Database Foundations for Business Analytics

# 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.’

u

# Data Understanding

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 betwwen 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 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 |
| {FLIGHT\_ID}->{AIRLINE\_NAME} | FLIGHT\_ID AND AIRLINE\_NAME ARE IN TABLE 5, FLIGHT\_ID IS KEY | INFERRED |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column Name** | **Mean** | **Min** | **Max** | **Range** | **Std Dev** | **Unique No of observations** |
| FL\_DATE | - | - | - | - |  | 1048575 |
| AIR\_ID | - | - | - |  |  |  |
| FL\_NUM | 2610 | 1 | 7909 | - | 1860 | 1879 |
| ORIGIN | - | - | - | - |  |  |
| DEST | - | - | - | - |  |  |
| PL\_DEP\_TIME | 1200 | 0001 | 2400 | - | 491 | 440 |
| DEP\_TIME | 1200 | 0001 | 2400 | - | 505 | 440 |
| DEP\_DELAY | 9.97 | -122 | 2710 | - | 44.8 |  |
| PL\_ARR\_TIME | 1200 | 0001 | 2400 | - | 518 | 56 |
| ARR\_TIME | 1200 | 0001 | 2400 | - | 538 |  |
| ARR\_DELAY | 5.05 | -120 | 2690 | - | 49.6 | 196 |
| CANCELLED | - | 0 | 1 | - |  |  |
| AIR\_TIME | 112 | 7 | 696 | - | 71.1 | 3624 |
| DISTANCE | 800 | 31 | 4980 |  | 598 |  |
| AIRPORT | - | - | - | - |  |  |
| CITY | - | - | - | - |  |  |
| STATE | - | - | - | - |  |  |
| COUNTRY | - | - | - | - |  |  |