Abstract:

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics provided the data that was taken into consideration. The performance of domestic flights run by the major airlines in terms of on-time arrivals is monitored by this branch. The monthly Air Travel Consumer Report from the DOT includes a summary of data pertaining to the number of flights that have been timely, delayed, cancelled, and diverted, as well as flight information and origin and destination airports.

Use case:

In civil aviation, flight delays are a significant issue. The main stake holders that get effected due to this flight delays are the passengers, airport authorities and the airline companies.

For the passengers:

* The passenger can avoid flights that get delayed or can schedule accordingly.
* There is chance to avoid booking flights in any particular airline company that has more flight delays.

For the airport authorities:

* Knowing about the delays the airport authorities can make arrangements for the passengers accordingly.
* Knowing if the delay is due to security delay or any other cause related to airport, they can improve their service and avoid such situations.

For the airline companies:

* Due to the delays the expenses increase on the airline company, they can avoid by taking measures for not delaying the flight
* Knowing the delay cause between any particular routes, they can change them priorly and avoid the delays.

Business Problem:

* Both direct and indirect costs are associated with flight delays, including expenses for gate maintenance, additional crew fees, food service, and accommodation. They also have an impact on passenger satisfaction.
* The operation of an airport will be greatly impacted by the late arrival of aircraft, with consequences including the relocation of parking gates, runways, ferries, and ground staff schedules.
* The analysis and forecast of flight delays are therefore very important to airlines, travellers, and airports. An airport is a hub for maintenance and travel where flights start and end.
* Inaccurate predictions of aircraft delays may cause losses to passengers and industries dependent on aviation, while delays will reduce the service capacity of the transportation network and cause delays in other airports.

Value of analysis done:

* Avoiding the flight delays, airline companies can increase their good will in the passengers as well as reduce the extra expenses they put and thus increase their profits.
* For the airport authorities avoiding the flight delays reduces the extra burden for them to maintain passengers in the airport by reducing the waiting period.

Data:

Data from each month of the year 2008 are included in this dataset. It has 29 columns and over 20 lakh records. This also includes details on the factors that contribute to flight delays of more than 15 minutes. The airline organisation is identified by its specific carrier code. Along with a timestamp, it provides the precise arrival and departure delay times. The CRS data is also taken into account.

Data Dictionary:

The attributes of the dataset are as follows:

**Year:**

* This is the year (2008)
* This is int datatype

**Month:**

* The month of flight movement
* Given in the form of 1-12 (Jan – Dec)
* This is integer datatype

**DayofMonth:**

* Day in month of flight movement
* Given in the form of 1-31 (Days of month)
* This is integer datatype

**Week:**

* Week in the day of month of flight movement
* Given in the form of 1-7 (Mon-Sun)
* This is integer datatype

**DepTime:**

* Actual departure time of the flight
* Given in the format of hhmm
* This is float datatype

**CRS Dep Time:**

* Scheduled departure time of the flight
* Given in the format of hhmm
* This is integer datatype

**ArrTime:**

* Actual Arrival time of the flight
* Given in the format of hhmm
* This is float datatype

**CRS ArrTime:**

* Scheduled arrival time of the flight
* Given in the format of hhmm
* This is integer datatype

**Unique Carrier Code:**

* 2-character IATA code of the airline company
* There are 20 unique carrier codes in the dataset
* This is string datatype

**FlightNum:**

* This is the Flight number
* This is integer datatype

**TailNum:**

* Flight registration number present at the tail of aircraft
* Unique to every aircraft
* This is string datatype

**ActualElapsedTime:**

* The actual time between origin to destination
* Given in the format of min
* This is float datatype

**CRS Elapsed Time:**

* The scheduled time between origin and destination
* Given in the format of min
* This is float datatype

**AirTime:**

* Time of flight in air
* Given in the format of min
* This is float datatype

**Arrdelay:**

* The time of delay for flight arrival
* Given in the format of min
* This is float datatype

**Depdelay:**

* The time of delay for flight departure
* Given in the format of min
* This is float datatype

**Origin:**

* The source airport code of the flight
* This is string datatype

**Dest:**

* The destination airport code of the flight
* This is string datatype

**Distance:**

* The distance between origin and destination
* Given in the format of miles
* This is integer datatype

**Taxi In:**

* The period of time between an aircraft arriving at terminal gate and actually landed in an airport.
* Given in the format of min
* This is float datatype

**Taxi Out:**

* The period of time between an aircraft leaving a terminal gate and actually taking off from an airport.
* Given in the format of min
* This is float datatype

**Cancelled:**

* This gives the status of flight cancellation
* Given in format like 0 (if flight not cancelled) and 1 (if flight is cancelled)
* This is integer datatype

**Cancellation code:**

* This gives the code for cause of cancellation
* Given in the format like A (for carrier delay)

B (for weather delay)

C (for NAS delay)

D (for security delay)

N (for flight not cancelled)

* This is string datatype

**Diverted:**

* This gives status of diverted flight
* Given in format like 1 (if flight is diverted) and 0 (if flight is not diverted)
* This is integer datatype

**CarrierDelay:**

* The delay caused due to maintenance issues for the aircraft
* Given in format of min
* This is float datatype

**WeatherDelay:**

* The delay caused due to extremely bad weather conditions
* Give in format of min
* This is float datatype

**NASDelay:**

* Expands to National Airspace System delay due to air traffic control, weather updates, etc
* Given in format of min
* This is float datatype

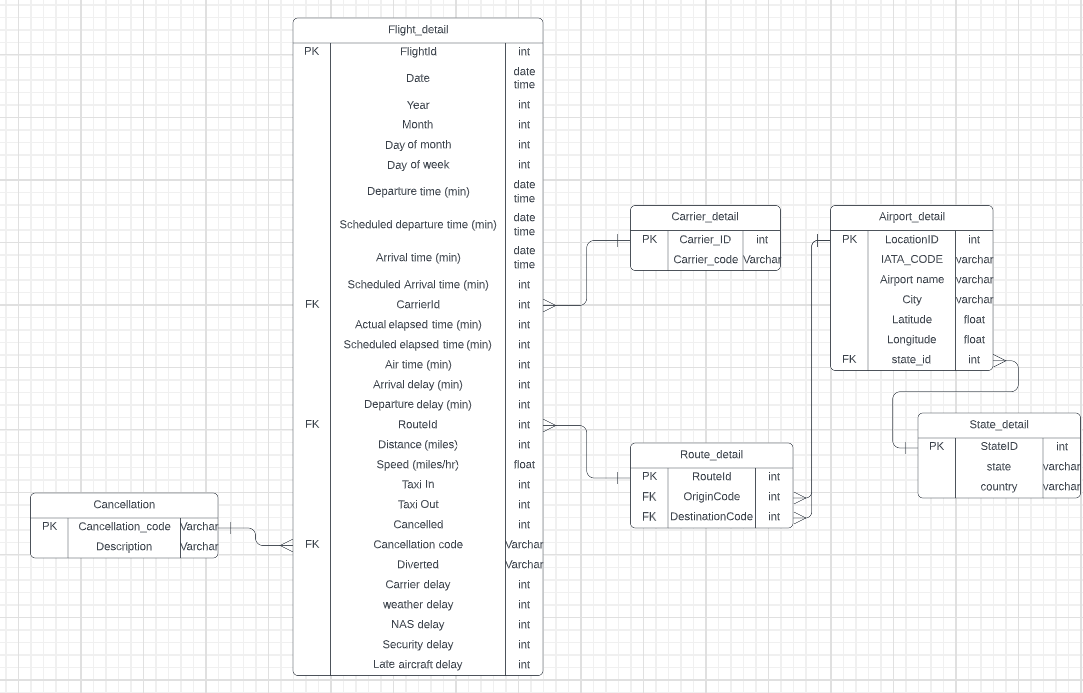
**SecurityDelay:**

* The delay caused due to security issues due to long queues, delay in checking, etc
* Given in format of min
* This is float datatype

**LateAircraftDelay:**

* The Arrival delay due to delay in previous stops.
* Given in format of min
* This is float datatype

**ER DIAGRAM**

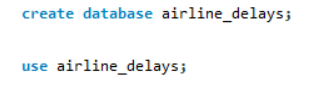
****

Questions that can be answered from data:

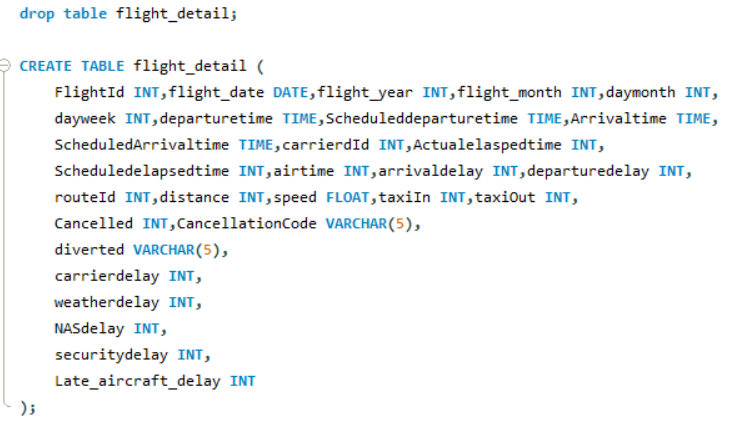
1. Which carrier flight gets more delayed?
2. Does origin and destination effect the delay?
3. Which delay cause occurs frequently?
4. At which part of the year delays are more?
5. Percentage of flights that are delayed compared to flights arrived on time.
6. Which delay cause is affecting for delay in each month?
7. Effect of distance between the airports for delay?
8. How accurate is the CRS estimation compared to actual time?
9. During which part of day delays are more?
10. Airtime vs distance
11. The wait period between arrival and departure delay?
12. Between which locations delay is more?
13. Descriptive statics for Taxi in and Taxi Out time
14. Effect of holidays in flight movements and delays
15. Any particular flight that gets delayed most frequently?
16. Which flight has maximum airtime?
17. Is delay more for more airtime or distance between locations?
18. What is the maximum delay period and when was it, location and cause?
19. Which delay cause is more on average for each carrier?
20. Which airport has maximum flight movement?

**HOW TO LOAD DATA INTO MYSQL**

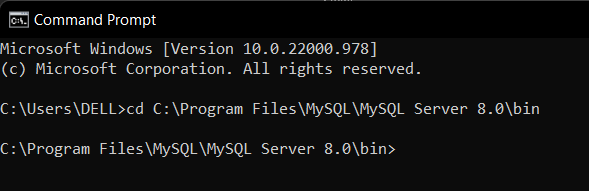
* Create a database in MySQL.



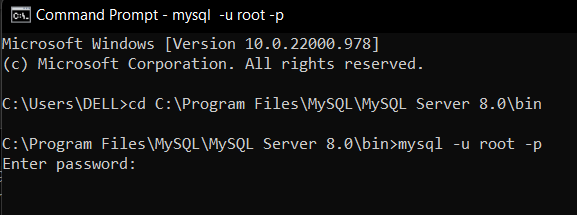
* Create a table in MySQL that has columns of the csv to be imported.
* The columns must be of the same datatype and in the same order as present in csv.



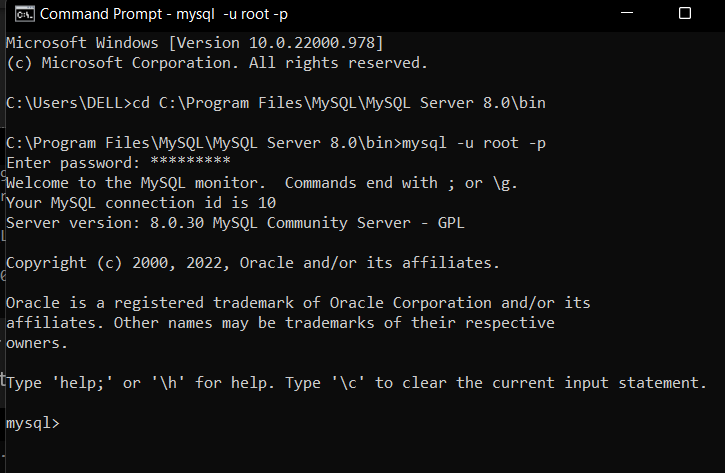
* Above, we use the drop command to avoid errors before creating the table.
* Next open command prompt by typing cmd in the windows search.
* Open the bin directory of the MySQL server.



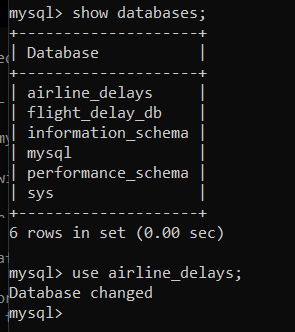
* Type the command “mysql -u root -p “
* It will prompt for the password enter it.



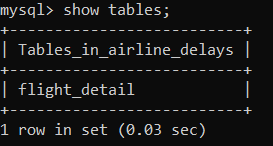
* Now MySQL server is started for us.



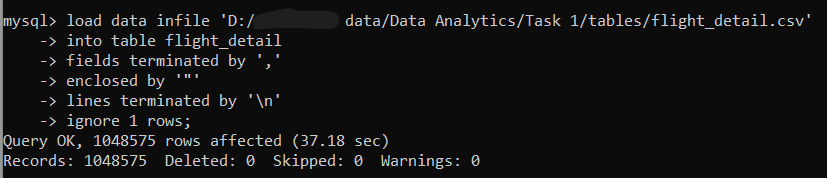
* See the databases available and use the required database.



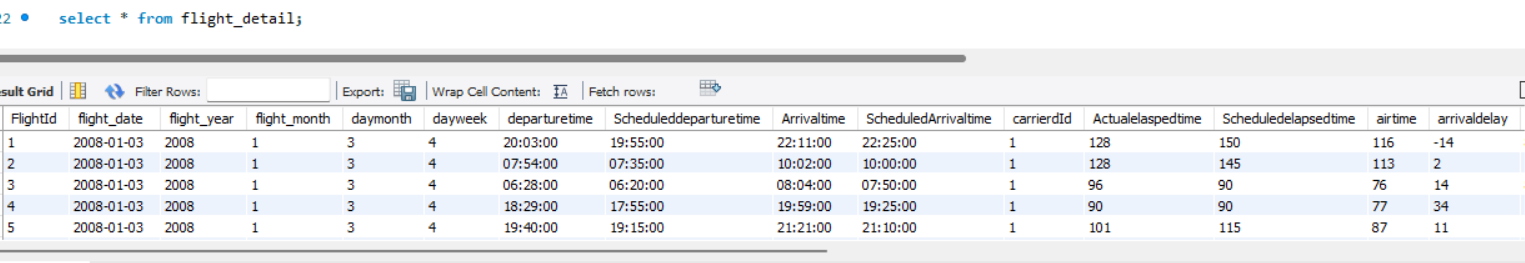
* Next check for the tables we created using show tables command



* Now type the following code to upload data from csv into the MySQL table.

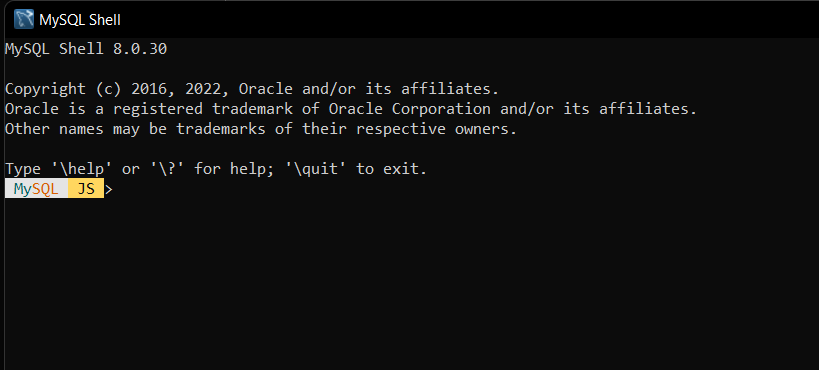


* **LOAD DATA INFILE** is to specify the csv file path that is to be imported.
* **INTO TABLE table\_name** to specify the destination table into which the data is to be imported.
* **FIELDS TERMINATED BY** to define the delimiter that separates the data in the file. In our case it is comma separated value file.
* **ENCLOSED BY** specifies that a double quote surrounds the values
* **LINES TERMINATED BY** specifies the code used as line break
* **IGNORE 1 ROWS** it tells MySQL to ignore the first row that usually contains the column names. As we already created the table with required column headings and data type.
* We can see file with almost 11 lakh records is imported with in 37sec, using this method to import data into MySQL is quick and easy.

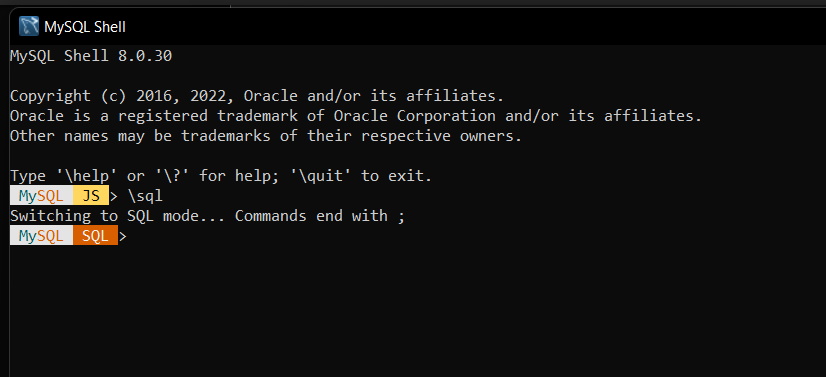


**HOW TO IMPORT SQL DATABASE INTO MYSQL**

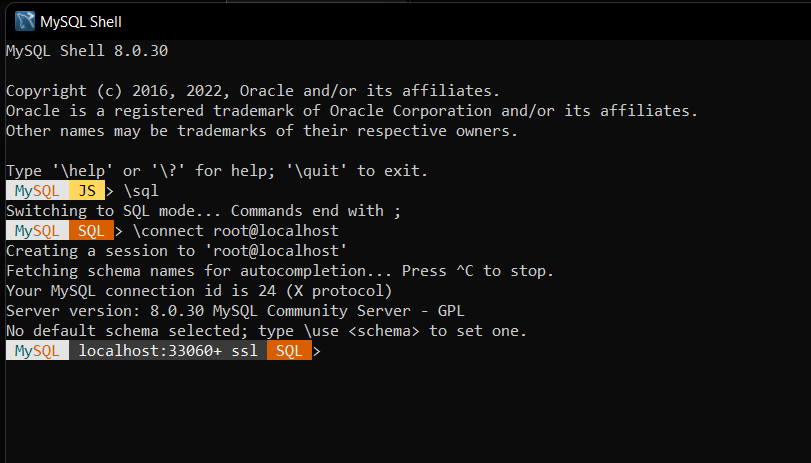
* In the windows search for MySQL shell and open it.



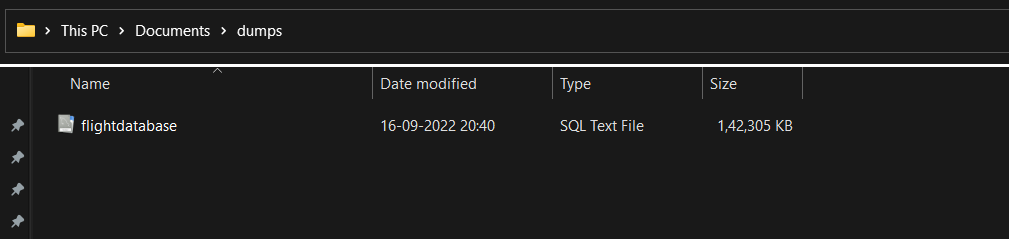
* Now type “ \sql “ for getting into sql mode in the shell



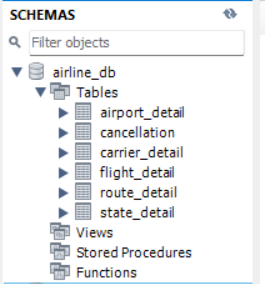
* Connecting to the server type “ \connect root@localhost” and enter. (It will prompt for password for first time type and hit enter, type y to save the password or n if not)



* Open the folder where the sql database file is located.



* Back in the shell type “ source (drag the SQL file into the shell window) “ it then gets the path of file and hit enter.
* The database starts uploading into the MySQL.
* We can see the database in the MySQL schemas



**SQL ASSIGNMENT QUESTIONS TO ANSWER**

The given data set has information regarding flight arrival and departure time given details about their airline company and as well as airport details of origin and destination.

1. Find out the airline company which has a greater number of flight movement.
2. Get the details of the first five flights that has high airtime.
3. Compute the maximum difference between the scheduled and actual arrival and departure time for the flights and categorize it by the airline companies.
4. Find the month in which the flight delays happened to be more.
5. Get the flight count for each state and identify the top 1.
6. A customer wants to book a flight under an emergency situation. Which airline would you suggest him to book. Justify your answer.
7. Find the dates in each month on which the flight delays are more.
8. Calculate the percentage of flights that are delayed compared to flights that arrived on time.
9. Identify the routes that has more delay time.
10. Find out on which day of week the flight delays happen more.
11. Identify at which part of day flights arrive late.
12. Compute the maximum, minimum and average TaxiIn and TaxiOut time.
13. Get the details of origin and destination with maximum flight movement.
14. Find out which delay cause occurrence is maximum.
15. Get details of flight whose speed is between 400 to 600 miles/hr for each airline company.
16. Identify the best time in a day to book a flight for a customer to reduce the delay.
17. Get the route details with airline company code ‘AQ’
18. Identify on which dates in a year flight movement is large.
19. Find out which delay cause is occurring more for each airline company.
20. Write a query that represent your unique observation in the database.

**COMPLETE DATA**

Cleaning up the dataset provided and converting it into sql normalised form ended up the data with the following columns.

Flight\_ID, date, year, month, day, week,

departure time, scheduled departure time, Arrival time, Scheduled arrival

time, carrier code, Actual elapsed time, scheduled elapsed time, Air time,

Arrival delay, Departure delay,

Origin, Origin airport name, Origin city, Origin latitude, Origin longitude,

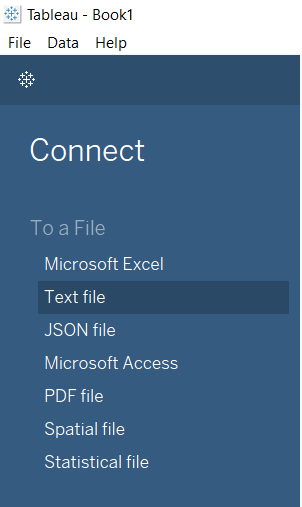
Destination, Destination airport name, Destination city, Destination latitude, Destination longitude,

Distance, Speed, taxi In, taxi Out, Cancelled, Cancellation code, Diverted,

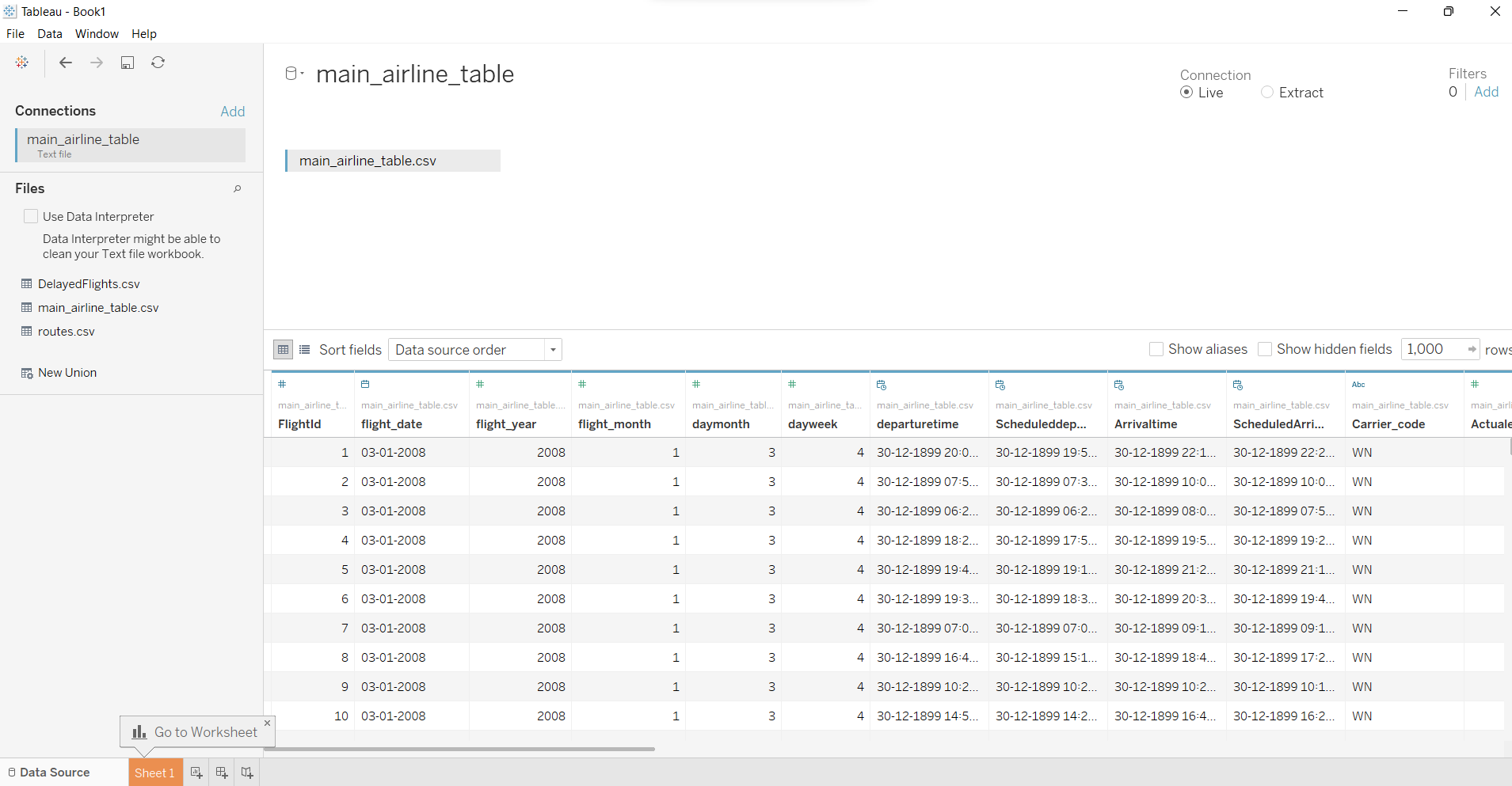
Carrier delay, Weather delay, NAS delay, Security delay, Late aircraft delay.

**IMPORT CSV INTO TABLEAU**

* To upload a csv file into tableau, open the tableau desktop app.
* Then from the side bar select text file under connect to a file.

****

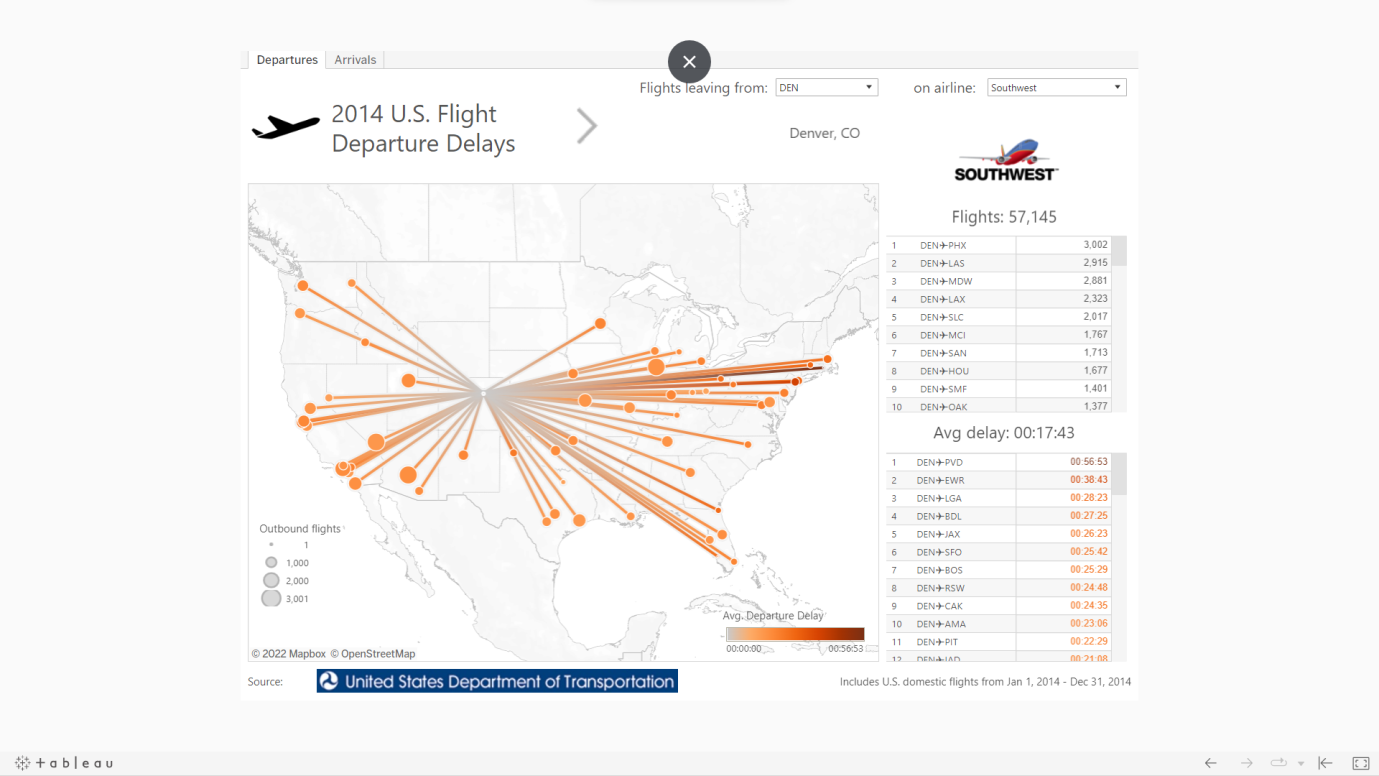
* From the pop-up window select the csv file and click open.
* The file will be uploaded into the data source and is visible as shown in the below picture.

****

**TABLEAU ASSIGNMENT**

* Analysing the different delay types for all the airline companies.
* Represent each delay cause for every airline.
* Get stats regarding percentage of each delay cause for entire data.
* Arrival and departure delay times with respect to month with delay cause as hue.
* Average of the delay times and delay cause times for each airline.
* Percentage of delayed flights with respect to origin and destination airports.
* Geographical map displaying the flight paths between origin to destination categorised by states.
* Getting dates on which, the delay average is maximum categorized by airlines.
* Analysis between the airtime, distance and speed for different airlines and routes.
* Stats regarding flight movement during weekends and weekdays.

**DASHBOARD IDEA**

****