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. 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?

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 the 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

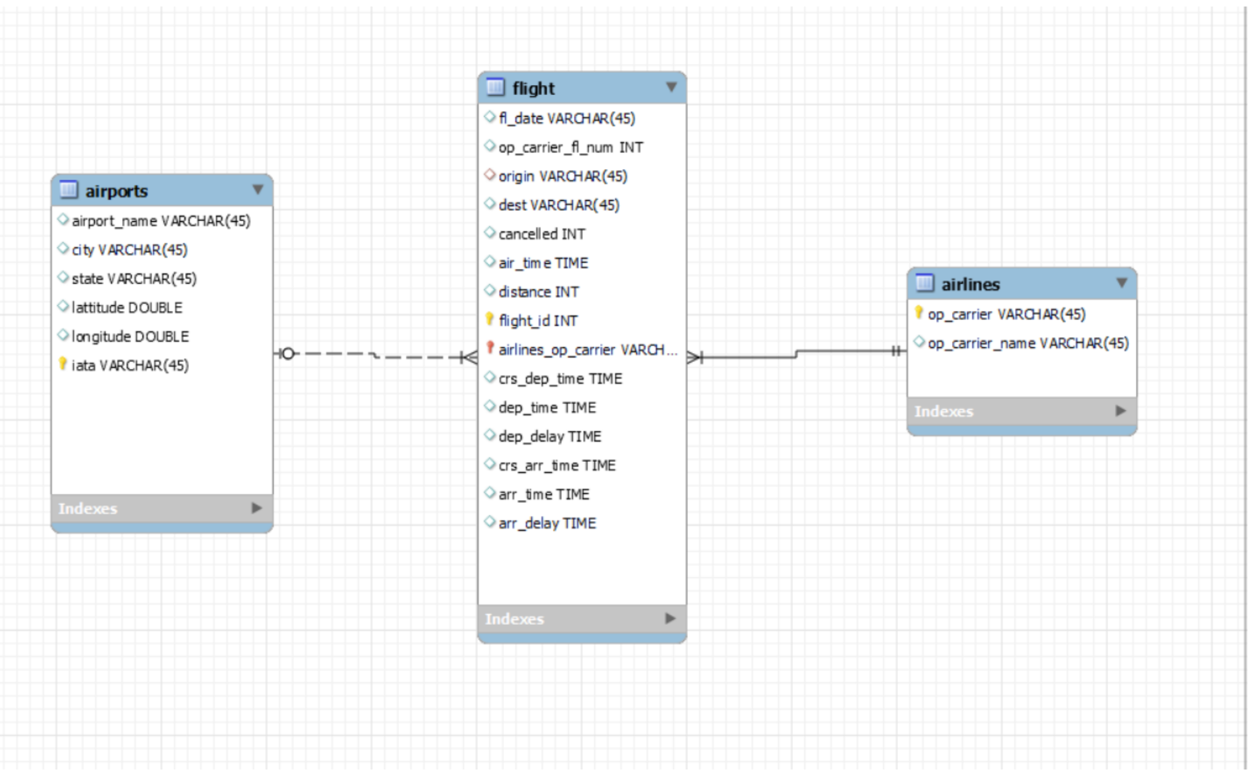
Column Name	Mean	Min	Max	Range	Std Dev
FL_DATE	-	-	-	-	
AIR_ID	-	-	-		
FL_NUM	2610	1	7909	-	1860
ORIGIN	-	-	-	-	
DEST	-	-	-	-	
PL_DEP_TIME	1200	0001	2400	-	491
DEP_TIME	1200	0001	2400	-	505
DEP_DELAY	9.97	-122	2710	-	44.8
PL_ARR_TIME	1200	0001	2400	-	518
ARR_TIME	1200	0001	2400	-	538
ARR_DELAY	5.05	-120	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 3 tables before BCNF

1. Flight
2. Airports
3. Airlines

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



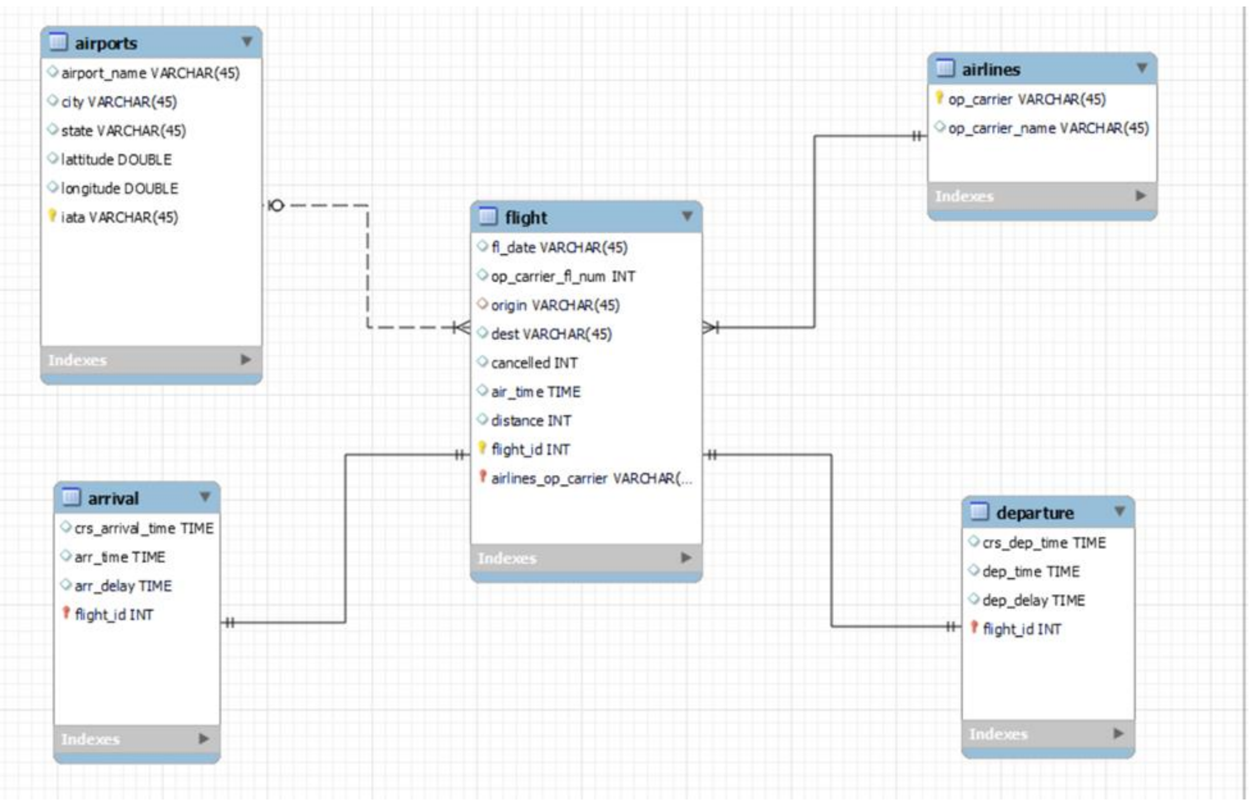
Next, we have performed checks for identifying whether our schema is in BCNF (Boyce-Codd Normal Form)

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

We have constructed total 3 tables before BCNF

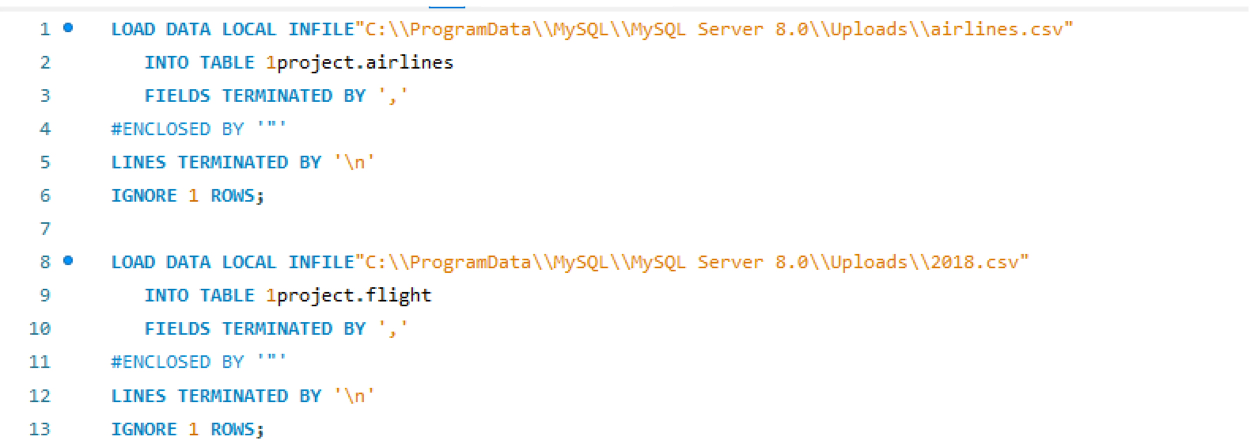
1. Flight
2. Airports
3. Airlines
4. Departure
5. Arrival

This is our E-R diagram after normalization,



Data loading

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



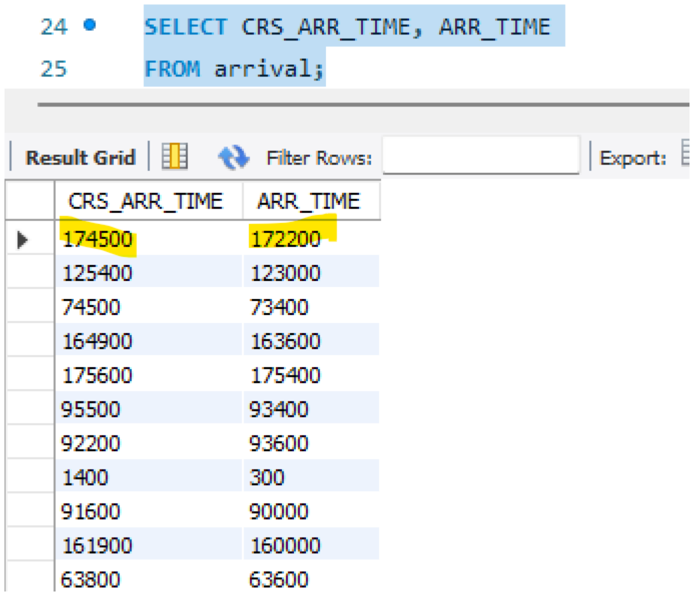
```
LOAD DATA LOCAL INFILE"C:\\ProgramData\\MySQL\\MySQL Server 8.0\\Uploads\\airlines.csv"
   INTO TABLE 1project.airlines
   FIELDS TERMINATED BY ','
#ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;

LOAD DATA LOCAL INFILE"C:\\ProgramData\\MySQL\\MySQL Server 8.0\\Uploads\\2018.csv"
   INTO TABLE 1project.flight
   FIELDS TERMINATED BY ','
#ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

Database cleaning

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.

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

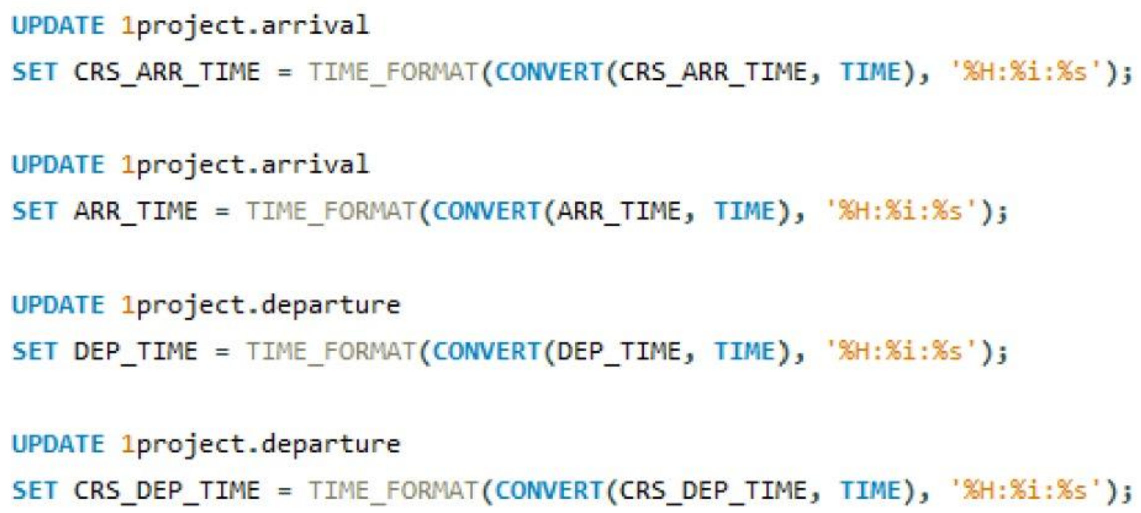
Below, attached picture displayed crs_arr_time and arr_time in INT datatype from arrival table before cleansing.



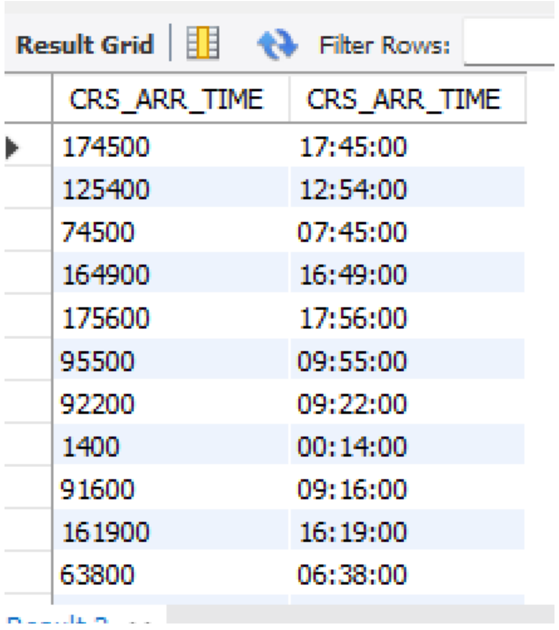
```
SELECT CRS_ARR_TIME, ARR_TIME
FROM arrival;
```

CRS_ARR_TIME	ARR_TIME
174500	172200
125400	123000
74500	73400
164900	163600
175600	175400
95500	93400
92200	93600
1400	300
91600	90000
161900	160000
63800	63600

Using the 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.

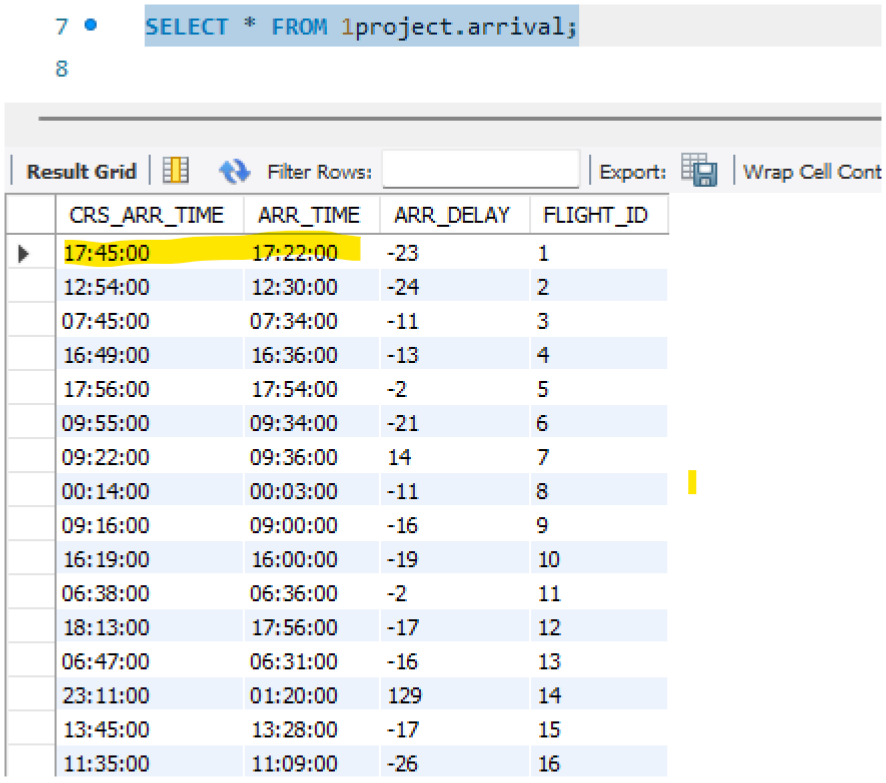


```sql
UPDATE 1project.arrival
SET CRS_ARR_TIME = TIME_FORMAT(CONVERT(CRS_ARR_TIME, TIME), '%H:%i:%s');

UPDATE 1project.arrival
SET ARR_TIME = TIME_FORMAT(CONVERT(ARR_TIME, TIME), '%H:%i:%s');

UPDATE 1project.departure
SET DEP_TIME = TIME_FORMAT(CONVERT(DEP_TIME, TIME), '%H:%i:%s');

UPDATE 1project.departure
SET CRS_DEP_TIME = TIME_FORMAT(CONVERT(CRS_DEP_TIME, TIME), '%H:%i:%s');
```



CRS_ARR_TIME	CRS_ARR_TIME
174500	17:45:00
125400	12:54:00
74500	07:45:00
164900	16:49:00
175600	17:56:00
95500	09:55:00
92200	09:22:00
1400	00:14:00
91600	09:16:00
161900	16:19:00
63800	06:38:00

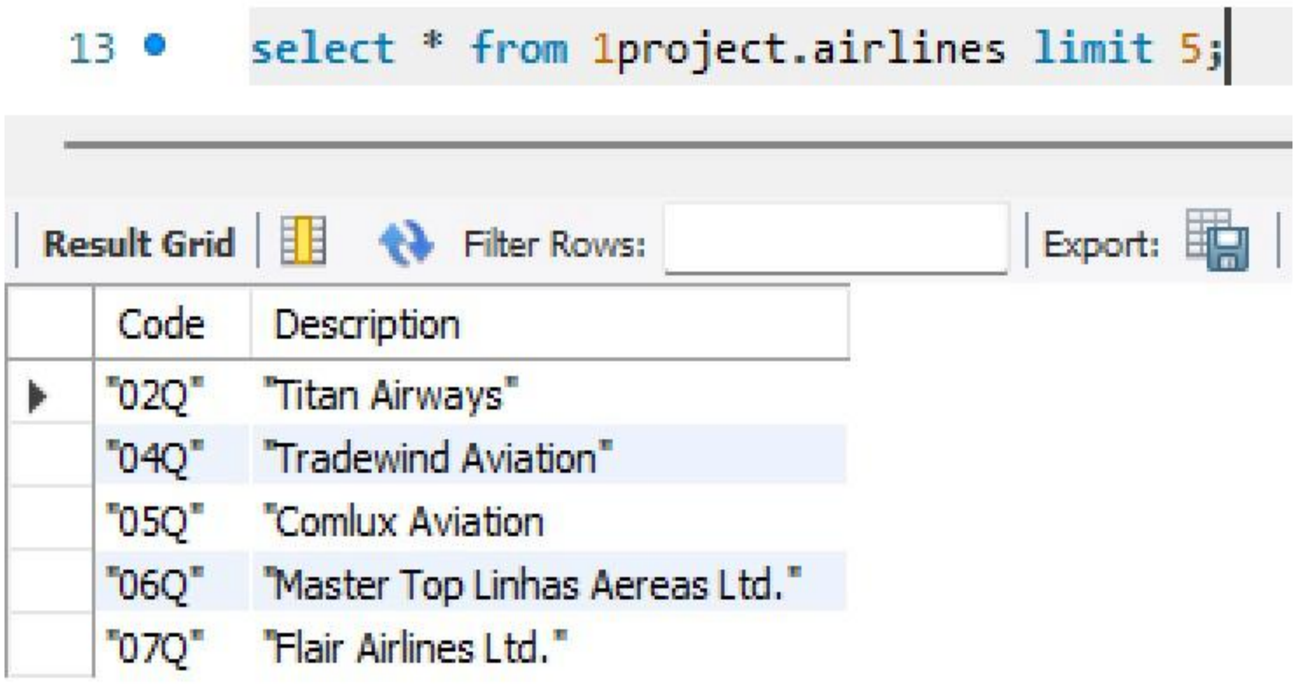
We have performed the following query to check whether datatype is correct or not for time variables (i.e, crs_arr_time, arr_time).



```sql
SELECT * FROM 1project.arrival;
```

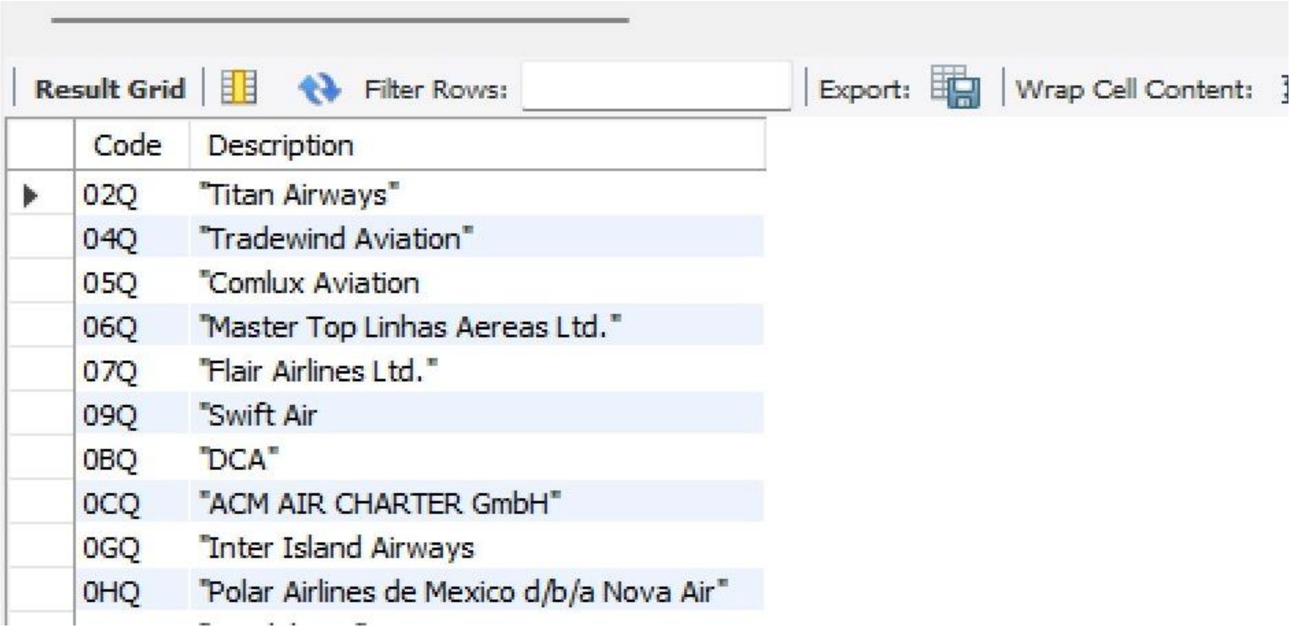
CRS_ARR_TIME	ARR_TIME	ARR_DELAY	FLIGHT_ID
17:45:00	17:22:00	-23	1
12:54:00	12:30:00	-24	2
07:45:00	07:34:00	-11	3
16:49:00	16:36:00	-13	4
17:56:00	17:54:00	-2	5
09:55:00	09:34:00	-21	6
09:22:00	09:36:00	14	7
00:14:00	00:03:00	-11	8
09:16:00	09:00:00	-16	9
16:19:00	16:00:00	-19	10
06:38:00	06:36:00	-2	11
18:13:00	17:56:00	-17	12
06:47:00	06:31:00	-16	13
23:11:00	01:20:00	129	14
13:45:00	13:28:00	-17	15
11:35:00	11:09:00	-26	16

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



```
select * from 1project.airlines limit 5;
```

Code	Description
"02Q"	"Titan Airways"
"04Q"	"Tradewind Aviation"
"05Q"	"Comlux Aviation
"06Q"	"Master Top Linhas Aereas Ltd."
"07Q"	"Flair Airlines Ltd."

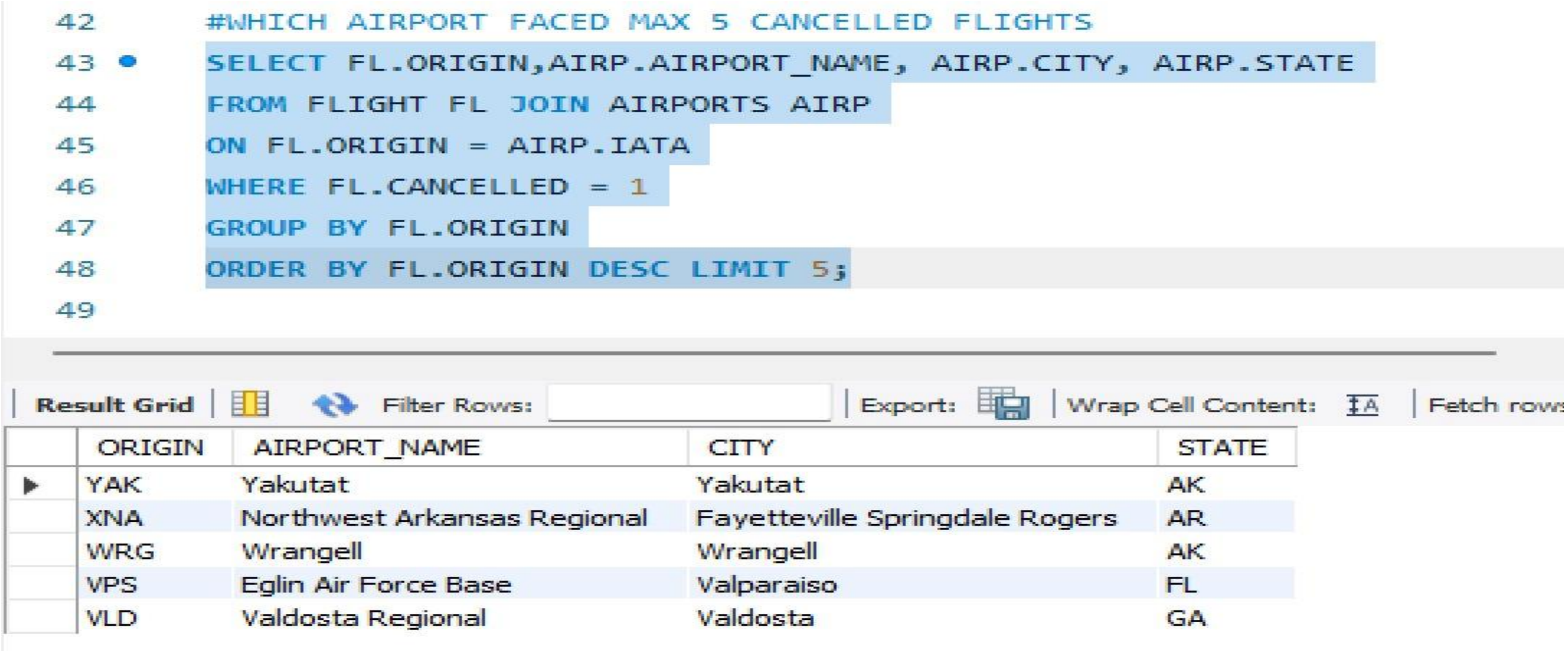


Code	Description
02Q	"Titan Airways"
04Q	"Tradewind Aviation"
05Q	"Comlux Aviation
06Q	"Master Top Linhas Aereas Ltd."
07Q	"Flair Airlines Ltd."
09Q	"Swift Air
0BQ	"DCA"
0CQ	"ACM AIR CHARTER GmbH"
0GQ	"Inter Island Airways
0HQ	"Polar Airlines de Mexico d/b/a Nova Air"

Database testing

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

1. Which airport faced top 5 maximum cancelled flights?

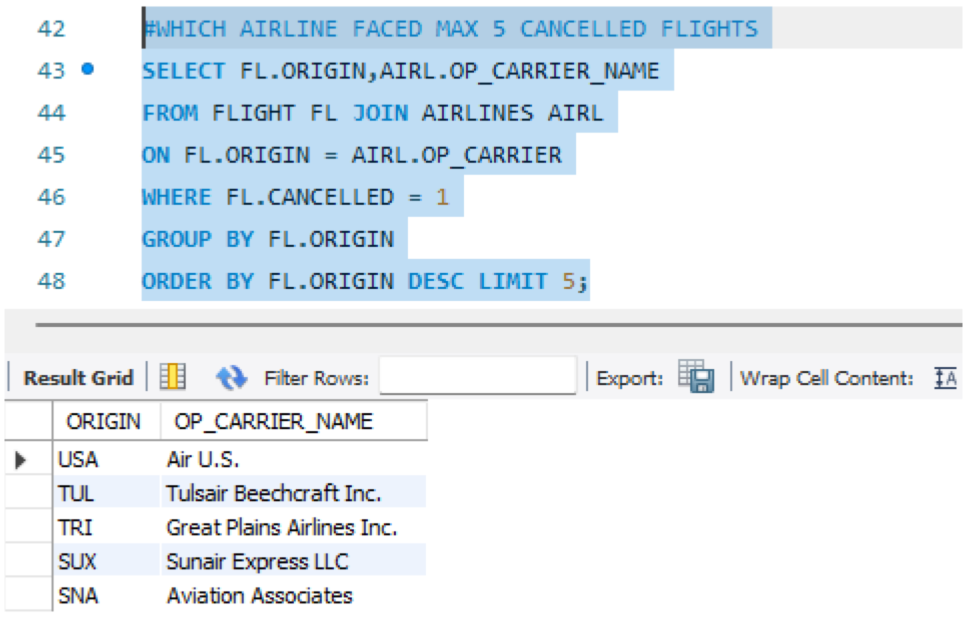


```
#WHICH AIRPORT FACED MAX 5 CANCELLED FLIGHTS
SELECT FL.ORIGIN,AIRP.AIRPORT_NAME, AIRP.CITY, AIRP.STATE
FROM FLIGHT FL JOIN AIRPORTS AIRP
ON FL.ORIGIN = AIRP.IATA
WHERE FL.CANCELLED = 1
GROUP BY FL.ORIGIN
ORDER BY FL.ORIGIN DESC LIMIT 5;
```

ORIGIN	AIRPORT_NAME	CITY	STATE
YAK	Yakutat	Yakutat	AK
XNA	Northwest Arkansas Regional	Fayetteville Springdale Rogers	AR
WRG	Wrangell	Wrangell	AK
VPS	Eglin Air Force Base	Valparaiso	FL
VLD	Valdosta Regional	Valdosta	GA

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

2. Which airline has faced the maximum number of cancellations?



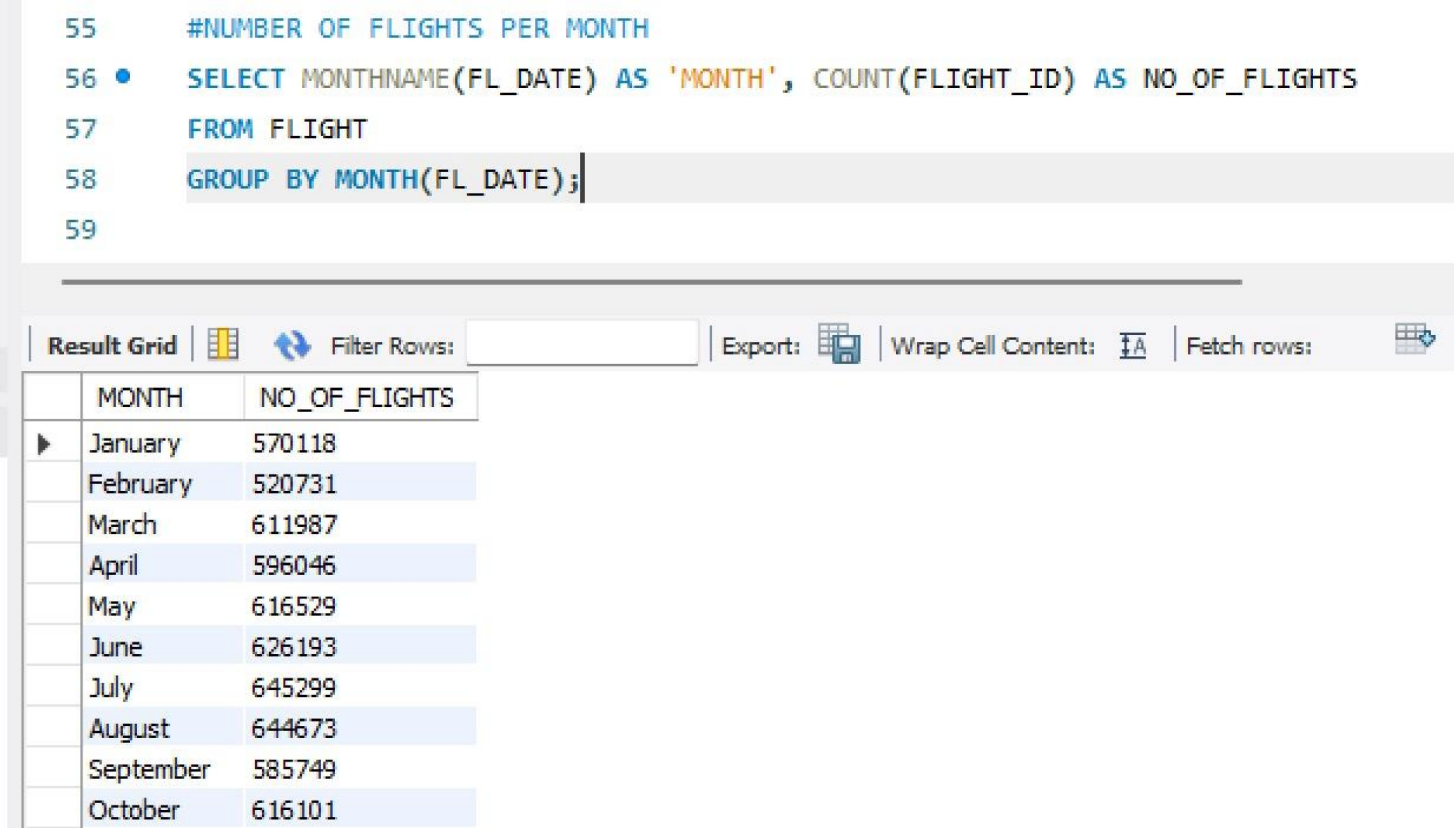
```
#WHICH AIRLINE FACED MAX 5 CANCELLED FLIGHTS
SELECT FL.ORIGIN,AIRL.OP_CARRIER_NAME
FROM FLIGHT FL JOIN AIRLINES AIRL
ON FL.ORIGIN = AIRL.OP_CARRIER
WHERE FL.CANCELLED = 1
GROUP BY FL.ORIGIN
ORDER BY FL.ORIGIN DESC LIMIT 5;
```

ORIGIN	OP_CARRIER_NAME
USA	Air U.S.
TUL	Tulsair Beechcraft Inc.
TRI	Great Plains Airlines Inc.
SUX	Sunair Express LLC
SNA	Aviation Associates

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

3. What is the number of flights per month?

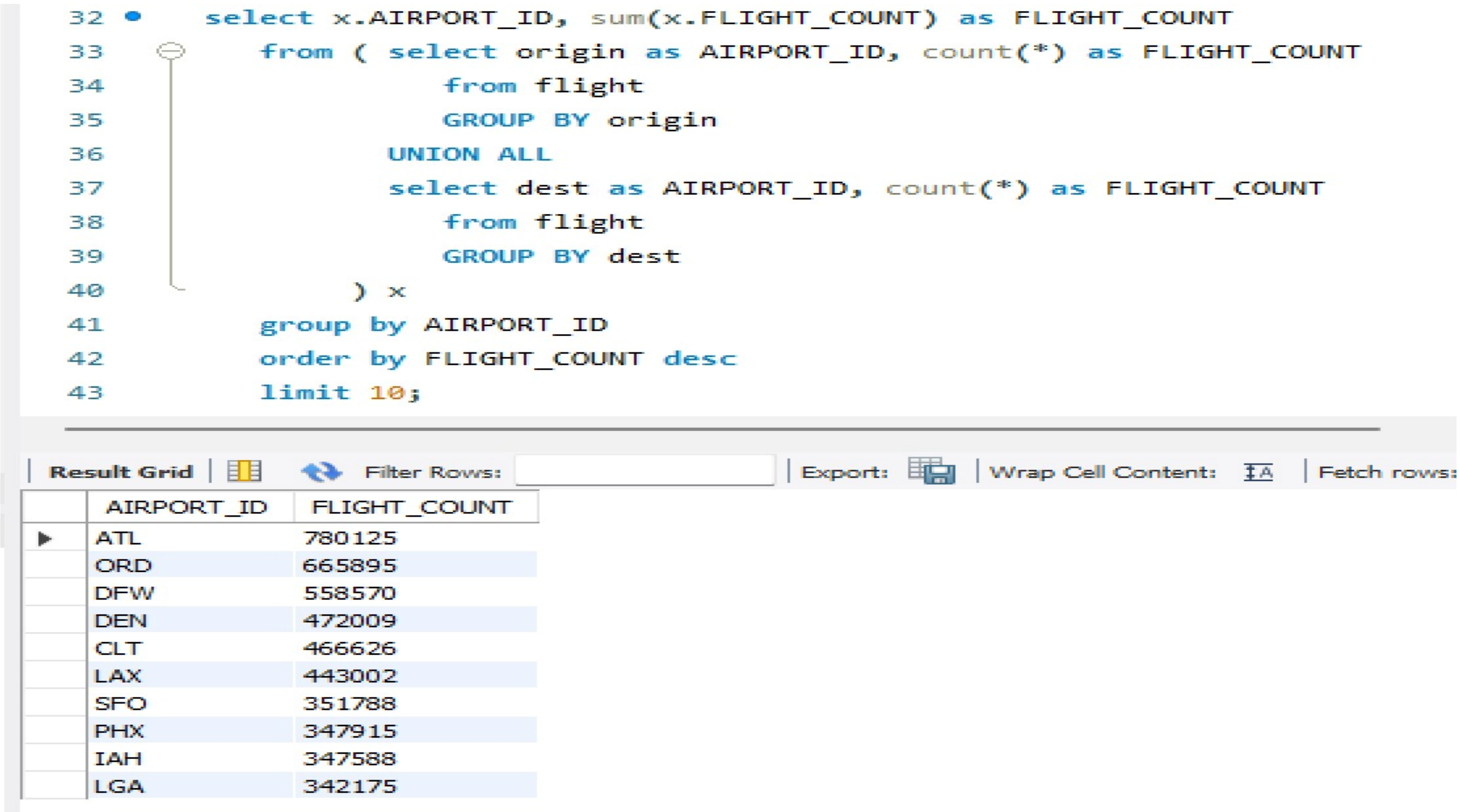
According to our query, we can see that January has 570118 number of flights. We can know the number of flights for January, February, March, April, May, June, July, August, September, and October as per our dataset.



```
#NUMBER OF FLIGHTS PER MONTH
SELECT MONTHNAME(FL_DATE) AS 'MONTH', COUNT(FLIGHT_ID) AS NO_OF_FLIGHTS
FROM FLIGHT
GROUP BY MONTH(FL_DATE);
```

MONTH	NO_OF_FLIGHTS
January	570118
February	520731
March	611987
April	596046
May	616529
June	626193
July	645299
August	644673
September	585749
October	616101

4. Which airport has the maximum number of flights?

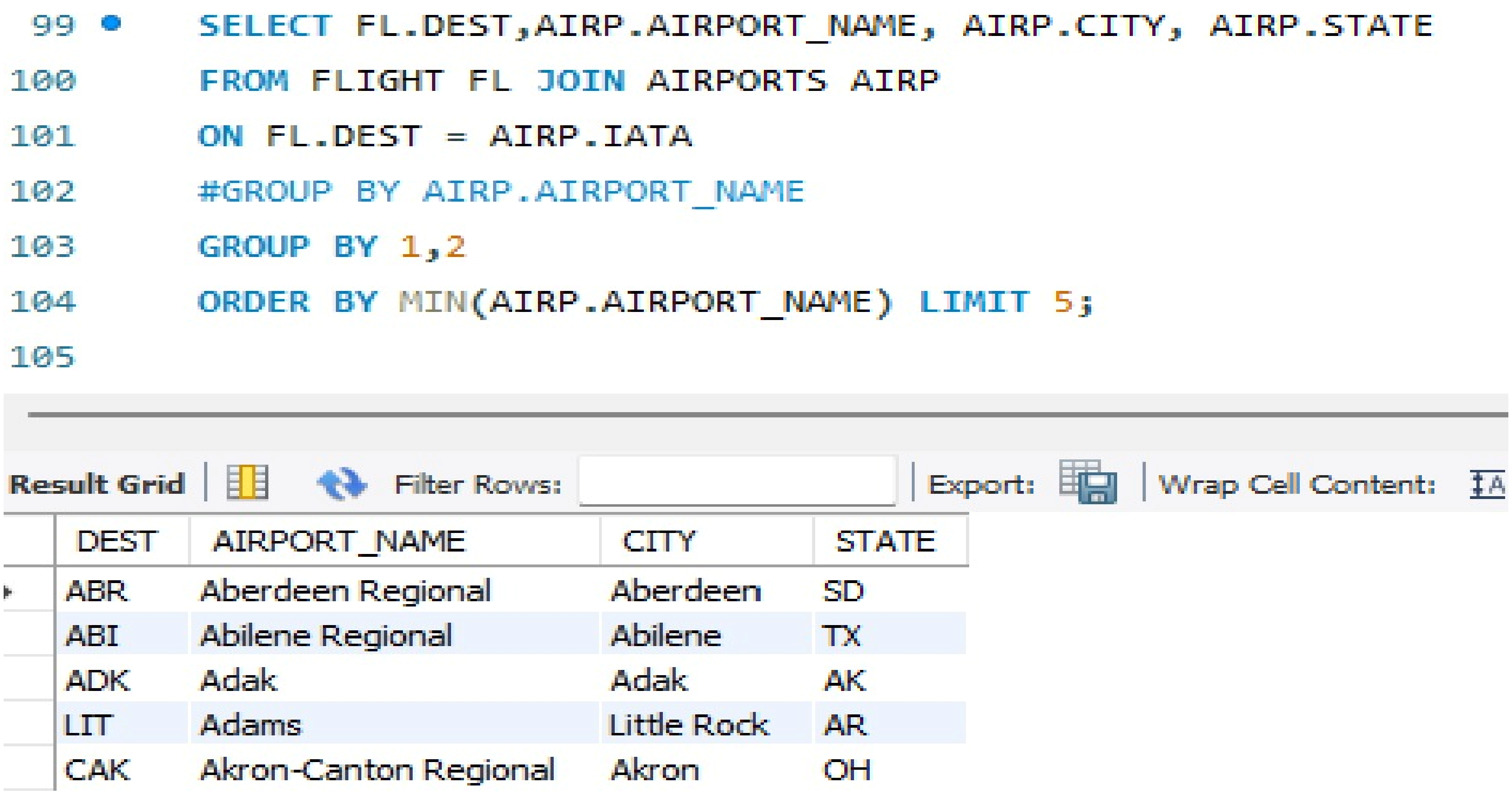


```
select x.AIRPORT_ID, sum(x.FLIGHT_COUNT) as FLIGHT_COUNT
    from ( select origin as AIRPORT_ID, count(*) as FLIGHT_COUNT
                from flight
                GROUP BY origin
            UNION ALL
            select dest as AIRPORT_ID, count(*) as FLIGHT_COUNT
                from flight
                GROUP BY dest
          ) x
    group by AIRPORT_ID
    order by FLIGHT_COUNT desc
    limit 10;
```

AIRPORT_ID	FLIGHT_COUNT
ATL	780125
ORD	665895
DFW	558570
DEN	472009
CLT	466626
LAX	443002
SFO	351788
PHX	347915
IAH	347588
LGA	342175

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

5. Which airport has the minimum traffic in the USA?



```
SELECT FL.DEST,AIRP.AIRPORT_NAME, AIRP.CITY, AIRP.STATE
FROM FLIGHT FL JOIN AIRPORTS AIRP
ON FL.DEST = AIRP.IATA
#GROUP BY AIRP.AIRPORT_NAME
GROUP BY 1,2
ORDER BY MIN(AIRP.AIRPORT_NAME) LIMIT 5;
```

DEST	AIRPORT_NAME	CITY	STATE
ABR	Aberdeen Regional	Aberdeen	SD
ABI	Abilene Regional	Abilene	TX
ADK	Adak	Adak	AK
LIT	Adams	Little Rock	AR
CAK	Akron-Canton Regional	Akron	OH

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