# 2022 U.S. Domestic Airport Delay & Cancellation Prediction

![Delayed (2)](https://user-images.githubusercontent.com/88692025/148868060-eeb992e9-a616-4918-9e6a-f6f33f5bbd5f.jpg)

## Group 10 Final Project - Segment 1

## Google Slides Presentation Link

- Access to Presentation -> https://docs.google.com/presentation/d/1fgzr0\_UD\_RiEuSSxB1nHUACtOhJ9jH2Yjc9kaCJDfmw/edit?usp=sharing

## Team Members & Roles

- Varun Rao (Circle)

- Nada Abubaker (Triangle)

- Herbert Han (Square)

- Saeed Husseini (X)

## Project Overview

The Airline Industry

- According to the Bureau of Transportation Statistics, in the year 2019 there were over 810 million passengers aboard U.S. scheduled airline flights. And that staggering figure is only accounting Domestic flights.

- In the same year of 2019, the US Domestic flight industry racked in almost 160 billion USD in revenue from these passenger flights.

- A quick look at the number of commercial air carriers in the US, with 22 current major carriers in 2021. And smaller air carriers with the 43 counted in 2021.

In 2019, there were 5,217 public airports in the U.S.

Selected Topic

- For our Data Analytics Boot Camp Final project with the University of Toronto, we have decided to take a look at the topic of domestic commercial flights in the US.

- More specifically the data we will examine includes all US domestic flights and information on the airline, how delayed and whether or not the flight was cancelled, the topic covers daily US domestic flights over the course of a 12-month period in 2018.

Reason for Selected Topic

- The reason we have chosen this topic is to help analyze and predict airline and airport data in the future. In order to determine potential variables that impact the delays and cancellations of certain airlines and airports.

- The topic is also truly relevant as it applies to many travelers in the world, with global tourism being the 8th largest industry in the world at 8.7 billion USD annually.

Use Case

- Our project use case would like to help travelers gain time in their planning for trips based on the power of prediction of machine learning. Based on the combination of airline and airport, travelers will receive a warning that their flight at this airport with this airline has an X chance of being delayed or cancelled by X amount of delay. Travelers can learn this ahead of time and plan accordingly. Whether its packing traffic or different arrangements at the destination, the use for this information is countless.

Primary Questions Data will answer

1. Is there a relationship between airline and delays/cancellations?

2. Are there specific airports that are more likely to be delayed/cancelled?

3. Based on the data, can our Supervised Machine Learning Model predict which airlines and airports suffer the most delays/cancellations?

## Description of Data & Source

- Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners.

- Kaggle which started in 2010 offers the ability to download Open Datasets on 1000s of Projects.

- Kaggle allows users to find and publish data sets, explore, and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

- Airline Delay and Cancellation Data, 2009 - 2018

- Flight info. of US domestic flights

- This dataset aims to incorporate multi-year data from 2009 to 2018 to offer additional time series insights.

- Airline Delay and Cancellation Data CSV file for the year of 2018

- All data files are downloaded from OST website, which stores flights on-time performance from 1987 to present.

- The OST contains information from the US Bureau of Transportation Statistics.

## Machine Learning Model

Data Pre-processing/Cleaning

Deciding the definition of a delayed flight, which variables to keep in the model, which variables to discard.

Supervised machine learning model aimed at predicting flight delays at the Origin Airport.

Using historical delay information from 2016-2018 from US domestic airports and flights.

- Random forest Regression

- Bagging Regressor

- Linear Regression

## Database

![Database](https://user-images.githubusercontent.com/88692025/149708413-5a7d9a7a-5cc1-4731-9e5f-f383848ded1c.PNG)

- The Database table structures accommodates the fields in the file

- In addition, we add fields for timestamps combining the data and time portions

- The table has a composite primary key for ensuring uniqueness and efficient data retrieval

- Based on the type of usage we plan to add additional indexes for efficient data retrieval

- The database technology used is PostgreSQL and it will be hosted on AWS

I am most proud of finding a solid purpose and use case for the project. It is important that the data has a valuable real-world implication that helps people improve the quality of their life. In this case the data will help travelers around the US in saving time and learning valuable delay and cancellation information about their airlines and airports.

Working as a team to find a timely manner to brainstorm for a solution, using group think to help narrow down options for the database challenge of the final project.