE3021 – SOCIAL AND INFORMATION NETWORKS PROJECT COMPONENT

Under the Guidance of

Dr. VIJAYASHERLY V

Department of Software Systems VIT University, Vellore

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT

Airports are vital to every country and a major relation between countries is often identified by the number of flights shared among them i.e. how many flights serve as a pathway between the two countries.

In this project we aim to identify the country with the highest relations i.e. Country which travels to and welcomes passengers from majority of countries and then perform a sentiment analysis of their residents who are actually confident with their answers and finally conclude with the impact Covid-19 had on airport travels.

Table of content

Introduction

The number of airports a country has is often considered as a direct measure of the flourishment and development of its travel and trade sector, both of which are invaluable pillars of a nation's economic progress and development. Hence by analysing a country's global air traffic, we can often get a fair idea of its financial and economic conditions. Thus, we would first map the nodes of the network on the world map to have a clear visualization. As all of us are well versed with the fact that social networks are seen to depict homophily, that is, the tendency of individuals to associate and bond with similar others. We will also be taking up a case study of certain countries, particularly some communist and antisocial countries like China, North Korea, etc. and attempt to draw parallels between their ideologies and airline networks, which can give us a further insight into their trade routes and the countries with which they have good and friend international relations.

It is a widespread notion 'the more the quantity of a product or service, the better the results yielded'. To challenge this thought, we wish to study the sentiments of the travellers belonging to one of the densest air networks in the world, the US. We would like to explore more about their satisfaction with their airlines even though they have an abundance of it.

Lastly, we would like to study the interrelation between the Covid 19 pandemic and the Aviation industry. It is well known that international travel has triggered the pandemic and 3 that the pandemic has in turn led to a drastic decline in air travel. We wish to emphasize this correlation.

Problem Statement

We plan to study the international air networks and answer the following questions:

Which country has the most airports?

How is a country's development related to the number of airports a country has?

Do any flights in anti-social countries? If yes, what is their source?

Does quantity (degree centrality) always mean quality? – Sentiment analysis of one of the biggest flight networks of the world

How has the international airline network led to a surge in Covid 19 infections across the globe?

In turn, how has the pandemic affected the aviation industry?

Which countries were responsible for fastest transmission of the virus due to their Closeness centrality?

Literature Survey

[1] Provides a contribution to analyse the topological property of complex airlines networks based on a content analysis of Lufthansa networks an integrated multidimensional approach was used in multicriteria analysis in order to evaluate all available data.

[2] Provides an early assessment of the impact of COVID19 on air transport: Just another crisis or the end of aviation as we know it? This assessment is done using the results of a series of in-depth consultations with senior industry executives. Semi-structured interviews were deemed the most appropriate method and focus on the long-term effects on the supply side, the potential long-term changes in passenger behaviour, and the potential effects of long-term control

[3] Conducted research for the recent use of complex network methods for the characterization of the structure of air transport and of its dynamics. It shows that most of the published research has focused on the topological and metric properties of flight 4 networks, where nodes represent airports, and links are made between the two of them, providing information on the presence and frequency of flights.

[4] Considers every airport as a node and the route can be considered as an edge connecting them. The work analyses the USA airport network using different centrality measures of social network analysis. It indicates that Chicago O'hare Intl airport plays an important role. Chicago is therefore one of the cities of Porsche that promotes economic growth in the country.

[5] Extensively investigated on how flight restrictions affect domestic and international travel to countries and continents by analysing time series and graph algorithms.

[6] Says that the world's major airports are directly connected to hundreds of airports without intermediate routes. This connectivity can be described as the network in which the airport becomes a node, and the route becomes a connection line. In this regard, this study analyses the air transport network of 1,060 airports using the social network analysis (SNA) methodology. We consolidated the data from three airline alliances and established a network of 1,060 airports and 5,580 routes in 173 countries

[7] Studies the spatial form of the networks of the 20 largest domestic airlines by analyzing their edge set, degree and betweenness distributions, clustering coefficients and network diameter, and discovers that the whole Chinese domestic airline network is not a scale-free but a small-world network. They analysed each network with two different metrics (i) fitness: minimize the diameter of networks and maximize the passenger flow for each edge; (ii) robustness: compute the change in average path length after removing predetermined hubs and account for the scale of network disintegration. Finally, we compare the robustness of the network of different airlines and find that some small domestic airlines have very good robustness properties

[8] Says that many countries have been implementing various control measures with different strictness levels to prevent the coronavirus disease 2019 (COVID-19) from spreading. With the great reduction in human mobility and daily activities, considerable impacts have been imposed on the global air transportation industry. This study applies a hybrid SARIMA-based intervention model to measure the differences in the impacts of different control measures implemented in China, the U.S. and Singapore on air passenger and air freight traffic

[9] Says that coronavirus outbreak has been highly disruptive for aviation sector, threatening the survival and sustainability of airlines. Apart from massive losses attributed to suspended operations, industry foresee a grim recession ahead. Restrictive movements, weak tourism, curtailed income, compressed commercial activities and fear psychosis are 5 expected to compress the passenger demand from 30 to 60%, endangering the commercial viability of airlines operation. Fragile to withstand the cyclic momentary shocks of oil price fluctuation, demand flux, declining currency, airlines in India warrants for robust structural changes in their operating strategies, business model, revenue, and pricing strategies to survive the long-lasting consequences of Covid-19

[10] Says the global airline networks play a key role in the global importation of emerging infectious diseases. Detailed information on air traffic between international airports has been demonstrated to be useful in retrospectively validating and prospectively predicting case emergence in other countries. In this paper, we use a well-established metric known as effective distance on the global air traffic data from IATA to predict COVID-19 times of arrival (ToA) for different countries as a consequence of direct importation from China. Using this model trained on official first reports from WHO, we provide estimated ToA for all other countries. By combining effective distance with a measure for the country's vulnerability (Infectious Disease Vulnerability Index (IDVI)), we propose a metric to rank vulnerable countries at immediate risk of case emergence.

Modules

1. **Visualize the Global Air route network**

The global airports will be visualized as nodes on the map and adding edges where a flight exists, forms the overall flight social network. We will create a tree map highlighting the nations with most of the airports in the world. We will create a Pie Chart of all the percentage of flights each airline has. The Countries will be sorted on the basis of the number of airports they have. This guarantees the ‘quantity’ aspect of the countries Air Network.

2. **Airline Sentiment Analysis for the most-dense Air Network in the World**

The data set we will use contains tweets in textual form and has their respective sentiments classified as positive, neutral, and negative. We will be doing the analysis of whether quantity always equate with quality i.e., whether the people are happy with the densest air network in the world. This module contains pre-processing of the tweet data, training of model, calculating the accuracy and finding of result. We will be doing a comparison between the share of each sentiment: namely, positive, negative and neutral. Sentimental analysis of each Airline. Sentiments and their relation with tweet length.

3. **Covid 19 and its interrelation with the Air network**

Plotting the Closeness Centrality Choropleth. Closeness centrality is a way of detecting the nodes that are able to spread any information very efficiently through a graph. Here the Closeness Centrality signifies how closely a nation is connected with the rest of countries. The more the closeness centrality of a country, the more closely it is associated with other nations, thus the more likely it is to be a significant contributor to the rise of the pandemic.

Then we will do a comparison of Covid 19 Cases and find the closeness centrality of countries with respect to covid-19 cases.

Finally, we will be finding out how did the pandemic affected the aviation Sector.

Dataset Description

1. **Statistics of Passengers for all carriers**

https://www.transtats.bts.gov/Data_Elements.aspx?Data=2

A1 | fx | Passengers (All Carriers - All Airports)

	A	B	C	D	E
1	**Passengers (All Carriers - All Airports)**				
2	Year	Month	DOMESTIC	INTERNATIONAL	TOTAL
3	2002	10	4,80,54,917	95,78,435	5,76,33,352
4	2002	11	4,48,50,246	90,16,535	5,38,66,781
5	2002	12	4,96,84,353	1,00,38,794	5,97,23,147
6	**2002**	**TOTAL**	**55,18,99,643**	**11,87,04,850**	**67,06,04,493**
7	2003	1	4,30,32,450	97,26,436	5,27,58,886
8	2003	2	4,11,66,780	82,83,372	4,94,50,152
9	2003	3	4,99,92,700	95,38,653	5,95,31,353
10	2003	4	4,70,33,260	83,09,305	5,53,42,565
11	2003	5	4,91,52,352	88,01,873	5,79,54,225
12	2003	6	5,22,09,516	1,03,47,900	6,25,57,416
13	2003	7	5,58,10,773	1,17,05,206	6,75,15,979
14	2003	8	5,39,20,973	1,17,99,672	6,57,20,645
15	2003	9	4,42,13,408	94,54,647	5,36,68,055
16	2003	10	4,99,44,935	96,08,358	5,95,53,293
17	2003	11	4,70,59,495	94,81,886	5,65,41,381
18	2003	12	4,97,57,124	1,05,12,547	6,02,69,671
19	**2003**	**TOTAL**	**58,32,93,766**	**11,75,69,855**	**70,08,63,621**
20	2004	1	4,38,15,481	1,02,52,443	5,40,67,924
21	2004	2	4,53,06,644	93,10,317	5,46,16,961
22	2004	3	5,41,47,227	1,09,76,440	6,51,23,667
23	2004	4	5,32,53,194	1,08,02,022	6,40,55,216

2. **Airports and Airlines data:** containing information on all airports

https://ourairports.com/data/

A1 fx id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	id	ident	type	name	latitude_d	longitude_	elevation_	continent	iso_countr	iso_region	municipali	scheduled_	gps_code	iata_code	local_code	home_link	wikipedia_	keywords
2	6523	00A	heliport	Total Rf He	40.0708	-74.9336	11	NA	US	US-PA	Bensalem	no	00A		00A			
3	323361	00AA	small_airp	Aero B Rar	38.70402	-101.474	3435	NA	US	US-KS	Leoti	no	00AA		00AA			
4	6524	00AK	small_airp	Lowell Fiel	59.94773	-151.693	450	NA	US	US-AK	Anchor Po	no	00AK		00AK			
5	6525	00AL	small_airp	Epps Airpa	34.8648	-86.7703	820	NA	US	US-AL	Harvest	no	00AL		00AL			
6	6526	00AR	closed	Newport H	35.6087	-91.2549	237	NA	US	US-AR	Newport	no						00AR
7	322127	00AS	small_airp	Fulton Airp	34.9428	-97.818	1100	NA	US	US-OK	Alex	no	00AS		00AS			
8	6527	00AZ	small_airp	Cordes Air	34.3056	-112.165	3810	NA	US	US-AZ	Cordes	no	00AZ		00AZ			
9	6528	00CA	small_airp	Goldstone	35.35474	-116.885	3038	NA	US	US-CA	Barstow	no	00CA		00CA			
10	324424	00CL	small_airp	Williams A	39.42719	-121.763	87	NA	US	US-CA	Biggs	no	00CL		00CL			
11	322658	00CN	heliport	Kitchen Cr	32.72737	-116.46	3350	NA	US	US-CA	Pine Valley	no	00CN		00CN			
12	6529	00CO	closed	Cass Field	40.6222	-104.344	4830	NA	US	US-CO	Briggsdale	no						00CO
13	6531	00FA	small_airp	Grass Patc	28.6455	-82.219	53	NA	US	US-FL	Bushnell	no	00FA		00FA			
14	6532	00FD	heliport	Ringhaver	28.8466	-82.3454	25	NA	US	US-FL	Riverview	no	00FD		00FD			
15	6533	00FL	small_airp	River Oak	27.2309	-80.9692	35	NA	US	US-FL	Okeechob	no	00FL		00FL			
16	6534	00GA	small_airp	Lt World A	33.7675	-84.0683	700	NA	US	US-GA	Lithonia	no	00GA		00GA			
17	6535	00GE	heliport	Caffrey He	33.88798	-84.737	957	NA	US	US-GA	Hiram	no	00GE		00GE			
18	6536	00HI	heliport	Kaupulehu	19.83288	-155.978	43	OC	US	US-HI	Kailua-Kon	no	00HI		00HI			
19	6537	00ID	small_airp	Delta Shor	48.1453	-116.214	2064	NA	US	US-ID	Clark Fork	no	00ID		00ID			
20	322581	00IG	small_airp	Goltl Airpo	39.72403	-101.396	3359	NA	US	US-KS	McDonald	no	00IG		00IG			
21	6538	00II	closed	Bailey Gen	41.6445	-87.1228	600	NA	US	US-IN	Chestertor	no						00II
22	6539	00IL	small_airp	Hammer A	41.9784	-89.5604	840	NA	US	US-IL	Polo	no	00IL		00IL			Radio Ranch
23	6540	00IN	heliport	St Mary M	41.5114	-87.2606	634	NA	US	US-IN	Hobart	no	00IN		00IN			
24	6541	00IS	small_airp	Hayenga's	40.0256	-89.1229	820	NA	US	US-IL	Kings	no	00IS		00IS			

3. This dataset contains information whether the sentiment of the tweets in this set was positive, neutral, or negative

https://www.kaggle.com/crowdflower/twitter-airline-sentiment

A1 fx tweet_id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	tweet_id	airline_ser	airline_ser	negativere	negativere	airline	airline_ser	name	negativere	retweet_c	text	tweet_coc	tweet_cre	tweet_loc	user_timezone		
2	5.7E+17	neutral	1			Virgin America		cairdin		0	@VirginAmerica Wha		########		Eastern Time (US & Canada)		
3	5.7E+17	positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plus		########		Pacific Time (US & Canada)		
4	5.7E+17	neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I didr		########	Lets Play	Central Time (US & Canada)		
5	5.7E+17	negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's re		########		Pacific Time (US & Canada)		
6	5.7E+17	negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica and i		########		Pacific Time (US & Canada)		
7	5.7E+17	negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@VirginA		########		Pacific Time (US & Canada)		
8	5.7E+17	positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes,		########	San Franci	Pacific Time (US & Canada)		
9	5.7E+17	neutral	0.634			Virgin America		pilot		0	@VirginAmerica Reall		########	Los Angele	Pacific Time (US & Canada)		
10	5.7E+17	positive	0.6559			Virgin America		dhepburn		0	@virginamerica Well,		########	San Diego	Pacific Time (US & Canada)		
11	5.7E+17	positive	1			Virgin America		YupitsTate		0	@VirginAmerica it wa		########	Los Angele	Eastern Time (US & Canada)		
12	5.7E+17	neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica did y		########	1/1 loner s	Eastern Time (US & Canada)		
13	5.7E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I <		########	NYC	America/New_York		
14	5.7E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica This i		########	NYC	America/New_York		
15	5.7E+17	positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @vir		########		Eastern Time (US & Canada)		
16	5.7E+17	positive	1			Virgin America		sjespers		0	@VirginAmerica Than		########	San Franci	Pacific Time (US & Canada)		
17	5.7E+17	negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon		0	@VirginAmerica SFO-		########	palo alto,	Pacific Time (US & Canada)		
18	5.7E+17	positive	1			Virgin America		ItzBrianHunty		0	@VirginAmerica So ex		########	west covin	Pacific Time (US & Canada)		
19	5.7E+17	negative	1	Bad Flight	1	Virgin America		heatherovieda		0	@VirginAmerica I fle		########	this place	Eastern Time (US & Canada)		
20	5.7E+17	positive	1			Virgin America		thebrandiray		0	I â¤ï¸ flying @VirginAr		########	Somewher	Atlantic Time (Canada)		
21	5.7E+17	positive	1			Virgin America		JNLpierce		0	@VirginAmerica you I		########	Boston | V	Quito		

4. **Closeness centrality of countries**

https://developing-trade.com/wp-content/uploads/2014/11/Work ing-Paper-DTC-2011- 7.pdf

Appendix 2: Alternative Connectivity/Centrality Measures

Country	ACI (a=100)	ACI (Nodal Dist.)	Traffic Share	No. of Links	Clustering Coefficient	Weighted Clustering Coefficient	Theil Index	Kullback-Leibler Distance	Closeness Centrality
Afghanistan	0.27%	5.09%	0.03%	9	0.83	0.96	1.63	3.10	0.49
Albania	0.61%	17.10%	0.11%	13	0.97	0.99	1.82	1.92	0.56
Algeria	0.58%	23.32%	0.20%	23	0.68	0.98	1.68	1.85	0.61
Angola	0.17%	9.93%	0.02%	13	0.49	0.52	2.15	3.68	0.53
Anguilla	0.36%	5.59%	0.04%	4	1.00	1.00	1.12	3.45	0.50
Antigua and Barbuda	0.46%	19.48%	0.15%	19	0.43	0.46	2.53	3.78	0.58
Argentina	1.25%	20.35%	0.43%	21	0.49	0.84	2.10	3.42	0.58
Armenia	0.35%	18.59%	0.08%	17	0.84	0.98	1.55	2.79	0.59
Aruba	0.44%	8.35%	0.12%	7	0.81	1.00	1.32	3.14	0.52
Australia	3.55%	22.18%	0.67%	32	0.34	0.76	2.53	2.61	0.59
Austria	1.34%	49.04%	1.59%	61	0.49	0.92	3.05	0.76	0.82
Azerbaijan	0.37%	25.41%	0.11%	30	0.66	0.91	2.68	2.06	0.64
Bahamas	0.82%	12.39%	0.55%	7	0.76	0.99	0.20	2.33	0.55
Bahrain	0.65%	27.85%	0.33%	34	0.62	0.91	2.75	2.21	0.65
Bangladesh	0.41%	16.95%	0.15%	21	0.72	0.93	2.50	2.35	0.57
Barbados	0.44%	15.19%	0.17%	15	0.51	0.81	2.24	3.56	0.58
Belarus	0.53%	19.83%	0.08%	16	0.87	0.97	2.32	2.07	0.58
Belgium	1.74%	48.51%	1.34%	64	0.41	0.94	3.06	0.51	0.81
Belize	0.31%	6.44%	0.04%	5	1.00	1.00	0.98	2.75	0.50
Benin	0.26%	4.97%	0.03%	14	0.58	0.61	2.32	4.93	0.50
Bermuda	0.46%	10.19%	0.07%	3	1.00	1.00	0.45	2.05	0.55
Bhutan	0.21%	1.39%	0.01%	3	1.00	1.00	0.90	4.71	0.43
Bolivia	0.23%	9.62%	0.05%	7	0.90	0.91	1.86	3.37	0.52
Bosnia and Herzegovina	0.67%	15.26%	0.06%	11	0.96	0.96	2.10	2.55	0.55

PROPOSED ARCHITECTURE AND FRAMEWORK

Basic Architecture

Identify the country with the highest relations i.e. Country which travels to and welcomes passengers from majority of countries and then perform a sentiment analysis of their residents who are actually confident with their answers and finally conclude with the impact Covid-19 had on airport travels.

Detailed Framework

To approach the above problem statements, we would be visualizing and analyzing various datasets. Through these visualizations, most of our questions would be answered to accuracy.

Firstly, we would preprocess the data from the data sets collected and then pIot aII the airports on the World map so as to visualize all the nodes of the sociaI network that we wish to be working with. Next, we would iterate through our data of the connected nodes. If two airports (nodes) have direct flight between them, they should be shown connected on the world map. Through such a visualization of the airline sociaI network, a clear picture about our analysis of both the nodes and the gIobaI airIine routes can be painted.

The next part of our study focused on anaIyzing the density and concentration of airports in all the countries and then ranking them from the highest to the Ieast number of airports. After social network analysis, we will conduct a case study between the countries with the most and least dense flight networks, namely the United States of America and North Korea.

Next, to carry out a quantitative vs qualitative analysis, we wish to perform a sentiment analysis based on tweets. Through such an exercise, we would be establishing an answer to ‘whether a country having a maximum number of airports equates to a good quantity of airlines. To further explore about the fact, we plan to implement a deep Iearning modeI and perform sentiment anaIysis on the Tweets dataset to get a generaI idea of how satisfied the flyers were with the airIines in their country.

This would give us an idea as to how satisfied the people actuaIly are with the services provided to them despite having the most robust and dense flight network in the world.

Finally, to correlate the Covid 19 Crisis with the Global Air Network, we'll plot maps depicting closeness centralities of all the countries and then compare the covid 19 cases in these countries in the initial period of the Pandemic. Further we'll plot a comparative graph between the number of flights scheduled and cancelled during the pandemic to emphasize the impact the pandemic had over the air network.

PROPOSED SYSTEM ANALYSIS AND DESIGN

A clear picture about the global airline network will be presented. With the latitude and longitude of the airport taken as coordinates, the airports across the globe will be visualized as a node on the world map. A comparison between countries conducted through visualizations, equating the number of airports each country houses will be done. The number of direct flights between different countries will be represented by graph.

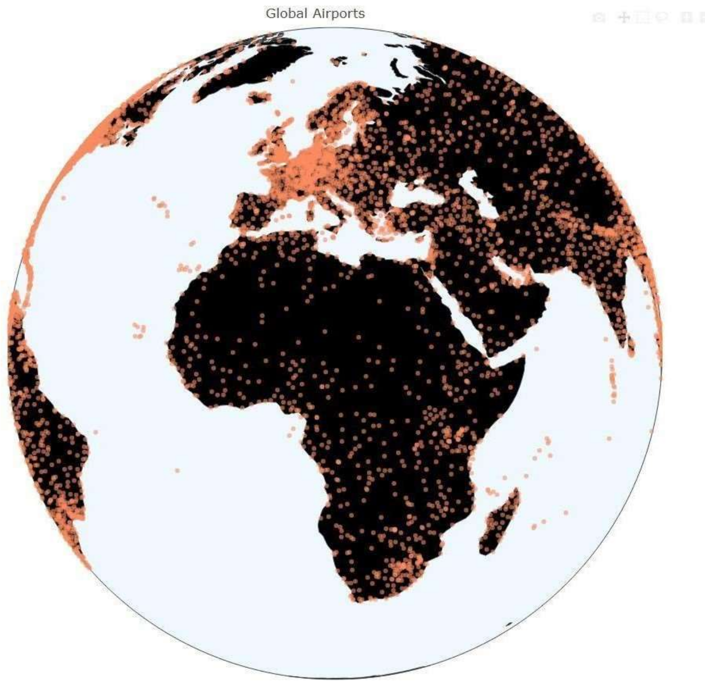
One of the biggest contributors to the rapid surge in Covid 19 cases across different nations was international air travel. This statement has been validated. The pandemic's effect on the aviation industry will be corroborated.

a. Screenshot and Demo along with visualization

- **Visualize the airports across the globe.**

The latitude and longitude of the airport are taken as coordinates and the airport is visualized as a node on the world map.

The data is taken from https://ourairports.com/data/ that provides latest data onairports globally.




```r
geo <- list(
  scope = "world",
  projection = list(type = "orthographic"),
  showland = TRUE,
  resolution = 10,
  landcolor = toRGB("black"),
  countrycolor = toRGB("black"),
  oceancolor = toRGB("aliceblue"),
  showocean = TRUE
)
plot_geo(locationmode = "Greenwich") %>%
  add_markers(data = airport %>%
                filter(Type == "airport"),
              x = ~Longitude,
              y = ~Latitude,
              text = ~paste('Airport: ', Airport_Name),
              alpha = .6, color = "red") %>%
  layout(
    title = "Global Airports",
    geo = geo,
    showlegend = FALSE
  )
print(paste("There are", airport %>%
              filter(Type == "airport") %>%
              nrow(),
            "airports around the world."))
```

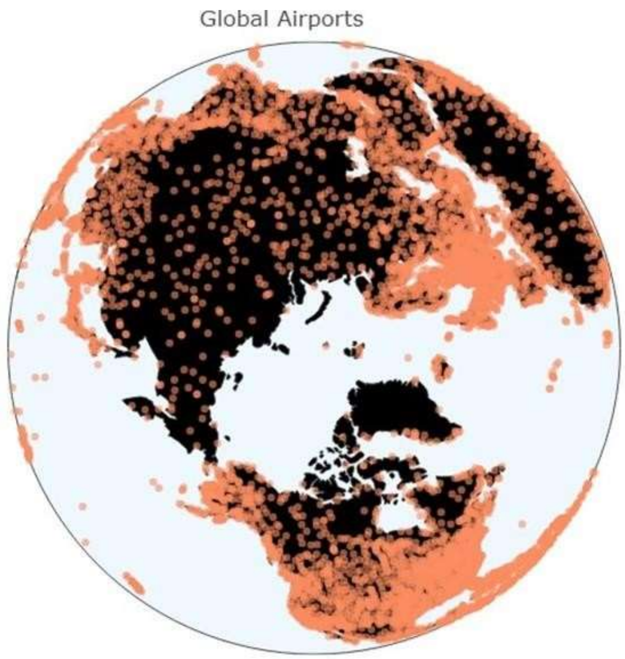


Fig 1 and 2: Marking every airport in the world

- **Visualize the Global Air route network**

The global airports are visualized as nodes on the map and adding edges where a flightexists, forms the overall flight social network.

```
route <- route %>% mutate(id = rownames(route))
route <- route %>% gather('source_airport', 'destination_airport', key = "Airport_type", value = "Airport")
gloabal.flight.route <- merge(route, airport %>% select(Airport_Name, IATA, Latitude, Longitude, Country, City),
                              by.x = "Airport", by.y = "IATA")
world.map <- map_data ("world")
world.map <- world.map %>%
  filter(region != "Antarctica")
ggplot() +
  geom_map(data=world.map, map=world.map,
          aes(x=long, y=lat, group=group, map_id=region),
          fill="white", colour="black") +
  geom_point(data = gloabal.flight.route,
            aes(x = Longitude, y = Latitude),
            size = .1, alpha = .5, colour = "blue") +
  geom_line(data = gloabal.flight.route,
           aes(x = Longitude, y = Latitude, group = id),
           alpha = 0.05, colour = "blue") +
  labs(title = "Global Airline Routes")
```

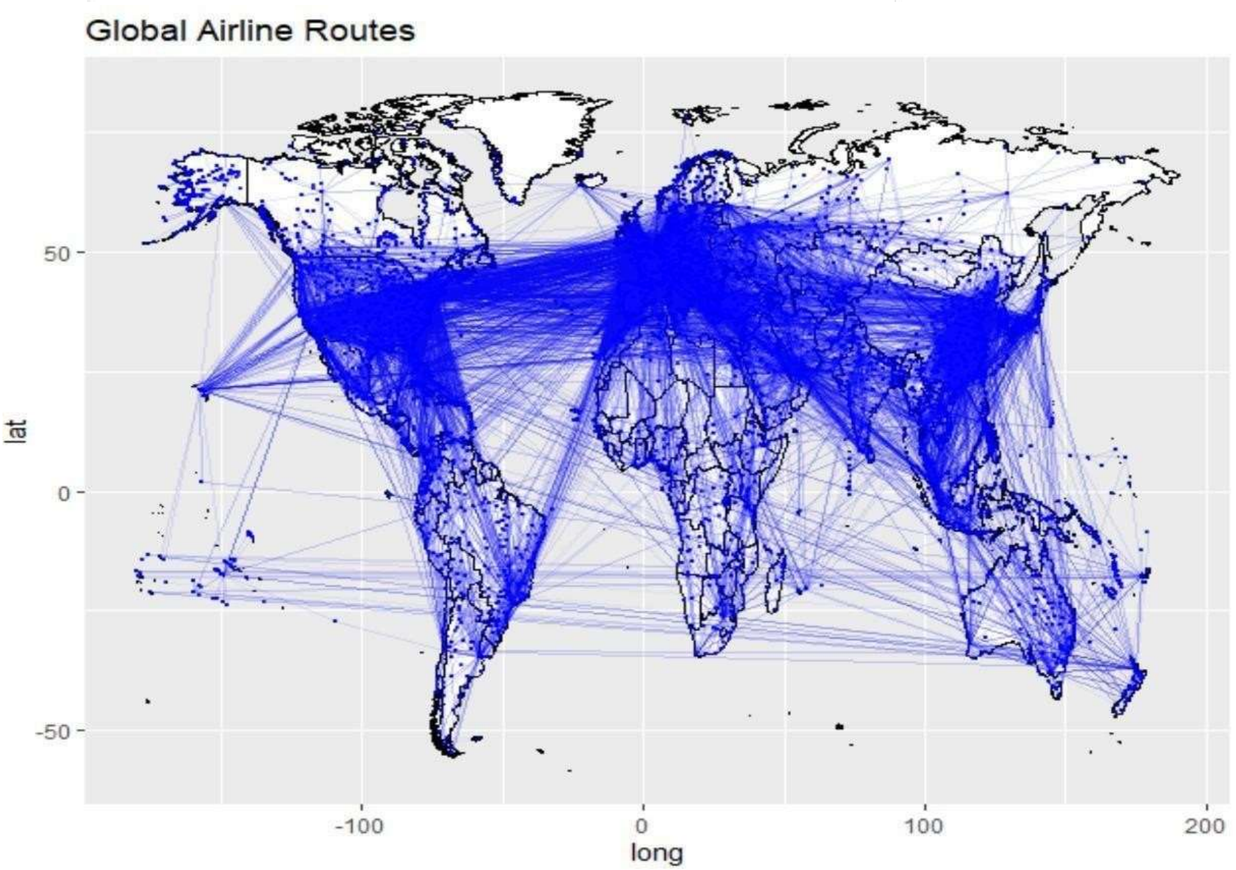


Fig 3: Mark the route of every airport route on Map

- **Sort Countries on the basis of the number of airports they have.**

The USA has the most number of airports as the visualization below shows. This guarantees the 'quantity' aspect of the United States Air Network.

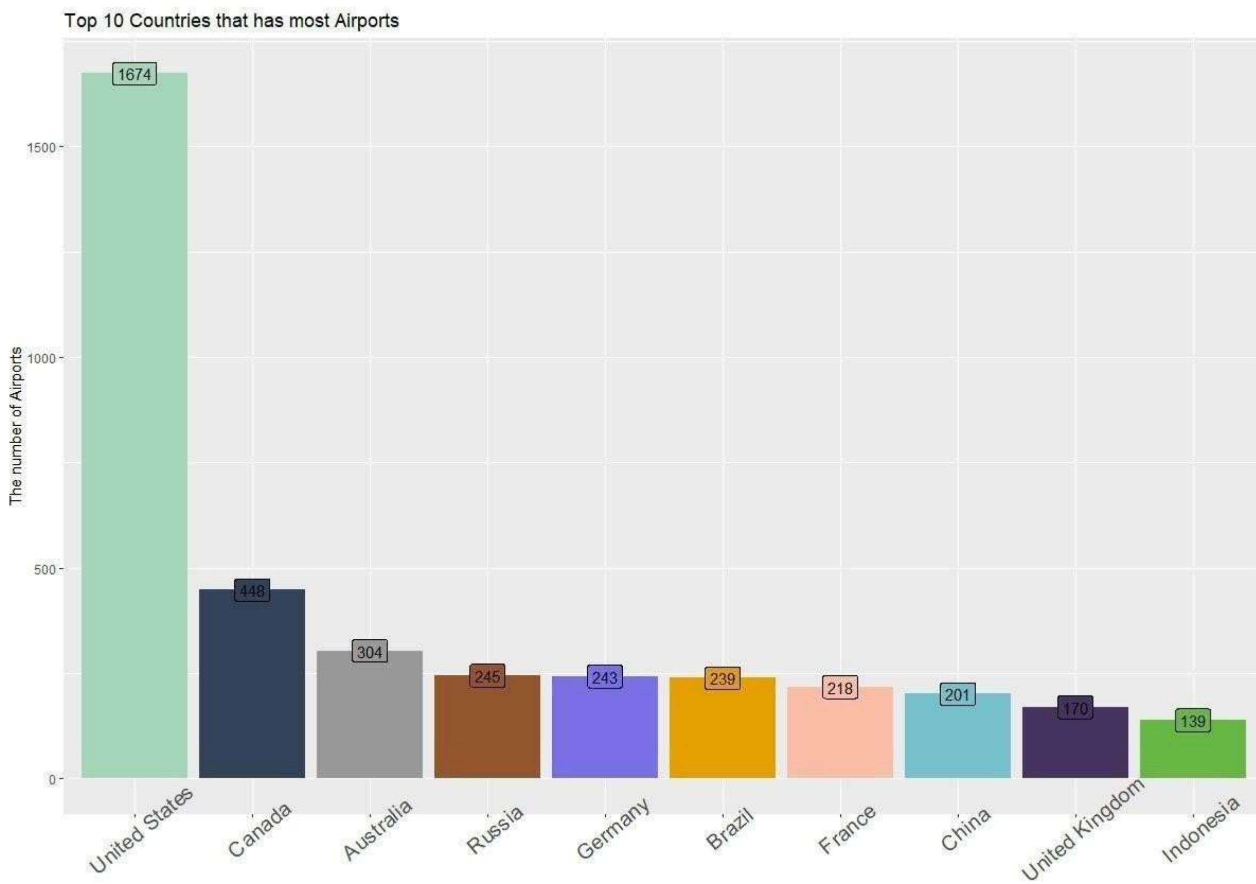


Fig 4: Ascending order of most number of flights in each country

```r
data.frame(table(airport$Country)) %>%
  arrange(desc(Freq)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(Var1, -Freq), y = Freq,fill = Var1,  label = Freq)) +
  geom_bar(stat = "identity", show.legend = F) +
  scale_fill_manual(values=c("#999999", "#E69F00", "#324359","#75c2cc","#fabea7","#7970e5","#66b743", "#92552c","#463360","#a4d7ba"))+
 labs(title = "Top 10 Countries that has most Airports", x = "Country", y = "The number of Airports") +
  geom_label(angle = 45, show.legend = F) +
  theme(axis.text.x = element_text(angle = 40, size = 15))
```

- **Tree Map highlighting the nations with most of the airports in the world.**

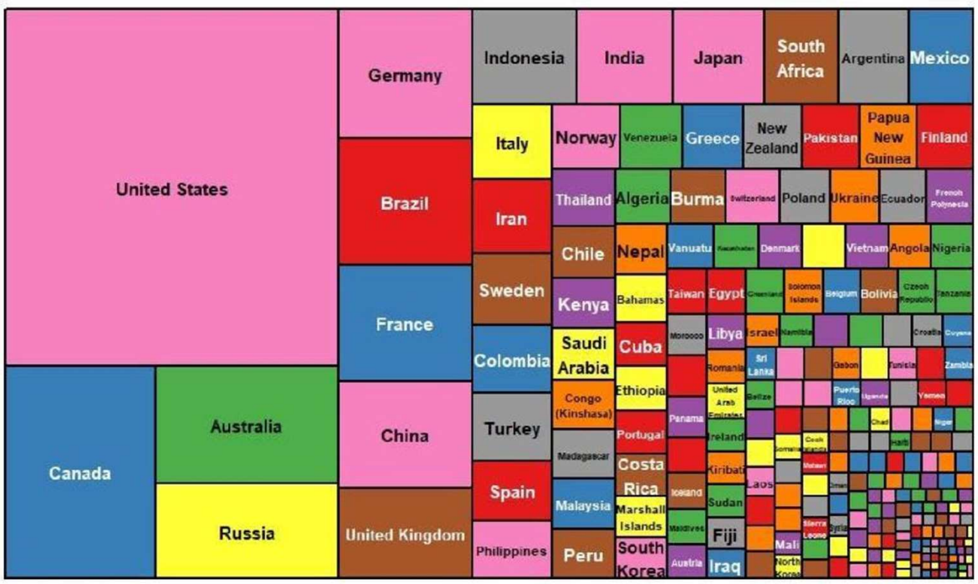


Fig 5: Tree Map of the Nations with most number of Airports.

The visualization shows the number of airports owned by each country with proportion to the area of a rectangle. We see that the developed and economically progressive countries occupy the majority of this area and hold almost 60% of the world's total airports.

Does quantity aIways equate with quality?

Airline Sentiment Analysis for the most-dense Air Network in the World:

The data set we used contains tweets in textual form and has their respective sentiments classified as positive, neutral, and negative.

Pre-Processing:

```
In [1]: import numpy as np
        import pandas as pd
        import re
        import nltk
        nltk.download('stopwords')
        import matplotlib.pyplot as plt
        %matplotlib inline

        [nltk_data] Downloading package stopwords to
        [nltk_data]     C:\Users\admin\AppData\Roaming\nltk_data...
        [nltk_data]   Package stopwords is already up-to-date!

In [2]: import seaborn as sns

In [3]:
        import emoji
        import tensorflow as tf
        from nltk.stem import PorterStemmer
        from tensorflow.keras.preprocessing.text import Tokenizer
        from tensorflow.keras.preprocessing.sequence import pad_sequences
        from sklearn.model_selection import train_test_split

In [4]: airline_tweets = pd.read_csv(r'G:\6th sem\SIN\project\final dbs\Tweets.csv')
        airline_tweets.head()

Out[4]:
```

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn

Fig 6.1 Pre-Processing of data

In [5]:
```
confidence_threshold = 0.6

airline_tweets = airline_tweets.drop(airline_tweets.query("airline_sentiment_confidence < @confidence_threshold").index, axis=(
```

In [6]:
```
tweets_df = pd.concat([airline_tweets['text'], airline_tweets['airline_sentiment']], axis=1)
tweets_df
```

Out[6]:

	text	airline_sentiment
0	@VirginAmerica What @dhepburn said.	neutral
1	@VirginAmerica I didn't today... Must mean I n...	neutral
2	@VirginAmerica it's really aggressive to blast...	negative
3	@VirginAmerica and it's a really big bad thing...	negative
4	@VirginAmerica seriously would pay $30 a fligh...	negative
...	...	...
14397	@AmericanAir right on cue with the delays👌	negative
14398	@AmericanAir leaving over 20 minutes Late Flig...	negative
14399	@AmericanAir Please bring American Airlines to...	neutral
14400	@AmericanAir you have my money, you change my ...	negative
14401	@AmericanAir we have 8 ppl so we need 2 know h...	neutral

14402 rows × 2 columns

In [7]:
```
tweets_df.isna().sum().sum()
```

Out[7]:
```
0
```

In [8]:
```
tweets_df['airline_sentiment'].value_counts()
```

Out[8]:
```
negative    9113
neutral     2997
positive    2292
Name: airline_sentiment, dtype: int64
```

In [9]:
```
sentiment_ordering = ['negative', 'neutral', 'positive']

tweets_df['airline_sentiment'] = tweets_df['airline_sentiment'].apply(lambda x: sentiment_ordering.index(x))
```

In [10]:
```
tweets_df
```

Out[10]:

	text	airline_sentiment
0	@VirginAmerica What @dhepburn said.	1
1	@VirginAmerica I didn't today... Must mean I n...	1
2	@VirginAmerica it's really aggressive to blast...	0
3	@VirginAmerica and it's a really big bad thing...	0
4	@VirginAmerica seriously would pay $30 a fligh...	0
...	...	...
14397	@AmericanAir right on cue with the delays👌	0
14398	@AmericanAir leaving over 20 minutes Late Flig...	0
14399	@AmericanAir Please bring American Airlines to...	1
14400	@AmericanAir you have my money, you change my ...	0
14401	@AmericanAir we have 8 ppl so we need 2 know h...	1

14402 rows × 2 columns

Fig 6.2 Pre-Processing of data (Assigning the general words such as good, bad etc. a value in the range of [-1,1].

```
In [11]: emoji.demojize('@AmericanAir right on cue with the delays👌')

Out[11]: '@AmericanAir right on cue with the delays:OK_hand:'

In [12]: ps = PorterStemmer()
         ps.stem('eating')

Out[12]: 'eat'

In [13]:

         def process_tweet(tweet):
             new_tweet = tweet.lower()
             new_tweet = re.sub(r'@\w+', '', new_tweet)
             new_tweet = re.sub(r'#', '', new_tweet)
             new_tweet = re.sub(r':', ' ', emoji.demojize(new_tweet))
             new_tweet = re.sub(r'http\S+', '',new_tweet)
             new_tweet = re.sub(r'\$\S+', 'dollar', new_tweet)
             new_tweet = re.sub(r'[^a-z0-9\s]', '', new_tweet)
             new_tweet = re.sub(r'[0-9]+', 'number', new_tweet)
             new_tweet = new_tweet.split(" ")
             new_tweet = list(map(lambda x: ps.stem(x), new_tweet))
             new_tweet = list(map(lambda x: x.strip(), new_tweet))
             if '' in new_tweet:
                 new_tweet.remove('')
             return new_tweet

In [14]: tweets = tweets_df['text'].apply(process_tweet)

         labels = np.array(tweets_df['airline_sentiment'])

In [15]: tweets

Out[15]: 0                                       [what, , said]
         1        [i, didnt, today, must, mean, i, need, to, tak...
         2        [it, realli, aggress, to, blast, obnoxi, enter...
         3         [and, it, a, realli, big, bad, thing, about, it]
         4        [serious, would, pay, dollar, a, flight, for, ...
                                        ...
         14397           [right, on, cue, with, the, delay, hand, ]
         14398    [leav, over, number, minut, late, flight, no, ...
         14399    [pleas, bring, american, airlin, to, blackberr...
         14400    [you, have, my, money, you, chang, my, flight,...
         14401    [we, have, number, ppl, so, we, need, number, ...
         Name: text, Length: 14402, dtype: object
```

Fig 6.3 Pre-Processing of data (Removal of extra characters such as @, # andthen putting each of remaining words in an array to preform sentimental analysis.

```
Name: text, Length: 14402, dtype: object

In [16]: # Get size of vocabulary
         vocabulary = set()

         for tweet in tweets:
             for word in tweet:
                 if word not in vocabulary:
                     vocabulary.add(word)

         vocab_length = len(vocabulary)

         # Get max length of a sequence
         max_seq_length = 0

         for tweet in tweets:
             if len(tweet) > max_seq_length:
                 max_seq_length = len(tweet)


         print("Vocab length:", vocab_length)
         print("Max sequence length:", max_seq_length)

         Vocab length: 11257
         Max sequence length: 90

In [17]: tokenizer = Tokenizer(num_words=vocab_length)
         tokenizer.fit_on_texts(tweets)

         sequences = tokenizer.texts_to_sequences(tweets)

         word_index = tokenizer.word_index

         model_inputs = pad_sequences(sequences, maxlen=max_seq_length, padding='post')
```

Fig 6.4 Pre- Processing of data (Tokenization of String, assigning each string a specific value in range of [-1,1])

```
In [19]: word_index

Out[19]: {'to': 1,
          '': 2,
          'the': 3,
          'number': 4,
          'i': 5,
          'flight': 6,
          'a': 7,
          'you': 8,
          'for': 9,
          'on': 10,
          'and': 11,
          'my': 12,
          'is': 13,
          'in': 14,
          'it': 15,
          'of': 16,
          'your': 17,
          'me': 18,
          'have': 19,

In [20]: model_inputs

Out[20]: array([[  49,    2,  218, ...,    0,    0,    0],
                [   5,  191,  102, ...,    0,    0,    0],
                [  15,  138, 2844, ...,    0,    0,    0],
                ...,
                [  76,  506,  434, ...,    0,    0,    0],
                [   8,   19,   12, ...,    0,    0,    0],
                [  37,   19,    4, ...,    0,    0,    0]])

In [21]: model_inputs.shape

Out[21]: (14402, 90)

In [22]: X_train, X_test, y_train, y_test = train_test_split(model_inputs, labels, train_size=0.7, random_state=22)
```

Fig 6.5 storing tweets as an array of integers

Training the Model:

Training

In [23]:
```python
embedding_dim = 32


inputs = tf.keras.Input(shape=(max_seq_length,))

embedding = tf.keras.layers.Embedding(
    input_dim=vocab_length,
    output_dim=embedding_dim,
    input_length=max_seq_length
)(inputs)


# Model A (just a Flatten layer)
flatten = tf.keras.layers.Flatten()(embedding)

# Model B (GRU with a Flatten layer)
gru = tf.keras.layers.GRU(units=embedding_dim)(embedding)
gru_flatten = tf.keras.layers.Flatten()(gru)

# Both A and B are fed into the output
concat = tf.keras.layers.concatenate([flatten, gru_flatten])

outputs = tf.keras.layers.Dense(3, activation='softmax')(concat)


model = tf.keras.Model(inputs, outputs)

tf.keras.utils.plot_model(model)
```

Fig 7.1 Training of Model

Out[23]:

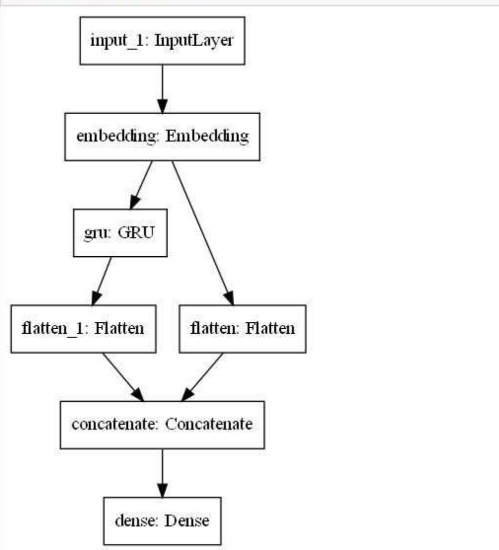


Fig 7.2: Training of Model Flow Chart

```
In [24]: model.compile(
             optimizer='adam',
             loss='sparse_categorical_crossentropy',
             metrics=['accuracy']
         )
```

```
In [25]: batch_size = 32
         epochs = 100

         history = model.fit(
             X_train,
             y_train,
             validation_split=0.2,
             batch_size=batch_size,
             epochs=epochs,
             callbacks=[
                 tf.keras.callbacks.EarlyStopping(
                     monitor='val_loss',
                     patience=3,
                     restore_best_weights=True,
                     verbose=1
                 ),
                 tf.keras.callbacks.ReduceLROnPlateau()
             ]
         )
```

```
Epoch 1/100
252/252 [==============================] - 12s 47ms/step - loss: 0.7798 - accuracy: 0.6674 - val_loss: 0.6592 - val_accuracy:
0.7263
Epoch 2/100
252/252 [==============================] - 10s 41ms/step - loss: 0.5190 - accuracy: 0.7990 - val_loss: 0.5420 - val_accuracy:
0.7833
Epoch 3/100
252/252 [==============================] - 11s 43ms/step - loss: 0.3725 - accuracy: 0.8695 - val_loss: 0.5062 - val_accuracy:
0.7967
Epoch 4/100
252/252 [==============================] - 10s 42ms/step - loss: 0.2773 - accuracy: 0.9081 - val_loss: 0.5073 - val_accuracy:
0.8022
Epoch 5/100
252/252 [==============================] - 10s 40ms/step - loss: 0.2097 - accuracy: 0.9343 - val_loss: 0.5188 - val_accuracy:
0.8071
Epoch 6/100
251/252 [============================>.] - ETA: 0s - loss: 0.1574 - accuracy: 0.9572Restoring model weights from the end of the
best epoch.
252/252 [==============================] - 11s 42ms/step - loss: 0.1578 - accuracy: 0.9571 - val_loss: 0.5350 - val_accuracy:
0.8042
Epoch 00006: early stopping
```

Results

```
In [26]: model.evaluate(X_test, y_test)

         136/136 [==============================] - 2s 12ms/step - loss: 0.4883 - accuracy: 0.8102

Out[26]: [0.4883204698562622, 0.8102291226387024]
```

Fig 8: Results of first analysis

Thus, we could successfully train our model for predicting the sentiment of a tweetwith an accuracy of 81%.

Now we'll be visualising what this data represents and indicates and thus see ifthis abundance also implies satisfaction among the travellers.

1. **Airlines present in the US and their market Share:**

```
In [30]: airline_tweets.airline.value_counts().plot(kind='pie',autopct=

Out[30]: <AxesSubplot:ylabel='airline'>
```

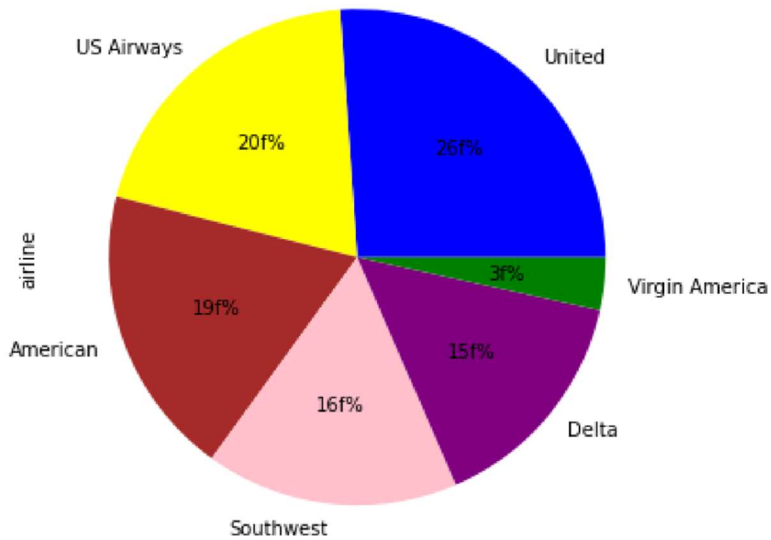


Fig 9: Pie Chart of all the percentage of flights each airline has in US.

2. **A comparison between the share of each sentiment: namely, positive, negative and neutral.**

Overall sentiments in our United States flight dataset (independent of airline) – Post sentiment analysis we see that despite the biggest and the most robust flight network, the overall sentiment of the fliers is still **negative.**

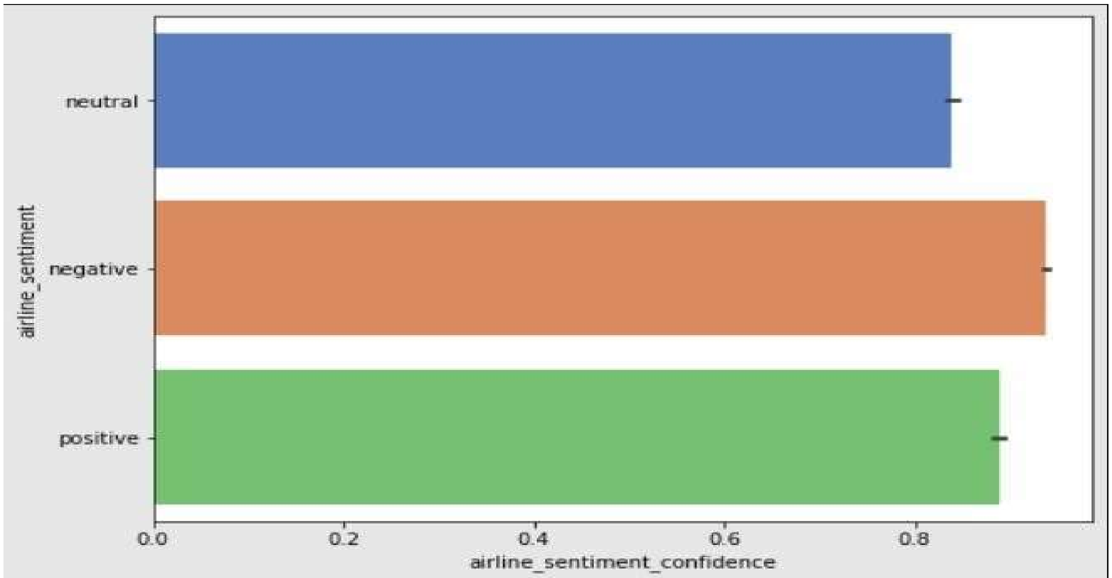


Fig 10: Comparison of total positive, negative and neutral tweets of UScitizens on the Airlines

3. Share Percentage of each sentiment in the entire dataset.

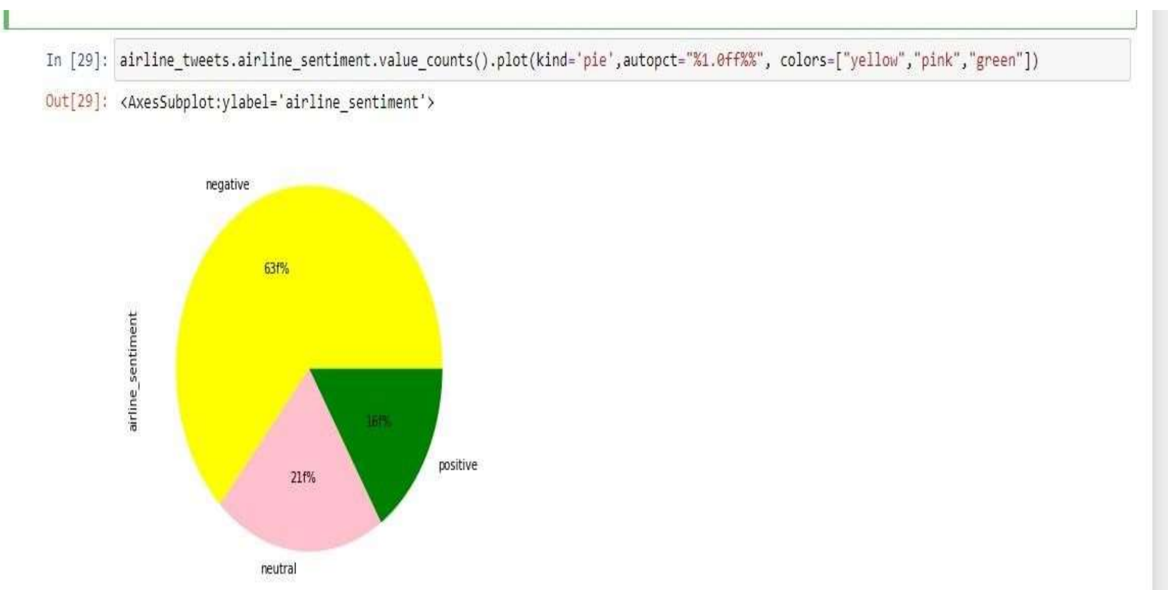


Fig 11. Pie chart representing the percentage of positive, negative and neutral tweets

4. **Airline vs the kinds of sentiment associated with them.- Some of the popular US airlines and the general sentiment the fliers have towards them.**

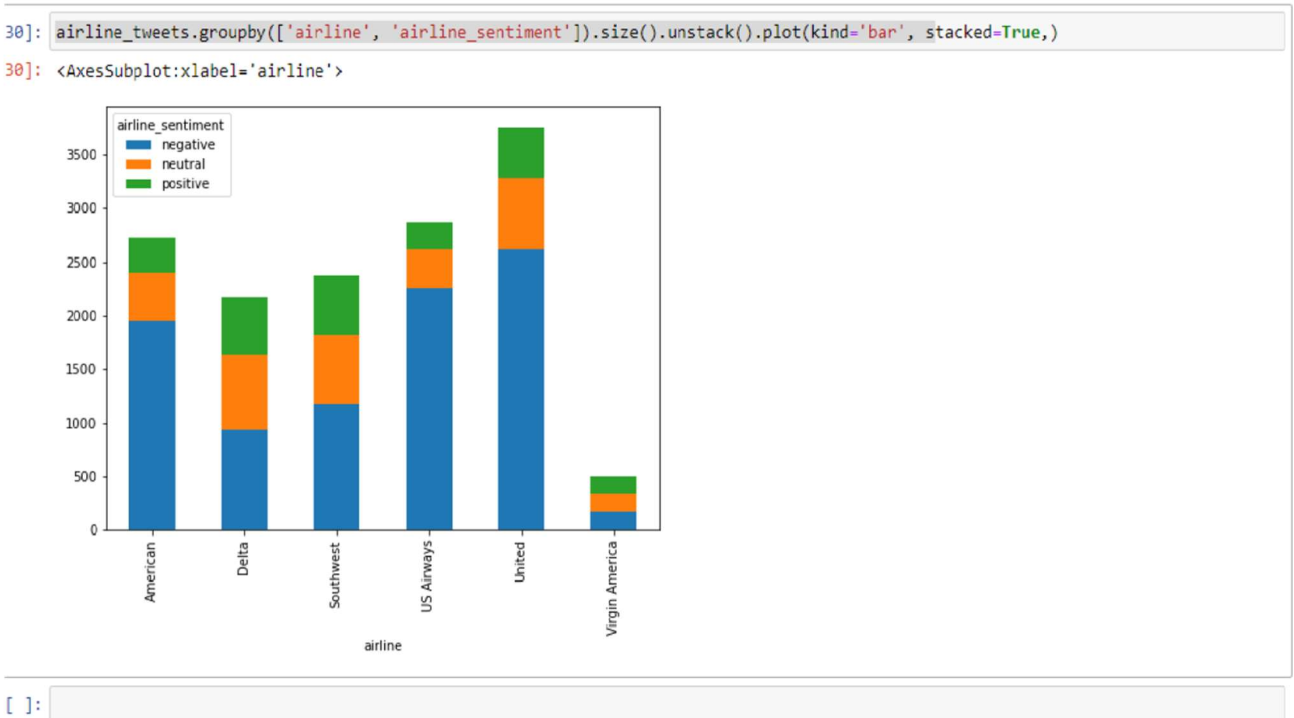


Fig 12. Sentimental analysis of each Airline

5. **Sentiments and their relation with tweet length.** We see, longer the tweet, more likely it is to be negative

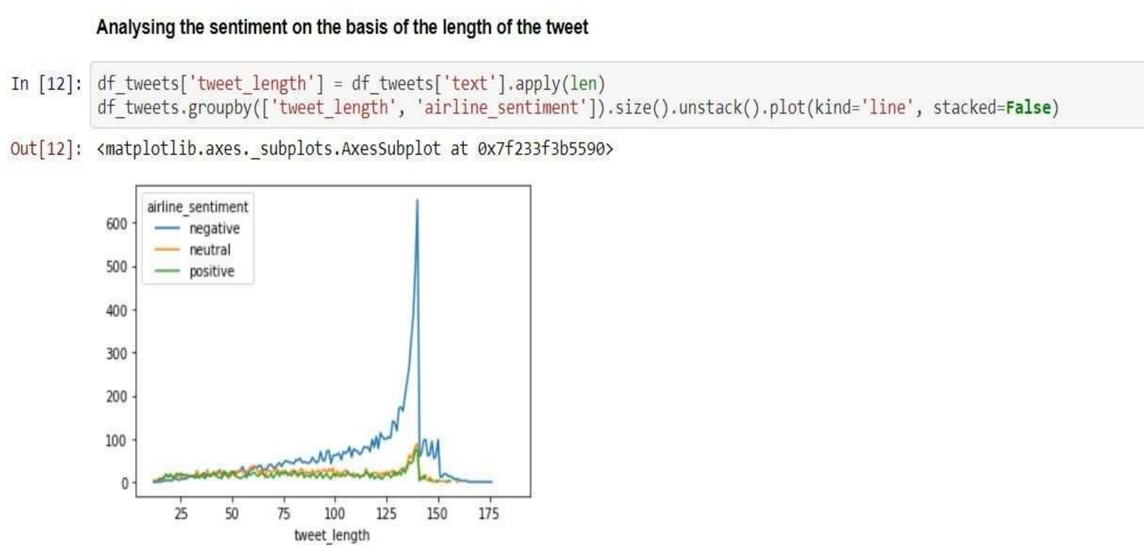
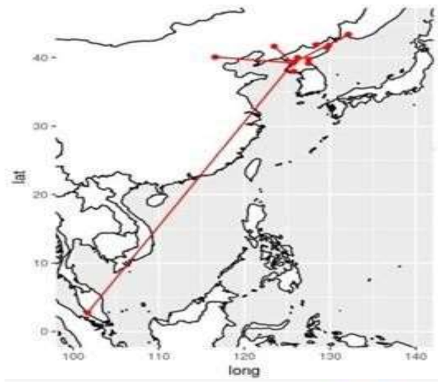


Fig 13: Relation of tweets with its tweet length

Analyzing the flight network of antisocial countries like North Korea




```
NK.gloabal.flight.route.id %>%
  filter(Country != "North Korea") %>%
  select(Airport_Name, Country, City, Latitude, Longitude) %>%
  distinct(City, .keep_all = T) %>%
  DT::datatable(options = list(
    lengthMenu = c(4,1)
  ))
```

	Airport_Name	Country	City	Latitude	Longitude
1	Kuala Lumpur International Airport	Malaysia	Kuala Lumpur	2.745579957962	101.70999908447
2	Beijing Capital International Airport	China	Beijing	40.0801010131836	116.584999084473
3	Taoxian Airport	China	Shenyang	41.6398010253906	123.483001708984
4	Vladivostok International Airport	Russia	Vladivostok	43.398998260498	132.147994995117

Fig: 14.1, 14.2 and 14,3 Analysis of Flights in North Korea and their destinations

Info of airports that have flight access to North Korea. We see that only 4 airports in 3 countries have airline access to North Korea (which is an antisocial country). These countries themselves have high centrality and act as bridge nodes to keep North Korea connected to the rest of the world.

We see that most of these countries are Asian and predominantly those
Asian countries that have a large flight network themselves that keeps
North Korea connected to the rest of the subcontinent and these nodes act as bridge nodes. Let us now analyse the network of these Asian countries and see which are the most prominent bridge nodes that connect North Korea to the rest of the Asian subcontinent and helps maintain its trade

routes without having direct flights to many countries.

From the graph generated from the below code snippet, it is safe to infer that Malaysia and China act as prominent bridge nodes, having a very high degree of centrality and hence holding a high magnitude of importance in the flight network of Asian countries. We see that almost every country has a direct flight to China, hence our social network is seen to obey typical SNA principles like Power Law, by virtue of which if a new country were to be formed, it is more likely to have a direct flight to China or Malaysia than to countries like Bhutan, Bangladesh, etc.

The graph also shows us how weakly North Korea is connected to the network of Asian Countries, having an edge cut of just two, the edge cut of a country can be used as a safe measure to judge how social a particular country is, in the case of a flight social network at least.Usually, the more flights a country has, the better international relations it has.

Shown below is the Social Network showing the international flight network of North Korea with respect to other Asian Countries. With an edge cut of just two, we see that it is an antisocialcountry which does not have international flights to many countries. We also see that Malaysia and China act as bridge nodes for North Korea.

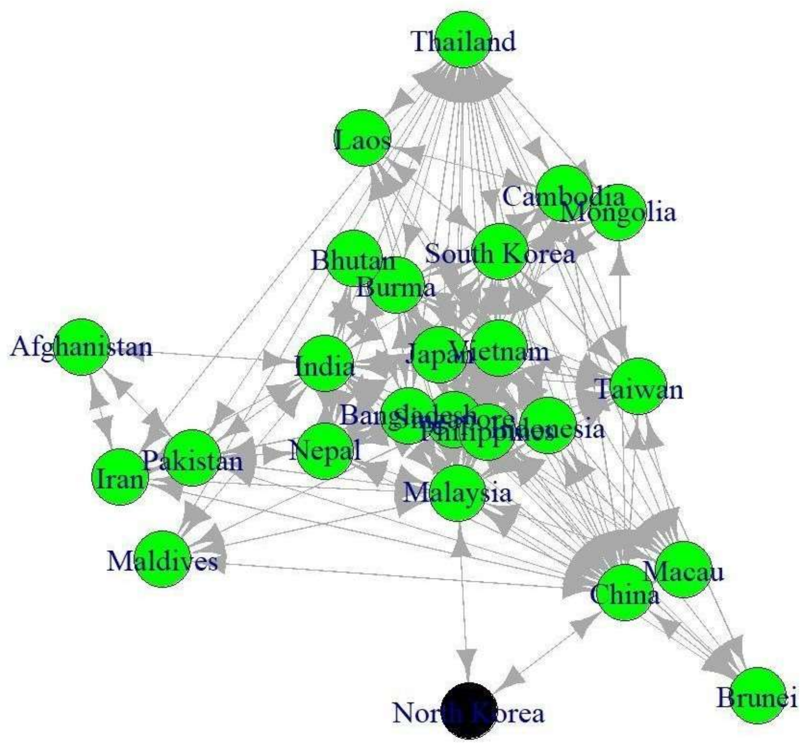


Fig 15: Graph representing the destinations of North Korea and their paths.

Covid 19 and its interrelation with the Air network:

Plotting the Closeness Centrality Choropleth

Following is a choropleth of the closeness centralities of all the nations in the world.

Closeness centrality is a way of detecting the nodes that are able to spread any information very efficiently through a graph. The closeness centrality of a node measures its average farness (inverse distance) to and from all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes.

Here the Closeness Centrality signifies how closely a nation is connected with the rest of countries. The more the closeness centrality of a country, the more closely it is associated with other nations, thus the more likely it is to be a significant contributor to the rise of the pandemic.

In [34]:

```
import plotly.graph_objects as go
import pandas as pd


df = pd.read_csv(r"G:\6th sem\SIN\project\final dbs\country.csv")


fig = go.Figure(data=go.Choropleth(
    locations = df['CODE'],
    z = df['Closeness Centrality'],
    text = df['COUNTRY'],
    colorscale = 'Blues',
    autocolorscale=False,
    reversescale=False,
    marker_line_color='darkgray',
    marker_line_width=0.5,
    colorbar_tickprefix = '',
    colorbar_title = 'Closeness Centrality',
))

fig.update_layout(
    title_text='Closeness Centralities of Countries',
    geo=dict(
        showframe=False,
        showcoastlines=False,
        projection_type='equirectangular'
    )

)

fig.show()
```

Here we see, China has a considerable closeness centrality, which is the reason why the pandemic spread so fast through the globe.

Again, countries like the USA, Spain, Italy too have high centralities, all of which were the worst affected initially.

However, the African countries seem to have lesser centralities thus they we not that tremendously affected.

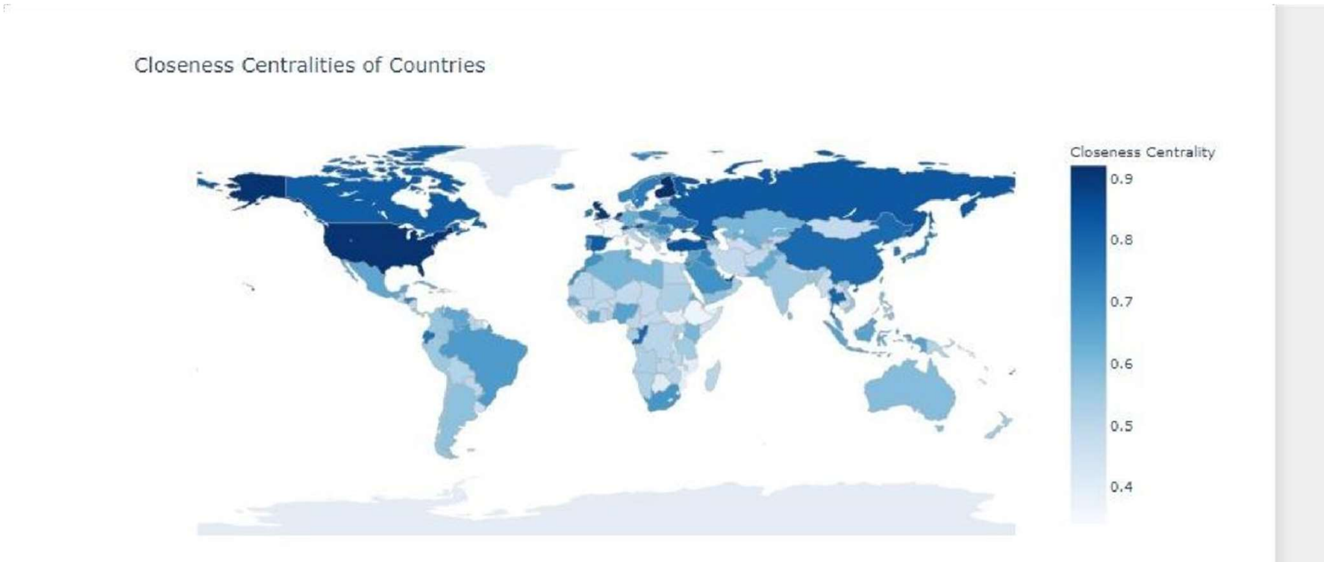


Fig 16: Closeness Centrality of The world map indicating that African Countries were less affected compared to other countries.

Comparison of Covid 19 Cases

```
In [35]: # Library
import matplotlib.pyplot as plt
from matplotlib.patches import Patch
import pandas as pd
# Create bars
covid_cases= pd.read_excel(r"G:\6th sem\SIN\project\final dbs\covid_cases.xlsx")


barwidth = 0.9
```

```
In [36]: plt.rcParams["figure.figsize"] = [16, 10]
colours = {1: "#273c75", 0: "#44bd32"}
ax =covid_cases['Cases_per_million'].plot(kind="bar",color=covid_cases['centrality'].replace(colours),width=barwidth ).legend(
    [
        Patch(facecolor=colours[1]),
        Patch(facecolor=colours[0])
    ], ["Closeness Centrality <= 0.40", "Closeness Centrality >= 0.80"]
)

plt.title("Covid 19 cases reported for each countty on 01/04/2020")
plt.xlabel("Countries with Closeness Centrality <= 0.40 or >= 0.80")
plt.ylabel("Reported Covid 19 Cased per Million people")

plt.xticks([r for r in range(len(covid_cases['Country']))],covid_cases['Country'] , rotation=90)
plt.show()
```

Fig 17: Finding the closeness centrality of countries with respect to covid-19 cases

Here, the countries with CC less than, equal to 0.4 are plotted with Blue color, whereas the ones with CC >= 0.8 are plotted with green. Clearly, the latter had a very significant rise in the number of Covid 19 cases as on 1st April 2020.

We Chose 1st April 2020 as this was an initial period of the pandemic and international travel started to ban eventually thereafter.

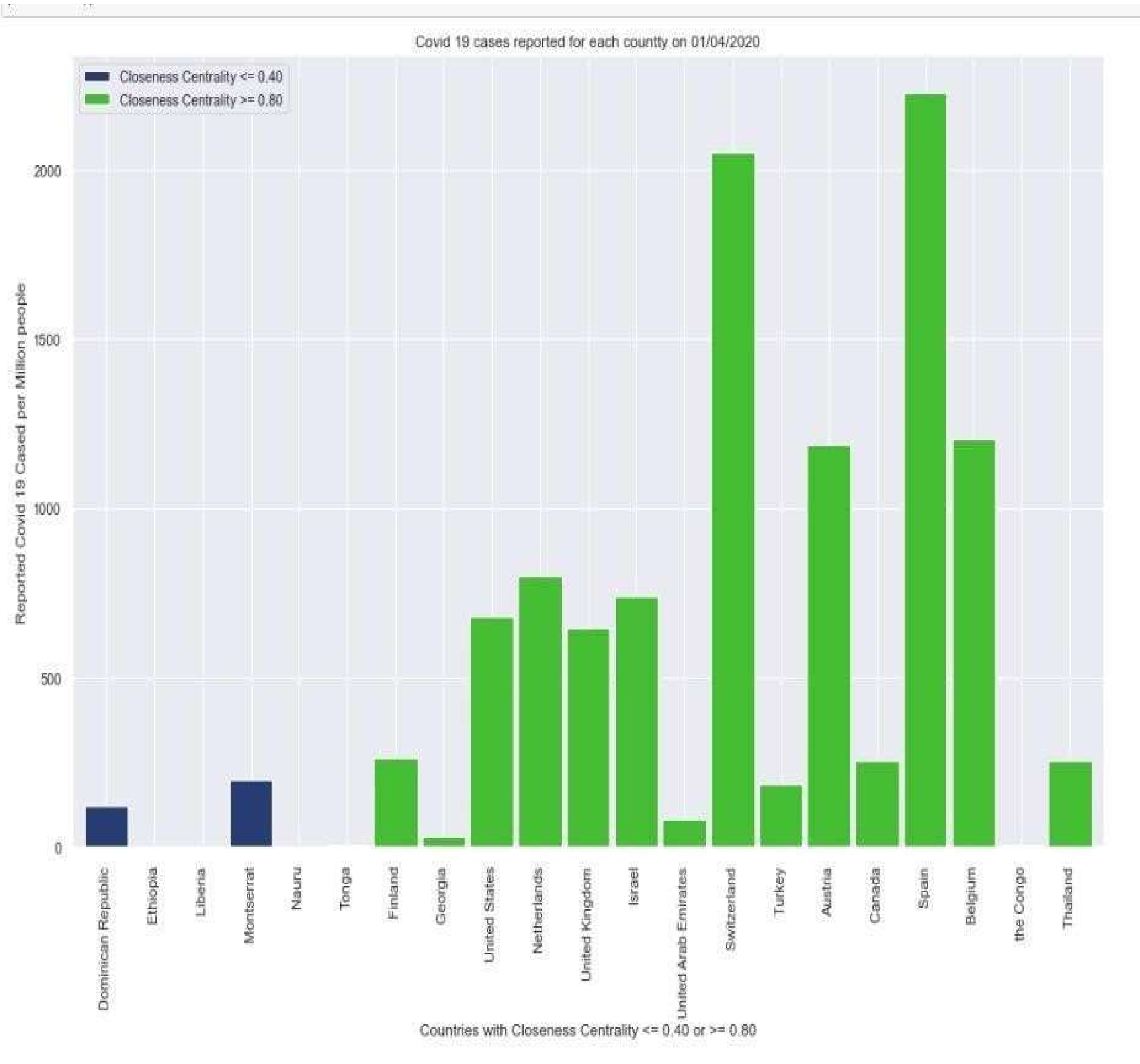


Fig 18: Bar chart representing the closeness centrality of Countries and Number of Cases.

How did the pandemic affect the aviation Sector?

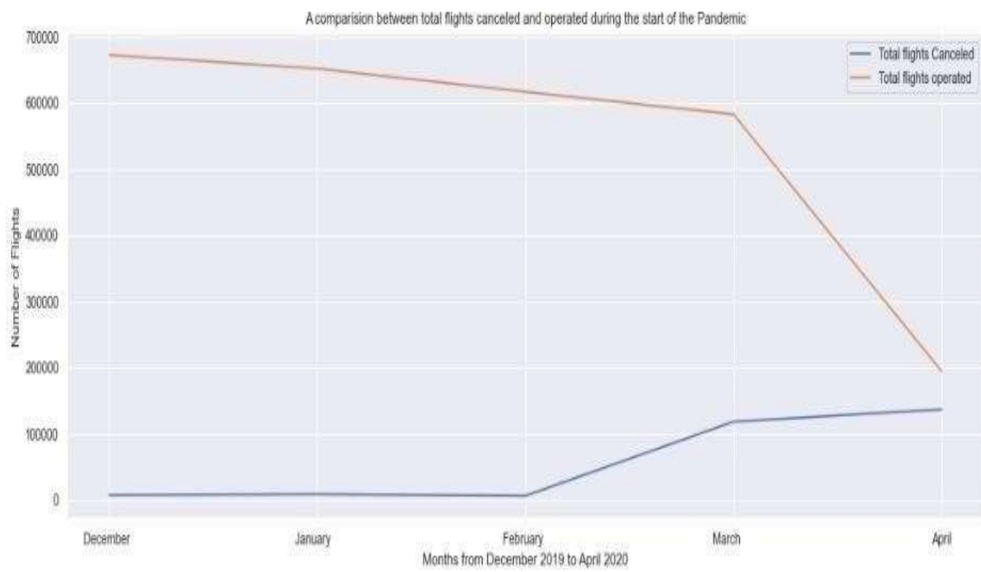
Due to the pandemic, the entire global economy came to a standstill, but the aviation sector remains one of the worst affected.

The number of flights scheduled during the month of March - April 2020 had drastically declined while the number of flights cancelled drastically increased as can be seen.

```
In [37]: flight_status=pd.read_excel(r'G:\6th sem\SIN\project\final dbs\flight_status.xlsx')
         flight_status.head()
         plt.rcParams["figure.figsize"] = [18, 6]
```

```
In [38]: plt.plot(flight_status['Month'],flight_status['Total Cancellations'],label='Total flights Canceled')
         plt.plot(flight_status['Month'],flight_status['Total Flights Operated'],label='Total flights operated')
         plt.ylabel("Number of Flights")
         plt.xlabel("Months from December 2019 to April 2020")
         plt.title("A comparision between total flights canceled and operated during the start of the Pandemic")
         plt.grid(linewidth=1)
         plt.legend()
```

```
Out[38]: <matplotlib.legend.Legend at 0x29c66789b20>
```




```
In [ ]:
```

Fig 19: Line graph comparing number of flights scheduled vs Flights cancelled.

6. Results and discussion

- A clear picture about the global airline network has been presented.
- With the latitude and longitude of the airport taken as coordinates, the airports across the globeshall be visualized as a node on the world map.
- A comparison between countries conducted through visualizations, equating the number of airports each country houses. The number of direct flights between different countries represented via a graph.

- A sentiment analysis stating the tweets about the US airlines as positive, negative and neutral will be carried out. Since the US houses the maximum number of airline services throughout the world, one can expect that most of the US population is happy with their Airline services. This expectation has been validated. The Sentiment Analysis proves, quantity doesn't imply quality always as most of the US travellers seem to be unhappy with the air travel services in their country.
- One of the biggest contributors to the rapid surge in Covid 19 cases across different nations wasinternational air travel. This statement has been validated.
- The pandemic's effect on the aviation industry has been corroborated. The pandemic has led toan exponential decline for the air travel industry in the last 1 year.

REFERENCES

[1] Connectivity and Concentration in Airline Networks: A Complexity Analysis of Lufthansa's Network Aura Reggiani, 1 Peter Nijkamp2 and Alessandro Cento3,

2011 online available - https://papers.tinbergen.nl/11111.pdf

[2] Suau-Sanchez, Pere & Voltes-Dorta, Augusto & Cugueró-Escofet, Natàlia. (2020). An early assessment of the impact of COVID-19 on air transport: Just another crisis or the end of aviation as we know it?. Journal of Transport Geography. 86. 102749. 10.1016/j.jtrangeo.2020.102749.

Online available-https://www.researchgate.net/publication/341905801_An_early_assessment_of_the_impact_of_COVID-19_on_air_transport_Just_another_crisis_or_the_end_of_aviation_as_we_know_it

[3] Modelling the Air Transport with Complex Networks: a short review Massimiliano Zanina Fabrizio Lillob March 1, 2013

Online available - https://arxiv.org/pdf/1302.7017.pdf

[4] Saleena, P. & Swetha, P.K. & D, Radha. (2018). Analysis and visualization of airport network to strengthen the economy. International Journal of Engineering and Technology(UAE). 7. 708-713. 10.14419/ijet.v7i2.9915.

Online available-https://www.researchgate.net/publication/326148275_Analysis_and_visualization_of_airport_network_to_strengthen_the_economy

[5] The Impact of COVID-19 on Flight Networks Toyotaro Suzumura, Hiroki Kanezashi, Mishal Dholakia, June 2020

Online available-https://arxiv.org/abs/2006.02950#:~:text=We%20found%20that%20the%20number,density%20decreased%20during%20this%20period

[6] The Asian Journal of Shipping and Logistics Volume 33, Issue 3, September 2017, Pages 117-125 Online: https://www.sciencedirect.com/journal/the-asian-journal-of-shippingand-logistics

[7] Analysis and Optimization of airline networks: A Case Study of China Gergana Bounova 1, Huang Yan 2, Julia Silvis 3, Qu Li 2, and Li Jianghai 2 https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.527.832&rep=rep1&type=pdf

[8] Impact of different control policies for COVID-19 outbreak on the air transportation industry: A comparison between China, the U.S. and Singapore Fanyu Meng, Wenwu Gong, Jun Liang, Xian Li, Yiping Zeng, Lili Yang

https://doi.org/10.1371/journal.pone.0248361

[9] Sustainability of airlines in India with Covid-19: Challenges ahead and possible way-outs Anshu Agrawal https://link.springer.com/article/10.1057/s41272-020-00257-z

[10] Adiga, Aniruddha & Venkatramanan, Srinivasan & Peddireddy, Akhil & Telionis, Alex & Dickerman, Allan & Wilson, Amanda & Bura, Andrei & Warren, Andrew & Vullikanti, Anil & Klahn, Brian & Mao, Chunhong & Xie, Dawen & Machi, Dustin & Raymond, Erin & Meng, Fanchao & Barrow, Golda & Baek, Hannah & Mortveit, Henning & Schlitt, James & Barrett, Chris. (2020). Evaluating the impact of international airline suspensions on COVID-19 direct importation risk. 10.1101/2020.02.20.20025882.

https://www.researchgate.net/publication/339452693_Evaluating_the_impact_of_internat ional_airline_suspensions_on_COVID-19_direct_importation_risk