|  |
| --- |
| TECHNICAL REPORT TEMPLATE |

|  |
| --- |
| **Distributed and Scalable Data Engineering (DSCI-6007)** |

|  |  |
| --- | --- |
| SPRING 23 |  |



Table of Contents

[AIRLINE DELAY ANALYSIS 2](#_Toc133433980)

[EXECUTIVE SUMMARY 2](#_Toc133433981)

[TEAM 3](#_Toc133433982)

[HIGHLIGHTS OF PROJECT 4](file:////Users/trinadhnandamuri/Downloads/Technical%20Report-%20Team10.docx#_Toc133433983)

[SUBMITTED ON:03/22/2023 4](file:////Users/trinadhnandamuri/Downloads/Technical%20Report-%20Team10.docx#_Toc133433984)

[ABSTRACT 5](#_Toc133433985)

[METHODOLOGY 5](#_Toc133433986)

[DATA UNDERSTANDING 5](#_Toc133433987)

[DATA PIPELINE ARCHITECTURE 7](#_Toc133433988)

[RESULTS 8](#_Toc133433989)

[DATA VISUALIZATION 8](#_Toc133433990)

[MODELING 9](#_Toc133433991)

[DEPLOYMENT 11](#_Toc133433992)

|  |
| --- |
| AIRLINE DELAY ANALYSIS |

|  |  |  |
| --- | --- | --- |
| EXECUTIVE SUMMARY Flight delays are one of the most widespread issues on the planet. Given the importance of time in daily life, a flight schedule delay is the most frustrating thing someone can encounter. Missing a planned function or an urgent meeting is possible. Unavoidable flight delays have a big impact on the airlines' profits and losses. Every country, not just the US, has been struggling with this issue for generations. These days, a lot of businesses rely on various airlines to connect them with other parts of the globe, and the aviation sector is crucial to the global transportation sector as well. Statistics show that 20% of airline flights are postponed or cancelled each year, costing passengers $20 billion in lost time and money. Precise flight delay forecasting allows airlines to address the root causes of the delays beforehand and allows passengers to be completely prepared for the disruption caused to their journey in order to tackle this problem. Accurately estimating or projecting an airline's time delay can increase customer satisfaction and income for airline agencies.  The objective of this project is to investigate the approaches used to develop models for forecasting flight delays caused by various causes. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**   * **HARI VISHAL REDDY ANEKALLU**   APPLICATION DEVELOPER   * **VIJAYA BHARGAVI PEDINEEDI**   DATA SCIENTIST.   * **TRINADH NANDMURI**   DATA VISUALIZATION.   * **SRI SAI SINGALA-**   DATA MODELING. | **Questions?**  Contact:  [hanek1@unh.newhaven.edu](mailto:hanek1@unh.newhaven.edu)  [vpedi1@unh.newhaven.edu](mailto:vpedi1@unh.newhaven.edu)  [tnand3@unh.newhaven.edu](mailto:tnand3@unh.newhaven.edu)  [ssing39@unh.newhaven.edu](mailto:ssing39@unh.newhaven.edu) |  |

|  |
| --- |
| TEAM |

|  |  |
| --- | --- |
| **AIRLINE DELAY ANALYSIS** |  |
| HIGHLIGHTS OF PROJECT We'll be employing a variety of supervised machine-learning techniques for this research to get the most precise forecast of whether there will be a flight delay. In this classification issue, the dependent variable (Delay) might have one of two values—0 or 1—where 0 denotes the absence of a delay and 1 denotes the presence of a delay. This information only offers general, static flying information. Information that can change up until takeoff, like as the weather, is not included in this data because it is typically unknown until takeoff**.** SUBMITTED ON: 04/30/2023. |

## ABSTRACT

It is possible to skip a scheduled event or an important meeting due to flight delays. The profitability and losses of the airlines are significantly impacted by unavoidable flight delays. This problem has been in all nations, not just the US, for generations. Many companies now depend on different airlines to connect them to other parts of the world, and the aviation industry is essential to the global transportation industry as well. In order to combat this, airlines can address the underlying causes of the delays with precise flight delay forecasting, which also enables passengers to be fully prepared for the disruption to their journey.

**GitHub link is attached here-** [**https://github.com/Team10-Unicorn/AIRLINE\_DELAY\_ANALYSIS.git**](https://github.com/Team10-Unicorn/AIRLINE_DELAY_ANALYSIS.git)

## METHODOLOGY

This analysis aims to anticipate a delay. To forecast a delay, we will use the airline data from the airline dataset. First, we'll look at the data and perform some simple analysis. Python will be used by us to interact with, access, and visualize our data. Additionally, we'll train machine learning models utilizing the data. Later, an application will be developed and released utilizing the trained machine learning model, allowing any user to provide input to predict whether the flight will be delayed or not.

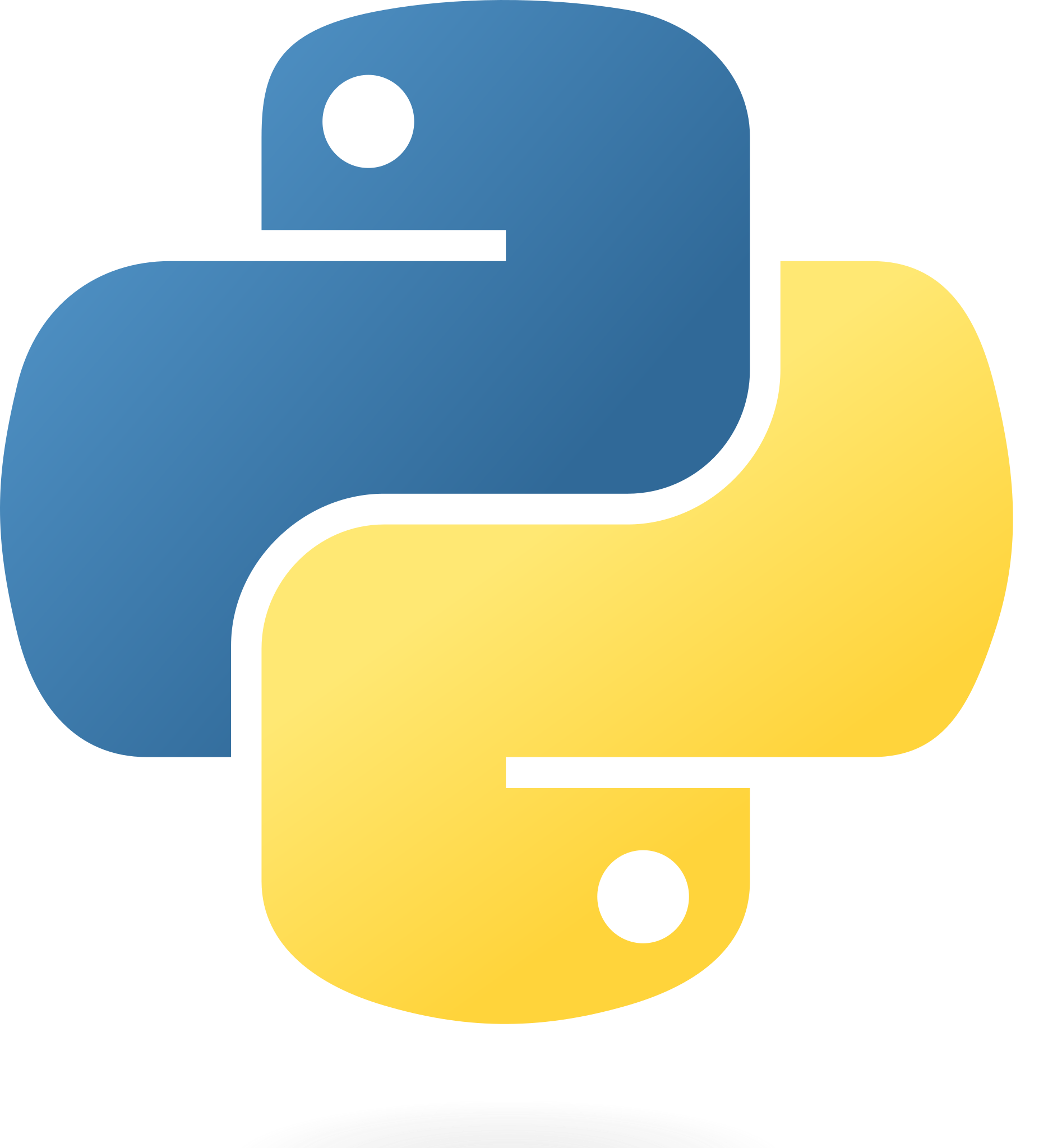
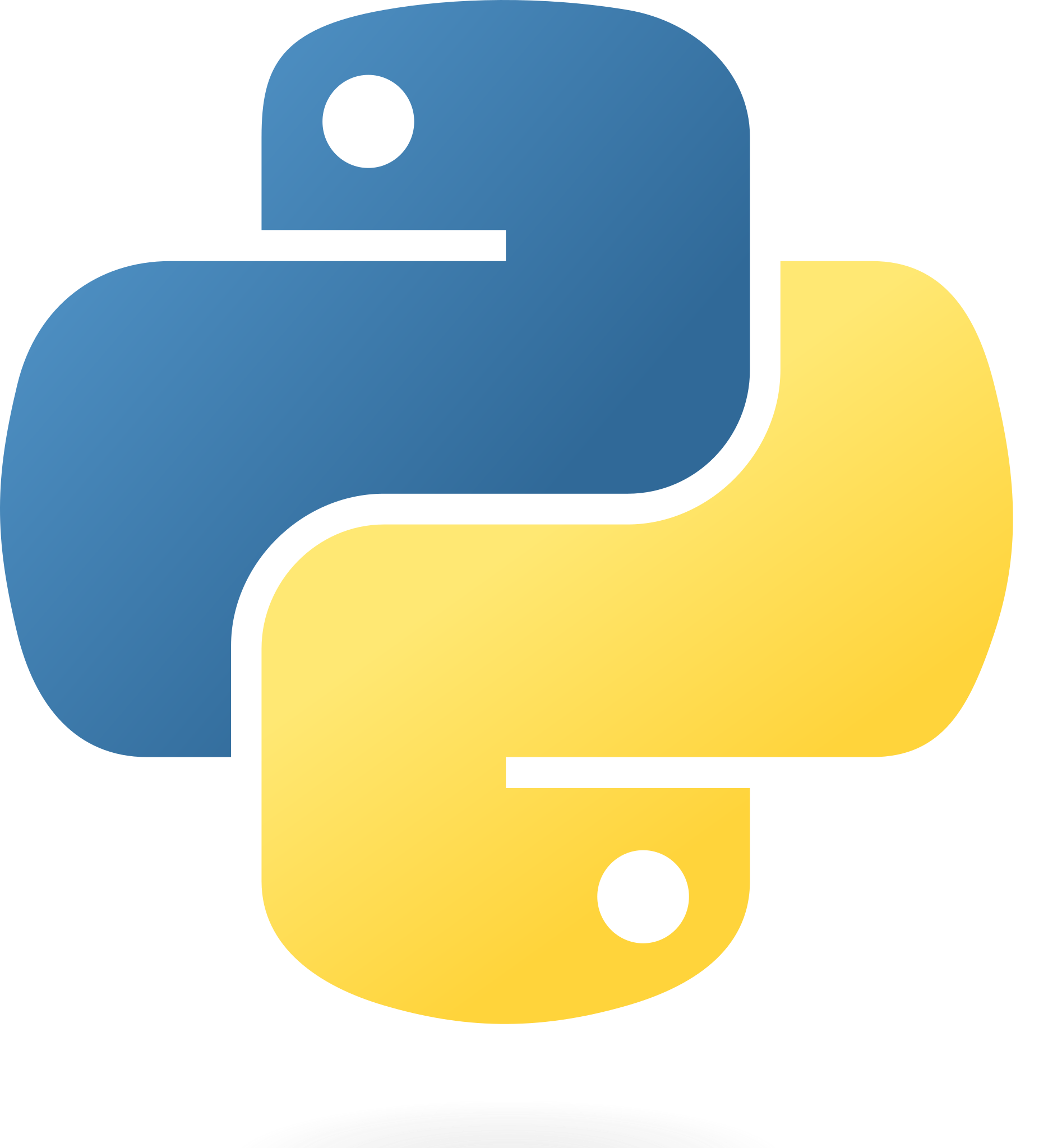
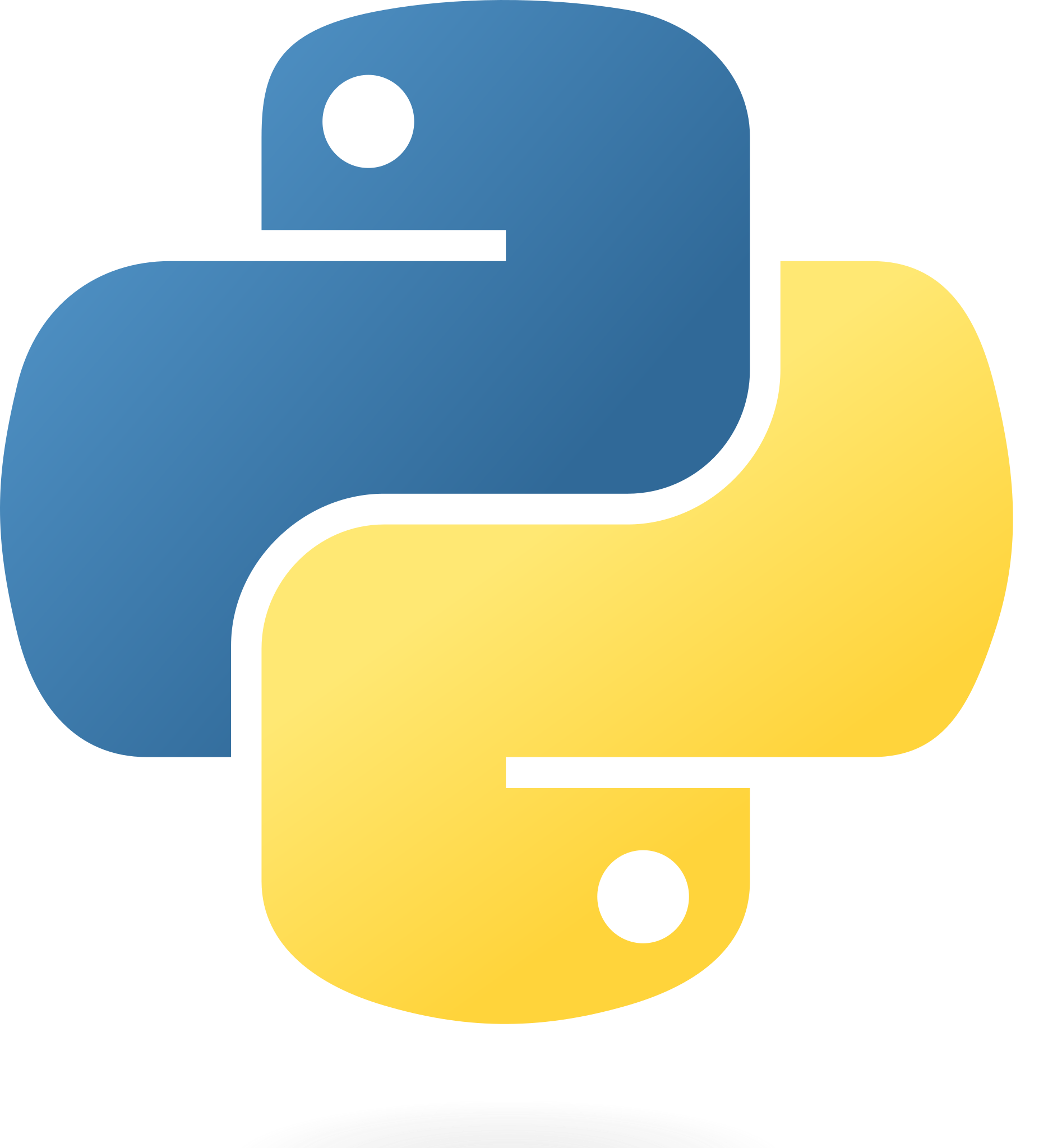
## DATA UNDERSTANDING

The information comes from the Bureau of Transportation Statistics(BTS) of the U.S. Department of Transportation (DOT) . The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the domestic flight schedule performance of major airlines (BTS). The DOT's monthly Air Travel Consumer Report, which is normally released 30 days after the month's end, also includes a summary of the number of on-time, delayed, canceled, and diverted flights in addition to the summary tables provided on this website. BTS began gathering data on the causes of flight delays in June 2003. The general public has access to raw data and summary statistics when the Air Travel Consumer Report is created.

Table

Description automatically generated

## DATA PIPELINE ARCHITECTURE



Snowflake

Snowflake.py

Model.py

Streamlit\_app.py Hosted

on AWS

Model created from Decision Tree

This data pipeline gets the data from snowflake database. The basic dataset is taken from Kaggle. The data pipeline is made of multiple python scripts.

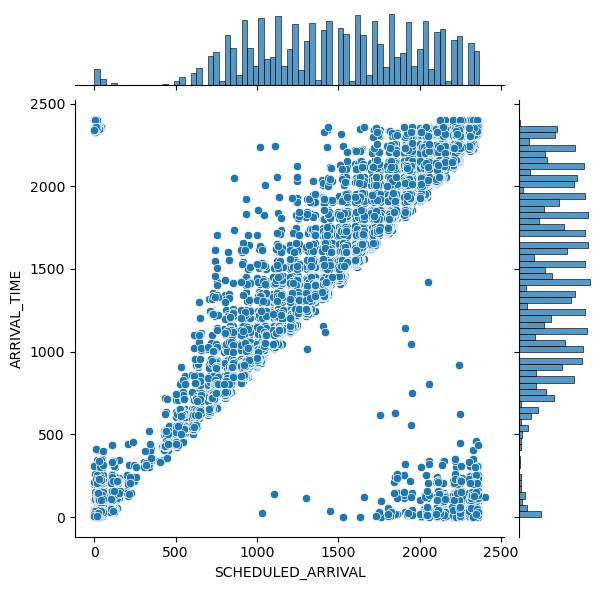
1. Snowflake\_ingestion.py connects to a database in snowflake and downloads the table named data to a local csv file.
2. Model.py uses the downloaded csv file to create a model using decision trees and pickles the model and saves on the local disk.
3. Streamlit\_app.py is used for deployment. This creates a web app where the user can enter the required inputs to get the prediction and to insert the record back into the snowflake.

**RESULTS**

**DATA VISUALIZATION**

The scatter plot between Scheduled arrival time and arrival time. From this graph we can understand weather the flights are arriving on correct time, delay or early.





**Heat Map:** This will explain us how the variables have correlated each other. If the value is 1 or near to 1 then both the variables are highly correlated. If the value is near to 0 that means, there is no correlation between the variables.

Text

Description automatically generated

A picture containing qr code

Description automatically generated

**MODELING**

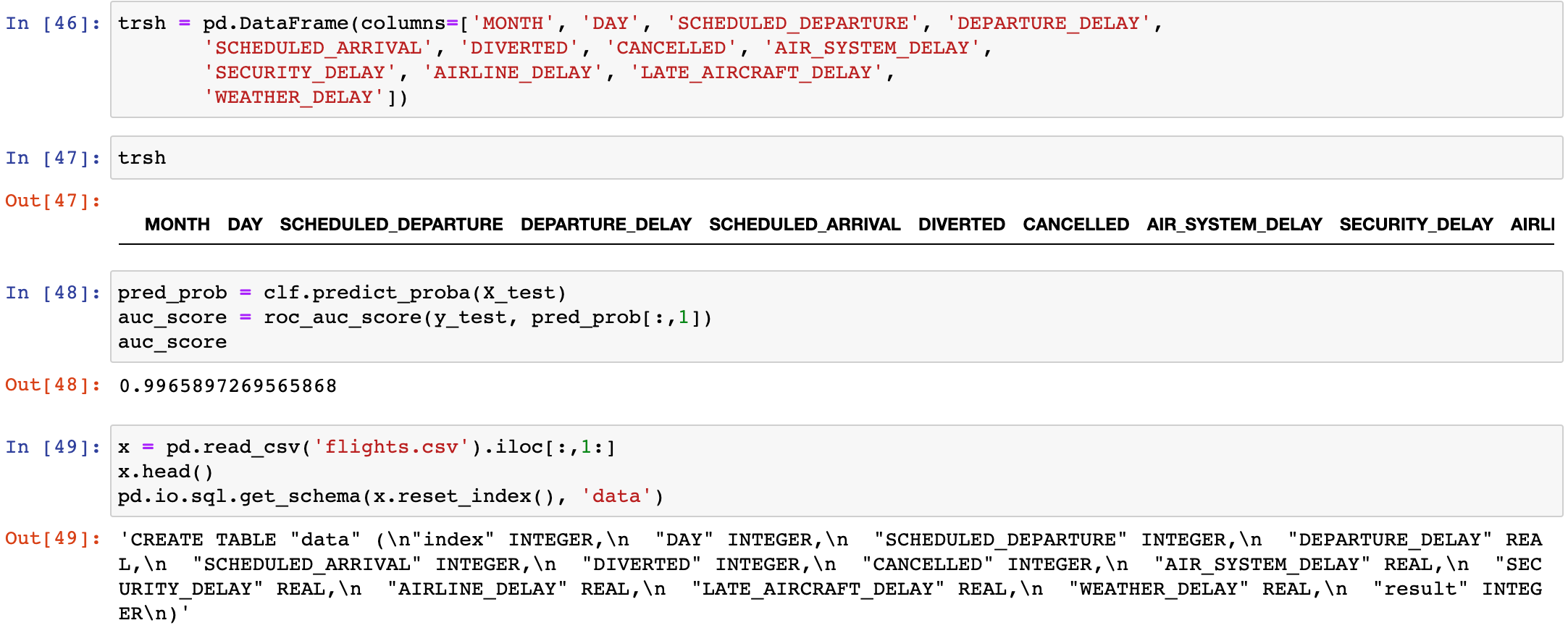
With the help of the sklearn library's train test split () method, we first divided the data into train and test data sets. Pandas are used for data cleaning and data wrangling. The data generated during data preparation is used to train the machine learning models.

Splitting the data into testing and training then applied to decision tree classifier. The data is splitted into 70% of training and 30% for testing.

Table

Description automatically generated

After the model get trained with training data will check the accuracy of the model with testing data. We got the accuracy of the model is 99.52%



**DEPLOYMENT**

Application deployment is done using the flask server. The web application is very helpful for the real time experience. We must fill the values for the variables then it will predict the outcome based on the given values. The values that we have predicted can also be inserted into the database.

