Josh Wattay
CareerFoundry
Exercise 6.1 Task

Data Source

This data set was sourced from Kaggle.com (https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022). The source of the raw data provided by the Kaggle.com user is from the Bureau of Transportation Statistics (link below): https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr

The Bureau of Transportation Statistics (BTS) is a federal agency within the United States Department of Transportation (DOT). Its primary role is to collect, analyze, and disseminate transportation-related data and statistics to support decision-making, policy development, and research within the transportation sector. BTS provides a wide range of information, including statistics on air, maritime, rail, and highway transportation, as well as on topics such as safety, energy, and the economy. The agency serves as a valuable resource for policymakers, researchers, businesses, and the general public interested in understanding various aspects of transportation in the United States.

This data set was selected because it contains tens of thousands of records of both geospatial, temporal, and continuous quantitative variables that can be utilized for the purposes of this achievement. I am also interested in the performance of different airlines and given the modern ease of access to airline travel as a means of relatively affordable domestic transportation in the United States. While there are certainly challenges, it is fascinating to me that airlines are able to accommodate large volumes of travelers.

Data Profile

Data Cleaning

```
[38]: #search for missing values

      null_counts = samples_combined.isnull().sum()

[39]: null_counts

[39]: FlightDate               0
      Airline                  0
      Origin                   0
      Dest                     0
      Cancelled                0
                              ...
      ArrDel15             15443
      ArrivalDelayGroups   15443
      ArrTimeBlk               0
      DistanceGroup            0
      DivAirportLandings       3
      Length: 61, dtype: int64
```

There were three (3) null values discovered in the DivAirportLandings column and 15443 null values discovered in the ArrDel15 and ArrivalDelayGroups columns. Further inspection demonstrated that these null values were for flights that were cancelled and therefore were not in a delay group. Therefore, these values will be included in the clean data set for analysis.

```
[40]: #identify columns with mixed variable types

      def detect_types(series):
          return series.apply(type).nunique()

      mixed_type_columns = [col for col in samples_combined.columns if detect_types(samples_combined[col]) > 1]

      print("Columns with mixed types:")
      print(mixed_type_columns)

      Columns with mixed types:
      ['Tail_Number']
```

Josh Wattay
CareerFoundry
Exercise 6.1 Task

There was one column with mixed variable type, which was the Tail_Number column. This was expected, as tail numbers of aircraft typically contain a combination of letters and numbers. This column had the data type changed to object in order to remove the mixed data type for future analysis.

[41]:
```
#check counts

for column in samples_combined.columns:
    print(f"Counts for column {column}:")
    print(samples_combined[column].value_counts())
    print("\n")
```

```
Counts for column FlightDate:
FlightDate
2022-07-01    553
2022-06-05    548
2022-06-08    547
2022-07-21    543
2022-04-28    539
             ...
2020-05-12    107
2020-05-30    104
2020-05-17    103
2020-05-16    100
2020-05-19     87
Name: count, Length: 1673, dtype: int64
```

Counts were inspected with the code pictured adjacent to this paragraph. There were no issues discovered with the counts and the totals resulted in 500000, consistent with the sample size taken for analysis (100,000 records per year (2018, 2019, 2020, 2021, and 2022)).

[42]:
```
#remove duplicates and rename dataframe to more appropriate title

flights_clean = samples_combined.drop_duplicates()
print(flights_clean)
```

	FlightDate	Airline	Origin	Dest	Cancelled	Diverted	\
2079139	2018-12-02	Comair Inc.	DCA	BTV	False	False	
4054924	2018-06-30	Spirit Air Lines	ATL	LAS	False	False	
2486050	2018-02-11	Frontier Airlines Inc.	MCI	RSW	False	False	
4020576	2018-06-22	United Air Lines Inc.	DEN	CID	False	False	
715981	2018-10-19	SkyWest Airlines Inc.	BGM	DTW	False	False	
...	...	...	...	...	...	...	
4016104	2022-03-08	Southwest Airlines Co.	DSM	DEN	False	False	
3347015	2022-01-16	Republic Airlines	DCA	ATL	False	False	
640944	2022-02-03	Southwest Airlines Co.	DEN	ORF	False	False	
3329891	2022-01-09	JetBlue Airways	MCO	PSE	False	False	
2702925	2022-06-22	Southwest Airlines Co.	DEN	SAT	False	False	

[500000 rows x 61 columns]

No duplicate records were detected.

Josh Wattay
CareerFoundry
Exercise 6.1 Task

Understanding the Data

This data set contains 61 columns and I have created a sample of 500,000 records (100,000 per year for the time period of 2018 – 2022). The columns are listed and defined more fully below:

Column Name	Data Type	Defined
Flight Date	Date	Date the Flight is scheduled to depart from the destination of origin
Airline	Nominal	The name of the Airline servicing the flight
Origin	Geospatial	The location of the origin of the flight AKA where the flight takes off
Dest (destination)	Geospatial	The location of the destination of the flight AKA where the flight will land
Cancelled	Boolean	Was the flight cancelled (true or false)
Diverted	Boolean	Was the flight diverted or redirected to a different destination after taking off (true or false)
CRSDepTime	Numerical (discrete)	Computer Reservation System departure time for the flight
DepTime	Numerical (discrete)	Actual departure time for the flight
DepDelayMinutes	Numerical (discrete)	Sum of minutes that the flight was delayed in departing from the airport, calculated by taking the difference between the CRSDepTime and DepTime columns
DepDelay	Numerical (discrete)	Sum of minutes that the flight was delayed in departing from the airport, calculated by taking the difference between the CRSDepTime and DepTime columns
ArrTime	Numerical (discrete)	Arrival time or time the flight actually landed (based on the 24 hour clock)
ArrDelayMinutes	Numerical (discrete)	Sum of minutes that the flight was delayed in arriving to its destination
AirTime	Numerical (discrete)	Sum of minutes the flight spent in the air during the flight
CRSElapsedTime	Numerical (discrete)	Computer Reservation System estimate for the amount of time between take-off and landing
ActualElapsedTime	Numerical (discrete)	The actual amount of time between take-off and landing
Distance	Numerical (continuous)	The distance in miles between the origin and destination
Year	Numerical (discrete)	The year the flight occurred
Quarter	Numerical (discrete)	The quarter the flight occurred

Josh Wattay
CareerFoundry
Exercise 6.1 Task

Month	Numerical (discrete)	The month the flight occurred
DayofMonth	Numerical (discrete)	The date of the Month the Flight occurred
DayOfWeek	Numerical (discrete)	The day of the week the flight occurred
Marketing Airline Network	Nominal	The abbreviation for the marketing network affiliated with the airline providing the flight
Operated_Or_Branded_Code_Share_Partners	Nominal	Reporting Carrier Operated or Branded Code Share Partners
DOT_ID_Marketing_Airline	Numerical (discrete)	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
IATA_Code_Marketing_Airline	Nominal	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
Flight_Number_Marketing_Airline	Numerical	Flight Number
Operating_Airline	Nominal	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
DOT_ID_Operating_Airline	Numerical	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
IATA_Code_Operating_Airline	Nominal	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
Tail_Number	Nominal	Tail Number
Flight_Number_Operating_Airline	Numerical	Flight Number
OriginAirportID	Numerical	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.

Josh Wattay
CareerFoundry
Exercise 6.1 Task

OriginAirportSeqID	Numerical	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
OriginCityMarketID	Numerical	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
OriginCityName	Nominal	Origin Airport, City Name
OriginState	Nominal	Origin Airport, State Code
OriginStateFips	Numerical	Origin Airport, State Fips
OriginStateName	Nominal	Origin Airport, State Name
OriginWac	Numerical	Origin Airport, World Area Code
DestAirportID	Numerical	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
DestAirportSeqID	Numerical	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
DestCityMarketID	Numerical	Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
DestCityName	Nominal	Destination Airport, City Name
DestState	Nominal	Destination Airport, State Code
DestStateFips	Numerical	Destination Airport, State Fips
DestStateName	Nominal	Destination Airport, State Name
DestWac	Numerical	Destination Airport, World Area Code
DepDel15	Binary Numerical	Departure Delay Indicator, 15 Minutes or More (1=Yes)
DepartureDelayGroups	Numerical	Departure Delay intervals, every (15 minutes from <-15 to >180)
DepTimeBlk	Numerical Range	CRS Departure Time Block, Hourly Intervals
TaxiOut	Numerical	Taxi Out Time, in Minutes
WheelsOff	Numerical	Wheels Off Time (local time: hhmm)
WheelsOn	Numerical	Wheels On Time (local time: hhmm)
TaxiIn	Numerical	Taxi In Time, in Minutes
CRSArrTime	Numerical	CRS Arrival Time (local time: hhmm)

ArrDelay	Numerical	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDel15	Numerical Binary	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
ArrivalDelayGroups	Numerical	Arrival Delay intervals, every (15-minutes from <-15 to >180)
ArrTimeBlk	Numerical Range	CRS Arrival Time Block, Hourly Intervals
DistanceGroup	Numerical	Distance Intervals, every 250 Miles, for Flight Segment
DivAirportLandings	Numerical	Number of Diverted Airport Landings

Limitations and Ethics

Analyzing flight data from 2018 to 2022 can offer valuable insights, but it comes with several limitations and ethical considerations.

1. Data Completeness and Accuracy:

- Not all airlines and airports may consistently report flight data.

- Data Errors: Mistakes in data entry or transmission can affect the accuracy of the analysis.

2. Data Standardization:

- Different Formats: Flight data from various sources may be in different formats, requiring extensive preprocessing to standardize.

- Varied Definitions: Terms and metrics (e.g., delays, cancellations) might be defined differently by different entities, complicating comparisons.

3. Temporal Changes:

- Pandemic Impact: The COVID-19 pandemic significantly disrupted air travel, making 2020-2021 data atypical. This period may not be representative of normal flight patterns and can skew trend analysis.

- Regulatory Changes: Changes in aviation regulations and policies over these years can affect flight operations and reporting.

4. Technological and Operational Changes:

- Improvements in Technology: Advances in aircraft technology and air traffic management systems can influence flight efficiency and safety, affecting trends over time.

- Operational Changes: Shifts in airline operations, such as route changes or fleet updates, can impact the data.

Josh Wattay
CareerFoundry
Exercise 6.1 Task
5. External Factors:

- Weather Events: Natural disasters and extreme weather can cause fluctuations in flight data, adding variability that might be hard to account for.

- Economic Factors: Economic conditions, fuel prices, and geopolitical events can also influence air travel patterns.

Ethical Considerations

1. Privacy:

- Passenger Data: If the analysis involves detailed passenger information, it raises significant privacy issues. Ensuring anonymization and compliance with data protection laws is crucial. Care must be taken not to expose sensitive information that could identify individuals or proprietary airline operations.

2. Bias and Fairness:

- Bias in Data: The data might reflect inherent biases, such as socioeconomic disparities in travel patterns. Others need to be aware of these biases to avoid misleading conclusions.

- Equity Issues: The findings could impact policy decisions that may disproportionately affect certain groups, such as less affluent travelers or smaller regional airlines.

3. Transparency and Accountability:

- Methodology Disclosure: Clear documentation of data sources, methodologies, and assumptions is necessary to ensure the analysis can be scrutinized and replicated.

- Responsible Use: The results of the analysis should be used responsibly, avoiding sensationalism or misuse that could lead to public distrust or panic.

4. Impact on Stakeholders:

- Airline and Airport Operations: Findings might influence operational decisions, impacting employees and stakeholders. Analysts should consider the broader implications of their work.

- Public Perception: Misinterpretation of data can affect public confidence in air travel safety and reliability. Clear communication of findings and their limitations is essential.

Mitigating Limitations and Ethical Risks

- Data Validation: Implement rigorous data validation and cleaning processes to improve data quality.

- Standardization Protocols: Develop protocols for standardizing data from different sources.

- Contextual Analysis: Incorporate contextual factors like the pandemic, regulatory changes, and technological advancements into the analysis to provide more nuanced insights.

Josh Wattay
CareerFoundry
Exercise 6.1 Task
- Privacy Safeguards: Ensure data is anonymized and stored securely, and comply with relevant data protection regulations.

- Bias Mitigation: Use statistical techniques to identify and correct for biases in the data.

- Ethical Frameworks: Establish ethical guidelines for data analysis and usage, ensuring transparency, accountability, and responsible communication of results.

Questions to Explore

How has the total number of flights (domestic and international) changed from 2018 to 2022? What are the trends in passenger numbers over this period?

What are the trends in on-time performance for flights over these years?

Which airports and airlines have the highest and lowest rates of delays?

How have flight cancellation rates changed over the years?

What are the primary causes of cancellations during this period?

Are there noticeable seasonal patterns in flight operations and delays?

How do holiday seasons and major events affect flight performance?

How did the COVID-19 pandemic impact flight volume and passenger numbers in 2020 and 2021 compared to 2018 and 2019?

What recovery trends are observed in 2022?

How did airlines and airports adjust their operations during the pandemic?

What long-term changes in airline operations and passenger behavior are attributable to the pandemic?

Which airports have seen the most significant changes in flight volume and on-time performance?

How do major hubs compare to regional airports in terms of delays and cancellations?

Are there regional differences in flight performance?

How do weather-related delays vary across different parts of the country?

How do different airlines compare in terms of on-time performance, cancellation rates, and passenger volume?

What are the trends in market share among major airlines?

How has the fleet composition and utilization changed for major airlines over these years?

What innovations or changes in operations have airlines implemented?

What trends can be observed in safety incidents and their causes from 2018 to 2022?