Flight On-time Analysis for Year 2005 - 2006

3 April 2023

Lei Yuhe

Registration Number: 210406713

Contents

Introduction

Flight on-time performance is a crucial metric for the aviation industry, as it directly affects customer satisfaction, airline revenue, and operational efficiency. It is essential for airlines to maintain a high level of punctuality to ensure a smooth travel experience for passengers, reduce delays and cancellations, and minimize costs.

In recent years, flight on-time performance has become increasingly important due to the rising number of air travellers and the growing competition in the aviation industry. Customers expect flights to depart and arrive on time, and delays can have a significant impact on their travel plans and overall experience. Moreover, the increasing demand for air travel has led to higher pressure on airport infrastructure and air traffic control systems, resulting in more frequent delays and congestions.

In this report, we analyse the US domestic flight data recorded in 2005 and 2006. We will evaluate various metrics, such as departure and arrival time (actual and scheduled), departure and arrival delay, to provide insights into the industry's overall performance and trends. We will also consider external factors, such as air traffic control issues, that may impact flight operations and contribute to delays.

The objective of this report is to provide a comprehensive overview of the flight on-time performance, identify the key factors influencing their performance, and assess on different modelling techniques for flight delay prediction.

Data Wrangling

The original flight data files are extracted from the compressed .bz2 files into .csv files to get the complete data observations. After choosing 2005 and 2006 flight data for analysis, along with the plane data, carrier data and airport data, they are imported into R and Python as data frames. Thereafter, the data frames will be manipulated for various exploratory analysis.

Data cleaning has been done in R to filter out rows with null values and remove any duplicated data. In Python, the original data with null values has been used for analysis. There will be differences in certain parts of the results which will be explained in the later part.

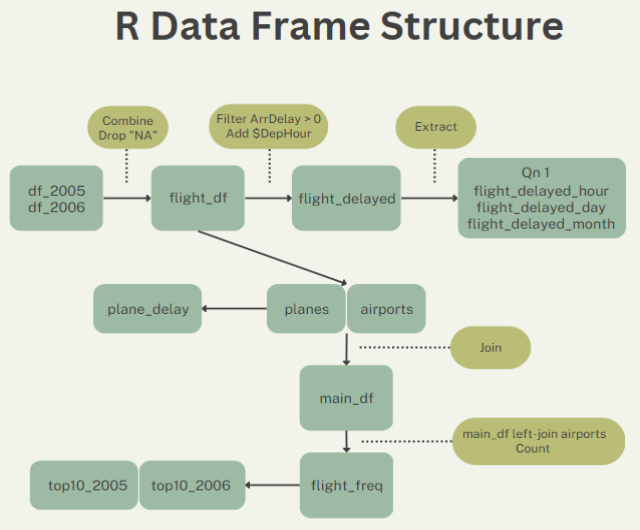

Figure 1. Data Frame Structure in R

Database Setup

In Python, we use SQL queries for data manipulation. Hence, a database called “airline.db” has been created in DB Browser for SQLite. We will have to connect to the database, read the data files into it as tables and execute SQL queries to manipulate the data. In R, we use dplyr package for data manipulation instead.

Name	Type	Schema
Tables (4)		
airports		CREATE TABLE "airports" ("iata" TEXT, "airport" TEXT, "city" TEXT, "state" TEXT, "country" TEXT, "lat" REAL, "long" REAL)
carriers		CREATE TABLE "carriers" ("Code" TEXT, "Description" TEXT)
flight		CREATE TABLE "flight" ("Year" INTEGER, "Month" INTEGER, "DayofMonth" INTEGER, "DayOfWeek" INTEGER, "DepTime" REAL
planes		CREATE TABLE "planes" ("tailnum" TEXT, "type" TEXT, "manufacturer" TEXT, "issue_date" TEXT, "model" TEXT, "status" TEXT

Figure 2. Tables in airline.db

Data Analysis

As one of the objectives of this report to find out key factors influencing the on-time performance, we analysed the relationship between departure time and arrival delay, as well as the relationship between plane age and arrival delay. We will also analyse flight frequencies between different locations to identify the trend of passenger flow. Finally, cascading failures due to late aircraft delay will also be discussed.

1. Relationship between Departure Time and Arrival Delay

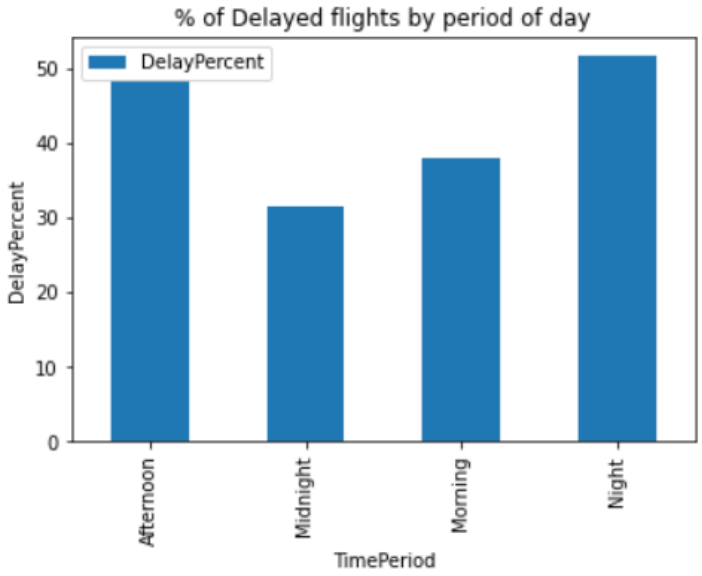

Figure 3. Percentage of delay by period (Python)

Departure time has been categorised into time periods, such that Midnight represents 12am to 6am, Morning represents 6am to 12pm, Afternoon represents 12pm to 6pm and Night represents 6pm to 12am. First, we visualise the percentage of delayed flights per period according to Arrival Delay, where ArrDelay > 0 refers to delayed flights. We can see from the bar chart that flights departed in the Midnight period (12am – 6am) causes the lowest percentage of delayed flights at around 30%.

Besides the percentage of delayed flights, we would also like to find out what is the average delay time for the delayed flights. The line plot generated in R on hourly average delay will give us some insights.

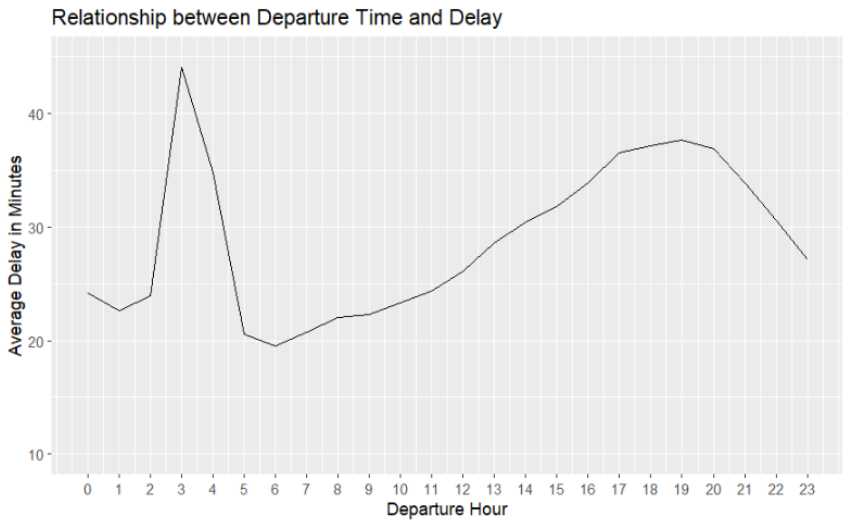

Figure 4. Average delay per hour of delayed flights (R)

We will look at the average delay from 12am to 6am first, since this period comes with the lowest percentage of arrival delay. Within this period, the average delay time per hour is also relatively low except 3am and 4am, where average delay time is the highest at 3am. This implies that the probability of being caught up by delayed flights is lower, but it comes with higher risks of longer delay time.

The best time to fly to minimise delay is at 6am because there both the delay percentage and average delay time is lower at that time. After 6am, the average delay time continues to increase and peaks at 7pm.

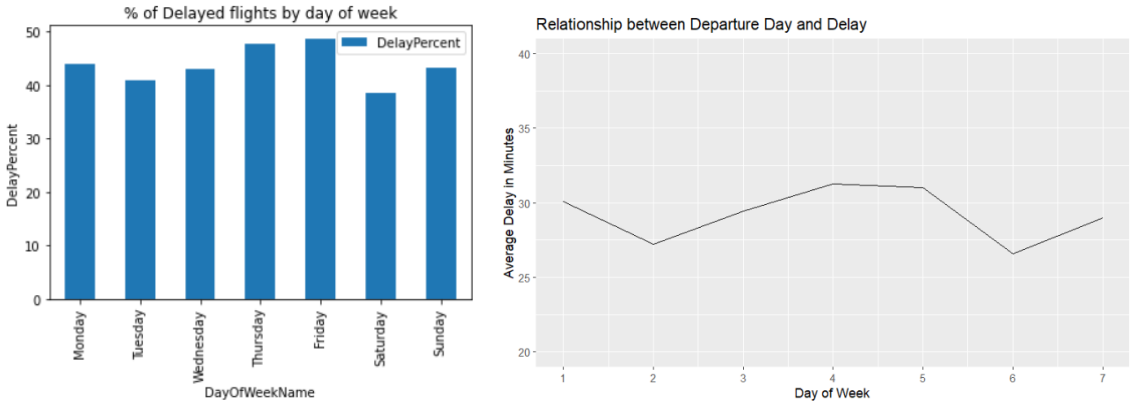

Figure 5. Delay by day of week (Python and R)

The percentage of delayed flights by day of week shows that people should choose Saturday to fly to minimise delays. The average delay time in Saturday is also lower than other days of the week. During the weekdays, Tuesdays and Wednesdays are better choices for passengers to minimise delays.

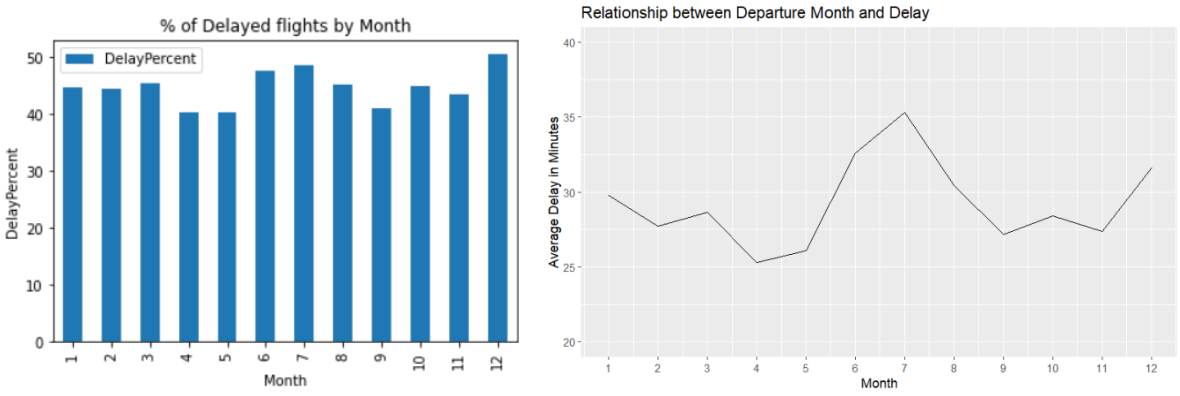

Figure 6. Delay by month (Python and R)

By grouping the flight data by month, we get the monthly delayed flight percentage. There are 2 seasons suitable to fly to minimise delays. One season is from April to May, another season is from

September to November. During both seasons, the delay percentage and average delay time are relatively lower.

2. Relationship between Plane Age and Arrival Delay

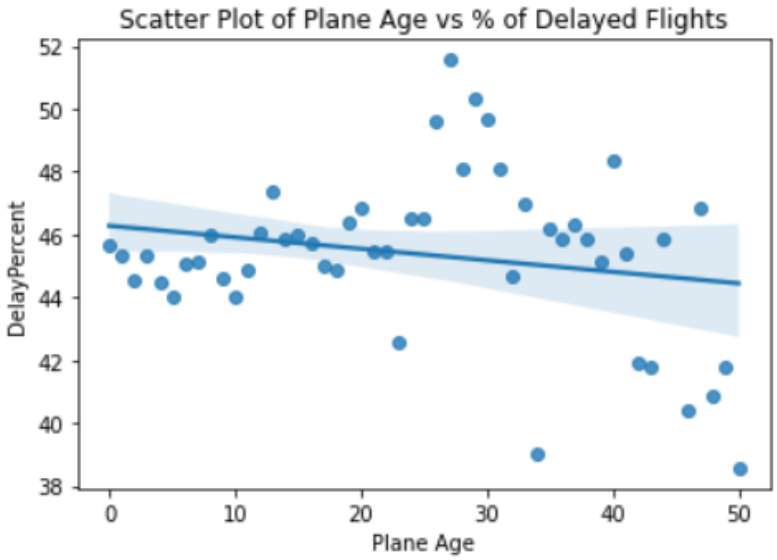

Figure 7. Correlation between plane age and percentage of delayed flights (Python)

To find the relationship between age of plane and flight delay percentage, we first calculated the age of the planes by subtracting the plane's manufacture year which is available in the planes data frame from the year of flight (either 2005 or 2006). Because we use the original data for data manipulation in Python, there are some outliers due to null values, such as plane age is 2005 or 2006. Hence, we filtered out those data by limiting the plane age between 0 and 60 years.

From the above regression plot generated in Python, it shows that there is a weak negative relationship between plane age and flight delay percentage. The older the plane, the lesser the flight delay percentage.

Besides the delay percentage, we also generated a regression plot to visualise the relationship between average delay time and age of planes.

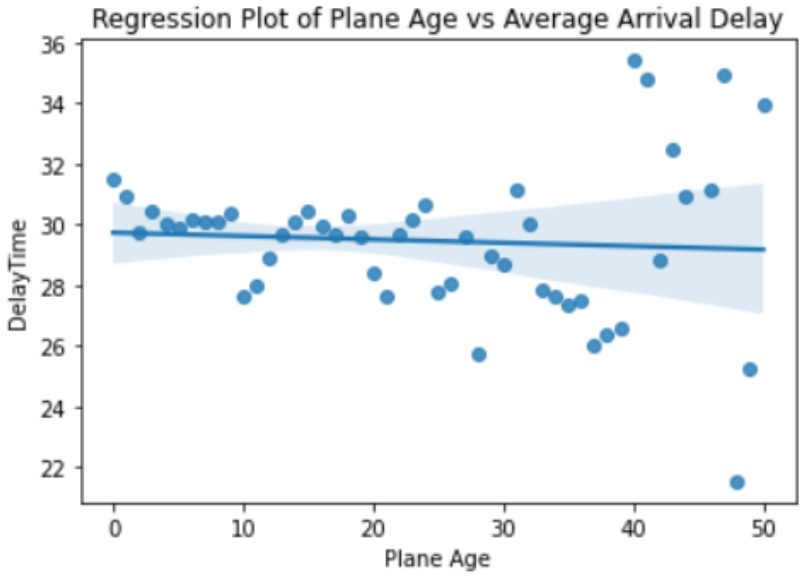

Figure 8. Correlation between plane age and average delay (Python)

We can see a slight negative correlation, but with a flatter slope compared to the previous regression plot. The flight delay percentage and the average delay time are concentrated for planes aged between 0 to 20, but the data are more dispersed when the planes get older.

Hence, the older planes may not suffer more delay. But it's less predictable because the data are dispersed and there is a possibility for plane delay to be very high.

3. Trend on Passenger Flow

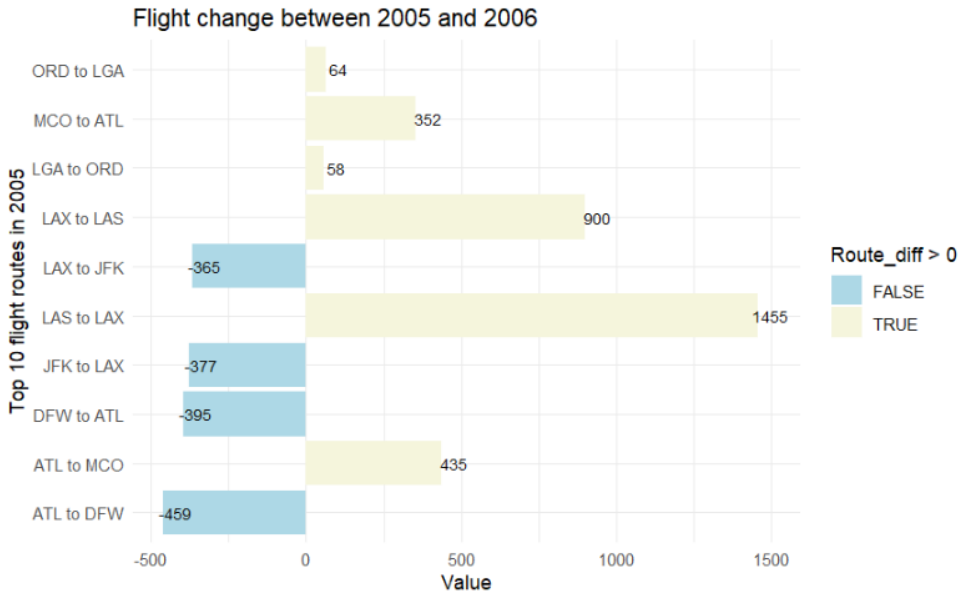

Figure 9. Top 10 flight routes in 2005 (R)

Since the passenger data is not provided, flight frequency is used to estimate the passenger flow, with the assumption that each flight carries same number of passengers.

We calculated the flight frequencies by counting the flight origin, then sorted it to get the top 10 flight routes in 2005. We also extracted out the flight routes in 2006 to match the top 10 routes in 2005 and calculated their differences. Therefore, we can see how the flight frequencies change from 2005 to 2006.

From the bar chart above, 4 of the top 10 routes in 2005 decreased in flight frequencies in 2006 while the other 6 increased. LAS – LAX and LAX – LAS are the 2 routes with the highest increase in flight frequencies.

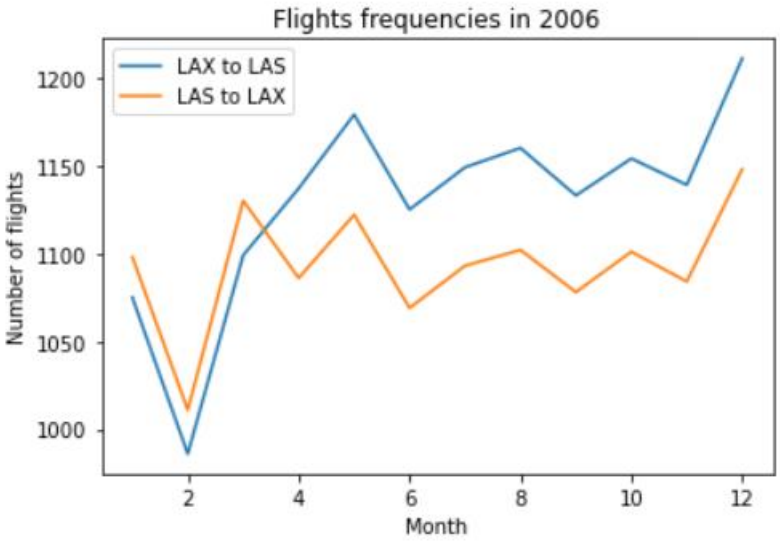

Figure 10. Change of flight frequencies across year (Python)

We did the same analysis in Python, but the ranking of the flight frequencies is different from the results in R because of the different data cleaning process. We decided to analyse the monthly flight frequencies on the 2 routes with highest increasing rate, which are LAS – LAX and LAX – LAS. These two also appeared in the top 10 routes generated both in Python and R.

From the line plot above, we can see that passengers travelling between LAX and LAS hit the bottom in February. Since February, the number of flights increases with fluctuations and reaches the top in December.

4. Cascading Failures Detection

According to the US Bureau of Transportation Statistics, late-arriving aircraft[1] refers to a previous flight with same aircraft arrived late, causing the present flight to depart late. From the data provided, there is a LateAricraftDelay variable, recording the flight delay time caused by the previous flight of the same aircraft.

Hence, to detect the flights with cascading delays, we first limit the LateAricraftDelay to 2 hours (120mins), the output is as shown in the table below.

	Year	Month	DayofMonth	TailNum	Origin	Dest	LateAircraftDelay
0	2005	1	3	N462UA	SFO	SNA	119
1	2005	1	4	N352UA	ORD	LGA	119
2	2005	1	26	N542UA	LAX	IAD	119
3	2005	1	5	N488UA	PHX	SFO	119
4	2005	1	4	N601AU	CLT	FLL	119
...	...	...	...	...	...	...	...
1276495	2006	12	27	N676DL	JAC	ATL	1

Figure 11. Late Aircraft Delay (Python)

Then, we select one of those aircrafts and visualise the flights by the same aircraft in the same day and we have the below table telling us about the cascading delays caused by late-arriving aircraft. For example, the plane with tail number N462UA has a late aircraft delay of 119 mins, it executed 5 flight tasks on 3rd January 2005, 4 of those are with cascading delays where the plane arrives late in one airport causing the subsequent flight departure delay in the same airport.

	Year	Month	DayofMonth	DepDelay	ArrDelay	TailNum	Origin	Dest	LateAircraftDelay
0	2005	1	3	2.0	47.0	N462UA	DEN	LAX	0
1	2005	1	3	146.0	134.0	N462UA	LAX	SFO	36
2	2005	1	3	126.0	126.0	N462UA	SFO	SNA	119
3	2005	1	3	119.0	114.0	N462UA	SNA	SFO	114
4	2005	1	3	81.0	85.0	N462UA	SFO	BUR	81

Figure 12. Cascading delay for N462UA (Python)

Therefore, we are able to detect cascading failures based on the late aircraft delay time. The flights with high late aircraft delay time are likely to cause cascading failures.

[1] Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics

Data Modelling

The objective is to build models for effective arrival delay prediction. Hence, we take ArrDelay as the response variable in the modelling process. For predictor variables, we did a correlation analysis in Python on the variables in the flight data. The result returned the correlation coefficient for each continuous variable, with 6 variables correlated to ArrDelay.

We take the response variable ArrDelay as a continuous variable, so we will use regression under the supervised learning. Linear Regression, Ridge Regression and Lasso Regression are tested for their fit. The dataset used for modelling are divided into training and testing sets. 70% of data are used for training and 30% used for testing. Below table lists out the modelling results, including Mean Squared Error (MSE) and R-Squared value. Both are suitable measures for regression tasks.

We can see that the results are very close together. Lasso regression performed the best in Python with lower MSE and higher R-Squared value, whereas Linear regression performed the best in R with lower MSE.

	Mean Squared Error (R)	Mean Squared Error (Python)	R-Squared (Python)
Linear Regression	65.7329	65.27	0.9504776721215474
Ridge Regression	65.73912	65.27	0.9504776721327222
Lasso Regression	65.76232	65.23	0.9505094421398566

Conclusion

To conclude, this report analysed the key factors influencing the flight arrival delay, such as flight departure time and plane age. An overview of the trend on passenger flow is provided and cascading delayed is discussed where late arriving time causes late departure time of subsequent flight. Lastly, three regression models are tested for their fit to predict arrival delay.

From the analysis, flights that departure in the early morning at around 5-6am cause lower delay percentage and average delay time. Tuesday and Saturday would be better choices to fly to minimise delays. April – May, September – November are the two seasons with lesser delay. It also shows that older plane's performance is less consistent compared to new planes. The flights between LAX and LAS increased significantly in 2006.

For modelling tasks to predict arrival delay, it's recommended to use linear and lasso regression model.

References

(23 February, 2023). Retrieved from DataScience Made Simple: https://www.datasciencemadesimple.com/join-in-r-merge-in-r/

(23 February, 2023). Retrieved from DataScience Made Simple: https://www.datasciencemadesimple.com/remove-duplicate-rows-r-using-dplyr-distinct-function/

(18 February , 2023). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/drop-columns-by-name-from-a-given-dataframe-in-r/

(20 February, 2023). Retrieved from R : KEEP / DROP COLUMNS FROM DATA FRAME: https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html

(27 February, 2023). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/how-to-perform-a-countif-function-in-r/

(21 February, 2023). Retrieved from Bureau of Transportation Statistics: https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations

case_when: A general vectorised if. (20 February, 2023). Retrieved from https://www.rdocumentation.org/packages/dplyr/versions/1.0.10/topics/case_when

R Remove Data Frame Rows with NA Values. (25 February, 2023). Retrieved from https://www.youtube.com/watch?v=O_gPPrezk5o