Week 13 Final Project Diary

Wong Zi Xin

2023-11-21

Week 9 Diary

1. What is the topic that you have finalised?

The final topic that I have chosen is to create a data story and analyse air transportation data and information in the region of Europe. In the context of my analysis, the definition of Europe refers to the countries in the European Union to date: EU-27 countries. In particular, I will look at EU-27 countries with the most number of passengers and analyse which regions do most people travel to, and observe trends over the years. I chose this topic as I am very interested in travelling, and am curious to derive insights about how the air transport industry was impacted and has picked up again in a post COVID world.

2. What are the data sources that you have curated so far?

I mainly curated my data sources from Kaggle, using datasets containing information about the number of air transport passengers carried by country and datasets containing informaion on the list of airports and airlines globally.

Global datasets:

https://www.kaggle.com/datasets/tjkyner/global-air-transport-data

https://www.kaggle.com/datasets/thedevastator/global-air-transportation-network-mapping-the-wo

https://www.kaggle.com/datasets/johnmwega/trends-and-insights-of-global-tourism

Datasets specifically looking at Europe:

https://www.kaggle.com/datasets/gpreda/passengers-air-transport-in-europe

https://data.europa.eu/data/datasets/38mt9yvqp2fhg7wwgqf13q?locale=en

Week 10 Diary

1. What is the question you are going to answer?

How have air transportation trends in the EU-27 changed over time?

2. Why is this an important question?

According to the International Air Transport Association (IATA), air travel is one of the most important modes of transportation as the aviation industry contributes significantly to global GDP by facilitating global trade, business, tourism and more. With the outbreak of the COVID-19 pandemic IATA revealed the aviation industry suffered a loss of $118 billion in 2020, but with the gradual revival of air travel post-COVID, insights into air travel can better inform strategies to foster economic recovery. Europe was selected as the focus region as according to the United Nations World Tourism Organisation (UNWTO), Europe is the world's top tourist destination.

Sources: https://www.iata.org/en/iata-repository/publications/economic-reports/aviation-economic-benefits/

https://www.iata.org/en/iata-repository/publications/economic-reports/understanding-the-pandemics-impact-on-the-aviation-value-chain/

https://www.unwto.org/impact-assessment-of-the-covid-19-outbreak-on-international-tourism

3. Which rows and columns of the dataset will be used to answer this question?

Columns that are useful to answer this question will be geo (for the country's name), TIME_PERIOD (to represent the corresponding year) and OBS_VALUE (to represent the total number of passengers). All rows are useful as they represent unique data entries of each country by year.

4. Include the challenges and errors that you faced and how you overcame them.

The main dataset that I am using, as seen in the screenshot provided when answering the above question, is not displayed in a very organised and tidy format. The data for each country for each year are all displayed as separate rows. This would make it difficult to create visualisation plots on R, therefore, the first thing I did was to tidy the dataset.

Firstly, I copied over only variables needed (geo, TIME_PERIOD and OBS_VALUE) into a new Excel sheet. Then, I used the pivot table function in Excel to reorganise the data to make it tidy.

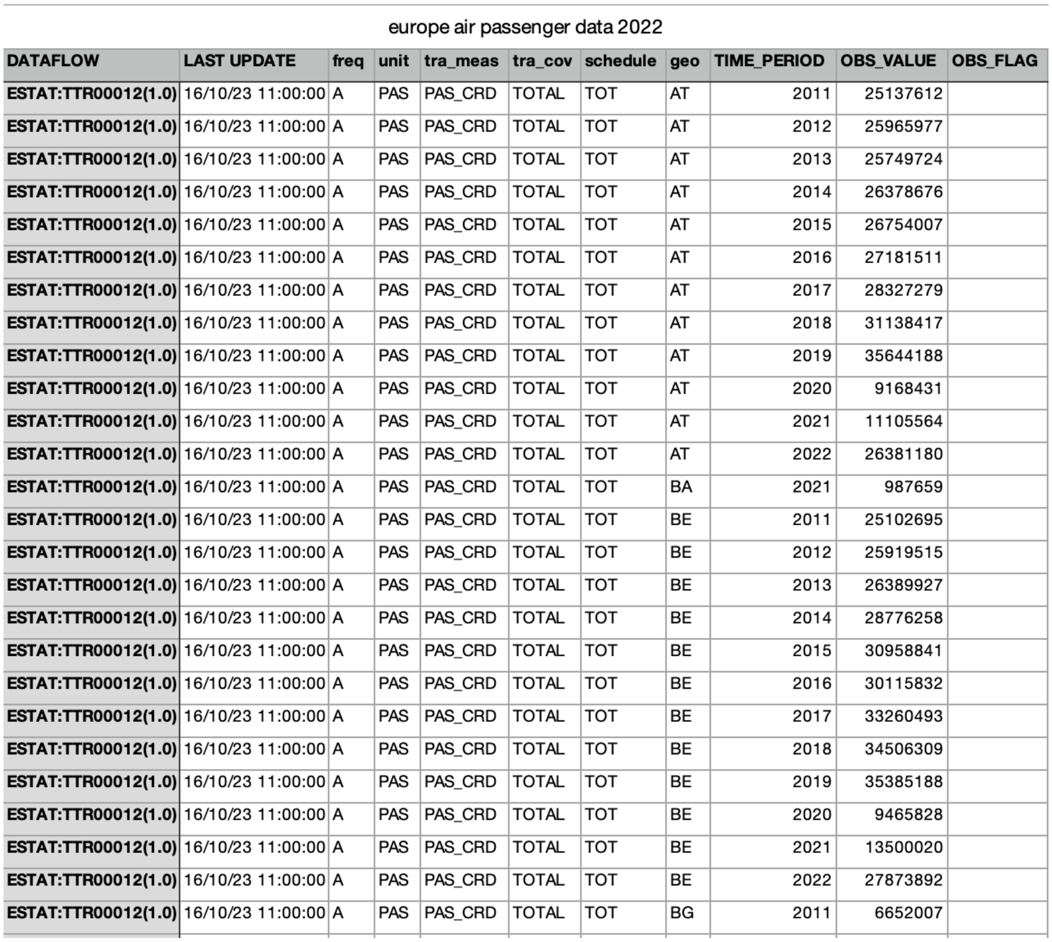
After doing so, I highlighted data for countries which did not belong in the EU-27, as I would need to exclude these data for the purpose of my analysis. Furthermore, some of the countries highlighted contain missing values, therefore I removed these countries from the final dataset to be used before I began my analysis

After cleaning, tidying and filtering out the data, this is the final dataset I used in my preliminary data analyses.

Week 11 Diary

1. List the visualisations that you are going to use in your project

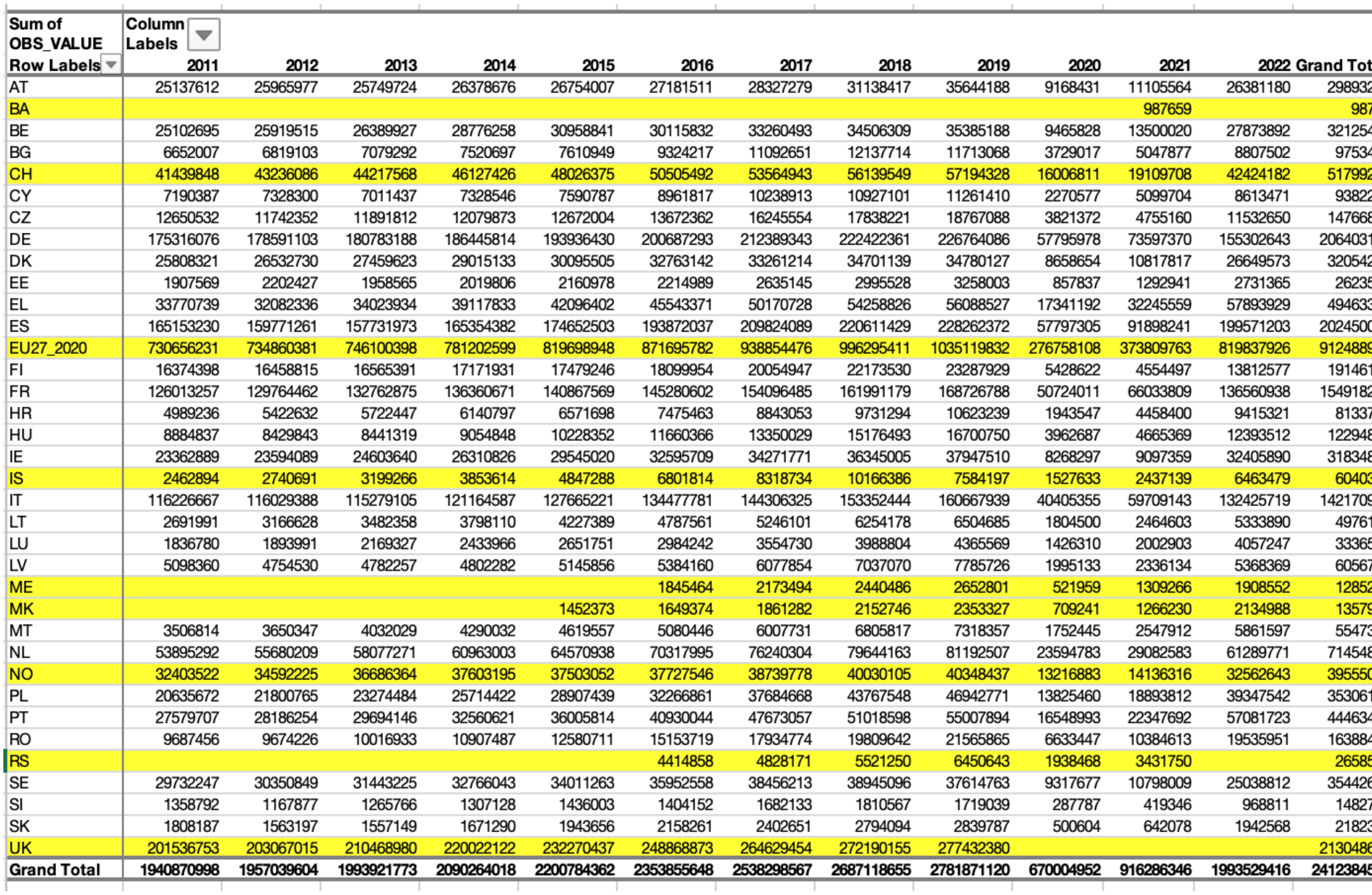
The variables I will be plotting include geo (for the country's name), TIME_PERIOD (to represent the corresponding year) and OBS_VALUE (to represent the total number of passengers). This will help me answer the larger question of how air transportation trends in the EU-27 changed over time, as it provides insight as to how the number of passengers has changed over the years. I will create general visualisation plots in the form of bar plots, looking at the data at a macro level, comparing the total number of passengers across the years and the total number of passengers by country. Given that the dataset contains data from 2011 to 2022, I will compare how the total number of passengers in EU-27 has changed from 2011 to 2022, and the total number of passengers across the 12 years by country. This will reveal broad trends of which



europe air passenger data 2022

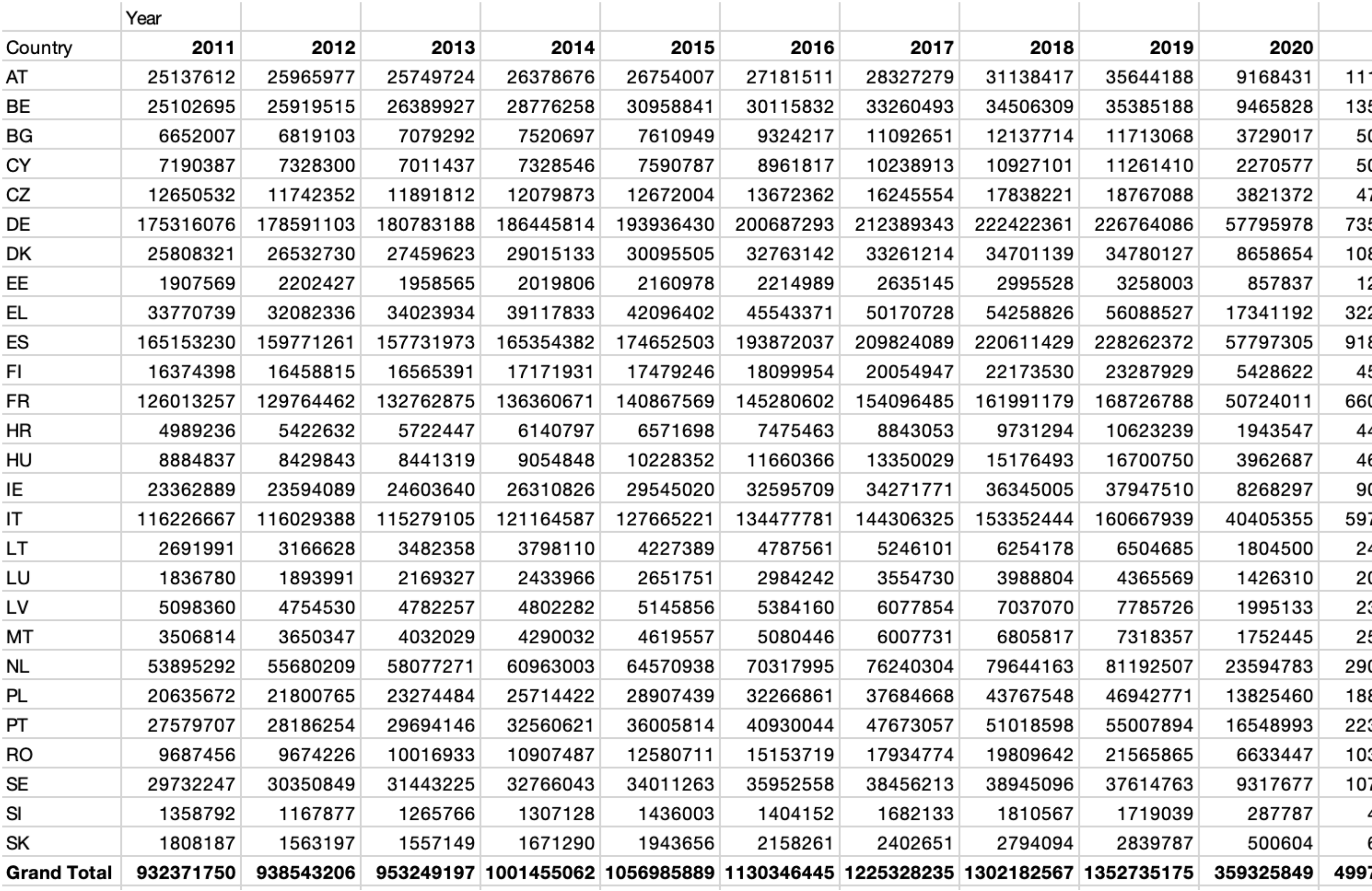
DATAFLOW	LAST UPDATE	freq	unit	tra_meas	tra_cov	schedule	geo	TIME_PERIOD	OBS_VALUE	OBS_FLAG
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2011	25137612	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2012	25965977	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2013	25749724	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2014	26378676	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2015	26754007	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2016	27181511	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2017	28327279	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2018	31138417	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2019	35644188	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2020	9168431	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2021	11105564	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	AT	2022	26381180	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BA	2021	987659	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2011	25102695	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2012	25919515	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2013	26389927	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2014	28776258	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2015	30958841	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2016	30115832	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2017	33260493	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2018	34506309	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2019	35385188	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2020	9465828	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2021	13500020	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BE	2022	27873892	
ESTAT:TTR00012(1.0)	16/10/23 11:00:00	A	PAS	PAS_CRD	TOTAL	TOT	BG	2011	6652007	

Figure 1: Screenshot of a portion of dataset used



Sum of OBS_VALUE	Column Labels												
Row Labels	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Grand Tot
AT	25137612	25965977	25749724	26378676	26754007	27181511	28327279	31138417	35644188	9168431	11105564	26381180	298932
BA											987659		987
BE	25102695	25919515	26389927	28776258	30958841	30115832	33260493	34506309	35385188	9465828	13500020	27873892	321254
BG	6652007	6819103	7079292	7520697	7610949	9324217	11092651	12137714	11713068	3729017	5047877	8807502	97534
CH	41439848	43236086	44217568	46127426	48026375	50505492	53564943	56139549	57194328	16006811	19109708	42424182	517992
CY	7190387	7328300	7011437	7328546	7590787	8961817	10238913	10927101	11261410	2270577	5099704	8613471	93822
CZ	12650532	11742352	11891812	12079873	12672004	13672362	16245554	17838221	18767088	3821372	4755160	11532650	147668
DE	175316076	178591103	180783188	186445814	193936430	200687293	212389343	222422361	226764086	57795978	73597370	155302643	2064031
DK	25808321	26532730	27459623	29015133	30095505	32763142	33261214	34701139	34780127	8658654	10817817	26649573	320542
EE	1907569	2202427	1958565	2019806	2160978	2214989	2635145	2995528	3258003	857837	1292941	2731365	26235
EL	33770739	32082336	34023934	39117833	42096402	45543371	50170728	54258826	56088527	17341192	32245559	57893929	494633
ES	165153230	159771261	157731973	165354382	174652503	193872037	209824089	220611429	228262372	57797305	91898241	199571203	2024500
EU27_2020	730656231	734860381	746100398	781202599	819698948	871695782	938854476	996295411	1035119832	276758108	373809763	819837926	9124889
FI	16374398	16458815	16565391	17171931	17479246	18099954	20054947	22173530	23287929	5428622	4554497	13812577	191461
FR	126013257	129764462	132762875	136360671	140867569	145280602	154096485	161991179	168726788	50724011	66033809	136560938	1549182
HR	4989236	5422632	5722447	6140797	6571698	7475463	8843053	9731294	10623239	1943547	4458400	9415321	81337
HU	8884837	8429843	8441319	9054848	10228352	11660366	13350029	15176493	16700750	3962687	4665369	12393512	122948
IE	23362889	23594089	24603640	26310826	29545020	32595709	34271771	36345005	37947510	8268297	9097359	32405890	318348
IS	2462894	2740691	3199266	3853614	4847288	6801814	8318734	10166386	7584197	1527633	2437139	6463479	60403
IT	116226667	116029388	115279105	121164587	127665221	134477781	144306325	153352444	160667939	40405355	59709143	132425719	1421709
LT	2691991	3166628	3482358	3798110	4227389	4787561	5246101	6254178	6504685	1804500	2464603	5333890	49761
LU	1836780	1893991	2169327	2433966	2651751	2984242	3554730	3988804	4365569	1426310	2002903	4057247	33365
LV	5098360	4754530	4782257	4802282	5145856	5384160	6077854	7037070	7785726	1995133	2336134	5368369	60567
ME						1845464	2173494	2440486	2652801	521959	1309266	1908552	12852
MK					1452373	1649374	1861282	2152746	2353327	709241	1266230	2134988	13579
MT	3506814	3650347	4032029	4290032	4619557	5080446	6007731	6805817	7318357	1752445	2547912	5861597	55473
NL	53895292	55680209	58077271	60963003	64570938	70317995	76240304	79644163	81192507	23594783	29082583	61289771	714548
NO	32403522	34592225	36686364	37603195	37503052	37727546	38739778	40030105	40348437	13216883	14136316	32562643	395550
PL	20635672	21800765	23274484	25714422	28907439	32266861	37684668	43767548	46942771	13825460	18893812	39347542	353061
PT	27579707	28186254	29694146	32560621	36005814	40930044	47673057	51018598	55007894	16548993	22347692	57081723	444634
RO	9687456	9674226	10016933	10907487	12580711	15153719	17934774	19809642	21565865	6633447	10384613	19535951	163884
RS						4414858	4828171	5521250	6450643	1938468	3431750		26585
SE	29732247	30350849	31443225	32766043	34011263	35952558	38456213	38945096	37614763	9317677	10798009	25038812	354426
SI	1358792	1167877	1265766	1307128	1436003	1404152	1682133	1810567	1719039	287787	419346	968811	14827
SK	1808187	1563197	1557149	1671290	1943656	2158261	2402651	2794094	2839787	500604	642078	1942568	21823
UK	201536753	203067015	210468980	220022122	232270437	248868873	264629454	272190155	277432380				2130486
Grand Total	**1940870998**	**1957039604**	**1993921773**	**2090264018**	**2200784362**	**2353855648**	**2538298567**	**2687118655**	**2781871120**	**670004952**	**916286346**	**1993529416**	**24123845**

Figure 2: Tidy dataset



	Year										
Country	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	
AT	25137612	25965977	25749724	26378676	26754007	27181511	28327279	31138417	35644188	9168431	111
BE	25102695	25919515	26389927	28776258	30958841	30115832	33260493	34506309	35385188	9465828	135
BG	6652007	6819103	7079292	7520697	7610949	9324217	11092651	12137714	11713068	3729017	50
CY	7190387	7328300	7011437	7328546	7590787	8961817	10238913	10927101	11261410	2270577	50
CZ	12650532	11742352	11891812	12079873	12672004	13672362	16245554	17838221	18767088	3821372	47
DE	175316076	178591103	180783188	186445814	193936430	200687293	212389343	222422361	226764086	57795978	735
DK	25808321	26532730	27459623	29015133	30095505	32763142	33261214	34701139	34780127	8658654	108
EE	1907569	2202427	1958565	2019806	2160978	2214989	2635145	2995528	3258003	857837	12
EL	33770739	32082336	34023934	39117833	42096402	45543371	50170728	54258826	56088527	17341192	322
ES	165153230	159771261	157731973	165354382	174652503	193872037	209824089	220611429	228262372	57797305	918
FI	16374398	16458815	16565391	17171931	17479246	18099954	20054947	22173530	23287929	5428622	45
FR	126013257	129764462	132762875	136360671	140867569	145280602	154096485	161991179	168726788	50724011	660
HR	4989236	5422632	5722447	6140797	6571698	7475463	8843053	9731294	10623239	1943547	44
HU	8884837	8429843	8441319	9054848	10228352	11660366	13350029	15176493	16700750	3962687	46
IE	23362889	23594089	24603640	26310826	29545020	32595709	34271771	36345005	37947510	8268297	90
IT	116226667	116029388	115279105	121164587	127665221	134477781	144306325	153352444	160667939	40405355	597
LT	2691991	3166628	3482358	3798110	4227389	4787561	5246101	6254178	6504685	1804500	24
LU	1836780	1893991	2169327	2433966	2651751	2984242	3554730	3988804	4365569	1426310	20
LV	5098360	4754530	4782257	4802282	5145856	5384160	6077854	7037070	7785726	1995133	23
MT	3506814	3650347	4032029	4290032	4619557	5080446	6007731	6805817	7318357	1752445	25
NL	53895292	55680209	58077271	60963003	64570938	70317995	76240304	79644163	81192507	23594783	290
PL	20635672	21800765	23274484	25714422	28907439	32266861	37684668	43767548	46942771	13825460	188
PT	27579707	28186254	29694146	32560621	36005814	40930044	47673057	51018598	55007894	16548993	223
RO	9687456	9674226	10016933	10907487	12580711	15153719	17934774	19809642	21565865	6633447	103
SE	29732247	30350849	31443225	32766043	34011263	35952558	38456213	38945096	37614763	9317677	107
SI	1358792	1167877	1265766	1307128	1436003	1404152	1682133	1810567	1719039	287787	4
SK	1808187	1563197	1557149	1671290	1943656	2158261	2402651	2794094	2839787	500604	6
Grand Total	**932371750**	**938543206**	**953249197**	**1001455062**	**1056985889**	**1130346445**	**1225328235**	**1302182567**	**1352735175**	**359325849**	**4997**

Figure 3: Final dataset used after tidying and filtering out certain countries' data

years had the most or least number of passengers, and which countries are generally most or least popular among travellers.

To obtain more specific analyses, I will break down the data and plot the total number of passengers for each country from 2011 to 2022 in a time series plot. This allows me to see trends in air travel for each country in EU-27. To derive further insights, I could compare the total number of passengers between countries for a specific year using a bar graph. Other plots that I am intending to create to help with the visualisation of the whole data story include a heatmap of the total number of passengers from 2011 to 2022 by country. With colour gradients on the heatmap, it will make visualising the data easier and more comprehensible.

With these visualisation plots, I could then research on possible major events that resulted in an increase or decrease in the number of passengers. A major event I have in mind includes the outbreak of the pandemic, where I could use the visualisation plots to analyse how much air travel decreased and picked up again.

2. How do you plan to make it interactive?

To make the story interactive, I intend to make numbers appear over the bar plots for the general visualisations when users hover over each bar representing a specific year or country. Based on research, I can do this by using ggplotly from the plotly package and incorporate it onto the Shiny app by utilising the "text" aesthetic in my ggplot code. To generate the output, I will then convert the ggplot object to a plotly object using ggplotly().

I am also intending to use Shiny widgets to allow users to select which country's data they would like to look at from the sidebar, such that they are able to navigate between the data for different countries and explore countries they are more interested in. While I am unsure of how to achieve this now, I recall doing something similar in Week 8's Code Alone and Challenge using the "10_download" example. Except this time, instead of providing options to choose a dataset to download, the options will be the EU-27 countries' data that users can view. Therefore, I will try to do something similar and rely on online resources to adapt my learning.

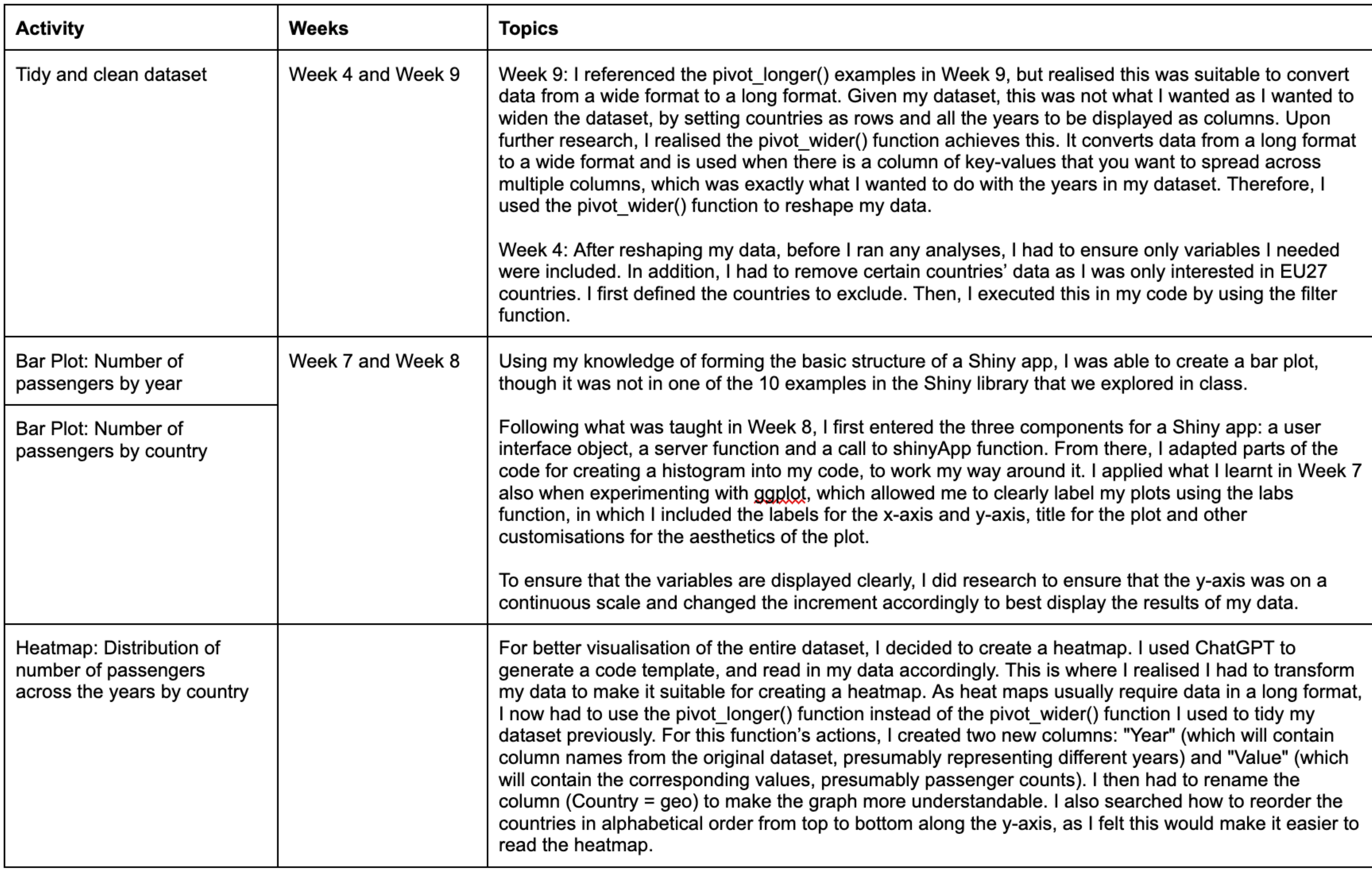
3. What concepts incorporated in your project were taught in the course and which ones were self-learnt?

For the plots I have created so far, these are the concepts incorporated that were taught in the course and self-learnt.

Previously, I had used Excel to clean and tidy my dataset. From there, I created multiple csv files for each plot I wanted to generate to write my code. However, I realised I could have approached this in a more efficient manner, by applying what I have learnt in class in Weeks 4 and 9. Further elaboration will be stated in the table.

Specifically when rendering the heatmap, I encountered the error message "error: [object Object]" and no output graph is generated. I googled and apparently, errors can be due to the Shiny environment or the way it interacts with ggplot2. Therefore, I followed their suggestion to render the heatmap outside of the Shiny app, in an R markdown file, to see if the heatmap is being generated correctly. This will help me isolate if the issue is with the Shiny environment or within my code and data itself. Given the error message is quite generic, I also referred to the console for more detailed error messages, which told me the problem was while computing aesthetics as object 'Country' was not found. From there, I realised that I had missed out the column for 'Country' when writing my code for data_melted, the dataset I am using for my heatmap, which was why the error had occurred. I refined this part of my code accordingly by ensuring the column 'Country' was defined properly, and managed to resolve the issue.

Generally for the plots I've generated thus far, I am intending to also work on inserting a legend for the country code to make the plots clearer as not everyone may know which country code stands for which country.



Activity	Weeks	Topics
Tidy and clean dataset	Week 4 and Week 9	Week 9: I referenced the pivot_longer() examples in Week 9, but realised this was suitable to convert data from a wide format to a long format. Given my dataset, this was not what I wanted as I wanted to widen the dataset, by setting countries as rows and all the years to be displayed as columns. Upon further research, I realised the pivot_wider() function achieves this. It converts data from a long format to a wide format and is used when there is a column of key-values that you want to spread across multiple columns, which was exactly what I wanted to do with the years in my dataset. Therefore, I used the pivot_wider() function to reshape my data. Week 4: After reshaping my data, before I ran any analyses, I had to ensure only variables I needed were included. In addition, I had to remove certain countries' data as I was only interested in EU27 countries. I first defined the countries to exclude. Then, I executed this in my code by using the filter function.
Bar Plot: Number of passengers by year	Week 7 and Week 8	Using my knowledge of forming the basic structure of a Shiny app, I was able to create a bar plot, though it was not in one of the 10 examples in the Shiny library that we explored in class. Following what was taught in Week 8, I first entered the three components for a Shiny app: a user interface object, a server function and a call to shinyApp function. From there, I adapted parts of the code for creating a histogram into my code, to work my way around it. I applied what I learnt in Week 7 also when experimenting with ggplot, which allowed me to clearly label my plots using the labs function, in which I included the labels for the x-axis and y-axis, title for the plot and other customisations for the aesthetics of the plot. To ensure that the variables are displayed clearly, I did research to ensure that the y-axis was on a continuous scale and changed the increment accordingly to best display the results of my data.
Bar Plot: Number of passengers by country		
Heatmap: Distribution of number of passengers across the years by country		For better visualisation of the entire dataset, I decided to create a heatmap. I used ChatGPT to generate a code template, and read in my data accordingly. This is where I realised I had to transform my data to make it suitable for creating a heatmap. As heat maps usually require data in a long format, I now had to use the pivot_longer() function instead of the pivot_wider() function I used to tidy my dataset previously. For this function's actions, I created two new columns: "Year" (which will contain column names from the original dataset, presumably representing different years) and "Value" (which will contain the corresponding values, presumably passenger counts). I then had to rename the column (Country = geo) to make the graph more understandable. I also searched how to reorder the countries in alphabetical order from top to bottom along the y-axis, as I felt this would make it easier to read the heatmap.

Figure 4: Final dataset used after tidying and filtering out certain countries' data

4. Include the challenges and errors that you faced and how you overcame them.

My main difficulty this week was creating the plots for specific countries, and presenting them in a way that users can select the countries from the sidebar and then explore the plots I have generated for each country. I had first intended to adapt the “10_Download” example on Shiny, but soon realised I did not know how to modify the code. To overcome this issue, I first double checked on Google to make sure that I did not need to create individual plots for every country one by one, which made sense as this would be too time-consuming. Online resources directed me to make use of Shiny’s reactive programming framework to generate plots based on user input, which I will attempt within this week.

Week 12 Diary

1. Include the challenges and errors that you faced and how you overcame them.

Working on where I left off, I tried to create bar plots for specific countries based on user’s input. I tried using Shiny’s reactive programmin framework to generate plots based on user input, but encountered error messages such as “Error: Object Object”. Based on chat_gpt’s response, a more specific error message could have been that line graphs require a data frame where each observation of the number of passengers (value) corresponds to a specific year for the selected country in long format. As the final dataset I used was reshaped into wide format for previous visualisations, I had to convert it back using long format.

This week, I was also working on creating stacked bar plots to complement the bar plot generated for number of passengers by year from 2011-2022. This allows users to see the unique contribution of each country for each year. I successfully generated a stacked bar plot after researching online, but I wanted to modify my plot such that the countries are ranked according to contribution. This would make the graph more readable and will help users easily identify who is the most versus the least contributor. To achieve this, I calculated a rank for each country within each year first and then arranged the data by this rank before plotting the graph.

I also worked on generating growth rate plots and predicted growth rate plots for each country. I used functions to create calculations for growth rates and predicted growth rates. Using functions found online, I attempted to write my own function. However, since functions usually rely on data from the previous year to be calculated, the first year could read in a NA value as there is no prior data to compare to. This NA value affected the rendering of the plots. Hence, I added in a line of code (filter(!is.na(GrowthRate))) to remove rows with ‘NA’ for growth rate and predicted growth rate.

Furthermore, I realised I had to transform my data from wide to long format again before I can plot growth rates based on year. The error message I initially got was “Year not found”, but there is a column called “Year” in the dataset I was calling. This could have been caused by the current format of the dataset, where the final dataset I used has year columns such as “2011,2012” etcetera instead of a single “Year” column. After reshaping my data, I managed to generate my intended plot.

Generally, I realised most of the time I encountered errors relating to calling the correct column names. This is because I make a lot of changes to the original dataset. From the beginning, I already cleaned, filtered and tidied data a few times, renaming certain columns in the process. Therefore, I have to ensure I am calling columns that exist in the dataset I am using currently. To resolve this issue, I check through each line of code and ensure the variables exist and are keyed in correctly (uppercase, lowercase) to ensure consistency. Most of the time, I was able to troubleshoot by checking each line of code.

Final Submission

What is the theme of your data story?

As a travel enthusiast, the onset of the pandemic and ensuing travel restrictions sparked my curiosity about their impact on air travel. This inspired me to explore the evolution of air transportation trends, with a focus on passenger numbers. I focused on the European region, specifically EU27 countries, as Europe is a popular travelling destination. My aim was to uncover the trends in passenger numbers, especially the effects of the pandemic on the industry. Thus, my final research question is: How have air transportation trends in EU27 countries evolved over the years?

Why is it important to address this question?

Understanding the evolution of air transportation trends is crucial given the aviation industry's substantial contribution to global economic activity. According to the International Air Transport Association (IATA) (2007), air travel is one of the most important modes of transportation, with the aviation industry contributing significantly to global GDP by facilitating global trade, business, tourism and more.The COVID-19 pandemic's onset brought unprecedented financial losses, with IATA (2022) reporting a staggering $118 billion deficit in 2020. As the industry recovers, detailed insights into air travel patterns are invaluable in better informing resilient strategies to foster economic recovery. Europe was selected as the focus region as according to the United Nations World Tourism Organisation (UNWTO), Europe is the world's top tourist destination. With borderless movement as its foundational principle, understanding data on the number of air passengers carried informs strategic policy development and infrastructure expansion to manage the flow of travellers effectively and sustaineably.

Why do you think the data sources that you have curated can help you answer the question?

I curated my data source from Kaggle, with the final dataset encompassing comprehensive records of air passenger traffic from 2011 to 2022.

Columns that are useful are geo (for the country's name), TIME_PERIOD (to represent the corresponding year) and OBS_VALUE (to represent the total number of passengers). This data is instrumental in charting the trends and drawing comparisons before, during, and predicting trends after the pandemic, offering a robust foundation for analysing the aviation industry's growth over the years.

What are the insights from the data and how are they depicted in plots?

To depict insights, a heatmap was generated to obtain a big picture of what the data represents. With deeper hues signifying higher passenger counts, Germany and Spain, followed by France and Italy, consistently outperform the rest of the EU27 countries in air travel density. The years 2020 and 2021 are markedly distinguished by a pronounced dip in color intensity, corresponding to a significant contraction in air travel, serving as a visual quantifier of the pandemic's impact.

A bar plot depicted the total number of air passengers carried by each country from 2011 to 2022. Countries with higher peaks correspond to a higher number of air passengers carried and vice versa. Countries with higher peaks include France, Germany, Italy, the Netherlands and Spain whereas countries with lower peaks include Estonia, Slovakia and Slovenia.

Another bar plot was used to compare year and total number of passengers (irrespective of country). This showed a steady increase in the total number of passengers carried from 2011 to 2019, with 2019 having the most number of passengers carried before a sharp drop in 2020. To enhance the interpretation of the data, a year-on-year growth rate line plot was generated to convert raw passenger numbers to percentage growth rates. This is crucial for grasping the actual scale of change in air passenger traffic over time. With the outbreak of COVID-19, the growth rate from 2019 to 2020 fell by a sharp -73.44%. With the gradual lifting of lockdowns and travel restrictions, the growth rate from 2021 to 2022 increased to 117.73%.

To complement my understanding of the bar plot on the number of passengers by year, a stacked bar plot using colour-coded schemes to differentiate each country was created for a stratified depiction of annual contributions, facilitating an immediate comprehension of each country's share.

Finally, to predict future trends, a line plot on predicted growth rates was generated. This forecasts future passenger growth based on historical data, providing a perspective on expected trends.

For individual country analyses, the same types of plots offered tailored views of each nation's passenger data, growth rates, and predicted future trends, changing dynamically based on user-selected country input.

How did you implement this entire project? Were there any new concepts that you learnt to implement some aspects of it?

The implementation of this project followed a structured approach that included several key phases:

Firstly, research and planning. This included setting the research question and gathering data on passenger numbers from reputable sources.

Secondly, data preparation to ensure accuracy and usability. The dataset came in an untidy form, with data for each country for each year displayed as separate rows. It also contained columns not needed in my analysis, so I had to filer those columns out. I also had to create a new column for country names, as countries were listed in their country codes, which are less intuitive.

Then, I started planning the types of visualisation plots to use. I determined the usefulness of each plot to see which would best present my data in a comprehensible form. Thereafter, I created each visualisation plot. This was where I decided to include elements of interactivity, such as having values shown while hovering over the data point. This makes it easier to read the data from my graph, especially since the values tend to span a higher range. To enhance user engagement, another interactive element was allowing users to select different countries and view corresponding data and projections.

Finally, I synthesised all findings into a coherent narrative, telling the story of how the EU27's air travel trends have evolved and potentially will continue to evolve.

Some new concepts I learnt included the interactive elements mentioned above, such as learning about reactive programming to allow users to view selected countries' plots based on user input. I learnt how to generate predicted growth rates using predictive modelling, such as linear regression models, within a function. I also had to weigh the pros and cons of different predictive models to see which will provide the best representation. With primary visualisation plots generated, I then worked on improving the aesthetics and readability of my graphs. This involved new learning to make my plots look cleaner by utilising theme_minimal(), using the scales library's "scales::comma" function to format the axis labels in a way that large numbers are more readable and adjusting the scales of the x-axis and y-axis using log functions to ensure data spanning a wide value range can be displayed clearly.

References

International Air Transport Association. (2007). Aviation economic benefits. Retrieved November 15, 2023, from https://www.iata.org/en/iata-repository/publications/economic-reports/aviation-economic-benefits/

International Air Transport Association. (2022). Understanding the pandemic's impact on the aviation value chain. Retrieved November 15, 2023, from https://www.iata.org/en/iata-repository/publications/economic-reports/understanding-the-pandemics-impact-on-the-aviation-value-chain/

Preda, G. (2021). Passengers air transport in Europe [Data set]. Kaggle. Retrieved October 16, 2023, from https://www.kaggle.com/datasets/gpreda/passengers-air-transport-in-europe

United Nations World Tourism Organization. (n.d.). Impact assessment of the COVID-19 outbreak on international tourism. Retrieved November 15, 2023, from https://www.unwto.org/impact-assessment-of-the-covid-19-outbreak-on-international-tourism