Database Foundations for Business Analytics

***BUAN 6320***

**PROJECT 1**

# Group Members

1. Moksh Mehta- mxm220009

2. Urvi Shah- uxs220005

3. Vishwa Shukla- vks210004

4. Anjali Patel- axp220017

5. Lakshmi Priya Darshini S P- lxs220009

# Dataset

The on-time performance of domestic flights run by significant airlines is monitored by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics. This dataset of 2018 obtained from Kaggle contains the number of on-time, delayed, canceled, and diverted flights.

# Business Understanding

Airlines face high costs due to delays and cancellations, including expenses on compensation to stuck travelers and maintenance. Domestic flight delays put a $32.9 billion dent in the U.S. economy, and about half that cost is borne by airline passengers, according to a study led by UC Berkeley researchers. They also found that airlines with high delay rates also have higher operating costs overall, and the inefficiency adversely affects the U.S. economy.

Airport delays are a significant problem for airlines and passengers alike. In order to reduce delays, airlines and airports need to better understand the causes of delays and use data analytics to improve their operations.

The first step is to collect data on delays. This data can come from a variety of sources, including flight tracking websites, airport management systems, and even social media. Once this data is collected, it can be analyzed to identify patterns and trends.

There are several ways in which data can be used to reduce delays. For example, data can be used to improve flight planning and scheduling, identify potential problems with airport infrastructure, and even help predict future delays. By using data, airlines and airports can make more informed decisions that can help reduce delays. Additionally, the data can be used to identify which airlines are consistently performing well and which ones are not, which can be helpful for consumers when choosing an airline.

The questions that we are trying to answer by studying this dataset for the year 2018 are:

* What airline gets the most delayed?
* What airline has the best on time performance?
* Which airport has the highest on time arrivals?
* Which state has the highest incoming flights?
* Which months have the highest cancellations?
* Which airline has the maximum number of delays?
* Which airline has the maximum number of cancellations?
* Which is the busiest route?
* Which airline provides the maximum number of flights per month?

# Data Understanding

The overall size of the dataset is 800 mb. The data has approximately 7,000,000 rows, which can be identified uniquely by flight\_id. The original dataset had 18 columns. We removed country column from the original dataset because it was not related to our business understanding. Based on this relationship structure we could see that there are functional dependencies between these columns, therefore we broke the larger dataset into 5 tables to minimize the functional dependency and to bring it into 4th normal form (i.e., BCNF). We have used iata(which is starting 3 letter acronym airport code) as primary key in airport table. Using iata we can access all other columns like city, state, airport name, longitude, latitude from airport table. We have used flight\_id as foreign key in arrival and departure table which we have separated after normalization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Original Column Name** | **Modified Column Name** | **SQL Data Type** | **Description** | **Missing Values(Y/N)** |
| FL\_DATE | FL\_DATE | date (yy/mm/dd) | Date of departure of flight | N |
| OP\_CARRIER | AIR\_ID | varchar (45) | Two letter unique code to identify the airline | N |
| OP\_CARRIER\_FL\_NUM | FL\_NUM | int | Flight number | N |
| ORIGIN | ORIGIN | varchar (45) | Starting 3 Letter Acronym Airport Code | N |
| DEST | DEST | varchar (45) | Destination 3 Letter Acronym Airport Code | N |
| CRS\_DEP\_TIME | PL\_DEP\_TIME | time (hh:mm: ss) | Planned Departure Flight | N |
| DEP\_TIME | DEP\_TIME | time (hh:mm: ss) | Actual Departure Time | Y |
| DEP\_DELAY | DEP\_DELAY | time (mm: ss) | Total Delay on Departure in minutes | Y |
| CRS\_ARR\_TIME | PL\_ARR\_TIME | time (hh:mm: ss) | Planned Arrival Time | N |
| ARR\_TIME | ARR\_TIME | time (hh:mm: ss) | Actual Arrival Time | Y |
| ARR\_DELAY | ARR\_DELAY | time (mm: ss) | Total Delay on Arrival in minutes | Y |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Original Column Name** | **Modified Column Name** | **SQL Data Type** | **Description** | **Missing Values(Y/N)** |
| CANCELLED | CANCELLED | int | Flight Cancelled | N |
| AIR\_TIME | AIR\_TIME | time (mm: ss) | The time duration in air between arrival and departure | Y |
| DISTANCE | DISTANCE | int | Distance between two airports | N |
| AIRPORT | AIRPORT | varchar (255) | Airport full names derived from its identifier | N |
| CITY | CITY | varchar (50) | Airport situated in which US city | N |
| STATE | STATE | varchar (5) | Airport situated in which US state | N |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column Name** | **Mean** | **Min** | **Max** | **Range** | **Std Dev** |
| FL\_DATE | - | - | - | - |  |
| AIR\_ID | - | - | - |  |  |
| FL\_NUM | 2610 | 1 | 7909 | - | 1860 |
| ORIGIN | - | - | - | - |  |
| DEST | - | - | - | - |  |
| PL\_DEP\_TIME | 1200 | 0001 | 2400 | - | 491 |
| DEP\_TIME | 1200 | 0001 | 2400 | - | 505 |
| DEP\_DELAY | 9.97 | -122 | 2710 | - | 44.8 |
| PL\_ARR\_TIME | 1200 | 0001 | 2400 | - | 518 |
| ARR\_TIME | 1200 | 0001 | 2400 | - | 538 |
| ARR\_DELAY | 5.05 | -120 | 2690 | - | 49.6 |
| CANCELLED | - | 0 | 1 | - |  |
| AIR\_TIME | 112 | 7 | 696 | - | 71.1 |
| DISTANCE | 800 | 31 | 4980 |  | 598 |
| AIRPORT | - | - | - | - |  |
| CITY | - | - | - | - |  |
| STATE | - | - | - | - |  |

# Design a Database

We have constructed total five tables, including

1. Flight
2. Airports
3. Airlines
4. Departure
5. Arrival

Here below we are attaching a E-R diagram of five tables,

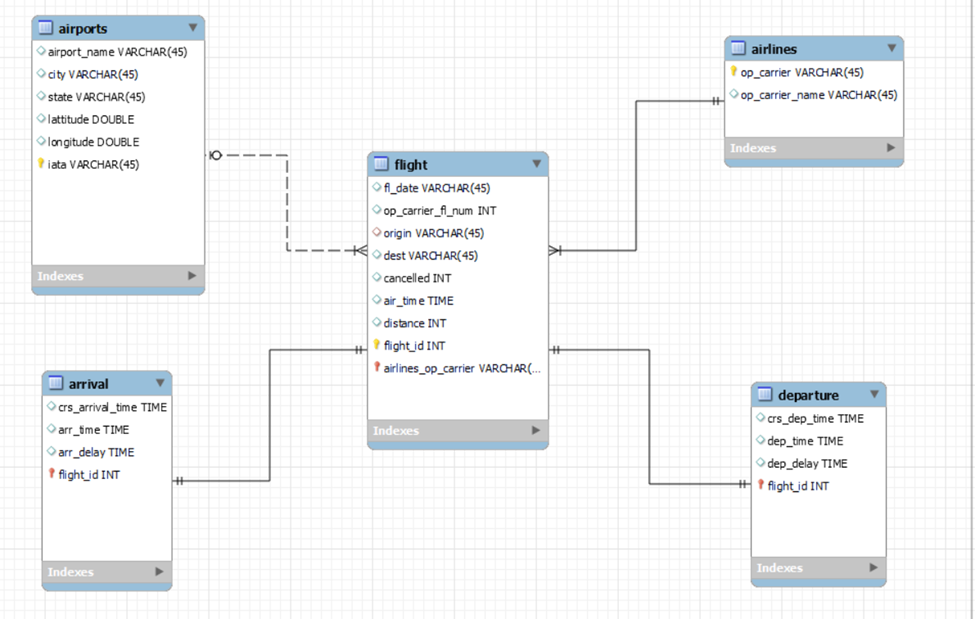
Graphical user interface, table, Excel

Description automatically generated

Next, we have performed checks for identifying whether our schema is in BCNF (Boyce-Codd Normal Form)

|  |  |  |
| --- | --- | --- |
| COLUMN DETAILS | CHECK FOR BCNF | FUNCTIONAL DEPENDENCY |
| {FLIGHT\_ID}->{ORIGIN} | FLIGHT\_ID AND ORIGIN ARE IN TABLE 1, FLIGHT\_ID IS KEY | BASED ON INITIAL DECOMPOSITION |
| {FLIGHT\_ID}->{ARR\_DELAY} | FLIGHT\_ID AND ARR\_DELAY ARE IN TABLE 3, FLIGHT\_ID IS KEY | BASED ON INITIAL DECOMPOSITION |
| {FLIGHT\_ID}->{DEP\_DELAY} | FLIGHT\_ID AND DEP\_DELAY ARE IN TABLE 4, FLIGHT\_ID IS KEY | BASED ON INITIAL DECOMPOSITION |
| {FLIGHT\_ID}->{OP\_CARRIER} | FLIGHT\_ID AND OP\_CARRIER ARE IN TABLE 5, FLIGHT\_ID IS KEY | BASED ON INITIAL DECOMPOSITION |
| {ORIGIN}->{CITY} | ORIGIN AND CITY ARE IN TABLE 2, ORIGIN IS KEY | INFERRED |
| {ORIGIN}->{STATE} | ORIGIN AND STATE ARE IN TABLE 2, ORIGIN IS KEY | INFERRED |
| {ORIGIN}->{AIRPORT} | ORIGIN AND AIRPORT ARE IN TABLE 2, ORIGIN IS KEY | INFERRED |
| {ORIGIN}->{LATTITUDE} | ORIGIN AND LATTITUDE ARE IN TABLE 2, ORIGIN IS KEY | INFERRED |
| {ORIGIN}->{LONGITUDE} | ORIGIN AND LONGITUDE ARE IN TABLE 2, ORIGIN IS KEY | INFERRED |
| {FLIGHT\_ID}->{CRS\_ARR\_TIME} | FLIGHT\_ID AND CRS\_ARR\_TIME ARE IN TABLE 3, FLIGHT\_ID IS KEY | INFERRED |
| {FLIGHT\_ID}->{CRS\_DEP\_TIME} | FLIGHT\_ID AND CRS\_DEP\_TIME ARE IN TABLE 3, FLIGHT\_ID IS KEY | INFERRED |
| {FLIGHT\_ID}->{ARR\_TIME} | FLIGHT\_ID AND ARR\_TIME ARE IN TABLE 4, FLIGHT\_ID IS KEY | INFERRED |
| {FLIGHT\_ID}->{DEP\_TIME} | FLIGHT\_ID AND DEP\_TIME ARE IN TABLE 4, FLIGHT\_ID IS KEY | INFERRED |
| {FLIGHT\_ID}->{OP\_CARRIER\_FL\_NUM} | FLIGHT\_ID AND OP\_CARRIER\_FL\_NUM ARE IN TABLE 1, FLIGHT\_ID IS KEY | INFERRED |
| {FLIGHT\_ID}->{OP\_CARRIER} | FLIGHT\_ID AND OP\_CARRIER ARE IN TABLE 5, FLIGHT\_ID IS KEY | INFERRED |

This is our E-R diagram after normalization,



We have loaded our data into MySQL server using MySQL program.

# Database cleaning and database testing

The dataset cleaning is performed by removing inconsistencies like a columns name sometimes entered in uppercase or lowercase. The data variables need to be in acceptable format by MySQL. Also, we have removed numerical data with comma, as MySQL truncates data at comma leading to incorrect data. We have loaded our dataset with the help of queries shown in the below picture.

Graphical user interface, text, application, email

Description automatically generated

Here are some pictures of our dataset before and after data cleaning using queries.

Below, attached picture displayed crs\_arr\_time and arr\_time in INT datatype from arrival table before cleansing.

Graphical user interface

Description automatically generated with low confidence

Using following query, we have performed cleaning of data variables from arrival table to convert INT datatype for crs\_arr\_time and arr\_time into TIME datatype.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

We have performed following query to check whether datatype is correct or not for time variables (i.e, crs\_arr\_time, arr\_time).

Table

Description automatically generated

We have also removed double quotes from dataset as a part of data cleaning using below attached queries. We are attaching pictures to show output before and after execution of MySQL query.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated

After checking for all the constraints, we ran queries to answer the business understanding questions:

1. Which airport faced maximum cancelled flights?

Graphical user interface, text, application

Description automatically generated

As per our result, we can say that airport Yakutat in state Arkansas has faced maximum cancelled flights followed by Northwest Arkansas regional.

1. Which airline has faced maximum number of cancellations?

Graphical user interface, text, application, email

Description automatically generated

We ran this query to know which airline has faced maximum number of cancellations, and we get Air U.S as our answer.

1. What are the number of flights per month?

According to our query, we can be able to know that January has 570118 number of flights. We can know number of flights for January, February, March, April, May, June, July, August, September, and October as per our dataset.

Graphical user interface, text, application, email

Description automatically generated

1. Which airport has maximum number of flights?

Graphical user interface

Description automatically generated with medium confidence

Airport having airport\_id ATL has maximum number of flights i.e., 780125 and hence busiest airport in United states of America.

1. Which airport has minimum traffic in USA?

Graphical user interface, text, application

Description automatically generated

As per result of query we ran, Aberdeen regional airport has minimum number of flights.