

Name- Nayan Raj

Reg no-2347240

Analysis Report

Data cleaning

- There were 2 missing values in the DelayMinute column. I omitted 0 in that column because I cannot tell about the delay without sufficient data. Otherwise it would mislead the analysis.
- The next thing I did was converting all the columns data type to a appropriate data type. Like Date as date type and time as time type.
- There were inconsistency in data like Flight no. AA1234 was flying twice in a same day at same time of departure which was not possible. So I addressed the issue, by keeping only one flight out of the two.

Data Normalization

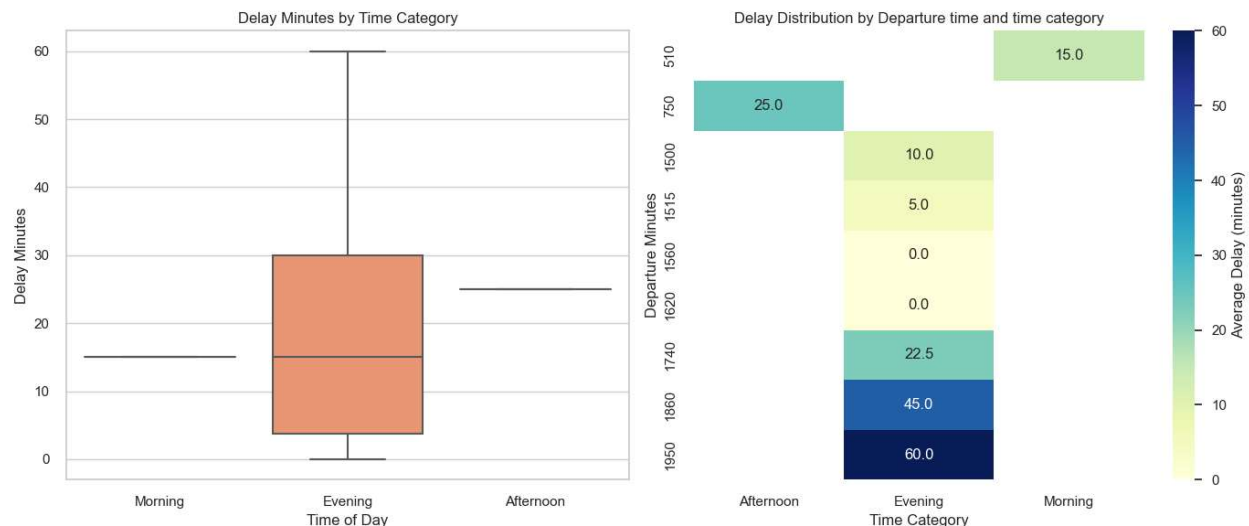
- In this step I converted the arrival date and the departure date in standard format as stated 'yyyy-mm-dd' format from 'mm-dd-yyyy' format.
- I also converted the Departure and the arrival time in 24hr format.

Insight Derived from the analysis

- Flights departing later in the day might experience slightly more delays compared to earlier flights, but this relationship is moderate.
- The data suggests that departure times have a moderate impact on flight delays, with later departures being associated with longer delays.
- There are noticeable differences in delays among airlines, which warrants further investigation into airline performance.

- Both American Airlines and United Airlines show significant delays in the evening, with American being the worst. This could be due to accumulated delays throughout the day, higher air traffic in the evenings, or other operational issues specific to these airlines.
- American Airlines is the only carrier showing data for morning flights, with a relatively low average delay of 15 minutes. This suggests better on-time performance in the mornings, possibly due to less congested airspace or fewer cascading delays from earlier flights.
- Delta appears to have the best overall performance, with only modest delays in the afternoon and no reported delays in other time slots. This could indicate better operational efficiency or possibly a smaller dataset for Delta in this analysis.
- The absence of data for certain time slots (e.g., Delta's morning and evening, United's morning) could indicate either excellent on-time performance or a lack of flights/data for those periods.
- The stark differences between airlines, particularly in evening performance, suggest that airline-specific factors (such as scheduling, maintenance practices, or hub locations) play a significant role in delay patterns.

Visualizations illustrating the key findings.

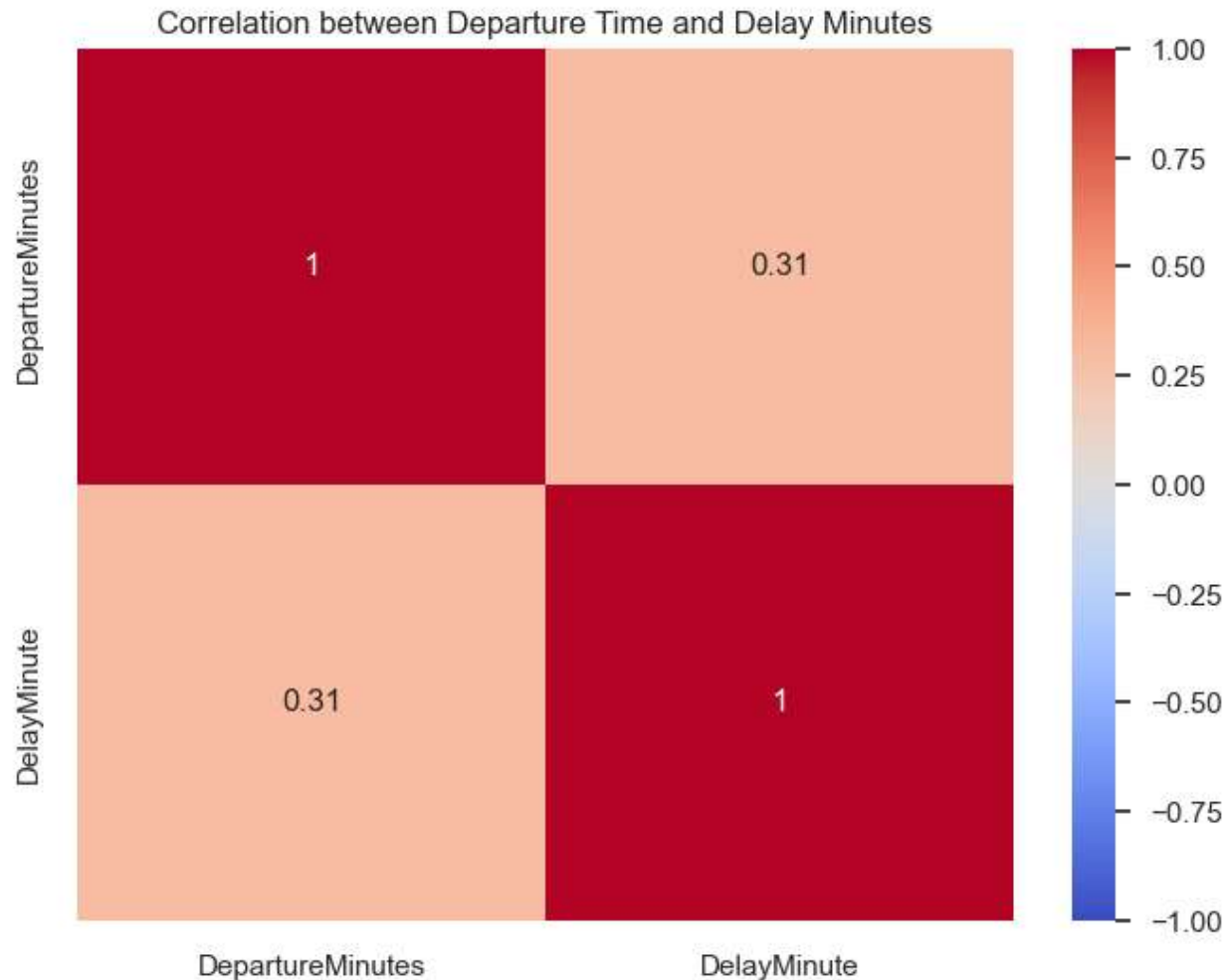


Observations boxplot:

The box plot reveals that evening flights tend to have higher delays on average compared to morning and afternoon flights.

Obsevation Heatmap:

- The heatmap provides a detailed view of how average delays vary with specific departure times.
- The heatmap shows that evening flights (17:00-19:00) tend to have higher
- For instance, the heatmap shows that flights departing in the morning have a lower average delay, while certain evening slots, especially those around 1500 to 1800 minutes (3:00 PM to 6:00 PM), may show significantly higher average delays.
- This visual format allows for quick identification of peak delay times, aiding in understanding how delays are distributed throughout the day.



Insights fom the plot

Diagonal (self-correlation)

The value of '1' along the diagonal ('DepartureMinutes' to 'DepartureMinutes' and 'DelayMinute' to 'DelayMinute') simply indicates that each variable is perfectly correlated with itself.

Off-diagonal correlation (0.31)

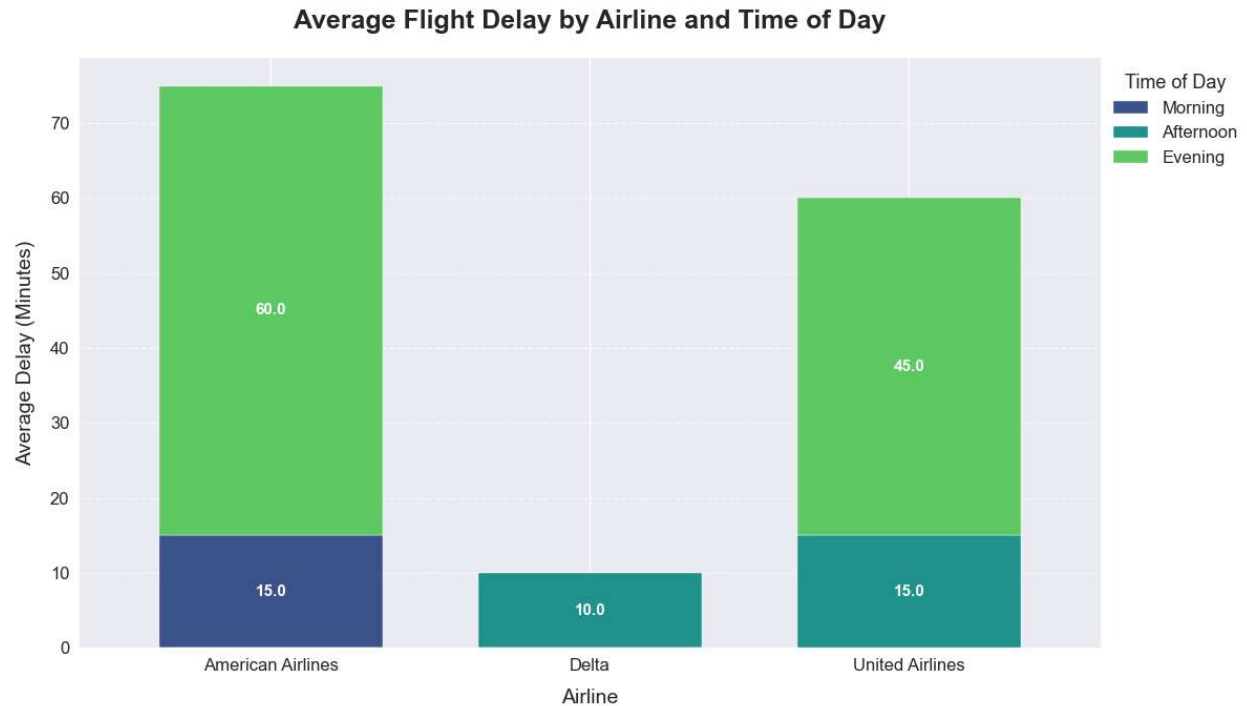
The correlation coefficient between 'DepartureMinutes' and 'DelayMinute' is ****0.31****. This means there is a **moderate positive correlation** between the time of departure and the delay duration.

- **Positive correlation:** As the 'DepartureMinutes' increases (i.e., as flights depart later in the day), the 'DelayMinute' also tends to increase.

- However, the correlation is not very strong, meaning that while there is some relationship between later flights and delays, it is not a very strong predictor.

Conclusion

Flights departing later in the day might experience slightly more delays compared to earlier flights, but this relationship is moderate.



Insights from this plot

Evening Delays:

Both American Airlines and United Airlines show significant delays in the evening, with American being the worst. This could be due to accumulated delays throughout the day, higher air traffic in the evenings, or other operational issues specific to these airlines.

Consistent Morning Performance:

American Airlines is the only carrier showing data for morning flights, with a relatively low average delay of 15 minutes. This suggests better on-time performance in the mornings, possibly due to less congested airspace or fewer cascading delays from earlier flights.

Delta's Efficiency:

Delta appears to have the best overall performance, with only modest delays in the afternoon and no reported delays in other time slots. This could indicate better operational efficiency or possibly a smaller dataset for Delta in this analysis.

Afternoon Variability:

All three airlines show different patterns for afternoon flights, ranging from no data (American) to moderate delays (United and Delta). This variability could be due to differences in route structures, hub operations, or airline-specific scheduling practices.

Data Gaps:

The absence of data for certain time slots (e.g., Delta's morning and evening, United's morning) could indicate either excellent on-time performance or a lack of flights/data for those periods.

Operational Differences:

The stark differences between airlines, particularly in evening performance, suggest that airline-specific factors (such as scheduling, maintenance practices, or hub locations) play a significant role in delay patterns.

Recommendations

For travelers: Consider booking morning flights, especially with American Airlines, to minimize potential delays.

For airlines: American and United should investigate the causes of their significant evening delays and implement strategies to mitigate them.

Further analysis: It would be beneficial to examine a larger dataset to fill in the gaps and confirm if these patterns are consistent over time.

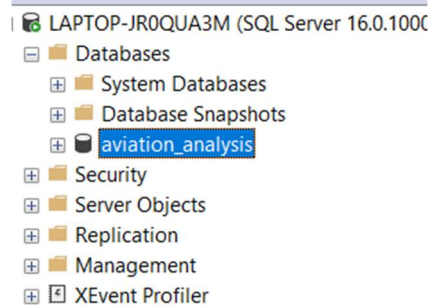
This visualization effectively highlights the variability in airline performance across different times of the day, providing valuable information for both travelers and airline management in understanding and addressing flight delays.

Airlines should investigate the reasons behind higher delays, which could include operational practices, scheduling, and external factors like weather or air traffic.

Further analysis could involve examining specific routes or times of day to provide more targeted insights.

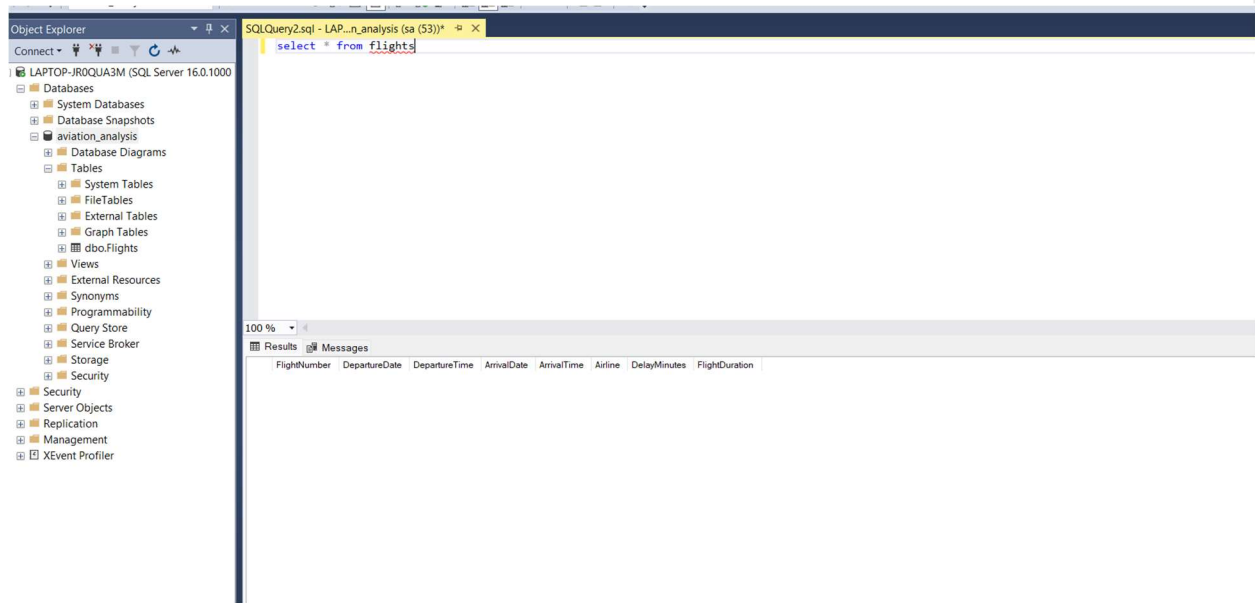
Documentation for creating a pipeline

Step1 – First I created a Database named aviation_analysis in MSSQL using the command create create Database aviation_analysis



Step2- I created a table schema by using the sql query

```
CREATE TABLE Flights (  
    FlightNumber VARCHAR(10) NOT NULL,  
    DepartureDate DATE NOT NULL,  
    DepartureTime TIME NOT NULL,  
    ArrivalDate DATE NOT NULL,  
    ArrivalTime TIME NOT NULL,  
    Airline VARCHAR(100) NOT NULL,  
    DelayMinutes INT,  
    FlightDuration FLOAT,  
);
```

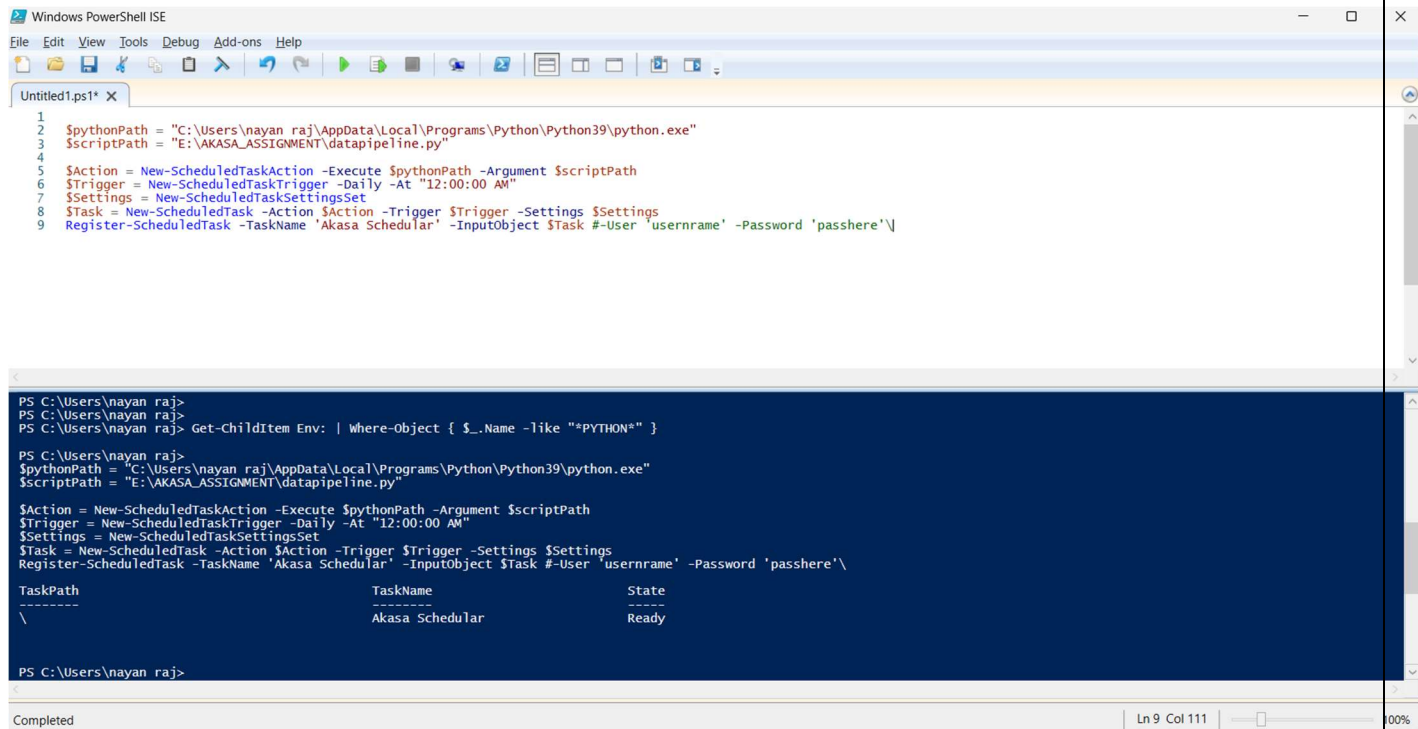


The table is created

Step 3- Created a python script for inserting the Data from formatted.csv to the table in MSSQL

https://github.com/rnayan123/Akasa_assignment/blob/main/datapipeline.py

Step 4- Created a Shell Script using Windows powershell ISE to create a task for automating the insertion in the database at a particular time. Here the task will run automatically at 12 AM



```
Windows PowerShell ISE
File Edit View Tools Debug Add-ons Help
Untitled1.ps1 X
1 $pythonPath = "C:\Users\nayan raj\AppData\Local\Programs\Python\Python39\python.exe"
2 $scriptPath = "E:\AKASA_ASSIGNMENT\datapipeline.py"
3
4
5 $Action = New-ScheduledTaskAction -Execute $pythonPath -Argument $scriptPath
6 $Trigger = New-ScheduledTaskTrigger -Daily -At "12:00:00 AM"
7 $Settings = New-ScheduledTaskSettingsSet
8 $Task = New-ScheduledTask -Action $Action -Trigger $Trigger -Settings $Settings
9 Register-ScheduledTask -TaskName 'Akasa Scheduler' -InputObject $Task #-User 'username' -Password 'passhere'\

PS C:\Users\nayan raj>
PS C:\Users\nayan raj>
PS C:\Users\nayan raj> Get-Childitem Env: | Where-Object { $_.Name -like "*PYTHON*" }

PS C:\Users\nayan raj>
$pythonPath = "C:\Users\nayan raj\AppData\Local\Programs\Python\Python39\python.exe"
$scriptPath = "E:\AKASA_ASSIGNMENT\datapipeline.py"

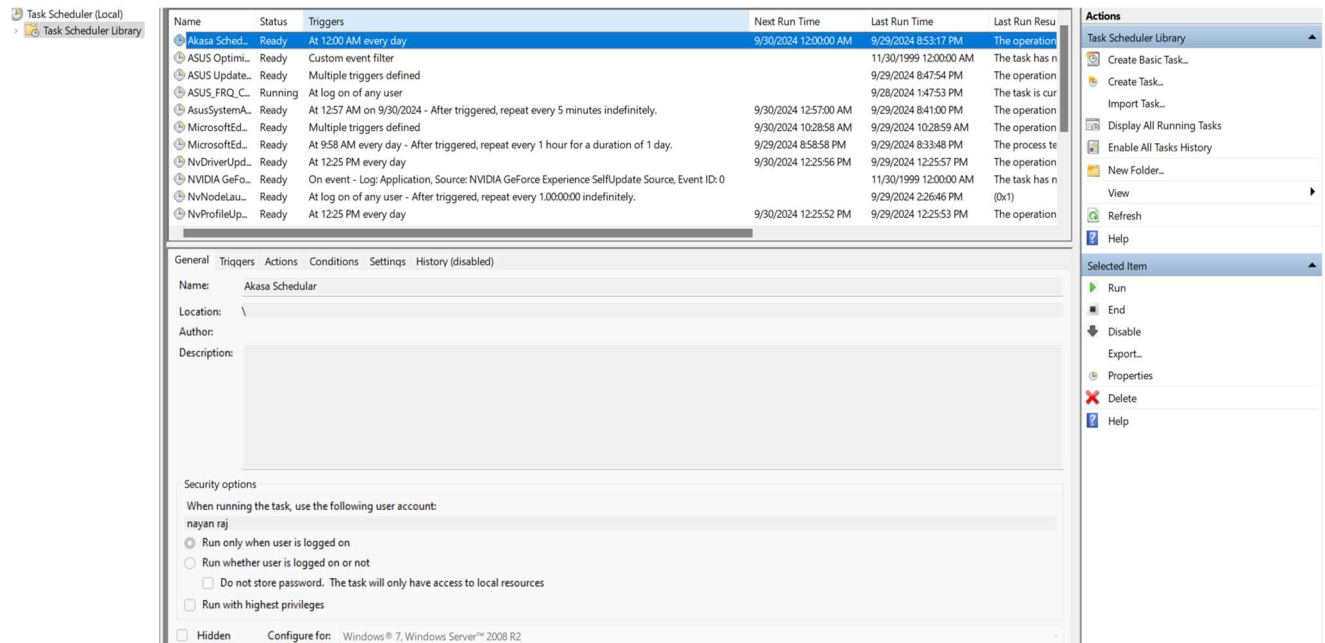
$Action = New-ScheduledTaskAction -Execute $pythonPath -Argument $scriptPath
$Trigger = New-ScheduledTaskTrigger -Daily -At "12:00:00 AM"
$Settings = New-ScheduledTaskSettingsSet
$Task = New-ScheduledTask -Action $Action -Trigger $Trigger -Settings $Settings
Register-ScheduledTask -TaskName 'Akasa Scheduler' -InputObject $Task #-User 'username' -Password 'passhere\'

TaskPath                TaskName                State
-----                -
\                        Akasa Scheduler        Ready

PS C:\Users\nayan raj>
```

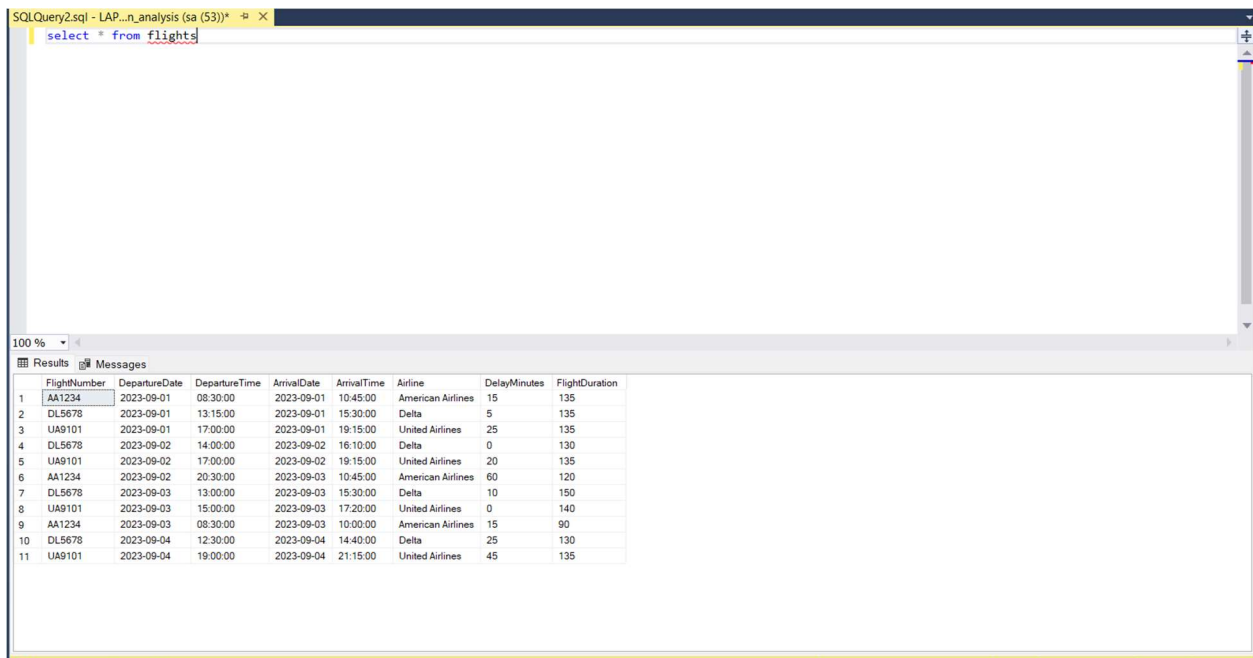
Completed | Ln 9 Col 111 | 00%

Step5- opening windows task scheduler



We can also manually run the task from here. So for now running the task from here to see if the data is inserted in the database or not.

Step6- checking the data in the table if it is inserted or not.



The screenshot shows a SQL query editor window with a query and its results. The query is:

```
select * from flights
```

The results are displayed in a table with the following columns: FlightNumber, DepartureDate, DepartureTime, ArrivalDate, ArrivalTime, Airline, DelayMinutes, and FlightDuration. The table contains 11 rows of data.

	FlightNumber	DepartureDate	DepartureTime	ArrivalDate	ArrivalTime	Airline	DelayMinutes	FlightDuration
1	AA1234	2023-09-01	08:30:00	2023-09-01	10:45:00	American Airlines	15	135
2	DL5678	2023-09-01	13:15:00	2023-09-01	15:30:00	Delta	5	135
3	UA9101	2023-09-01	17:00:00	2023-09-01	19:15:00	United Airlines	25	135
4	DL5678	2023-09-02	14:00:00	2023-09-02	16:10:00	Delta	0	130
5	UA9101	2023-09-02	17:00:00	2023-09-02	19:15:00	United Airlines	20	135
6	AA1234	2023-09-02	20:30:00	2023-09-03	10:45:00	American Airlines	60	120
7	DL5678	2023-09-03	13:00:00	2023-09-03	15:30:00	Delta	10	150
8	UA9101	2023-09-03	15:00:00	2023-09-03	17:20:00	United Airlines	0	140
9	AA1234	2023-09-03	08:30:00	2023-09-03	10:00:00	American Airlines	15	90
10	DL5678	2023-09-04	12:30:00	2023-09-04	14:40:00	Delta	25	130
11	UA9101	2023-09-04	19:00:00	2023-09-04	21:15:00	United Airlines	45	135

We can see here that the required data is correctly inserted in the table.

This was my basic implementation of pipeline.