**Code First Girls Nanodegree - Data 1 - Group Project**

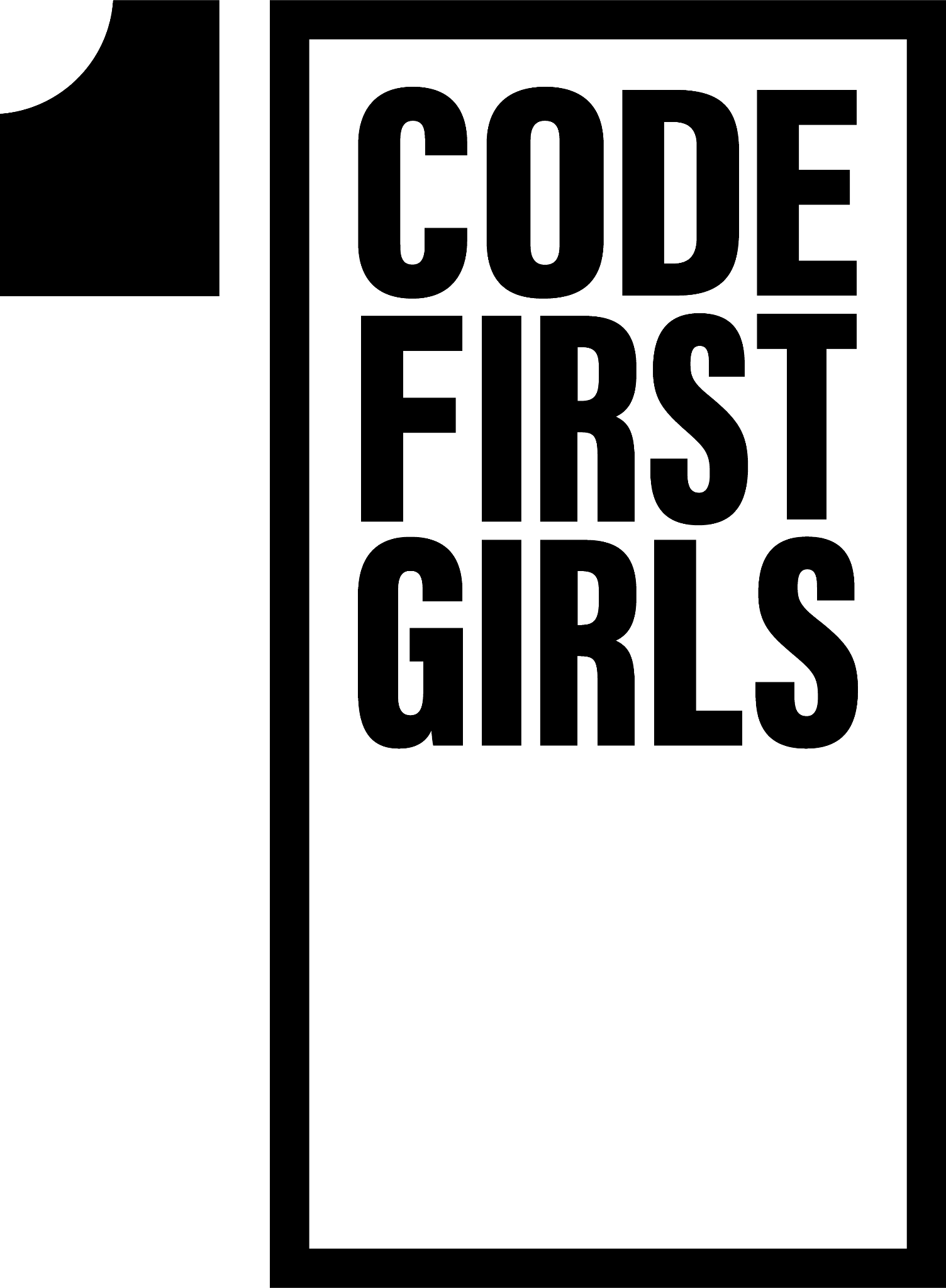
**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**The Environmental & Economical Impact**

**of USA’s Airlines**

**Data Analysis Project**

**Final Report**



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**By Saweena Dona, Tsamaija Forsythe-Gidharry, Geneva Kirwan, Yolanta Kwiatkowski,**

**Ashruti Rajesh and Zsoka Rati**

Submission: 3 January 2022

**Contents**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**INTRODUCTION Page 3**

Project Aims & Objectives

**BACKGROUND**

Data Analysis into USA Airlines **Page 3**

**STEPS SPECIFICATIONS Pages 3-5**

Detailed Data Analysis

Data Sources Used

**IMPLEMENTATION AND EXECUTION Page 5**

Development Approach

Tools and Libraries

Implementation Process and Challenges

**RESULT REPORTING Pages 5-7**

Key Findings

**CONCLUSION Page 7**

Final Thoughts

**APPENDIX Pages 7-11**

**INTRODUCTION**

**Project Aims & Objectives**

In today's world it is undeniable that climate change and concerns over the welfare of our environment are at the forefront of socio economic issues. The aviation industry, specifically air travel, is one component of the economical landscape that contributes to pollution levels, namely through the release of carbon dioxide. Every year, the United States of America (USA) has the highest number of airline passengers in comparison to any other country in the world. In 2019, their passengers accounted for one-fifth of all airline passengers worldwide, thus finding methods for airlines to reduce their emissions can reduce levels of pollution.

As a result, this project explores the original idea of comparing the environmental and economic impact of the USA’s main airlines, including the ten airlines which have almost a 90% share of the domestic flight market in the USA: American, Southwest, Delta, United, Spirit, Alaska, JetBlue, SkyWest, Frontier and Hawaiian Airlines. We explore different factors which may affect the levels of CO2 produced by each airline, such as fleet size, ticket prices and revenue, altogether not only investigating the way in which airlines can affect the environment but also its impact on aspects of the aviation economy.

**BACKGROUND**

**Data Analysis into USA Airlines**

As the USA has the largest number of airline passengers worldwide, it is not improbable to assume that air travel may also have a large impact on the environment and the economy. There is a broad spectrum of factors related to air travel that can contribute to environmental and economic changes, therefore we have focused on the following questions to lead our analysis:

1. How is a passenger’s carbon footprint impacted by the airline used?
2. What is the link between the total emissions of each airline via fuel combustion and the fleet profile/breakdown of each airline?
3. Is there a correlation between average ticket prices and emissions per airline?
4. Is there a relationship between company revenue, profit and emissions?

The purpose of our analysis has been to investigate how different factors alter the amount in which competitive airlines in the USA produce emissions and thus its impact on the environment, as well as finding out the factors which promote and predict low emissions and a reduced carbon footprint (with data between 2010-2019). We believe that this data can be informative in contemporary society where climate change and the health of the environment is a significant matter. Air travel will exist for years to come, therefore investigating ways in which airlines can be efficient and sustainable - fulfilling its purpose of transporting passengers across long distances while also making an effort to preserve the environment - can be one way in which we can contribute to a healthier world.

**STEPS SPECIFICATION**

**Detailed Data Analysis**

Our question was framed out of the collective observation that many factors go into the economic success and environmental impact of airlines. We held multiple meetings in which we brainstormed not only what factors appeared most important in understanding these impacts but also what potential trends interested us most hence ‘emissions’, ‘fleet’, ‘revenue’ and ‘ticket prices’ were selected as our avenues of analysis. It should also be noted that the availability of data also limited our choices of how our question would be framed as data quality was one of our priorities.

Within our team we were aware that having a high data quality was essential to outputting useful and valid trends, as if you begin with a poor data input you are guaranteed a poor data output. As a result, we chose a reputable source - as close to the survey data as possible - for our transport statistics. More detail on the specific sources can be found in the data sources section.

Since each of the datasets were slightly different, we had to adapt our pre-processing approach to suit each source. In most cases, we were presented with large quantities of redundant data that had to be removed so that we were left with only the data that pertained to the scope of our study. In some cases, data was provided in different time frames, for example the revenue data was provided quarterly, so to compare it with the monthly emissions data, an average was calculated using the data for each of the 3 months in a quarter. Where data was missing e.g. the 2010 Spirit Airlines revenue, it was dealt with appropriately, either by using averages, or dropping the non-existent data altogether. Where a data processing or transformation step was taken in the project, you will find a statement justifying the decision in the Project Code.

Further details of steps taken in data analysis can be seen below:

* **Emissions and Passengers:** Pandas data frames were used to convert data for fuel consumption and passengers carried into csv files which were merged to obtain a yearly carbon footprint per passenger calculation.
* **Fleet:** Inventory data was used for each airline which included every individual airline owned. The data was imported into a pandas data frame and only the useful data was retained. The data was further reformatted to allow the qualitative data to be represented in nested pie charts and histograms.
* **Revenues:** The data was collected using the Macrotrends website, which contained data about revenues for the different airlines studied except for Frontier Airlines and Spirit Airlines for which revenue data for 2010-2019 could not be found. Since the data for the airlines was available for each quarter of each year between 2010 and 2019, the CO2 emissions and the revenues were calculated to get an average for three months (eg. Quarter 1 = January, Feb, March).
* **Ticket prices:** Using the Sodapy library, API data was obtained from the website of the US Department of Transportation, and was cleaned and standardised, using Pandas methods. This was merged with a revenue versus emission csv file that was produced for another part of the study. The merged data frame was analysed and the results were visualised using Matplotlib and Seaborn.

**Data Sources Used**

A wealth of historical aviation data can be found online, provided by The United States Department of Transportation. Their Bureau of Transportation Statistics (BTS) collects data from airlines via survey forms and stores the responses in online data tables. This data has formed the basis for our data analysis project. The following sources are not an exhaustive list, but are significant sources we have discovered which have aided the development of our project:

* <https://www.transtats.bts.gov/> - the graph presented on this page has been used to pick our top 10 airlines to study. It shows the ten airlines that had the greatest share of the domestic flights market this year.
* <https://www.transtats.bts.gov/fuel.asp?pn=1> - this provides us with the fuel in gallons consumed each year by each airline e.g. Airline Fuel Cost and Consumption (American Airlines - Scheduled, January 1990 - June 2021).
* <https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=GDM&QO_fu146_anzr=Nv4%20Pn44vr4> - provides a wide range of downloadable csv data tables with customisable columns including information on Carriers (Airlines), Aircraft, Time periods, Flight origins/destinations and other useful summary fields such as flight time or distance between airports. Also provides the fleet details of American Airlines with information such as regarding the aircraft, the active or parked status of current fleet, details on stored or scrapped fleet, history and age.
* <https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-1-Top-1-000-Contiguo/4f3n-jbg2> - consumer airfare report of top 1,000 contiguous state city-pair markets. Information on contiguous state city-pair markets that average at least 10 passengers per day [here](https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-6-Contiguous-State-C/yj5y-b2ir) and domestic airline consumer airfare report [here](https://www.transportation.gov/policy/aviation-policy/domestic-airline-consumer-airfare-report).
* <https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-6-Contiguous-State-C/yj5y-b2ir> US domestic air fare data was loaded from the API on this site.
* <https://www.transtats.bts.gov/AverageFare/> - average domestic airline itinerary fares; data that is searchable by year and quarter but does not list by airlines.
  + Original source of data:

<https://www.bts.gov/topics/airlines-and-airports/origin-and-destination-survey-data> - Origin and Destination Survey (DB1B) is a 10% sample of airline tickets from reporting carriers. Data includes origin, destination and other itinerary details of passengers transported.

* <https://www.bts.gov/topics/airlines-and-airports/quick-links-popular-air-carrier-statistics> - different types of American aviation related statistics.
* <https://www.easa.europa.eu/domains/environment/icao-aircraft-engine-emissions-databank> - information on emissions produced per engine.
* <https://www.eesi.org/papers/view/fact-sheet-the-growth-in-greenhouse-gas-emissions-from-commercial-aviation> - fact sheet on the growth in greenhouse gas emissions from commercial aviation in 2019. Includes information such as airline energy intensity and emissions, aviation emissions, regulating aircraft emissions, carbon emissions from international aviation, historically resilient growth and projections and global economic growth in the aviation industry.
* <https://www.macrotrends.net/> - quarterly data for revenues 2010-2019 for the different airlines.

**IMPLEMENTATION AND EXECUTION**

**Development Approach**

Our team approach focused on recognising each other’s strengths and weaknesses by performing a SWOT analysis and delegating tasks in accordance to each other’s level of comfort and expertise on a specific task. Task delegation was carried out within a weekly scrum meeting in which we checked in on our individual progress, hurdles and plans for the following week. We also used our scrum meetings to set milestones and deadlines for the rest of the project.

To allow for online collaboration, our meetings were held through voice chat on Discord alongside daily messaging on Discord to ensure that all team members are on track to achieve the goals assigned during the weekly meetings. Alongside utilising the voice chat and messaging features on Discord, we used Slack as a formal reminder platform for meetings, as well as Google Colab for rough draft code and GitHub as we progressed our work on the project.

An Agile approach was taken not only in how we held our scrum meetings, but also in how we manipulated the code. Code reviews were carried out before separate sections of code were collated into a single coded document. Where necessary, before code was collated into the final document it was refactored to ensure it only included the information required to yield useful analysis within the project.

**Tools and Libraries**

Libraries have been used to enable data to be transformed and visualised. We used the python tools and libraries listed below:

* **Pandas:** this library was used to represent and manipulate data within a pandas dataframe. Use of a data frame allowed for columns to be added, removed and calculated with ease which was particularly valuable when doing data cleansing.
* **Matplotlib and Seaborn:** These libraries were used for data visualisation allowing line graphs, scatterplots, histograms and pie charts to be produced.
* **Sodapy:**  A python library for the Socrata Open Data API. Its Socrata package was used to access API data.

**Implementation Process and Challenges**

Since everyone in the group had competing work schedules that ran alongside the nanodegree, it was sometimes difficult to find the time to meet online, which sometimes slowed decision making. The use of discord to message proved to be a very effective method of keeping everyone in the group informed and on track as to what was going on. The complicated nature of GitHub meant that it was not used to its full extent as a collaboration tool for this project, however we found Google Colab to be a much more useful way of sharing code, as well as sending Jupyter notebook files

**RESULT REPORTING**

**Key Findings**

***CO2 emissions for the studied airlines***

Appendix 3.2 shows the yearly CO2 emissions from fuel combustion for each studied airline between 2010-2019. From this graph, it is evident that each airline generally produces different levels of emissions with emissions gradually increasing throughout the years. For example, Delta Air Lines increases steadily from (approximately) 30million kg of CO2 per year to 35 million kg and JetBlue Airways follows a similar upwards trend however its figures go from (approximately) 5 million to 9 million kg. Looking at the yearly passenger data in Appendix 1, it’s clear that the number of passengers increasing yearly may be contributing to the increase in CO2 emissions, although whilst there is a general upwards trend, there may be other factors influencing these differences in total CO2 emissions produced per airline, as the carbon footprint per passenger for each airline doesn’t follow the same upwards linear trend..

The second graph in Appendix 1 shows the yearly carbon footprint per person for each airline studied between 2010-2019. Unlike the yearly emissions graph, every airline shows a different trend; for the passengers of Alaska Airlines, the CO2 emissions produced per passenger has generally decreased from (approximately) 350 to 290 kg per person per flight, whereas for United Airlines it has stayed fairly constant throughout the years at around 120kg per person. Moreover, each airline has quite a different per passenger carbon footprint to what you’d expect looking at just the emissions or passenger data, leading to the idea that there are many factors which make up the profile of a lower emissions airline.

***Correlation between revenues and CO2 emissions for the studied airline***

Analysis was conducted to find correlations between revenue and emissions for years in the time frame from 2010 -2019 for the selected US Airlines. It was seen that there is a general increase in revenue for airlines each year (See Appendix figure 3.1). This shows that the number of passengers using US airlines generally increased each year, which aligns with the yearly total passenger data collected in the first section of the project. Also, it is seen that when revenue increased there was also an increase in emissions. This is an expected outcome because emissions and revenue are both directly related to passengers using the airline and frequency of operation for the airlines. (See Appendix 3.2).

Analysis was also conducted for an individual airline to see trends in revenue across the time period. The results showed a gradual increase in revenue for Alaska Airlines(basis for analysis for individual airlines)(See Appendix 3.3).

Finally, a correlation matrix was computed (see Appendix 3.4) to explore whether there is a correlation between airlines’ revenues and the CO2 emissions they produce. A strong positive correlation (Pearson’s correlation coefficient = 0.98) was found suggesting that the higher the airline’s revenue, the higher the CO2 emissions generated by the airline.

***Correlation between fleet and CO2 emissions for the studied airlines***

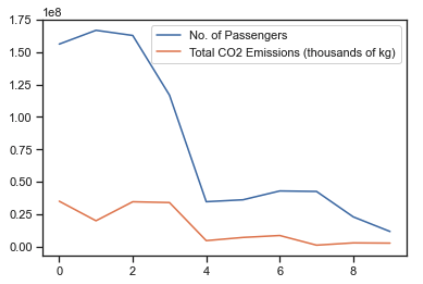
Analysis of the fleet data in conjunction with the emissions data showed trends in fleet characteristics and emissions. One pervasive trend was that the larger the fleet the greater the emissions that were produced in 2019 (see Appendix 2.1). This trend was unsurprising as it is to be expected that bigger airlines would have larger fleets, carry out more journeys and thus produce more CO2 emissions.

Additional more nuanced trends in fleet characteristics and emissions were also seen including:

* Airlines with older fleets tended to have higher emissions
* Simple fleets with fewer models appeared to produce lower emissions
* Boeing manufactured aircraft were potentially leading to higher emissions, but this required further investigation on individual aircraft to substantiate.

Another notable result was the outliers to some of the trends seen. In particular, SkyWest airline had a relatively large fleet (over 500 aircrafts) but it had the lowest emissions (see Appendix 2.2). The fleet model portfolio differed greatly from the other airlines as it featured aircrafts from manufacturers Embraer and Bombardier. Upon further review it was seen that SkyWest is a regional airline which mainly does short haul flight. This suggests that despite trends seen between flight characteristics and emissions, the differentia may not be based on what fleet an airline has, but how they use it.

***Correlation between total passengers carried and CO2 emissions for the studied airlines***



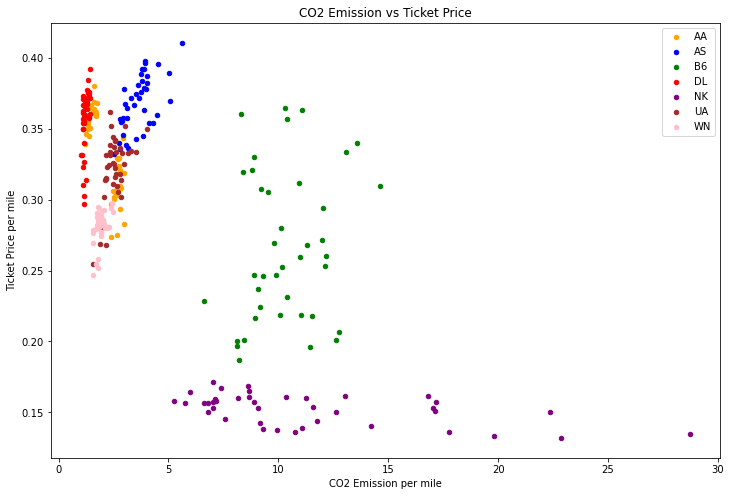
-for 2019, the higher the number of passengers, the higher the total CO2 emissions produced

-large gap between

***Correlation between ticket price and CO2 emissions for the studied airlines***

Using a correlation matrix, a weak negative correlation of -0.65 was found between average ticket price and CO2 emission data, on a per mile basis. The result implies that ticket price increases with reduced CO2 emission, and vice versa. This was further investigated with the help of a scatter plot, displaying CO2 emission versus ticket price for individual airline data points. The resulting plot shows that although the data points are largely on a negative slope, the values for individual companies do not tend to follow this pattern, their data points gather around the same area. The negative correlation is possibly due to different strategies of different companies.

As out of the 18 airlines in the API, only 7 had the necessary information to be investigated for these attributes, there is a possibility that these results may change if more data becomes available.



**CONCLUSION**

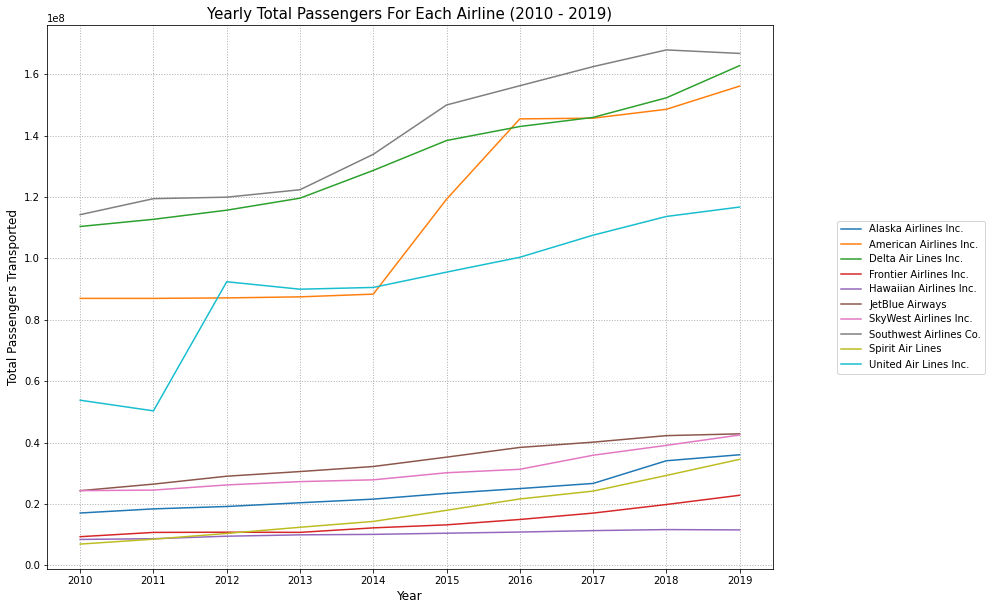
By carrying out data transformation, cleansing and analysis we have been able to reach the following conclusions:

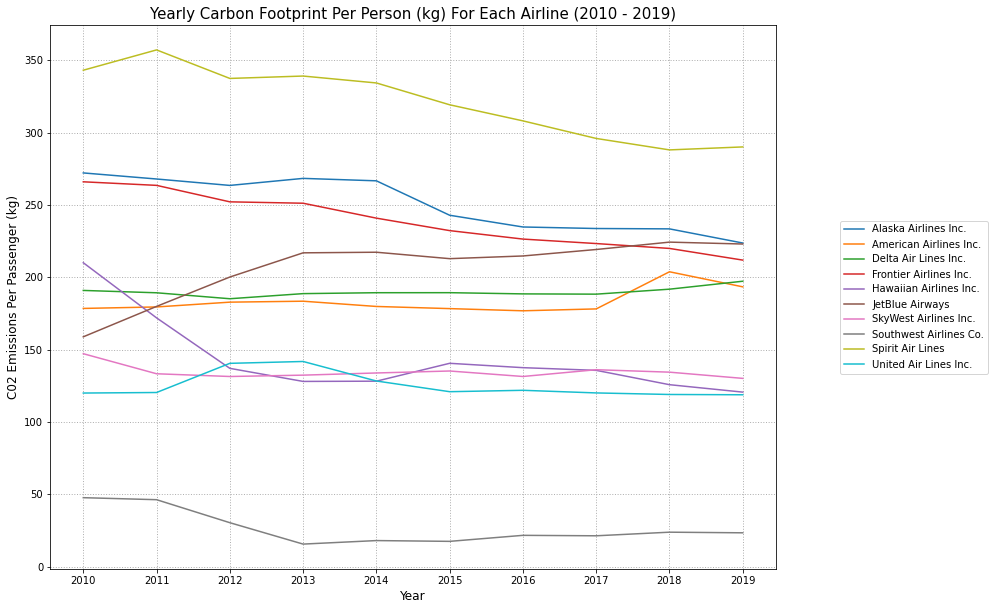
* The CO2 produced by each airline has gradually been increasing every year, signaling towards the popularity of air travel, with the carbon footprint per person showing different trends for each airline, thus denoting the presence of other factors which influence an airline’s ability to produce emissions and proven by our subsequent analysis.
* Fleet age, model distribution, and size all appear to play a role in airline annual emissions, however it is likely a greater role is played by how the fleet is utilised.
* Revenue for airlines showed a gradual increase from 2010 to 2019.There is a correlation between revenue of airlines and emissions per year. The higher the emissions the higher the revenue for that airline. A correlation of 0.98.
* Some of the higher emitting airlines actually have a lower carbon footprint per passenger. This may be due to the fact that higher emitting airlines have larger fleets, and the fleets tend to include newer planes that emit less CO2.
* Ticket price per mile showed slight increase for most airlines, however this trend is not universal, as in some cases prices have decreased between 2010 and 2019.
* There is a slight negative correlation between ticket prices and CO2 emission, but this is possibly due to different business strategies of individual airlines.

As we have demonstrated that there are correlations between factors such as fleet age and size, and number of passengers transported, we hope that In the future, our research in this project could be used to inform predictive models that would allow influential bodies to make more informed decisions about the future of air travel, whether that is in the USA or other countries in the world. Altogether, we hope our analysis contributes in decision making which can aid in the effort to decrease CO2 emissions and the overall preservation of our environment.

**APPENDIX**

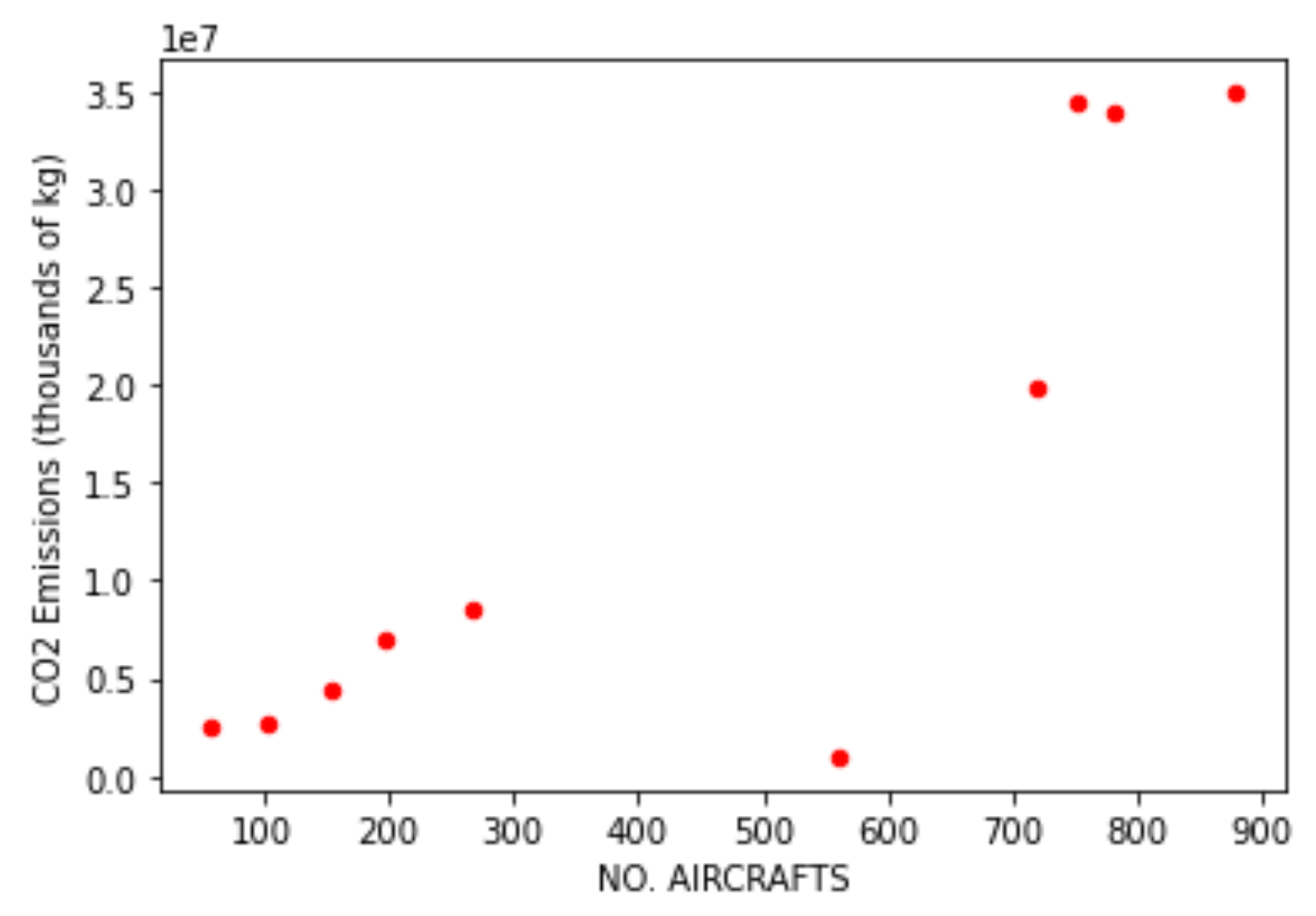
**Appendix 1: Emission Figures**

****

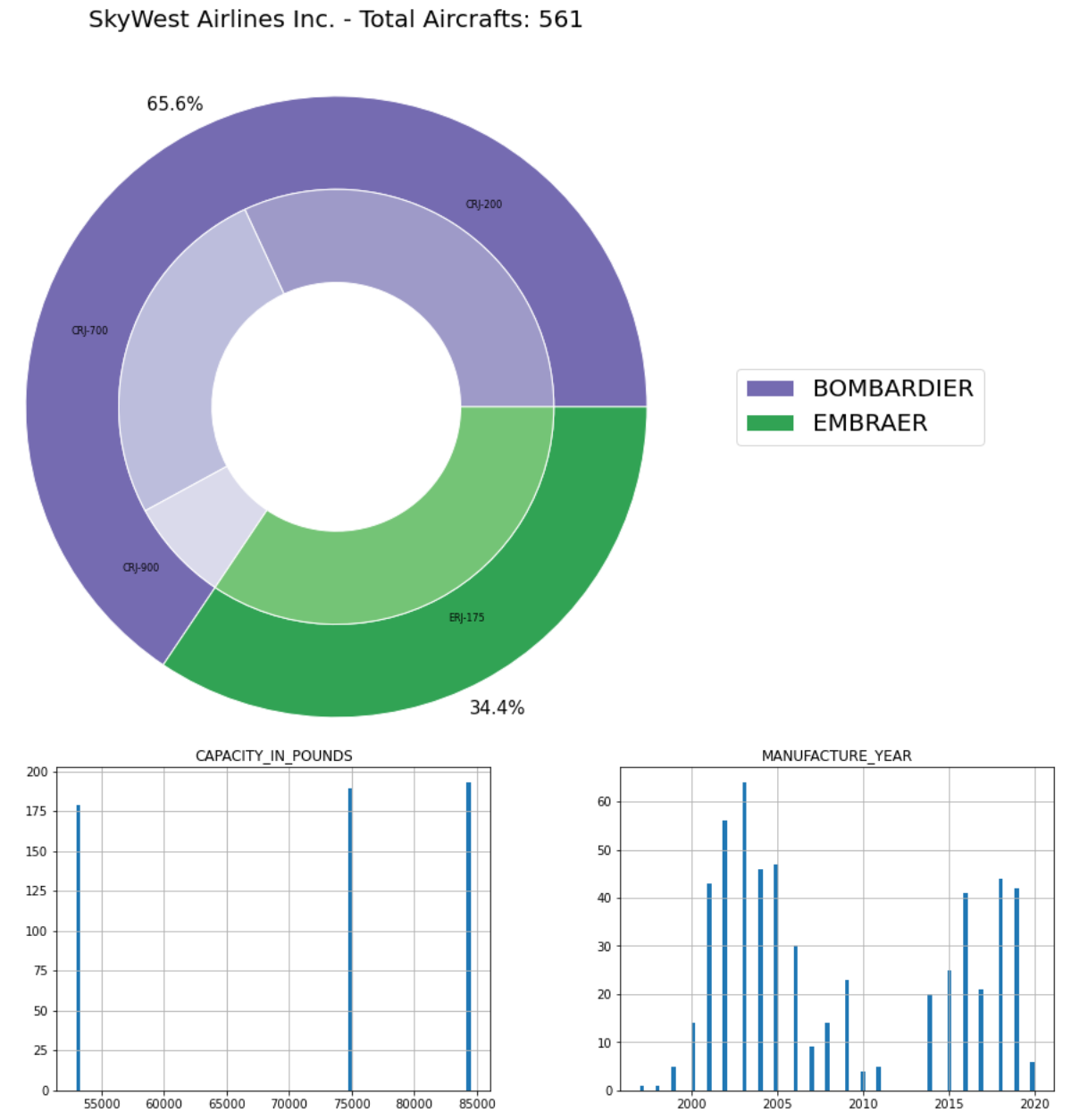
****

**Appendix 2: Fleet Figures**

**Appendix 2.1: Scatterplot of 2019 Emissions Compared to Fleet Size**

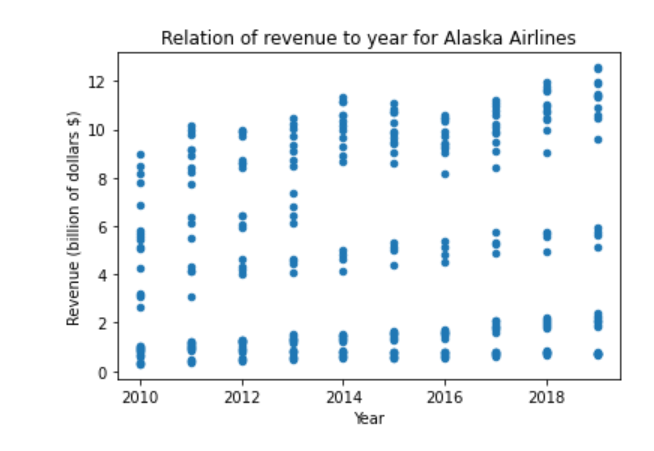
****

**Appendix 2.2: SkyWest Dashboard**

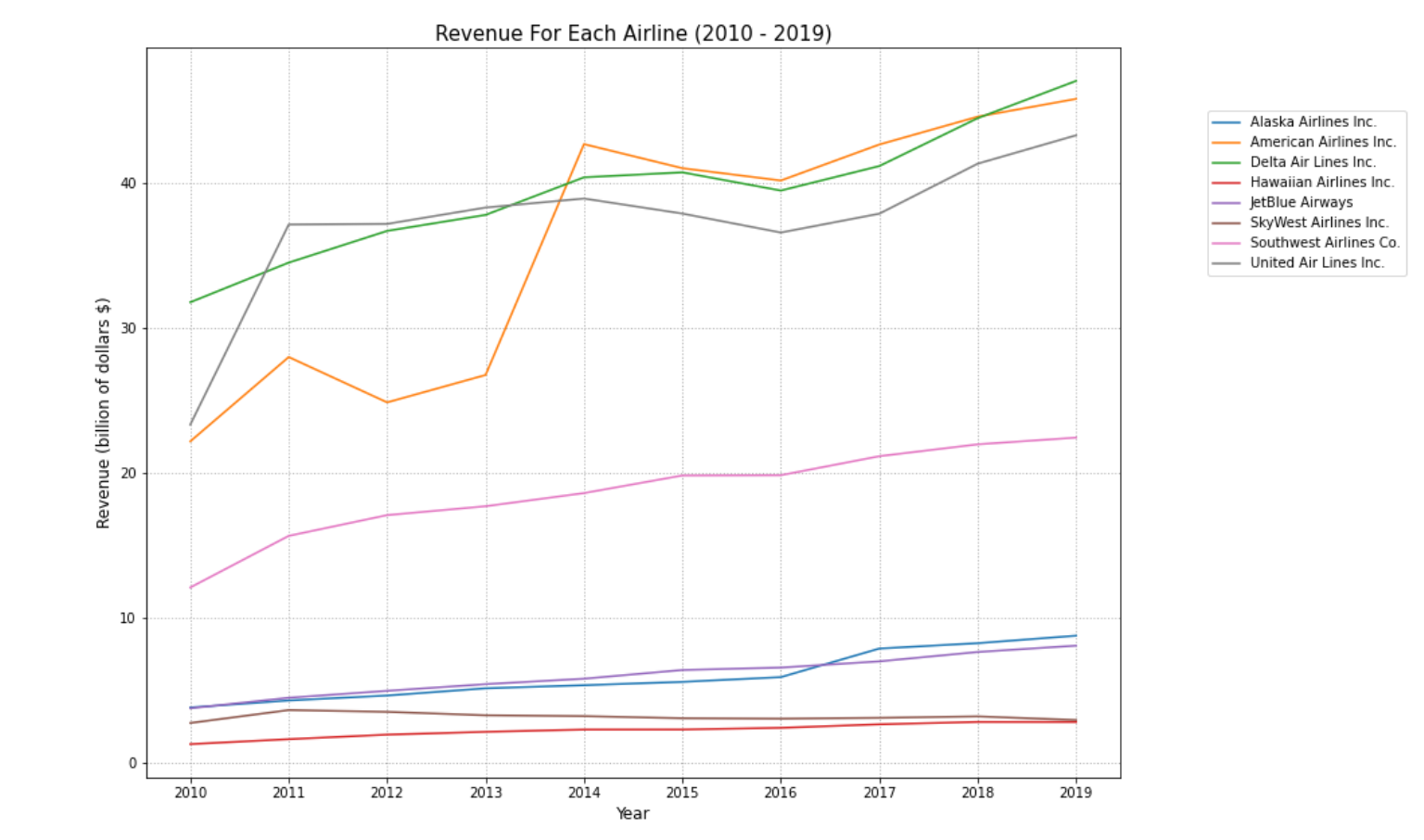
****

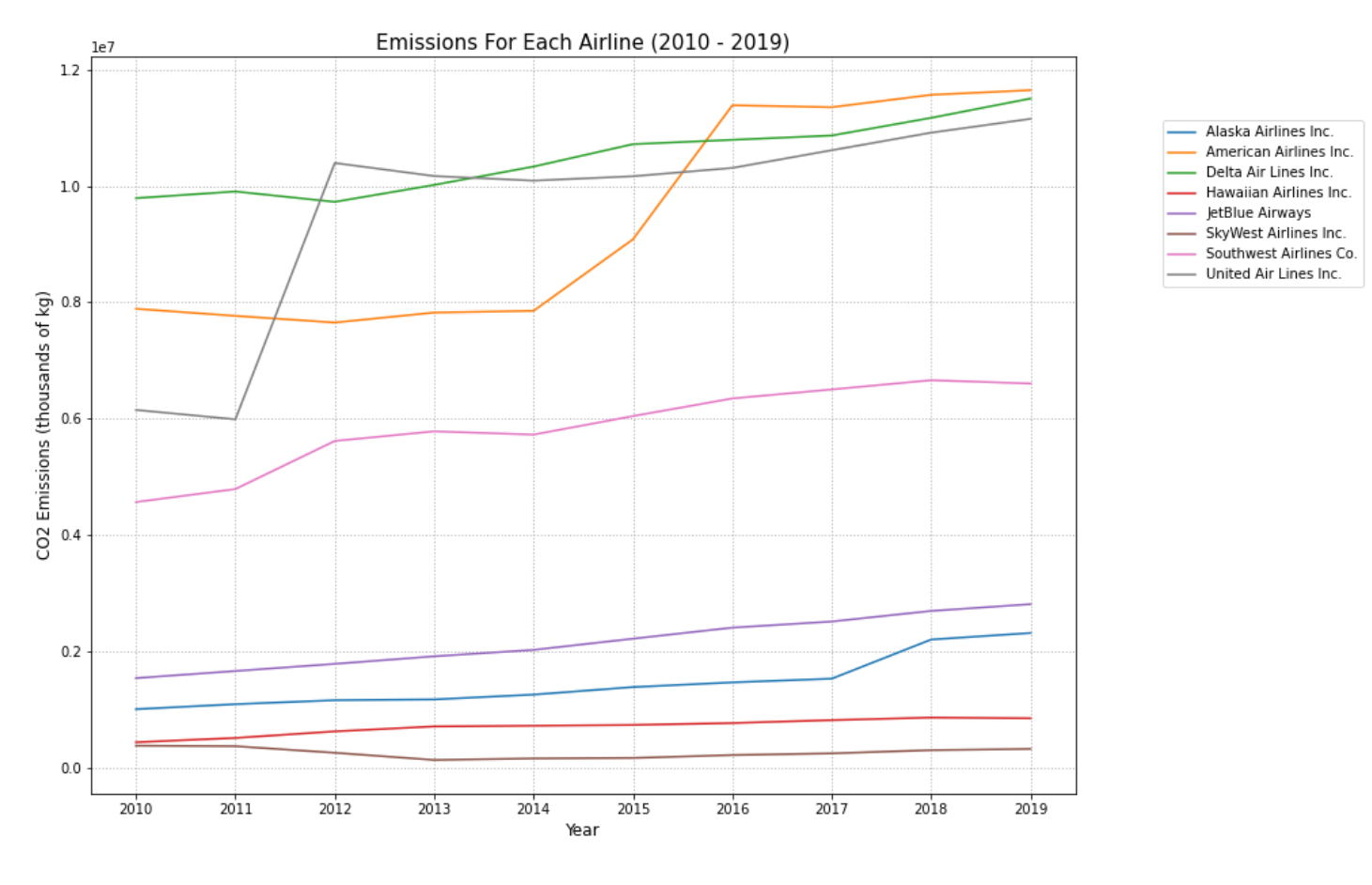
**Appendix 3:Revenue & Emissions figures**

**Appendix 3.1: Scatter plot to show relation between revenue for each year for US Airlines**

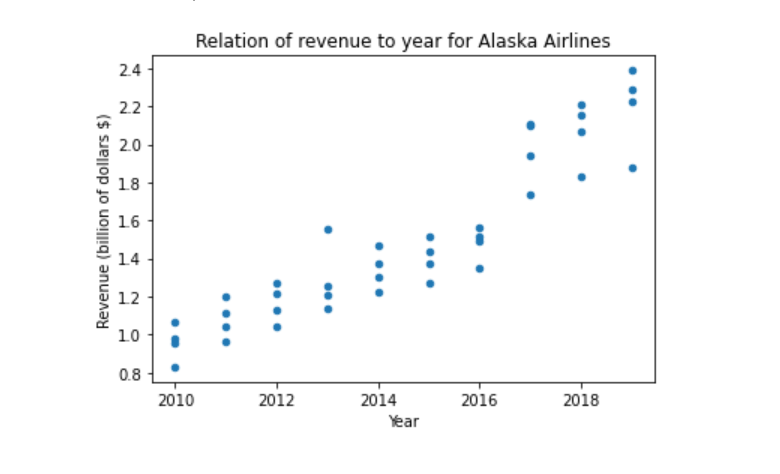
****

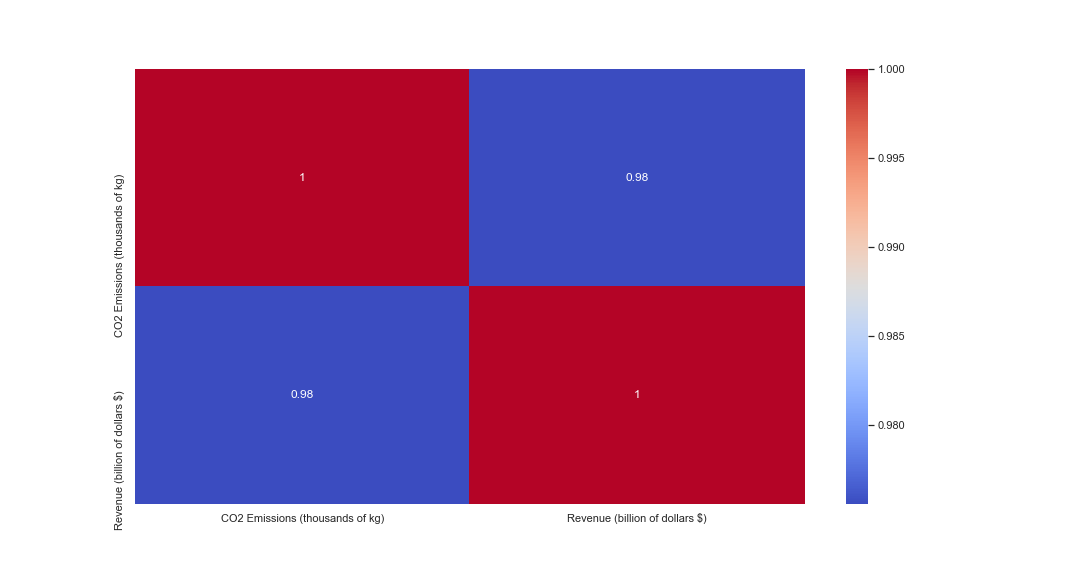
**Appendix 3.2:Plot to represent the emissions and revenue per year for each Airline**

****

****

**Appendix 3.3: Plot to represent revenue for Alaska Airlines**

****

**Appendix 3.4:Correlation matrix of airlines’ revenue and CO2 emissions**

**Appendix 4: Heatmap of Key Airline Factors for 2019**