1) Project Description

For my DSCI510 class project, I am interested in studying the relationship between the cost of flight tickets and airline performance in the US. Specifically, I would like to answer the following three questions regarding US budget airlines:

- Are flights operated by budget airlines more likely to be cancelled?
- Do budget airlines typically have longer flight delays?
- Are budget airlines less safe than other airlines?

To answer the above questions, I have collected data from three different sources: Kayak website, Aviation Edge API, and airplane crash data from Kaggle website. Since the questions are correlational in nature, the main method of analysis employed is the calculation of correlation coefficients between the different variables using Python.

2) Motivation

Recently, I was searching for flight tickets from Los Angeles to New York and I remember how overwhelming it was to find many different ticket prices offered by multiple US airlines. The cheapest price offered by "Spirit Airlines" immediately caught my attention, but then I wondered if these cheap airlines actually have a bad reputation. Airline performance is evaluated by many different aspects such as: cancellation rates, average delays, overall safety, quality of service, and luggage handling. Many people tend to have negative opinions about budget airlines, they often think that these airlines have poor on-time performance, high cancellations, and they're not too safe. However, how much of these negative perceptions are actually true? In this project, I would like to see if the bad reputation of budget airlines is actually true, by using real-world data and performing actual statistical analysis in order to discover the truth.

3) Data Sources

Dataset 1: Scraping Flights Information from "Kayak" Website
Link: https://www.kayak.com/flights

In order to find out what are the cheapest US airlines, I will use the *Kayak* website to scrap information about flight prices and their corresponding airline operators for multiple routes. Essentially, I will be inputting different one-way flight routes (non-stop) within the United States in order to discover a pattern of which airlines commonly offer the cheapest flights and which ones are typically more expensive. The output of this analysis will be a list of US airlines ranked by their typical ticket price, from cheapest to most expensive airlines.

Dataset 2: Flight Delay API – Current & Historical Cancellations and Delays
Link: https://aviation-edge.com/flight-delay-api/

To find out the statistics for how often the flights operated by each airline were either delayed or cancelled, the "Flight Delay API" by Aviation Edge can be used. This external API service provides historical flight schedule and timetables data of all airports and airlines around the world. I am planning to use this API to generate statistics regarding the flight delay or cancellation behavior for all the airlines in the list that I created from Dataset 1.

Dataset 3: Airplane Crash Data Since 1908 (Kaggle dataset)
Link: https://www.kaggle.com/datasets/cgurkan/airplane-crash-data-since-1908

To see if there is a relationship between cost of flight tickets and safety, I will use the dataset from *Kaggle* which summarizes all aircraft accidents from 1908 to 2019. Using the list of airlines generated from Dataset 1, I will create a new dataset that will have a column for the number of accidents that each airline was involved in.

4) Data Extraction

Scraping "Kayak" Website

Before integrating all the datasets together, a list of US airlines ranked by their typical ticket price must be generated first, by scraping flight ticket prices from "Kayak" website. The Python libraries Selenium, BeautifulSoup4, and Webdriver Manager were used. A list of 15 different one-way flight routes were entered in the "Kayak" website, as shown in Table 1, to scrape flight ticket prices for multiple US airlines. A single departure date was chosen as June 1st 2022 for all flight routes and no stops are allowed. The flights were entered in "Kayak" by generating a URL for each flight route and then looping through all of them. Each time, the Chrome web browser is automatically opened and the airline and ticket price information are scraped with BeautifulSoup4 and stored in a CSV file called "flight_prices.csv". The output CSV file is printed in terminal as can be seen in Figure 1.

Table 1: Flight routes searched in "Kayak"

Flight Routes	
JFK to LAX	JFK to MIA
LGA to ORD	HNL to LAX
ATL to MCO	DEN to LAS
ATL to FLL	HNL to SFO
DEN to PHX	EWR to MCO
LAX to SFO	ATL to LGA
LAS to LAX	DEN to LAX
LGA to MIA	

	Route	Price 1	Price 2	Price 3	Price 4	Price 5	Price 6
0	JFK to LAX	United Airlines: $230	JetBlue: $243	American Airlines: $269	NaN	NaN	NaN
1	LGA to ORD	United Airlines: $84	Delta: $84	American Airlines: $94	NaN	NaN	NaN
2	ATL to MCO	Frontier: $43	Spirit Airlines: $59	Delta: $109	NaN	NaN	NaN
3	ATL to FLL	Spirit Airlines: $40	Frontier: $41	Delta: $49	NaN	NaN	NaN
4	DEN to PHX	Frontier: $111	United Airlines: $141	American Airlines: $192	NaN	NaN	NaN
5	LAX to SFO	JetBlue: $44	United Airlines: $59	Alaska Airlines: $64	Delta: $79	American Airlines: $94	NaN
6	LAS to LAX	Spirit Airlines: $40	JetBlue: $44	Alaska Airlines: $44	Delta: $49	American Airlines: $94	United Airlines: $114
7	LGA to MIA	Spirit Airlines: $28	Delta: $74	American Airlines: $94	NaN	NaN	NaN
8	JFK to MIA	Delta: $74	JetBlue: $74	American Airlines: $94	NaN	NaN	NaN
9	HNL to LAX	United Airlines: $289	Alaska Airlines: $319	Hawaiian Airlines: $325	Delta: $360	American Airlines: $559	NaN
10	DEN to LAS	Frontier: $39	Spirit Airlines: $40	United Airlines: $173	NaN	NaN	NaN
11	HNL to SFO	Hawaiian Airlines: $215	United Airlines: $217	Alaska Airlines: $279	NaN	NaN	NaN
12	EWR to MCO	Spirit Airlines: $59	United Airlines: $129	NaN	NaN	NaN	NaN
13	ATL to LGA	Frontier: $54	Delta: $129	American Airlines: $149	JetBlue: $149	NaN	NaN
14	DEN to LAX	Delta: $139	United Airlines: $159	American Airlines: $260	NaN	NaN	NaN

Figure 1: Flight ticket prices scraped from "Kayak" (Dataset 1)

After the price information is extracted, the Python Statistics library is used to calculate the Z-score for each ticket price offered by an airline, for a single flight route. A dictionary of Z-scores is created for each airline, after which the average Z-score is computed. The airline ranking according to typical ticket price is obtained by ranking the airlines based on their average Z-score values. The results of this ranking are shown in Table 2, based on the information in the static dataset "flight_prices.csv", rank 1 means the airline is cheapest while rank 8 means it's most expensive. However, it is important to note that the flight ticket prices are dynamic in nature and that they change every day. Thus, the results of airline ranking shown in Table 2 (which were obtained on April 23rd) might not be the same if the price information is scraped again on another day, but it will generally be similar with few differences. Another thing to note is that the airline "Southwest" is not in the list, even though it is a well-known budget airline. The reason is that "Southwest" does not allow any third-party websites to display its flight ticket prices, including "Kayak".

Table 2: Airline price ranking based on Z-scores

Airline	Average Z-Score	Ranking
Frontier	-0.844	1
Spirit Airlines	-0.692	2
Hawaiian Airlines	-0.512	3
JetBlue	-0.410	4
Alaska Airlines	-0.042	5
United Airlines	-0.037	6
Delta	0.084	7
American Airlines	1.114	8

Scraping "Aviation Edge" API

Using the airline list that was generated in the previous section, statistics about the cancellation rates and delays can be obtained from the "Flight Delay" API offered by "Aviation Edge". This API can generate historical flight schedules for a given airport at a given date range, and the results can be filtered according to airline name. The free API key was obtained by contacting the "Aviation Edge" sales team, and they happily provided the key with a monthly call limit of 30,000. The Python library Requests was used to first obtain the airport hub code and the airline IATA code for each airline in the list, by utilizing the "Airline Database" API that's also offered by "Aviation Edge". Then, a URL was generated to make a GET request to retrieve historical flight schedules for each airline from the "Flight Delay" API. The URL consists of the API key, the airport hub code, the start date, the end date, and the airline IATA code. For the start and end dates, the API has a limit of around 1 week as a date range of a historical flight schedule, due to the high airport traffic in some hubs. Thus, flight schedules for 4 different weeks in the year 2021 were extracted and combined together for analysis. The four weeks correspond to: first week of July, first week of August, Thanksgiving week, and Christmas week. All of these weeks are considered busy travel weeks. Moreover, only the departure schedules were considered.

The output of the GET request is a JSON response from which the following information was extracted: 'flight date', 'flight number', 'airline name', 'departure airport', 'flight status', and 'departure delay'. The information was then stored in a SQL database "flight_data.db", with

the help of the SQLite Python library. A sample of the output printed as a Dataframe in terminal can be seen in Figure 2, where the total number of historical flights is 76,799.

	Flight_Number_IATA	Flight_Date	Airline_Name	Departure_Airport	Flight_Status	Delay_mins
0	aa2698	2021-06-30	American Airlines	DFW	scheduled	63
1	aa4173	2021-06-30	American Airlines	DFW	scheduled	62
2	aa1217	2021-06-30	American Airlines	DFW	scheduled	47
3	aa2226	2021-06-30	American Airlines	DFW	active	15
4	aa2448	2021-06-30	American Airlines	DFW	active	26
...	...	...	...	...	...	...
76794	nk1078	2021-12-27	Spirit	FLL	active	13
76795	nk977	2021-12-27	Spirit	FLL	active	39
76796	nk2157	2021-12-27	Spirit	FLL	active	23
76797	nk380	2021-12-27	Spirit	FLL	active	13
76798	nk2006	2021-12-27	Spirit	FLL	scheduled	55

Figure 2: Historical flight schedules from API (Dataset 2)

Airplane Crash Data (Kaggle)

The third data source is a ready-made CSV dataset called “Airplane Crashes and Fatalities Since 1908” found in “Kaggle” website. This dataset lists all fatal aircraft accidents (including ones that involved commercial airlines) from 1908 to 2019. However, it does not include any aircraft “incidents”, which do not involve any fatalities. The CSV data was imported into a SQL database “airplane_crashes.db” through the Python SQLite library. A sample of the output in terminal can be seen in Figure 3, where the total number of rows is 4,967.

	Date	Time	Location	Operator	...	Fatalities_Passangers	Fatalities_Crew	Ground
0	09/17/1908	17:18	Fort Myer, Virginia	Military - U.S. Army	...	1	0	0
1	09/07/1909		Juvisy-sur-Orge, France		...	0	0	0
2	07/12/1912	06:30	Atlantic City, New Jersey	Military - U.S. Navy	...	0	5	0
3	08/06/1913		Victoria, British Columbia, Canada	Private	...	0	1	0
4	09/09/1913	18:30	Over the North Sea	Military - German Navy	...	NULL	NULL	0
...	...	...	...	...	...	...	...	...
4962	04/16/2019	11:00	Puerto Montt, Chile	Archipelagos Service Aereos	...	5	1	0
4963	05/05/2019	18:30	Near Monclava, Mexico	TVPX Aircraft Solutions	...	11	2	0
4964	05/05/2019	18:30	Moscow, Russia	Aeroflot Russian International Airlines	...	40	1	0
4965	06/03/2019	13:00	Near Lipo, India	Military - Indian Air Force	...	5	8	0
4966	07/30/2019	02:00	Rawalpindi, India	Military - Pakistan Army	...	0	5	14

Figure 3: Aircraft crashes since 1908 (Dataset 3)

A workflow summarizing all the data extraction and data generation steps from the three different sources is as shown in Figure 4.

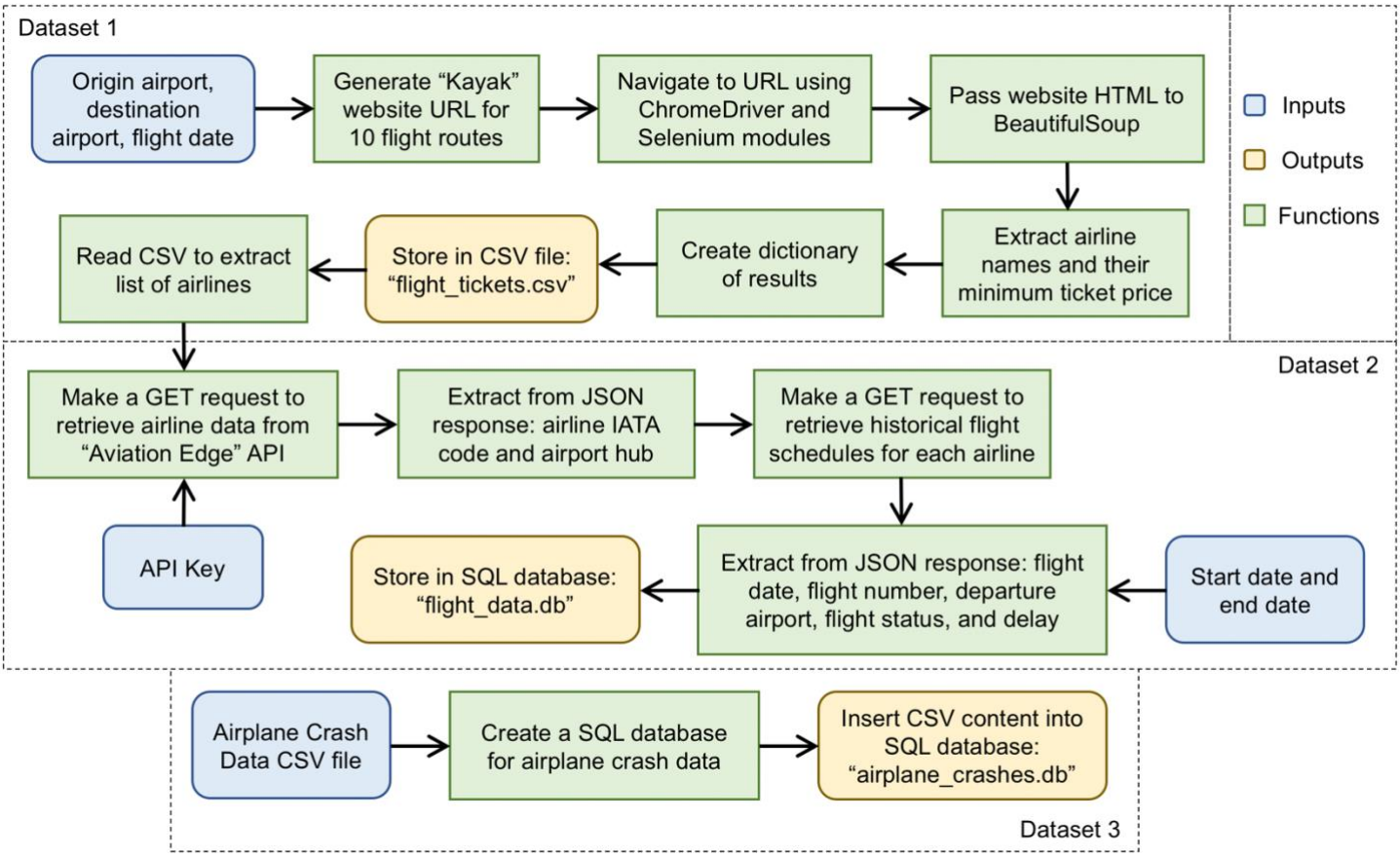

Figure 4: Datasets generation flowchart

5) Data Analysis

Combined Dataset

The following Table 3 presents the results after combining the datasets from the three different sources. The "Cancellation Rate" was calculated by dividing the number of cancelled flights by the total number of flights for each airline in the output file "flight_data.db", and multiplying the result by 100. The "Average Delay" was calculated by finding the average of all the departure delay values (in minutes) for each airline in "flight_data.db".

Table 3: Combined dataset for the list of US airlines

Airline	Price Rank	Price Z-Score (Avg)	Total Flights	Cancelled Flights	Cancellation Rate (%)	Average Delay (mins)	Number of Accidents
Frontier	1	-0.844	1,612	15	0.93	20.54	0
Spirit Airlines	2	-0.692	2,376	268	11.28	30.95	0
Hawaiian Airlines	3	-0.512	2,308	24	1.04	14.98	0
JetBlue	4	-0.410	5,877	109	1.85	38.50	0
Alaska Airlines	5	-0.042	8,165	256	3.14	24.44	9
United Airlines	6	-0.037	13,490	161	1.19	28.32	44
Delta	7	0.084	20,898	256	1.22	18.86	12
American Airlines	8	1.114	22,073	938	4.25	33.30	37

Analysis 1: Airline Price Ranking vs Cancellation Rate

In order to answer the question of whether cheaper airlines are more likely to have more flight cancellations, the correlation coefficient between the price ranking and cancellation ratio was calculated. Using Python's Scipy library, Pearson's r correlation coefficient can be easily obtained. The results are visualized as shown in the following Figure 5. Given that the coefficient value is -0.21, there's a very weak negative correlation between the variables. Therefore, there's no strong evidence to support the claim that cheaper airlines have higher cancellation rates. However, there's one exception to the rule as there's a very clear outlier. Looking at the bar chart in Figure 6, "Spirit Airlines" (second cheapest airline) had the highest percentage of cancelled flights at around 11%. Therefore, the hypothesis that budget airlines tend to have higher cancellation rates is not entirely true for all budget airlines, except for "Spirit Airlines".

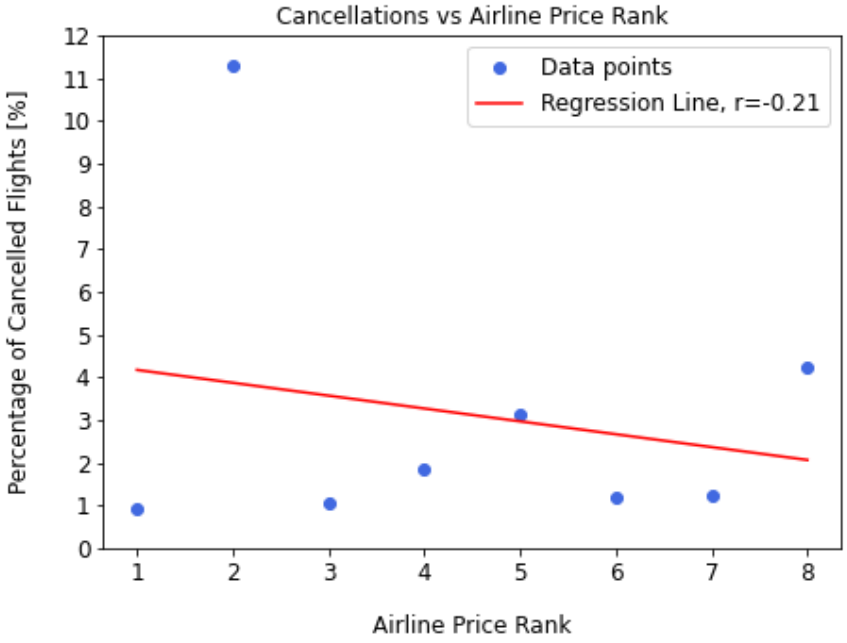

Figure 5: Scatter plot of cancellations vs airline price rank

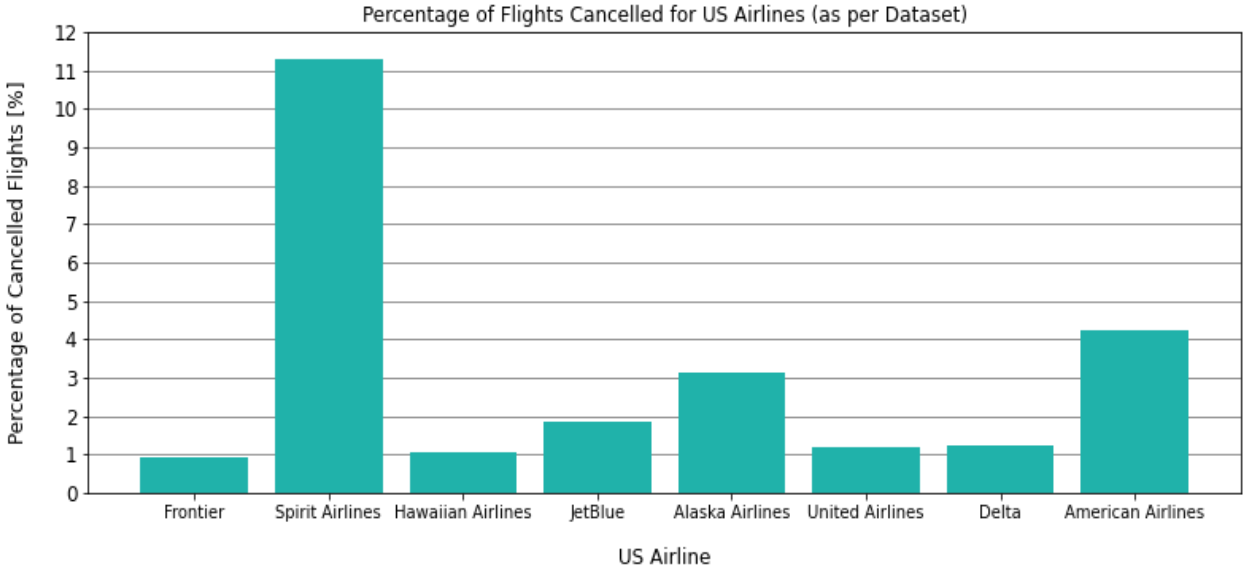

Figure 6: Bar chart of cancellations vs airline

Analysis 2: Airline Price Ranking vs Average Departure Delay

The second question is: do cheaper airlines have longer flight delays? To answer this, the Pearson's r correlation coefficient between airline price ranking and the average departure delay can also be calculated. The results are shown in the scatter diagram in Figure 7. Since the Pearson's r value is 0.20, the correlation between the variables is positively weak, which means there's likely no relationship between the variables. In fact, the airline with the highest average delay of 38 minutes is actually a non-budget airline "Jetblue", which stands somewhere in the middle in terms of the cost of its flight tickets.

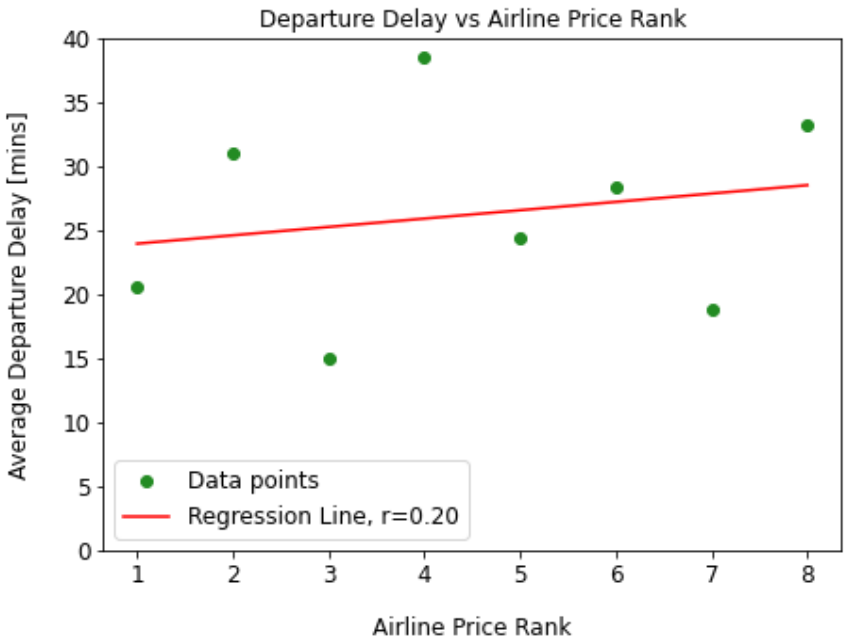

Figure 7: Scatter plot of average delay vs airline price rank

Analysis 3: Airline Price Ranking vs Aircraft Accidents

Finally, are budget airlines "less safe" than other airlines? Looking at the statistics from the accidents database, the answer is quite the opposite. As can be seen in Table 4, the top 4 cheapest airlines have never been involved in a fatal aircraft accident thus far. This is mainly due to the fact that the more expensive airlines have been around for much longer (since the 1930s). However, "Hawaiian Airlines" was also founded in 1929 but has never lost a single passenger. Figure 8 shows the correlation between the airline price ranking and number of aircraft accidents. As expected, since the top 4 budget airlines have never had an accident, there is a strong positive correlation between the variables and the Pearson's r coefficient value is 0.75. This proves that budget airlines are actually not less safe than any other airline.

Table 4: Airline accidents information

Airline	Rank	Year Founded	Number of Accidents
Frontier	1	1994	0
Spirit Airlines	2	1980	0
Hawaiian Airlines	3	1929	0
JetBlue	4	2000	0
Alaska Airlines	5	1932	18
United Airlines	6	1931	88
Delta	7	1928	24
American Airlines	8	1934	74

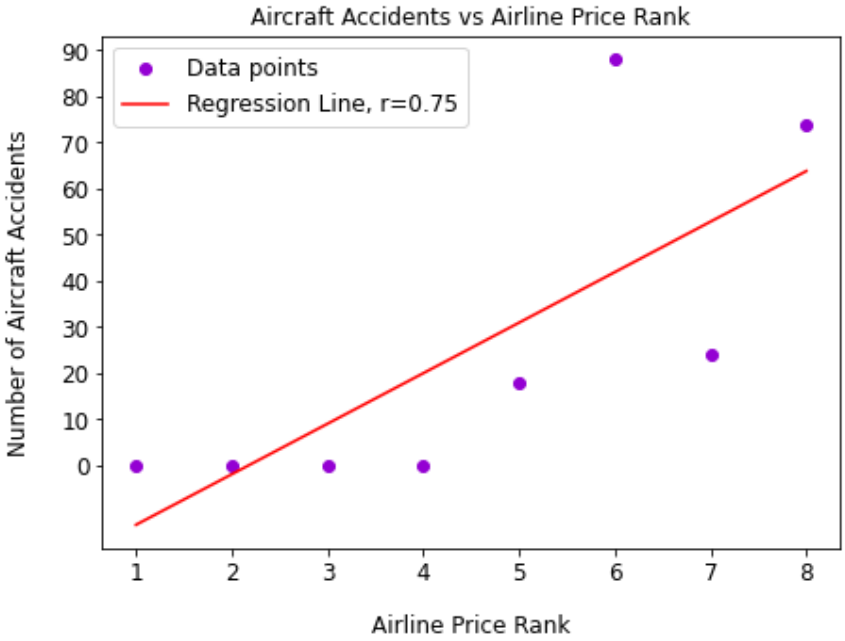

Figure 8: Scatter plot of accidents vs airline price rank

6) Conclusion

Going back to the original research questions, it can be concluded that the negative perceptions that many people have about budget airlines are not entirely true. The analysis results shown in this report proved that not all budget airlines have terrible delays or cancellations, and they are actually just as safe as any other airline.

The insights generated by this project can be used to create dashboards to help better educate the public about budget airlines and how do the different US airlines compare in general. Ultimately, these dashboards will help the user make better informed decision about which airline to book for their next trip.

Further research can be done by looking at additional airline performance metrics. For instance, the Twitter API can be utilized to mine people's opinions and complaints about US airlines. The tweets can also be further categorized into categories such as "cabin service" or "luggage handling" based on most popular keywords. This data can be used to answer the question of whether budget airlines have the worst customer satisfaction.

7) Extensibility and Maintainability

The code can be further extended or generalized in several ways. For example, in scraping the "Kayak" website, additional flight routes can be added to further confirm the trend of which airlines typically offer cheaper tickets and which ones are more expensive. As for using the API from "Aviation Edge", the code currently scrapes flight schedules for specific airports at specific periods in time. This can be further extended by modifying the code to get data from different airports or date ranges to make a more detailed and generalized airline study.

As for maintainability, since I was given a free developer API key by the "Aviation Edge" sales team, the API call limit is 30,000 calls, which is more than sufficient for the purposes of this project. However, due to high airport traffic in some busy airports, I can only extract historical flight schedules for up to 1 week in some cases. Additionally, in scraping the "Kayak" website, one should keep in mind that the prices of flight tickets are always changing and are never fixed. Thus, the flight fares in the static dataset "flight_tickets.csv" are most likely to be outdated.