Data Science Capstone

United States Airlines Analysis

Problem statement:

According to air travel consumer reports, a large proportion of consumer complaints are about frequent flight delays. Out of all the complaints received from consumers about airline services, 32% were related to cancellations, delays, or other deviations from the airlines' schedules. There are unavoidable delays that can be caused by air traffic, no passengers at the airport, weather conditions, and mechanical issues, passengers coming from delayed connecting flights, security clearance, and aircraft preparation.

Objective:

The objective of this project is to identify the factors that contribute to avoidable flight delays. You are also required to build a model to predict if the flight will be delayed.

Project Task: Week 1

Applied data science with Python

1. Import and aggregate data:

   a. Collect information related to flights, airports (e.g., type of airport and elevation), and runways (e.g., length_ft, width_ft, surface, and number of runways). Gather all fields you believe might cause avoidable delays in one dataset.

      Hint: In this case, you would have to determine the keys to join the tables. A data description will be useful.

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay	iata_code_source_airport	type_source_airport
0	1	CO	269	SFO	IAH	3	15	205	1	SFO	large_airport
1	2	US	1558	PHX	CLT	3	15	222	1	PHX	large_airport
2	3	AA	2400	LAX	DFW	3	20	165	1	LAX	large_airport
3	4	AA	2466	SFO	DFW	3	20	195	1	SFO	large_airport
4	5	AS	108	ANC	SEA	3	30	202	0	ANC	large_airport

In [80]:

```
combined_data.columns
```

Out[80]:

```
Index(['id', 'Airline', 'Flight', 'AirportFrom', 'AirportTo', 'DayOfWeek',
       'Time', 'Length', 'Delay', 'type_source_airport',
       'elevation_ft_source_airport', 'runway_count_source_airport',
       'type_dest_airport', 'elevation_ft_dest_airport',
       'runway_count_dest_airport'],
      dtype='object')
```

b. When it comes to on-time arrivals, different airlines perform differently based on the amount of experience they have. The major airlines in this field include US Airways Express (founded in 1967) Continental Airlines (founded in 1934), and Express Jet (founded in 19860. Pull such information specific to various airlines from the Wikipedia page link given below.

https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States.

Hint: Here, you should use web scraping to learn how long an airline has been operating.

In [32]:

```
airlines_wiki_list = []
for tab in tables_found:
    temp = pd.read_html(str(tab))
    temp = pd.DataFrame(temp[0])
    airlines_wiki_list.append(temp)
```

In [33]:

```
airlines_wiki = pd.concat(airlines_wiki_list)
```

c. You should then get all the information gathered so far in one place.

```
combined_data.head()
```

Out[29]:

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay	type_source_airport	elevation_ft_source_airpor
0	1	CO	269	SFO	IAH	3	15	205	1	large_airport	13.(
1	2	US	1558	PHX	CLT	3	15	222	1	large_airport	1135.(
2	3	AA	2400	LAX	DFW	3	20	165	1	large_airport	125.(
3	4	AA	2466	SFO	DFW	3	20	195	1	large_airport	13.(
4	5	AS	108	ANC	SEA	3	30	202	0	large_airport	152.(

d. The total passenger traffic may also contribute to flight delays. The term hub refers to busy commercial airports. Large hubs are airports that account for at least 1 percent of the total passenger enplanements in the United States. Airports that account for 0.25 percent to 1 percent of total passenger enplanements are considered medium hubs. Pull passenger traffic data from the Wikipedia page given below using web scraping and collate it in a table.

https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States

```
large_hub.head()
```

Out[39]:

	Rank(2021)	Hub_type	Airports (large hubs)	IATACode	Major cities served	State	2021[3]	2020[4]	2019[5]	2018[6]	2017[7]	201
0	1	large	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	GA	36676010	20559866	53505795	51865797	50251964	50501
1	2	large	Dallas/Fort Worth International Airport	DFW	Dallas & Fort Worth	TX	30005266	18593421	35778573	32821799	31816933	31283
2	3	large	Denver International Airport	DEN	Denver	CO	28645527	16243216	33592945	31362941	29809097	28267
3	4	large	O'Hare International Airport	ORD	Chicago	IL	26350976	14606034	40871223	39873927	38593028	37589
4	5	large	Los Angeles International Airport	LAX	Los Angeles	CA	23663410	14055777	42939104	42624050	41232432	39636

```
med_hub.head()
```

Out[40]:

	Rank(2021)	Hub_type	Airports (medium hubs)	IATACode	City served	State	2021[3]	2020[4]	2019[5]	2018[6]	2017[7]	2016[8]
0	31	medium	Dallas Love Field	DAL	Dallas	TX	6487563	3669930	8408457	8134848	7876769	7554596
1	32	medium	Daniel K. Inouye International Airport	HNL	Honolulu	HI	5830928	3126391	9988678	9578505	9743989	9656340
2	33	medium	Portland International Airport	PDX	Portland	OR	5759879	3455877	9797408	9940866	9435473	9071154
3	34	medium	William P. Hobby Airport	HOU	Houston	TX	5560780	3127178	7069614	6937061	6741870	6285181
4	35	medium	Southwest Florida International Airport	RSW	Fort Myers	FL	5080805	2947139	5144467	4719568	4461304	4350650

2. You should then examine the missing values in each field, perform missing value treatment, and justify your actions.

In [121]:

```
miss_val = {'US':1967, 'EV':1986, 'CO':1931}

for aline in miss_founded:
    combined_data_traffic.loc[(combined_data_traffic.Founded.isnull())& (combined_data_t
raffic.Airline ==aline),'Founded'] = miss_val[aline]
```

In [122]:

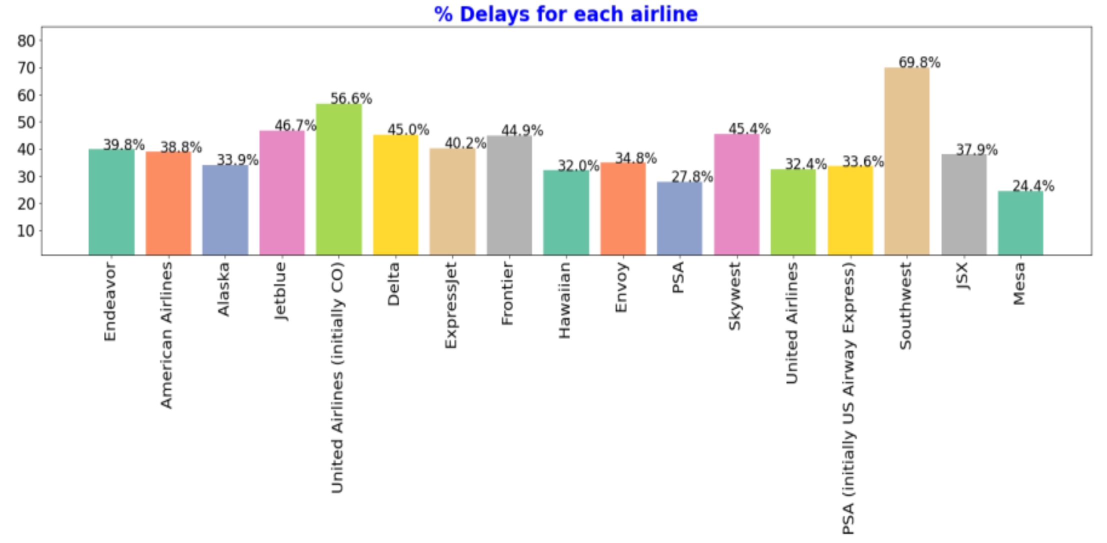
```
combined_data_traffic.isnull().sum()
```

Out[122]:

```
id                              0
Airline                         0
Flight                          0
AirportFrom                     0
AirportTo                       0
DayOfWeek                       0
Time                            0
Length                          0
Delay                           0
type_source_airport             0
elevation_ft_source_airport     0
```

3. Perform data visualization and share your insights on the following points:

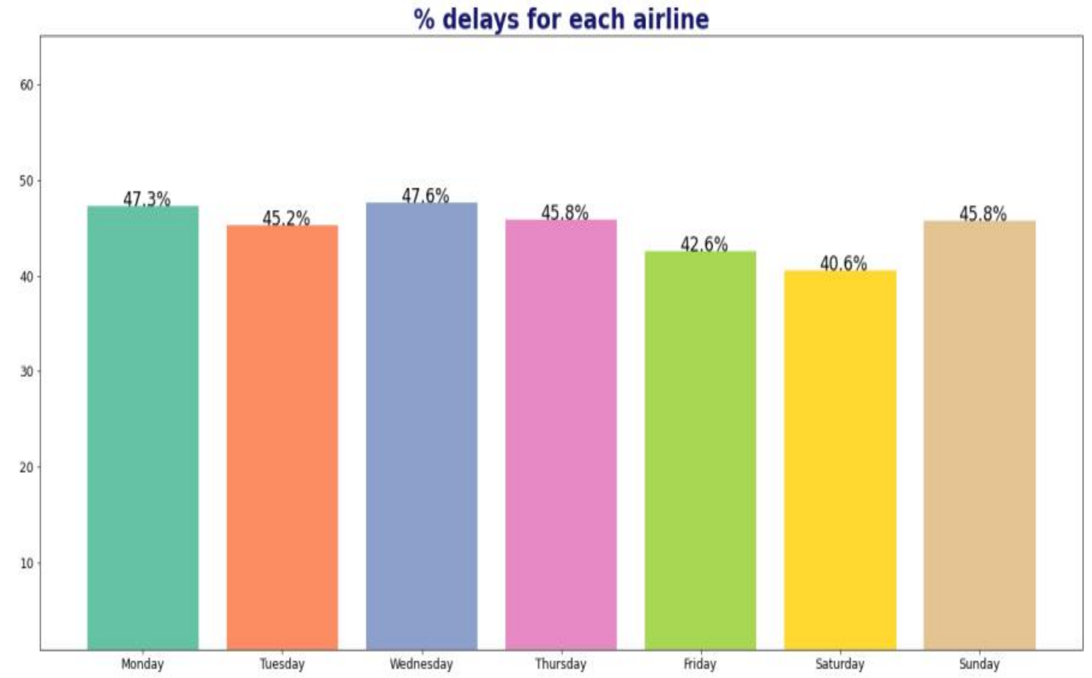
a. According to the data provided, approximately 70% of Southwest Airlines flights are delayed. Visualize it to compare it with the data of other airlines.



As we can see in the above bar chart, SouthWest airlines has the most delayed flgihts i.e. 69.8% to be exact.

Mesa is the airlines with least delays.

b. Flights were delayed on various weekdays. Which day of the week is the safest for travel?



From the bar chart above, we can see that Friday and Saturday have the shortest bars for % delays, hence these two days are the safest to travel.

c. Which airlines should be recommended for short-, medium-, and long-distance travel?

In [154]:

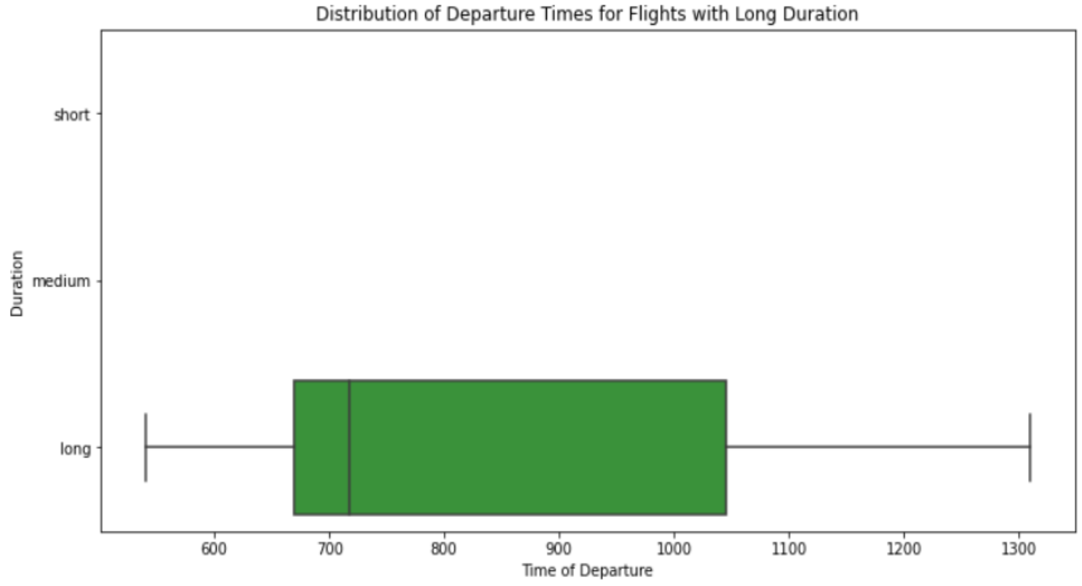
```
long = duration_grp[duration_grp.long == duration_grp.long.min()].Description.values.tol
ist()
print(len(long),'Airlines with minimum delays (0%) for long flights:\n',','.join(long))

medium= duration_grp[duration_grp.medium == duration_grp.medium.min()].Description.value
s.tolist()
print('\n',len(medium),'Airline with minimum delays (0%) for medium flights:\n',','.join(
medium))

short = duration_grp[duration_grp.short == duration_grp.short.min()].Description.values.
tolist()
print('\n', len(short),'Airline with minimum delays (24.37%) for short flights:\n',','.jo
in(short))
```

```
13 Airlines with minimum delays (0%) for long flights:
 Endeavor,Alaska,Jetblue,ExpressJet,Frontier,Hawaiian,Envoy,PSA,Skywest,PSA (initially US
Airway Express),Southwest,JSX,Mesa

 1 Airline with minimum delays (0%) for medium flights:
 Endeavor

 1 Airline with minimum delays (24.37%) for short flights:
 Mesa
```

Here are the recommended airlines with minimum delays that are safest to travel for each kind of travel distance:

- **Long flights (0% delay):** Endeavor, Alaska, Jetblue, ExpressJet, Frontier, Hawaiian, Envoy, PSA, Skywest, PSA (initially US Airway Express), Southwest, JSX, Mesa
- **Medium flights (0% delays):** Endeavor
- **Short flights (24.37% delays):** Mesa

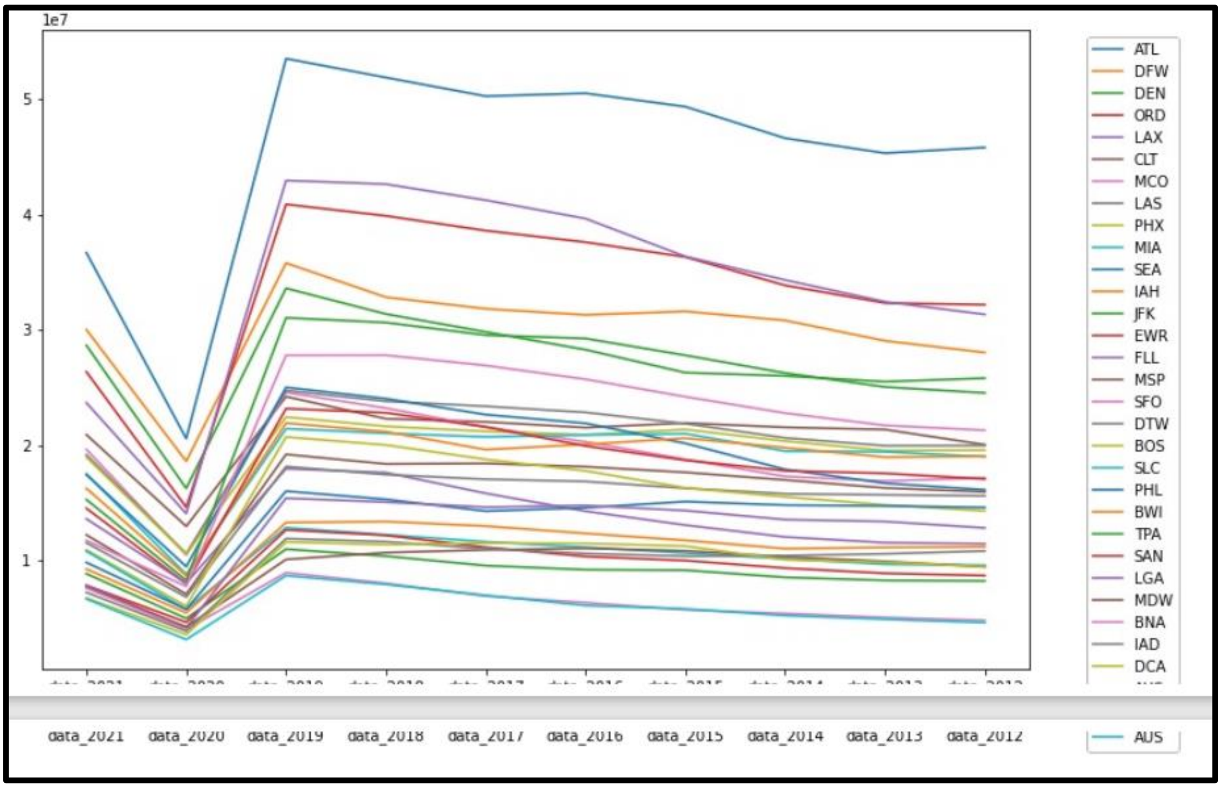
d. Do you notice any patterns in the departure times of long-duration flights?



Observations till now:

- **75% of all the long duration flights depart before 6:00 PM in the evening.**
- **50% of flights depart before 12:00 PM in the afternoon.**
- **The earliest long duration flight departs at 9:00 AM.**
- **The last long duration flight departs at 10:50 PM.**

4. How many flights were delayed at large hubs compared to medium hubs? Use appropriate visualization to represent your findings.



5. Use hypothesis testing strategies to discover:

a. If the airport's altitude has anything to do with flight delays for incoming and departing flights

For incoming flights.

In [198]:

```
# 2 sample t test for incoming flights with the following hypothesis.
# H0 : avg elevation for Delayed flights - avg elevation for not Delayed flights = 0
# Ha : avg elevation for Delayed flights - avg elevation for not Delayed flights != 0
```

In [199]:

```
sample1 = cdt[cdt.Delay == 1].elevation_ft_dest_airport
```

```
sample2 = cdt[cdt.Delay == 0].elevation_ft_dest_airport
```

In [200]:

```
t, p = stats.ttest_ind(sample1, sample2)
```

In [201]:

```
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'

print(result)
```

```
reject null
```

There is a statistically significant difference in the average elevation between delayed and not delayed flights. This suggests that the elevation of the destination airport may play a role in flight delays .

b. If the number of runways at an airport affects flight delays

```
s1 = cdt[cdt.Delay == 1].runway_count_source_airport
s2 = cdt[cdt.Delay == 0].runway_count_source_airport
```

In [205]:

```
t, p = stats.ttest_ind(s1, s2)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

In [206]:

```
s1 = cdt[cdt.Delay == 1].runway_count_dest_airport
s2 = cdt[cdt.Delay == 0].runway_count_dest_airport
```

In [207]:

```
t, p = stats.ttest_ind(s1, s2)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

The resulting output "reject null" for both tests suggests that there is evidence to support the alternative hypothesis, indicating that the average runway count for delayed flights is significantly lower than the average runway count for non-delayed flights in both the source and destination airports.

c. If the duration of a flight (length) affects flight delays

Hint: Test this from the perspective of both the source and destination airports

In [213]:

```
chi, p, df, ex = stats.chi2_contingency(cs)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

The result of the chi-square test was "reject null," it means that there is evidence to suggest a significant relationship between the duration of flights and flight delays.

In [214]:

```
# t test :
# H0 : avg duration for delayed filghts - avg duration for non delayed flights <= 0
# Ha : avg duration for delayed filghts - avg duration for non delayed flights > 0
```

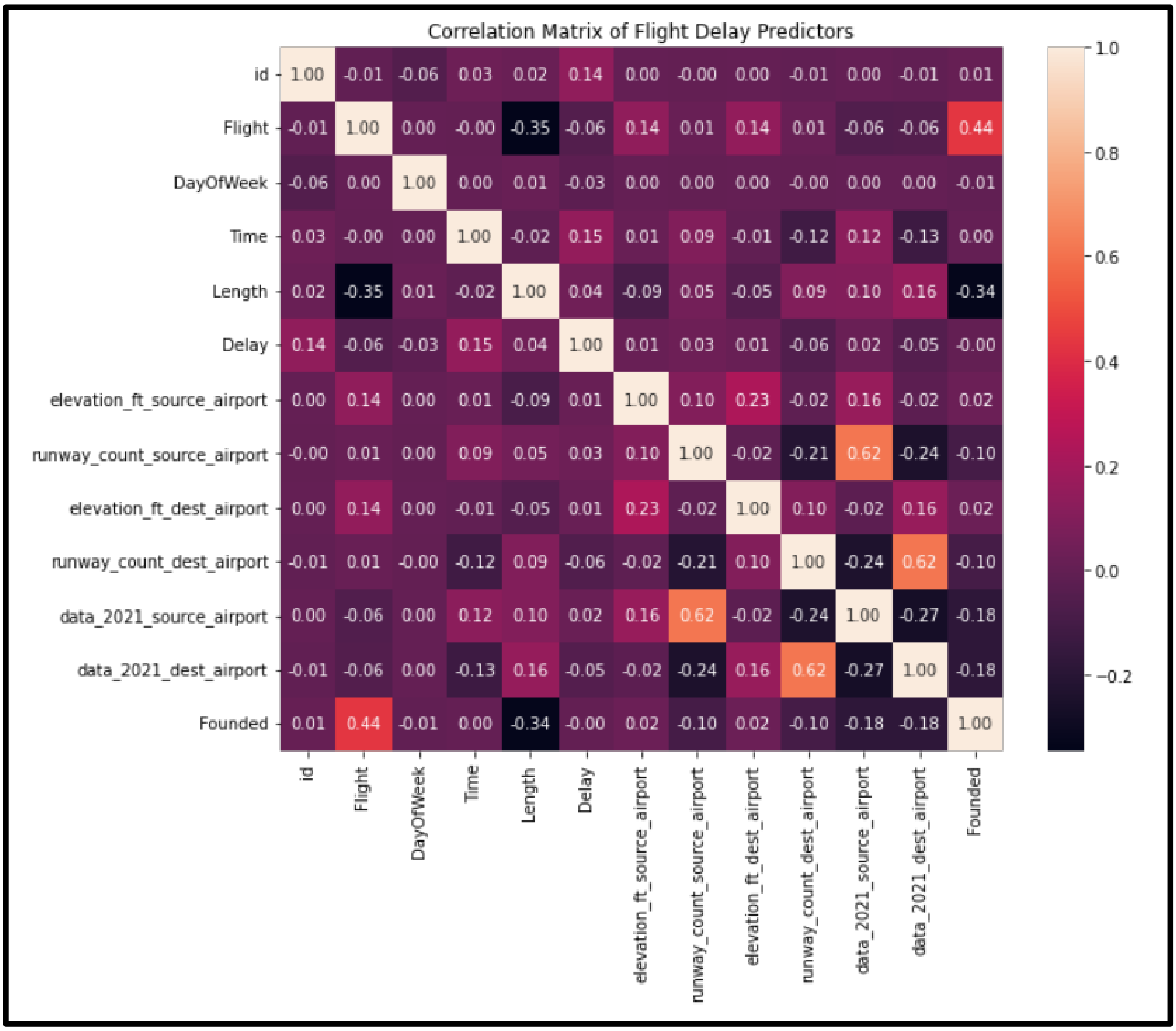
In [215]:

```
t, p = stats.ttest_ind(s1, s2)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

The result of the t-test was "reject null," it means that there is evidence to suggest a significant difference in the average duration between delayed flights and non-delayed flights. Specifically, the average duration of delayed flights is significantly greater than the average duration of non-delayed flights.

6. Find the correlation matrix between the flight delay predictors, create a heatmap to visualize this, and share your findings



Machine learning

1. Use OneHotEncoder and OrdinalEncoder to deal with categorical variables

```
dummy.Founded = 2023 - dummy.Founded
dummy.head()
```

Out[231]:

	DayOfWeek	Time	Length	Delay	type_source_airport	elevation_ft_source_airport	runway_count_source_airport	type_dest_
0	3	15	205	1	1.0	13.0	4.0	
1	3	15	222	1	1.0	1135.0	3.0	
2	3	20	165	1	1.0	125.0	4.0	
3	3	20	195	1	1.0	13.0	4.0	
4	3	30	202	0	1.0	152.0	3.0	

2. Perform the following model building steps:

   a. Apply logistic regression (use stochastic gradient descent optimizer) and decision tree models

   b. Use the stratified five-fold method to build and validate the models

Note: Make sure you use standardization effectively, ensuring no data leakage and leverage pipelines to have a cleaner code

c. Use RandomizedSearchCV for hyperparameter tuning, and use k-fold for crossvalidation

d. Keep a few data points (10%) for prediction purposes to evaluate how you would make the final prediction, and do not use this data for testing or validation

Note: The final prediction will be based on the voting (majority class by 5 models created using the stratified 5-fold method)

e. Compare the results of logistic regression and decision tree classifier

In [242]:

```
# compare results :
train_results = pd.DataFrame ({'sgd' : accuracy_train.values(), 'dt': dt_accuracy_train.values() },
                              index = ['Fold {}'.format(i) for i in range(1,6)])
train_results
```

Out[242]:

	sgd	dt
Fold 1	57.163	61.643
Fold 2	57.168	61.669
Fold 3	57.154	61.649
Fold 4	57.228	61.487
Fold 5	57.105	61.597

In [243]:

```
test_results = pd.DataFrame ({'sgd' : accuracy_test.values(), 'dt': dt_accuracy_test.values() },
                              index = ['Fold {}'.format(i) for i in range(1,6)])
test_results
```

Out[243]:

	sgd	dt
Fold 1	57.173	61.431
Fold 2	57.182	61.304
Fold 3	57.221	61.928
Fold 4	56.891	61.421
Fold 5	57.313	61.456

3. Use the stratified five-fold method to build and validate the models using the XGB classifier, compare all methods, and share your findings

In [257]:

```
train_results
```

Out[257]:

	sgd	dt	xgb
Fold 1	57.163	61.643	0.647
Fold 2	57.168	61.669	0.646
Fold 3	57.154	61.649	0.645
Fold 4	57.228	61.487	0.646
Fold 5	57.105	61.597	0.645

In [258]:

```
test_results
```

Out[258]:

	sgd	dt	xgb
Fold 1	57.173	61.431	0.643
Fold 2	57.182	61.304	0.643
Fold 3	57.221	61.928	0.646
Fold 4	56.891	61.421	0.642
Fold 5	57.313	61.456	0.645

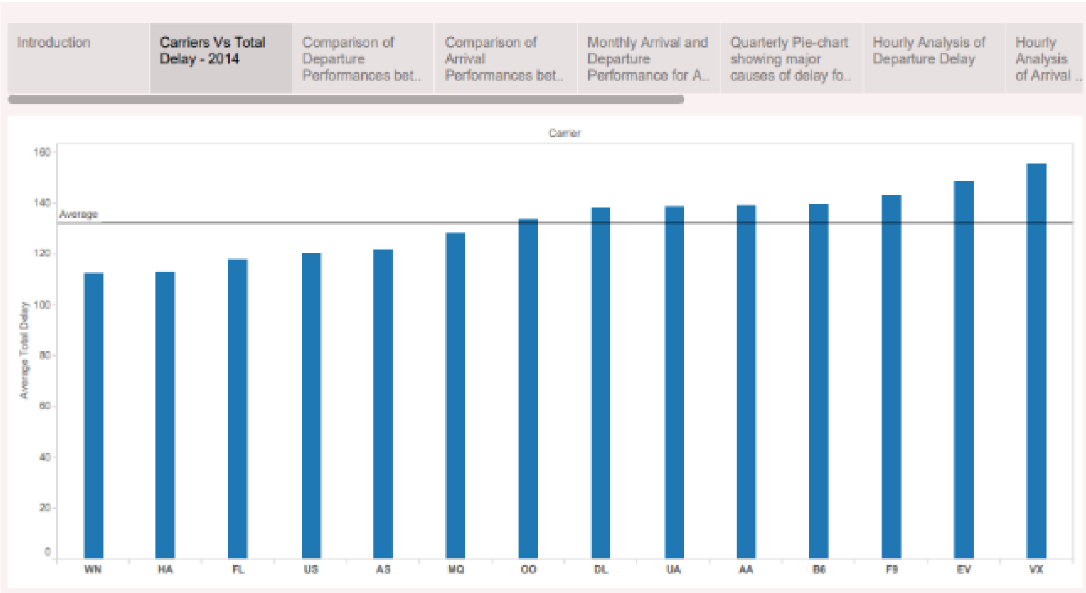
The logistic regression and decision tree models perform similarly and achieve higher accuracy compared to the XGBoost model.

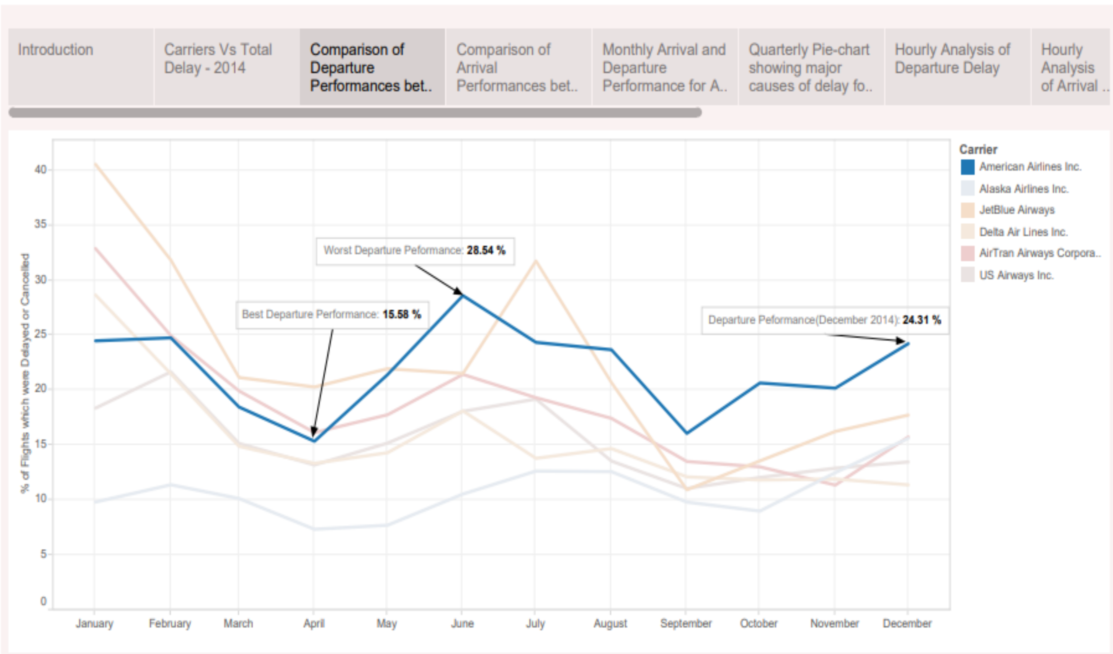
Project Task: Week 2

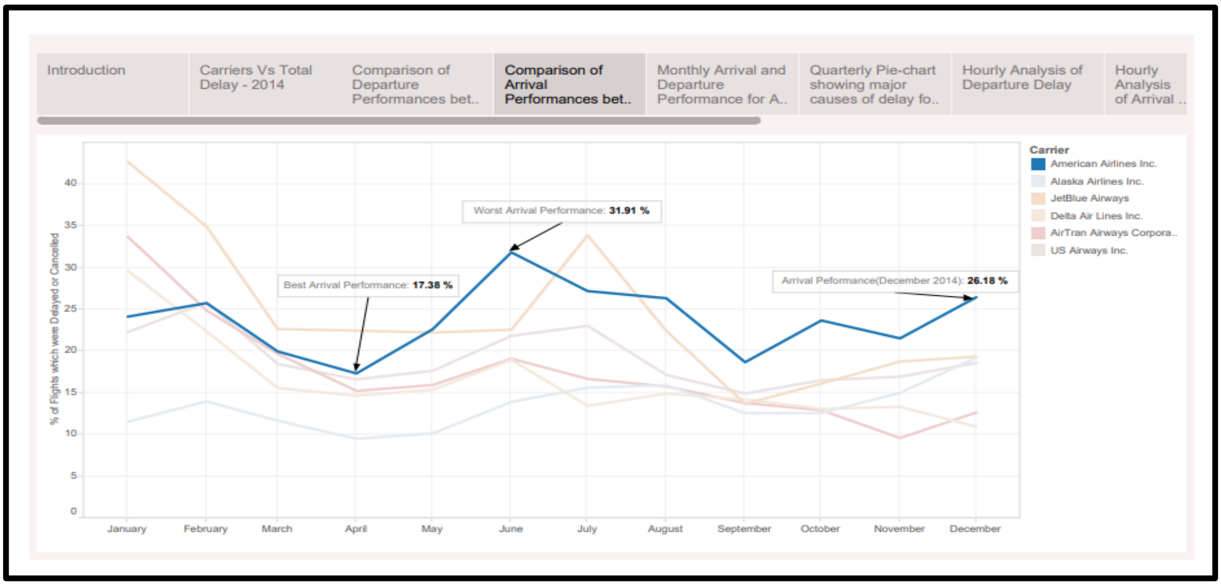
Tableau

1. Create a dashboard in Tableau by selecting appropriate chart types and metrics for the business

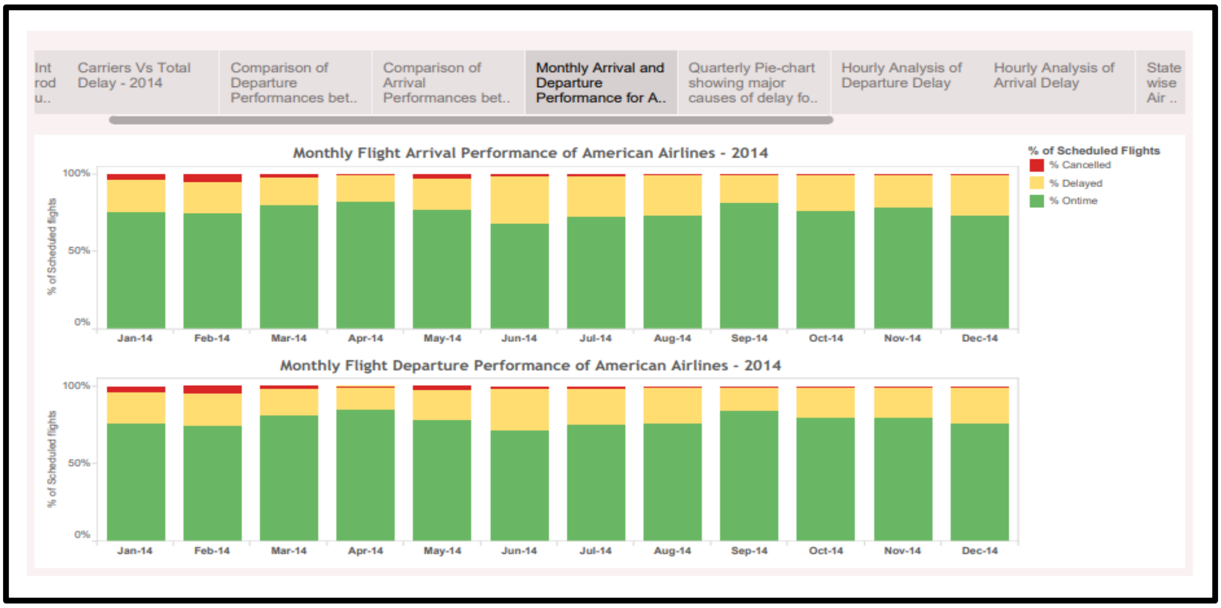
Note: Put more emphasis on data storytelling



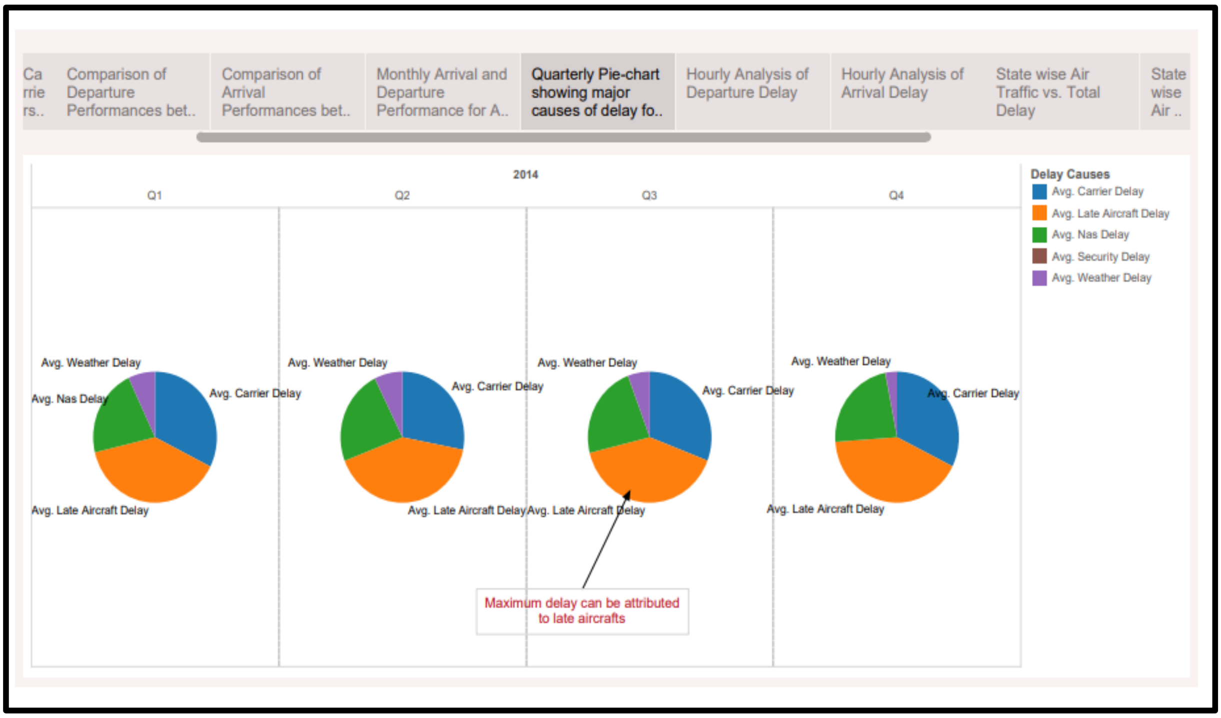




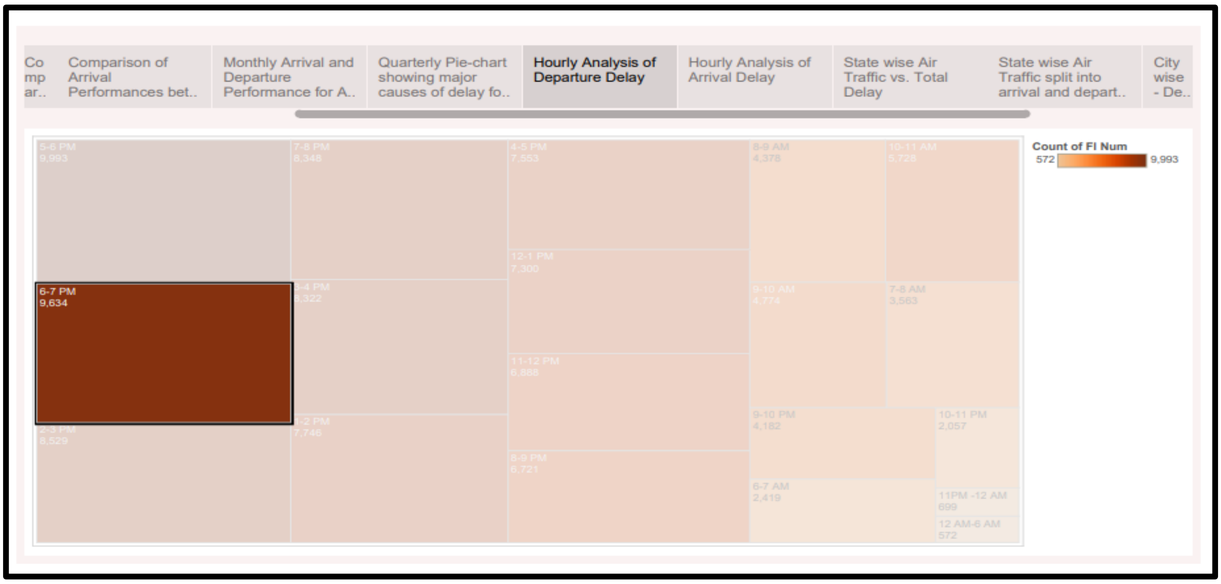
Introduction
Carriers Vs Total Delay - 2014
Comparison of Departure Performances bet..
Comparison of Arrival Performances bet..
Monthly Arrival and Departure Performance for A..
Quarterly Pie-chart showing major causes of delay fo..
Hourly Analysis of Departure Delay
Hourly Analysis of Arrival ..
Worst Arrival Performance: 31.91 %
Best Arrival Performance: 17.38 %
Arrival Peformance(December 2014): 26.18 %
% of Flights which were Delayed or Cancelled
January
February
March
April
May
June
July
August
September
October
November
December
Carrier
American Airlines Inc.
Alaska Airlines Inc.
JetBlue Airways
Delta Air Lines Inc.
AirTran Airways Corpora..
US Airways Inc.

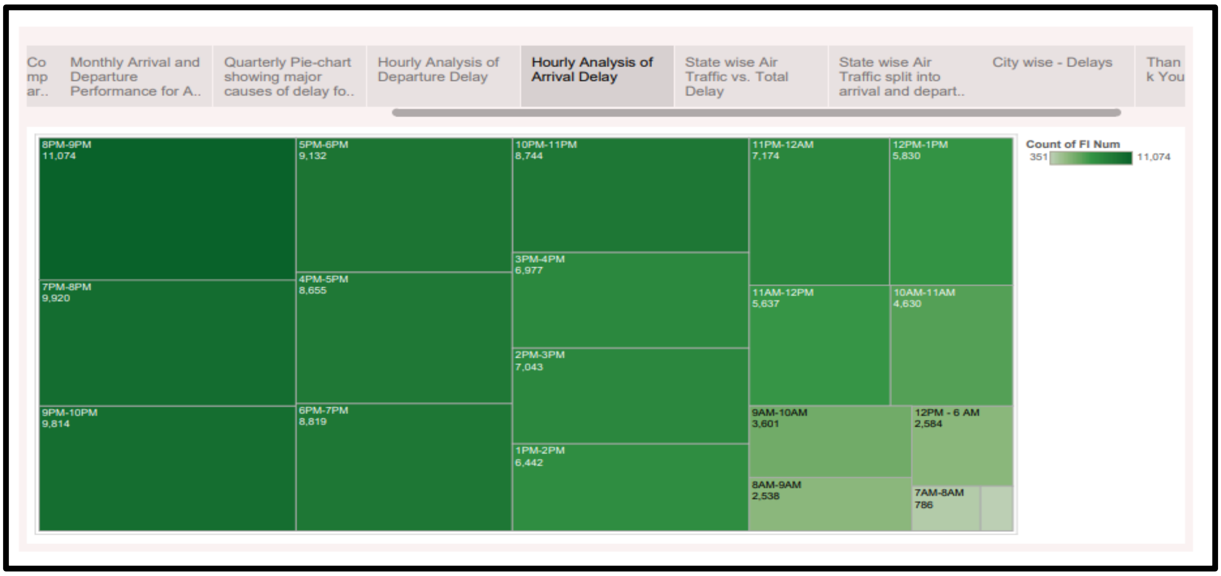


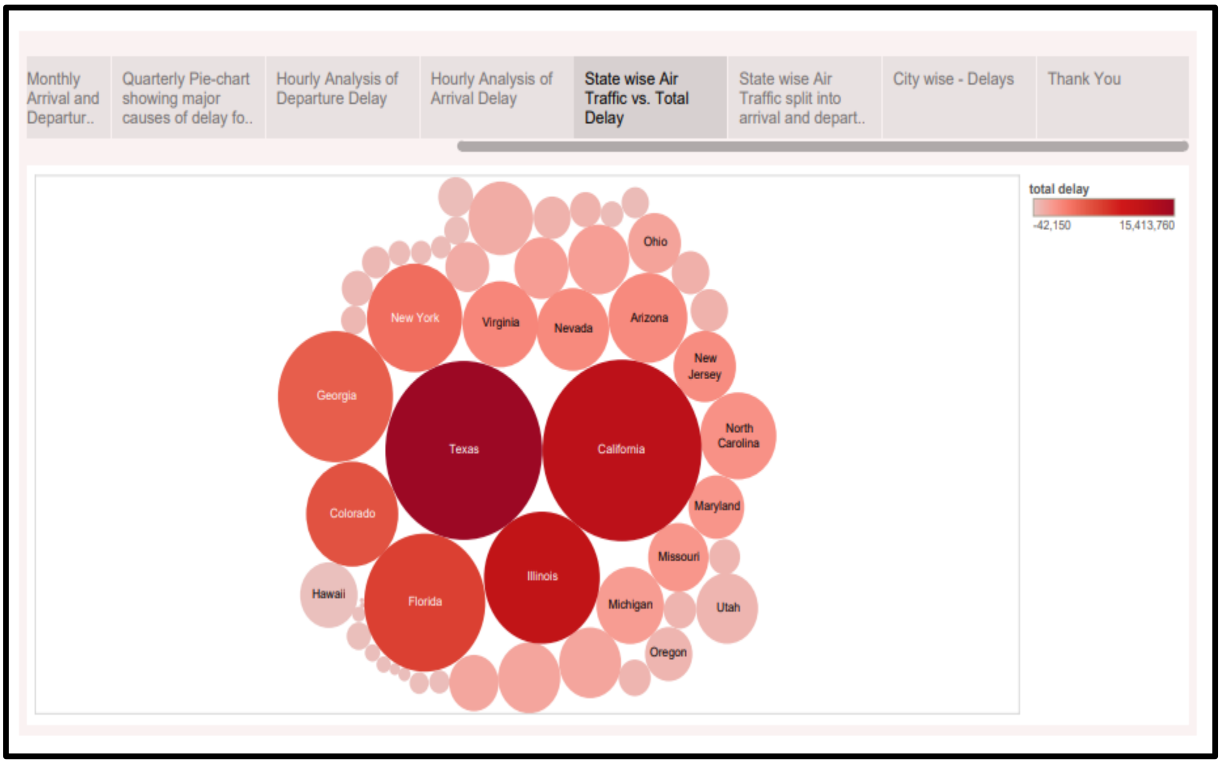
Monthly Flight Arrival Performance of American Airlines - 2014
Monthly Flight Departure Performance of American Airlines - 2014
% of Scheduled flights
% of Scheduled Flights
% Cancelled
% Delayed
% Ontime
Jan-14
Feb-14
Mar-14
Apr-14
May-14
Jun-14
Jul-14
Aug-14
Sep-14
Oct-14
Nov-14
Dec-14

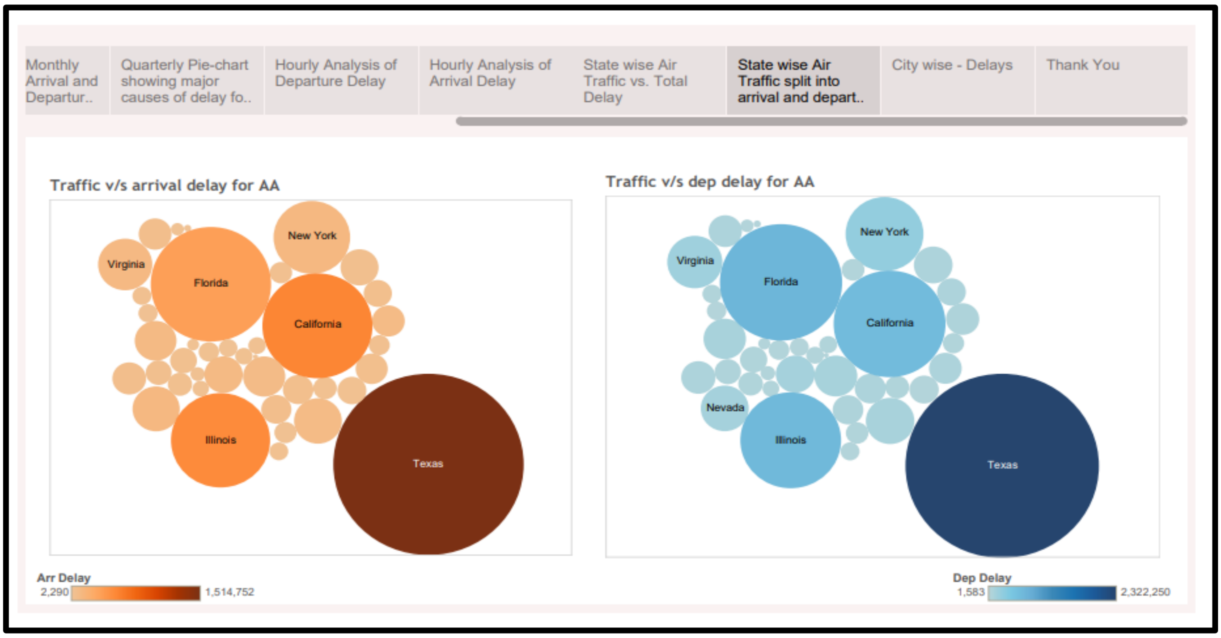


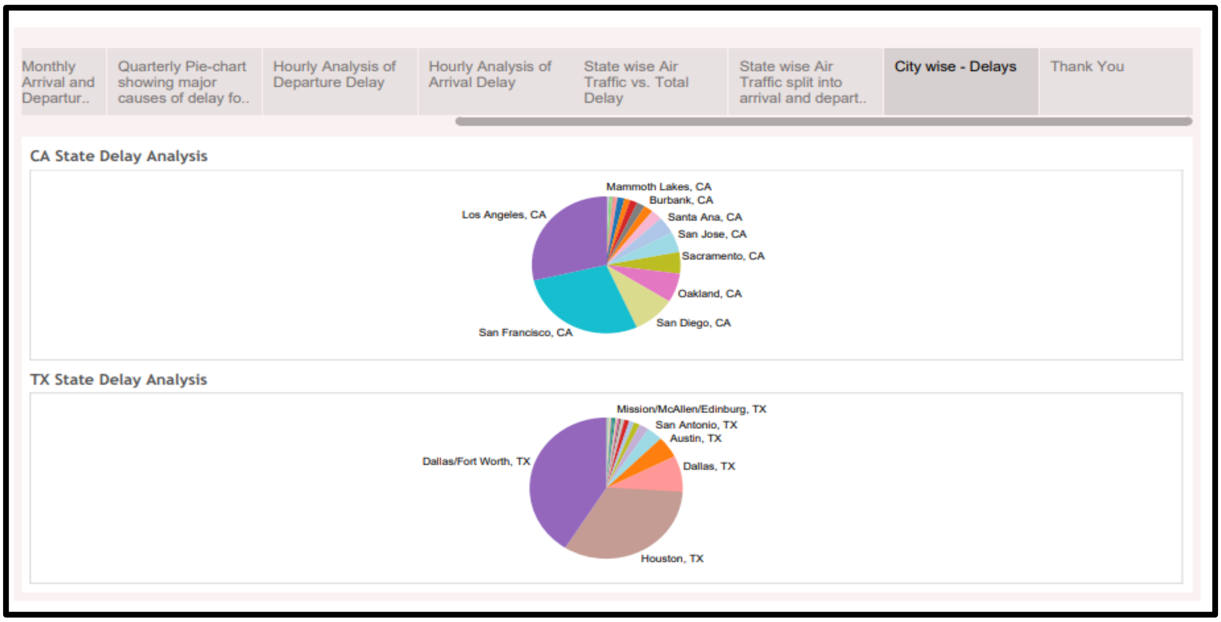
Quarterly Pie-chart showing major causes of delay fo..
2014
Q1
Q2
Q3
Q4
Delay Causes
Avg. Carrier Delay
Avg. Late Aircraft Delay
Avg. Nas Delay
Avg. Security Delay
Avg. Weather Delay
Maximum delay can be attributed to late aircrafts


Co mp ar..
Comparison of Arrival Performances bet..
Monthly Arrival and Departure Performance for A..
Quarterly Pie-chart showing major causes of delay fo..
Hourly Analysis of Departure Delay
Hourly Analysis of Arrival Delay
State wise Air Traffic vs. Total Delay
State wise Air Traffic split into arrival and depart..
City wise - De..
Count of Fl Num
572
9,993
6-7 PM
9,634


Co mp ar..
Monthly Arrival and Departure Performance for A..
Quarterly Pie-chart showing major causes of delay fo..
Hourly Analysis of Departure Delay
Hourly Analysis of Arrival Delay
State wise Air Traffic vs. Total Delay
State wise Air Traffic split into arrival and depart..
City wise - Delays
Thank You
Count of Fl Num
351
11,074
8PM-9PM 11,074
5PM-6PM 9,132
10PM-11PM 8,744
11PM-12AM 7,174
12PM-1PM 5,830
7PM-8PM 9,920
4PM-5PM 8,655
3PM-4PM 6,977
11AM-12PM 5,637
10AM-11AM 4,630
2PM-3PM 7,043
9PM-10PM 9,814
6PM-7PM 8,819
1PM-2PM 6,442
9AM-10AM 3,601
12PM - 6 AM 2,584
8AM-9AM 2,538
7AM-8AM 786


Monthly Arrival and Departur..
Quarterly Pie-chart showing major causes of delay fo..
Hourly Analysis of Departure Delay
Hourly Analysis of Arrival Delay
State wise Air Traffic vs. Total Delay
State wise Air Traffic split into arrival and depart..
City wise - Delays
Thank You
total delay
-42,150
15,413,760
Ohio
New York
Virginia
Nevada
Arizona
New Jersey
Georgia
Texas
California
North Carolina
Colorado
Maryland
Missouri
Illinois
Hawaii
Florida
Michigan
Utah
Oregon

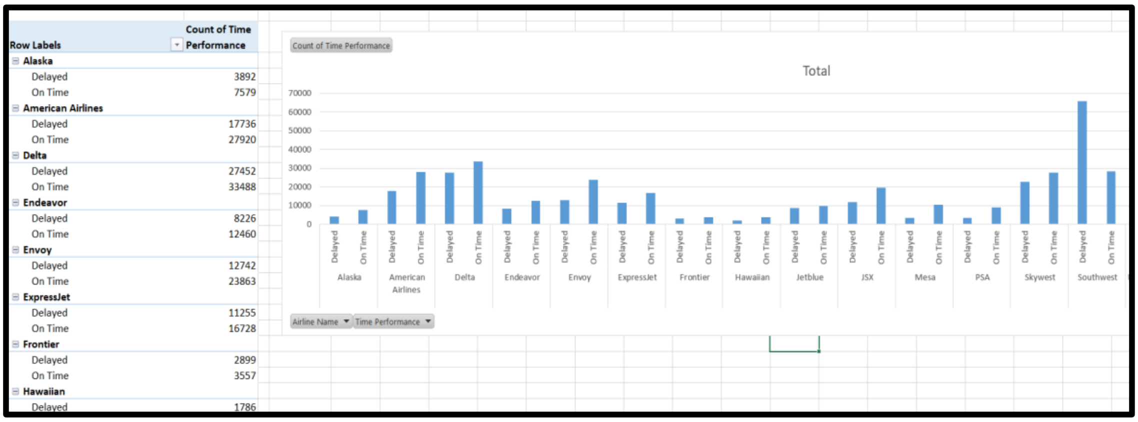




Excel

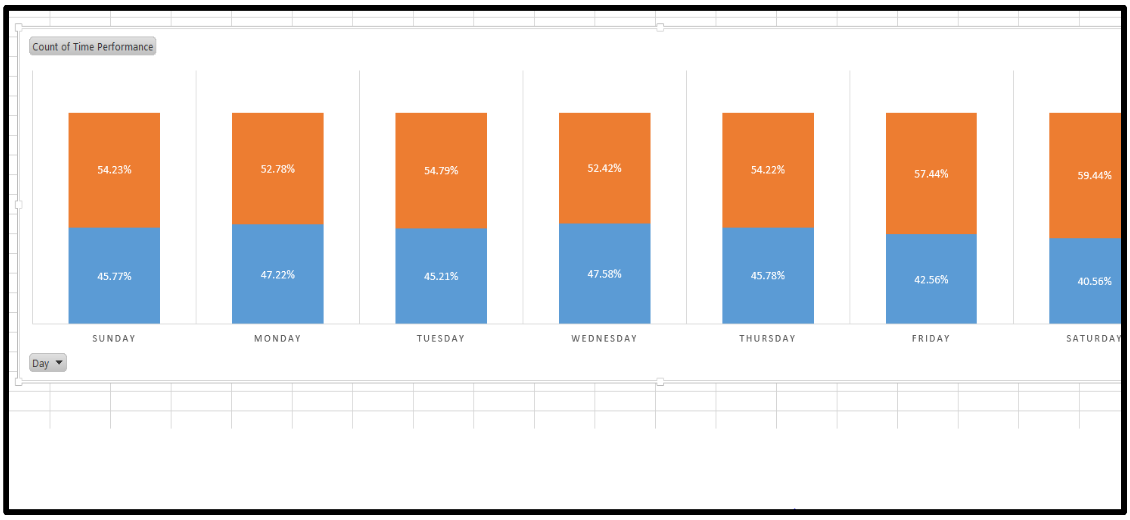
1. Create an Excel dashboard showcasing the following (use form controls to make a dynamic chart):

   a. Compare different airlines based on their on-time performance

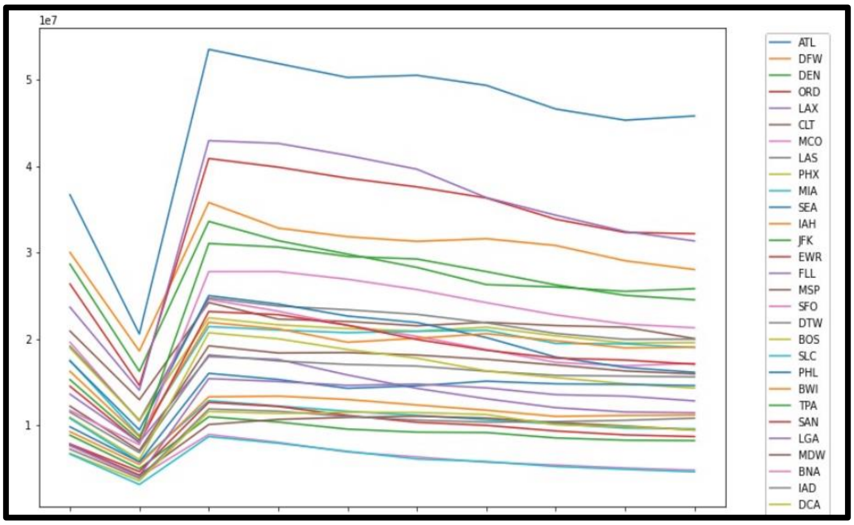


Row Labels	Count of Time Performance
Alaska	
Delayed	3892
On Time	7579
American Airlines	
Delayed	17736
On Time	27920
Delta	
Delayed	27452
On Time	33488
Endeavor	
Delayed	8226
On Time	12460
Envoy	
Delayed	12742
On Time	23863
ExpressJet	
Delayed	11255
On Time	16728
Frontier	
Delayed	2899
On Time	3557
Hawaiian	
Delayed	1786

b. Compare the percentage of delayed flights for different days of the week

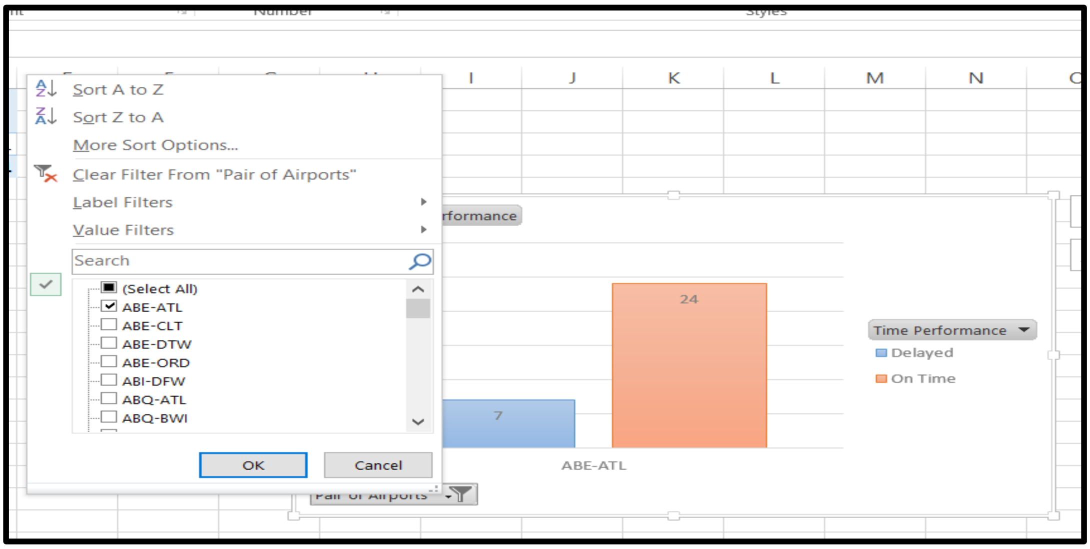


c. Create a trend chart for the number of passengers at large and medium hubs



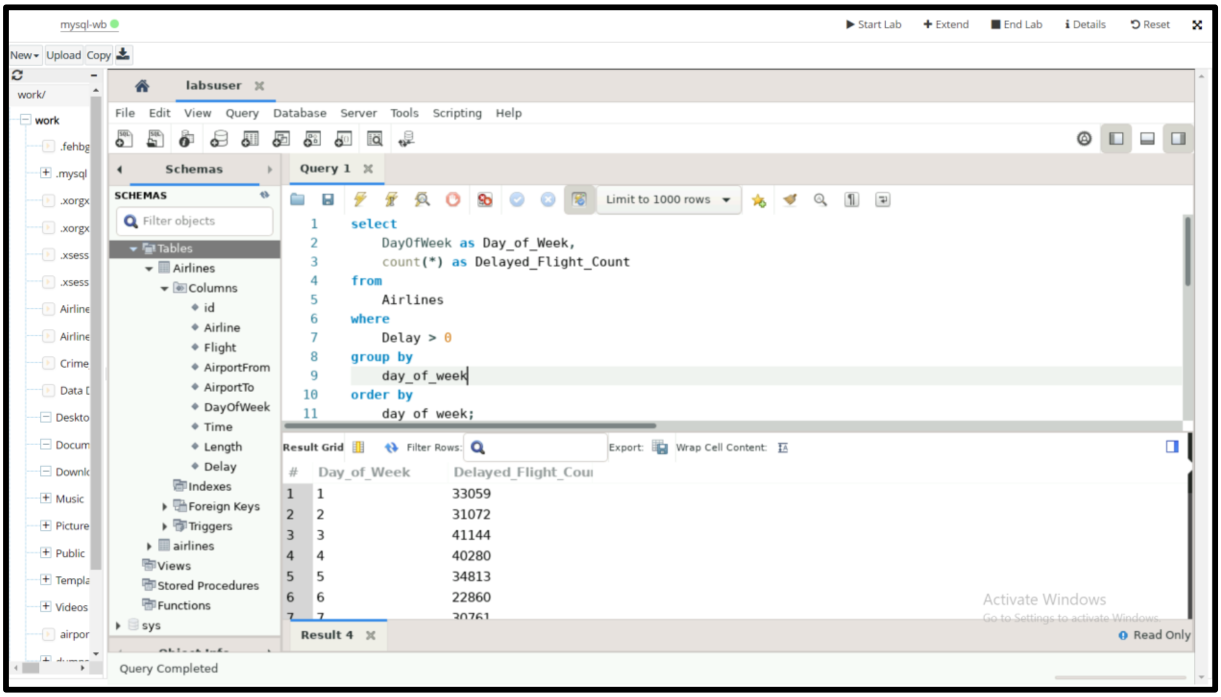
d. Visualize the count of delayed and on-time flights for different pairs of source and destination airports

• Create a dynamic chart that allows users to select a source and destination airport.

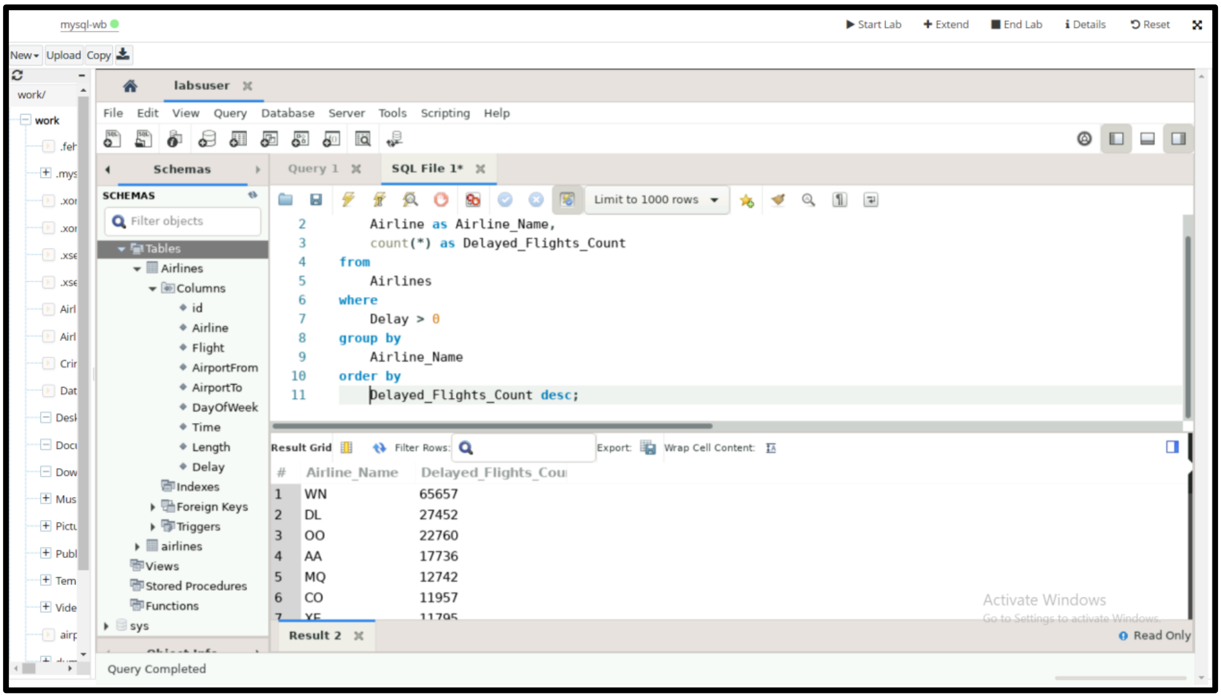


SQL

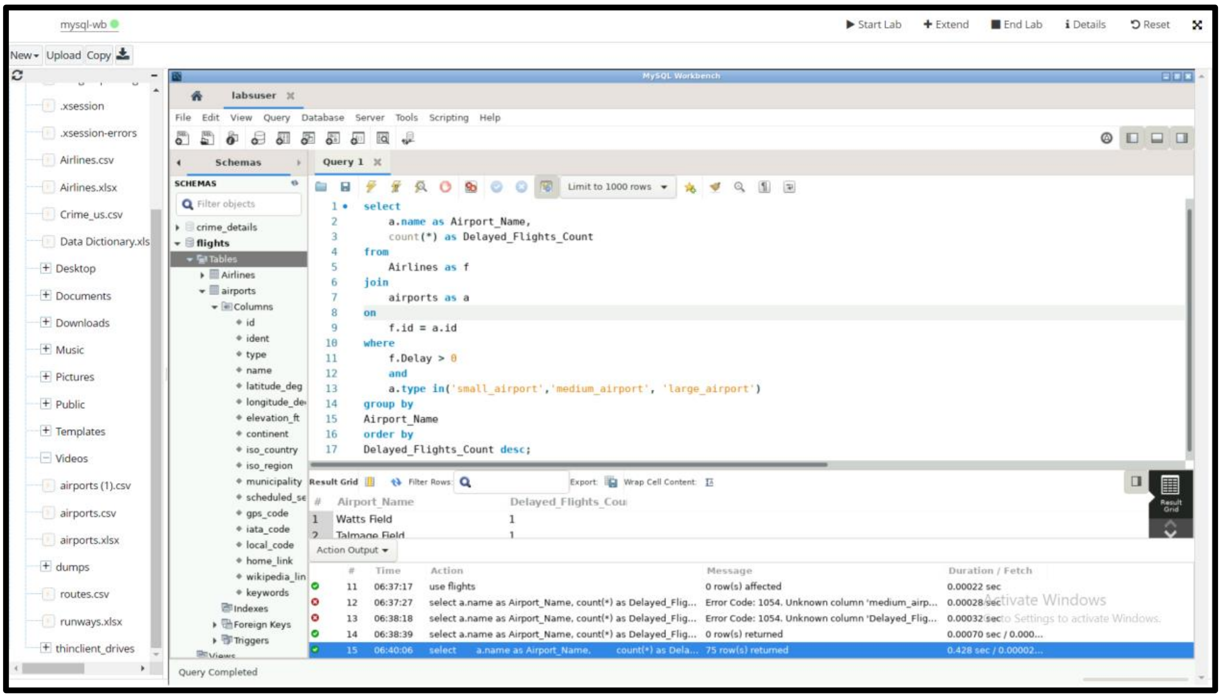
1. Determine the number of flights that are delayed on various days of the week



2. Determine the number of delayed flights for various airlines

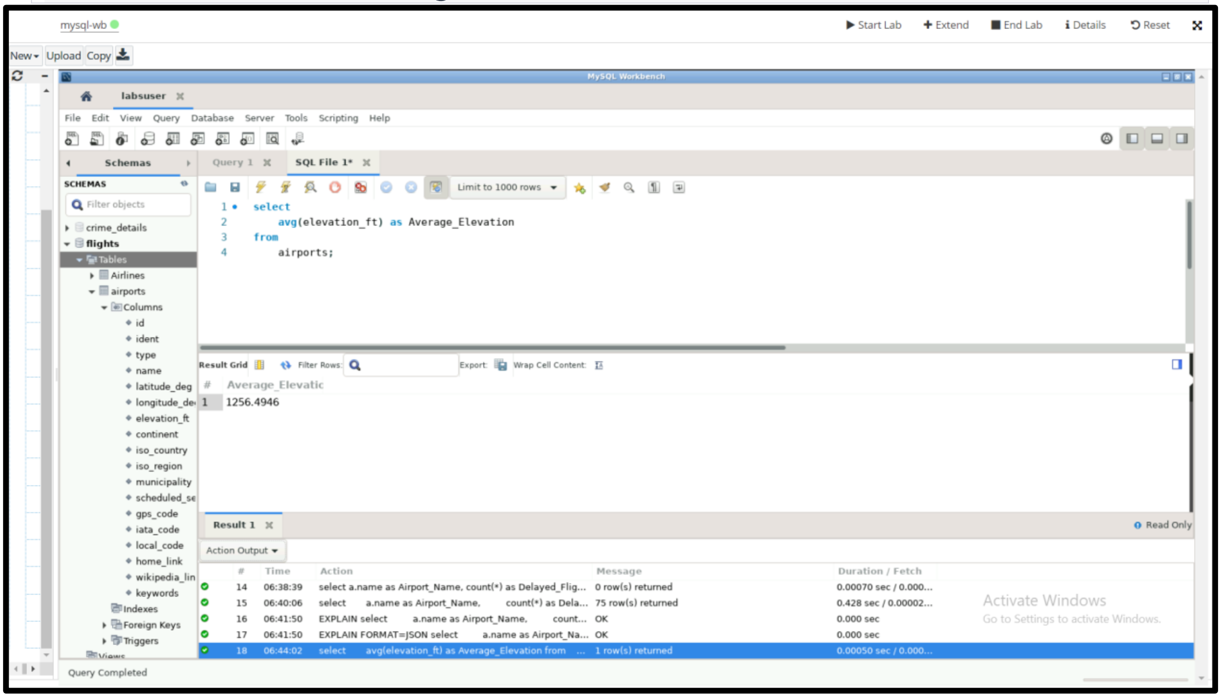


3. Determine how many delayed flights land at airports with at least 10 runways

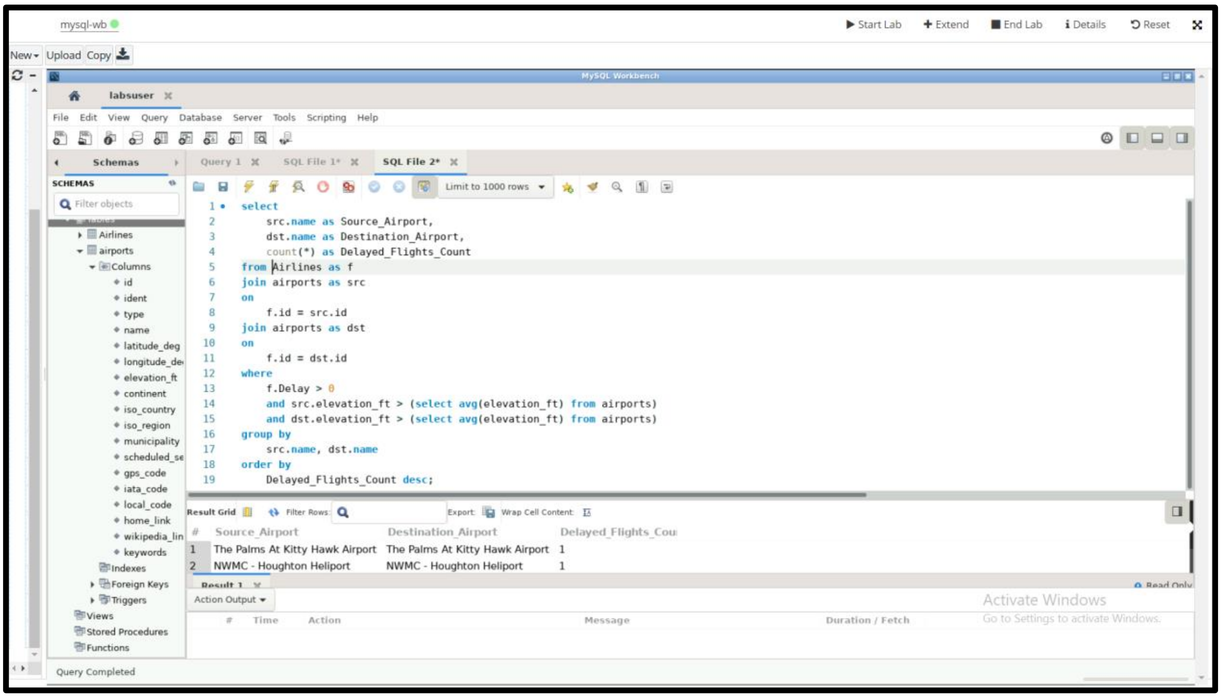


4. Compare the number of delayed flights at airports higher than average elevation and those that are lower than average elevation for both source and destination airports

- First, calculate the average elevation:



Then, use the calculated average elevation in the following query:


mysql-wb
Start Lab
Extend
End Lab
Details
Reset
New
Upload
Copy
labsuser
File Edit View Query Database Server Tools Scripting Help
Schemas
Query 1
SQL File 1*
SQL File 2*
SCHEMAS
Filter objects
Airlines
airports
Columns
id
ident
type
name
latitude_deg
elevation_ft
continent
iso_country
iso_region
municipality
gps_code
iata_code
local_code
home_link
keywords
Indexes
Foreign Keys
Triggers
Views
Stored Procedures
Functions
Limit to 1000 rows
select
src.name as Source_Airport,
dst.name as Destination_Airport,
count(*) as Delayed_Flights_Count
from Airlines as f
join airports as src
on
f.id = src.id
join airports as dst
on
f.id = dst.id
where
f.Delay > 0
and src.elevation_ft > (select avg(elevation_ft) from airports)
and dst.elevation_ft > (select avg(elevation_ft) from airports)
group by
src.name, dst.name
order by
Delayed_Flights_Count desc;
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
Source_Airport Destination_Airport Delayed_Flights_Cou
1 The Palms At Kitty Hawk Airport The Palms At Kitty Hawk Airport 1
2 NWMC - Houghton Heliport NWMC - Houghton Heliport 1
Result 1
Action Output
Time Action Message Duration / Fetch
Activate Windows
Go to Settings to activate Windows.
Query Completed