Task 6.1 Sourcing Open Data

Data Source

Analysis of US Airline Flight Delays and Cancellations (2019-2023)

Summary of Data Source

- **Provider**: US Department of Transportation, Bureau of Transportation Statistics
- **URL**: DOT On-Time Performance
- **Content**: Contains data on flight delays and cancellations for domestic flights in the US, covering flight routes, time ranges for events, and reasons for delays and cancellations.
- **Time Period**: January 2019 – August 2023
- **Format**: Monthly subsets, consolidated annually, processed using csvkit, Python, and Excel.

Explanation for Choosing This Data Set

I chose this dataset because:

- It provides comprehensive coverage of flight performance over several years, allowing for robust trend analysis.
- The dataset's granularity, with details on routes, time, and reasons for delays, supports in-depth examination of factors influencing flight punctuality.
- The data is sourced from a reputable government agency, ensuring its accuracy, reliability, and ethical collection.

Data Profile

Understanding the Data

There are 3000000 rows and 32 columns in the dataset.

Variables with description:

- FL_DATE: Flight date
- AIRLINE: Airline code
- AIRLINE_DOT: Airline DOT code
- AIRLINE_CODE: Airline code (alternative)
- DOT_CODE: Department of Transportation code

- FL_NUMBER: Flight number
- ORIGIN: Origin airport code
- ORIGIN_CITY: Origin city name
- DEST: Destination airport code
- DEST_CITY: Destination city name
- CRS_DEP_TIME: Scheduled departure time
- DEP_TIME: Actual departure time
- DEP_DELAY: Departure delay in minutes
- TAXI_OUT: Taxi out time in minutes
- WHEELS_OFF: Wheels off time
- WHEELS_ON: Wheels on time
- TAXI_IN: Taxi in time in minutes
- CRS_ARR_TIME: Scheduled arrival time
- ARR_TIME: Actual arrival time
- ARR_DELAY: Arrival delay in minutes
- CANCELLED: Cancellation indicator
- CANCELLATION_CODE: Cancellation code
- DIVERTED: Diversion indicator
- CRS_ELAPSED_TIME: Scheduled elapsed time
- ELAPSED_TIME: Actual elapsed time
- AIR_TIME: Air time in minutes
- DISTANCE: Distance between airports in miles
- DELAY_DUE_CARRIER: Delay due to carrier in minutes
- DELAY_DUE_WEATHER: Delay due to weather in minutes
- DELAY_DUE_NAS: Delay due to National Airspace System in minutes
- DELAY_DUE_SECURITY: Delay due to security in minutes
- DELAY_DUE_LATE_AIRCRAFT: Delay due to late aircraft in minutes

Data Cleaning and Consistency Checks

1. Identify and address missing or inconsistent data.
2. Ensure all variables have appropriate data types.
3. Remove duplicate entries if any.
4. Excluded Irrelevant Columns.
5. Summary Statical Analysis.

Variables	Missing Value	Handling Missing Values	Mixed datatype and handling	Missing Values after I	Duplicate Values
FL_DATE	0			0	0
AIRLINE	0			0	0
AIRLINE_DOT	0			0	0
AIRLINE_CODE	0			0	0
DOT_CODE	0			0	0
FL_NUMBER	0			0	0
ORIGIN	0			0	0
ORIGIN_CITY	0			0	0
DEST	0			0	0
DEST_CITY	0			0	0
CRS_DEP_TIME	0			0	0
DEP_TIME	77615	Dropped the rows that are not canceled.		0	0
DEP_DELAY	77644			0	0
TAXI_OUT	78806			0	0
WHEELS_OFF	78806			0	0
WHEELS_ON	79944		Mixed datatypes are found and handled by replacing the NAN with median.	0	0
TAXI_IN	79944		Mixed datatypes are found and handled by replacing the NAN with median.	0	0
CRS_ARR_TIME	0			0	0
ARR_TIME	79942	Dropped the rows that are not canceled.		0	0
ARR_DELAY	86198		Mixed datatypes are found and handled by replacing the NAN with median.	0	0
CANCELLED	0			0	0
CANCELLATION_CO	2920860	Replaced with 'Not canceled'		0	0
DIVERTED	0			0	0
CRS_ELAPSED_TIM	14	Replaced with Median Value		0	0
ELAPSED_TIME	86198		Mixed datatypes are found and handled by replacing the NAN with median.	0	0
AIR_TIME	86198		Mixed datatypes are found and handled by replacing the NAN with median.	0	0
DISTANCE	0			0	0
DELAY_DUE_CARRI	2466137	Replaced with 0 (No Delay)		0	0
DELAY_DUE_WEATI	2466137	Replaced with 0 (No Delay)		0	0
DELAY_DUE_NAS	2466137	Replaced with 0 (No Delay)		0	0
DELAY_DUE_SECUF	2466137	Replaced with 0 (No Delay)		0	0
DELAY_DUE_LATE_	2466137	Replaced with 0 (No Delay)		0	0

	FL_NUMBER	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	CRS_ARR_TIME	ARR_TIME	ARR_DELAY	CANCELLED	DIVERTED
count	2.920058e+06	2.920058e+06	2.920058e+06	2.920058e+06	2.920058e+06	2.920058e+06	2.920058e+06	2920058.0	2.920058e+06
mean	2.508223e+03	1.326088e+03	1.329660e+03	1.008084e+01	1.489969e+03	1.466511e+03	4.236733e+00	0.0	2.142423e-03
std	1.746049e+03	4.854828e+02	4.992750e+02	4.912334e+01	5.109341e+02	5.318383e+02	5.112263e+01	0.0	4.623672e-02
min	1.000000e+00	1.000000e+00	1.000000e+00	-9.000000e+01	1.000000e+00	1.000000e+00	-9.600000e+01	0.0	0.000000e+00
25%	1.049000e+03	9.150000e+02	9.160000e+02	-6.000000e+00	1.107000e+03	1.053000e+03	-1.600000e+01	0.0	0.000000e+00
50%	2.149000e+03	1.318000e+03	1.323000e+03	-2.000000e+00	1.515000e+03	1.505000e+03	-7.000000e+00	0.0	0.000000e+00
75%	3.791000e+03	1.730000e+03	1.739000e+03	6.000000e+00	1.918000e+03	1.913000e+03	7.000000e+00	0.0	0.000000e+00
max	9.562000e+03	2.359000e+03	2.400000e+03	2.966000e+03	2.400000e+03	2.400000e+03	2.934000e+03	0.0	1.000000e+00

There are 2920058 rows and 27 columns after cleaning the data.

Limitations and Ethical Considerations

Limitations:

1. **Data Completeness:**
    - There are missing values for certain variables, which could affect the accuracy of the analysis. Handling missing data appropriately is crucial to ensure reliable results.
2. **Data Lag:**
    - The dataset is updated periodically, and the most recent data available might not include the latest months. This lag can impact the relevance of findings for current decision-making.
3. **Scope Limitation:**
    - The dataset focuses on domestic flights within the US, which means international flight delays and cancellations are not covered. This limits the scope of the analysis to domestic air travel.
4. **Seasonal and External Factors:**
    - While the dataset includes reasons for delays and cancellations, it might not capture all external factors (e.g., economic conditions, major events) that could influence flight performance.
5. **Data Granularity:**
    - The dataset provides detailed information at the flight level, which can be both a strength and a limitation. The high granularity may require significant computational resources for processing and analysis.

Ethical Considerations:

1. **Privacy and Confidentiality:**
    - The data does not include personally identifiable information (PII) about passengers, ensuring privacy and confidentiality. However, it's important to continue to handle the data responsibly to maintain ethical standards.
2. **Bias and Representation:**
    - The dataset is collected and reported by airlines to the DOT, which could introduce reporting biases. Ensuring that the analysis accounts for potential biases and provides a balanced view is essential.
3. **Transparency:**
    - The data source is transparent and publicly available, which promotes ethical use. Ensuring transparency in the analysis process and clearly

communicating methodologies and findings is crucial for maintaining integrity.

4. **Data Usage:**
    - The data should be used solely for the purpose of improving operational efficiency, enhancing customer satisfaction, and informing policy decisions. Any misuse of the data for purposes not aligned with these goals would be unethical.

Questions to Explore

1. What are the trends in flight delays and cancellations over the period from 2019 to 2023?
2. How do delays vary by airline and airport?
3. What are the primary reasons for flight delays and cancellations?
4. Are there seasonal patterns in flight delays and cancellations?
5. How do external factors such as weather or holidays affect flight performance?
6. What are the average delay times for different flight routes?
7. How do weather-related delays vary by region or airport?