BIG DATA MANAGEMENT - ASSIGNMENT 2

MINING OVER DATASETS

August 12, 2024

Anastasia Drakou - 2022202304006
Anna Koutougera - 2022202304012
Despoina Angeliki Moisidou - 2022202304016

Contents

1 Dataset 1 : Airlines Dataset

Before delving into the details of the Airline dataset, let's provide some contextual information about the dataset to facilitate a more thorough analysis.

- DayofWeek: 1 (Monday) - 7 (Sunday)
- CRSDepTime: Scheduled departure time (local, hhmm)
- UniqueCarrier: unique carrier code, which means are two character designator for the carrier/airline.
- FlightNum: Flight number
- Origin: Origin IATA airport code
- Dest: Destination IATA airport code
- ArrDelay: Arrival delay, in minutes, which means the total number of delayed minutes due to delays.

After some research, it is found that:
A flight is counted as "on time" if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS).
The United States Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time as can be seen from here

1.1 Task 1 : Provide an overview of the dataset size, features, and distribution of feature values

The dataset at hand is an Airlines Dataset, drawing inspiration from Elena Ikonomovska's regression dataset. It provides useful information about airlines and flights for our analysis. Let's take a look at some data from the dataframe.

	DayofWeek	CRSDepTime	UniqueCarrier	FlightNum	Origin	Dest	ArrDelay
0	5	600	UA	899.0	ORD	IAD	-3
1	5	615	DL	357.0	IAD	ATL	5
2	5	615	UA	341.0	IAD	DEN	8

Figure 1: Airlines Dataset

The Airline dataset, as we can observe, contains 7 features and has in total 100,161 rows. Each feature has been previously explained, and each row in the dataset represents a specific flight. Now, let's delve deeper into our dataframe. For each feature, we have :

Table 1: Information for each feature

Feature Name	category	Type	Null Values.	Null Values (using isna)
DayofWeek	nominal	int64	0	0
CRSDepTime	numeric	int64	0	0
UniqueCarrier	nominal	object	0	0
FlightNum	numeric	float64	0	0
Origin	nominal	object	0	0
Dest	nominal	object	0	0
ArrDelay	numeric	int64	0	0

A statistical summary is crucial in providing valuable insights on how to approach a dataset. Below, you will find the statistical summary of the airline dataset.

	DayofWeek	CRSDepTime	UniqueCarrier	FlightNum	Origin	Dest	ArrDelay
count	100161.000000	100161.000000	100161	100161.000000	100161	100161	100161.000000
unique	NaN	NaN	9	NaN	58	59	NaN
top	NaN	NaN	UA	NaN	IAD	IAD	NaN
freq	NaN	NaN	63706	NaN	50229	49932	NaN
mean	3.952626	1357.918262	NaN	896.152654	NaN	NaN	4.638462
std	1.985074	460.022924	NaN	476.977517	NaN	NaN	23.397863
min	1.000000	5.000000	NaN	12.000000	NaN	NaN	-72.000000
25%	2.000000	925.000000	NaN	539.000000	NaN	NaN	-7.000000
50%	4.000000	1335.000000	NaN	1000.000000	NaN	NaN	0.000000
75%	6.000000	1735.000000	NaN	1264.000000	NaN	NaN	9.000000
max	7.000000	2359.000000	NaN	4007.000000	NaN	NaN	667.000000

Figure 2: Statistical Summary

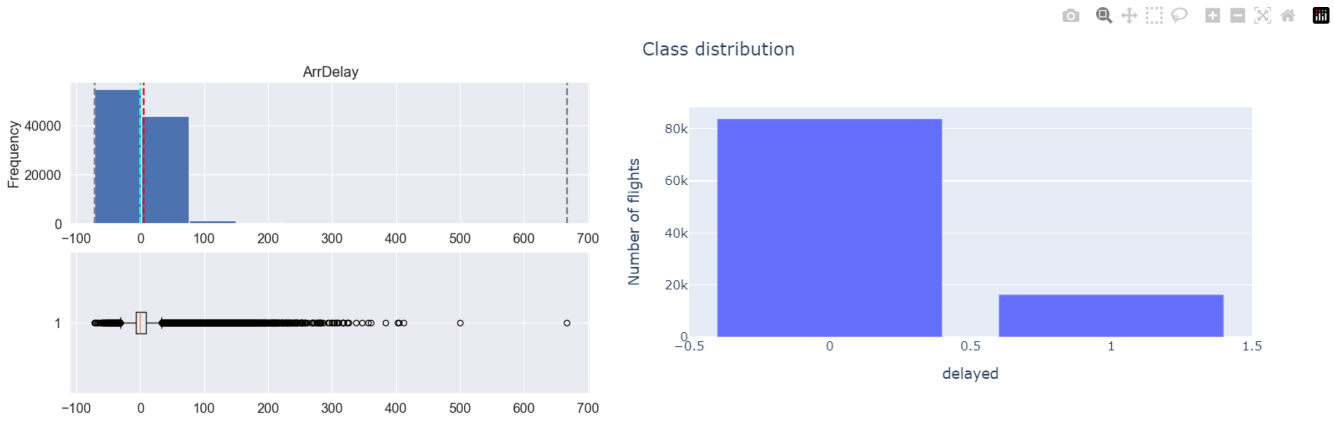
This dataset represents flight arrival delays for 9 airlines in the United States.It contains data points for delay prediction like Airline, Flight no, Origin and Destination Airport, Day of travel, time of departure and if that particular flight is delayed or on time.

```
Unique values in each column:
DayofWeek          7
CRSDepTime       424
UniqueCarrier      9
FlightNum        624
Origin            58
Dest              59
ArrDelay         350
dtype: int64
```

Figure 3: Unique values of each feature

In this section, we will present important and detailed information about each feature. We will depict distributions and some assumptions based on the graphs.

ArrDelay :



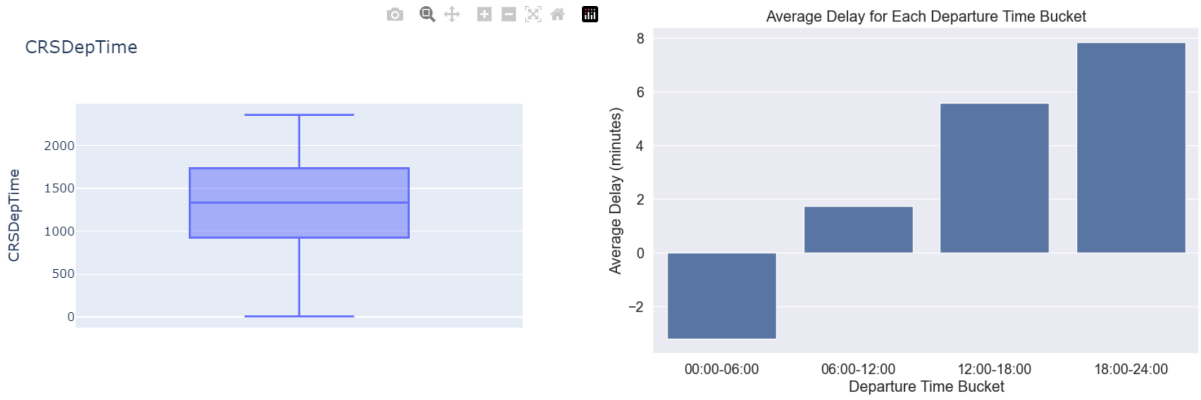
(a) Distribution and Box Plot

(b) Bar Plot

Figure 4: ArrDelay distribution

In the graph on the left, we observe that the majority of flights arrive on time, as indicated by the mode and the distribution. However, there is an average small delay typically around 4-5 minutes. The statistical summary reveals a minimum delay of -72 minutes, a maximum delay of 667 minutes, a mean delay of 4.64 minutes, a median delay of 0.0 minutes, and a mode of 0.
On the right side, flights are classified into two classes: Class 0 indicates no delay, and Class 1 indicates a delay. The graph illustrates that we have imbalanced data, with the minority class being the delayed class with a value of 1.

CRSDepTime :



(a) Box Plot **(b)** Bar Plot

Figure 5: CRSDepTime distribusion

In the left-side photo, the Box Plot of CSRDepTime is presented, revealing the absence of outliers in this feature. Additionally, on the right side of the Bar Plot, the average delay for each departure time is displayed in four buckets. Each bucket represents:

- 0 - DepTimeBucket: 00:00-06:00 ArrDelay: -3.222798
- 1 - DepTimeBucket: 06:00-12:00 ArrDelay: 1.749162
- 2 - DepTimeBucket: 12:00-18:00 ArrDelay: 5.572461
- 3 - DepTimeBucket: 18:00-24:00 ArrDelay: 7.847698

It appears that during 00:00-06:00 in the morning, there are no flight delays which seems understandable because this is the least busy time for an airport. In contrast, as we progress in time in the day, we see the delays increasing reaching peak during the afternoon, when it is most busy.

We observe that the majority of flights, as indicated by the mode value, depart at 7:00 in the morning. Additionally, popular departure times are typically distributed between 12:00 and 20:30. The statistical summary for the 'CRSDepTime' feature provides insights into the distribution of scheduled departure times.

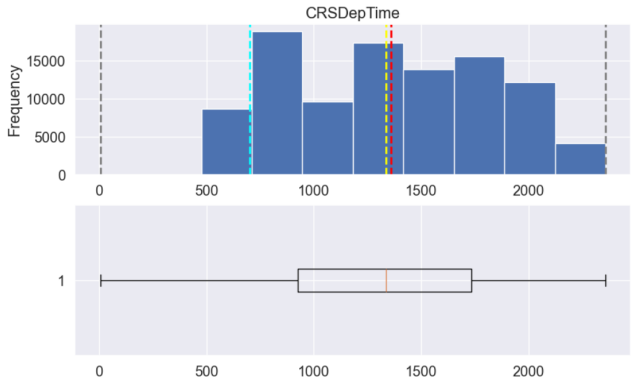
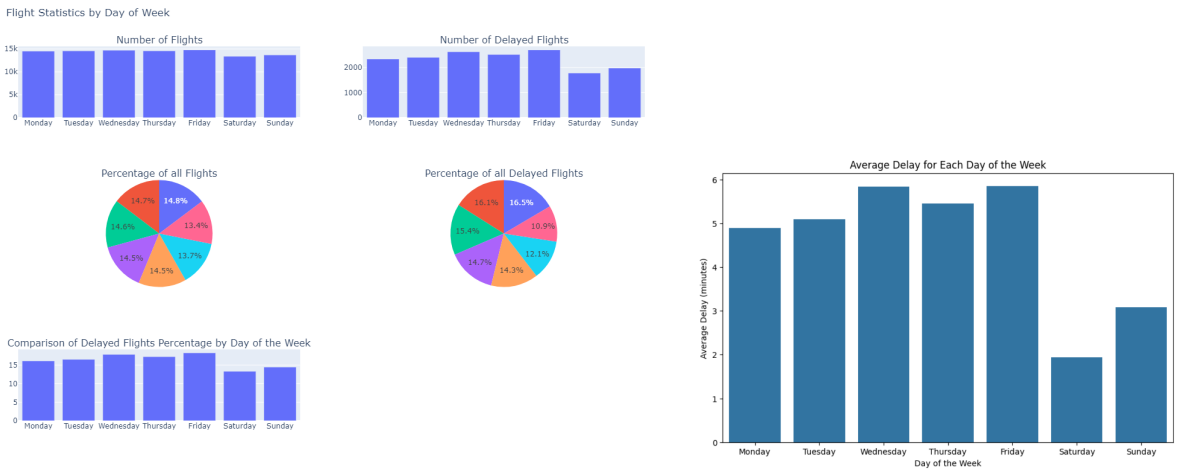


Figure 6: Unique values of each feature

The minimum departure time is 5, indicating the earliest scheduled departure, while the

maximum is 2359, signifying the latest scheduled departure. The mean departure time is 1357.92, and the median is 1335. The mode, representing the most frequently occurring departure time, is 700, suggesting a concentration of flights departing at 7:00 in the morning.

DayofWeek :



(a) Bar Plots and Graph Pie **(b)** Bar Plot

Figure 7: DayofWeek distribution

The data reveals a consistent number of flights on weekdays, with a noticeable decline during weekends, particularly on Saturdays—the day with the fewest flights. One plausible explanation is the potential higher labor costs associated with weekend operations. Airlines might have opted for a slightly reduced flight schedule on weekends to manage employee-related expenses.
Furthermore, the observation that Friday experiences the highest flight volume suggests a pattern where individuals embark on weekend getaways, departing on Friday and returning on Sunday. This could contribute to the dip in flights on Saturdays, followed by a gradual recovery in flight numbers on Sundays, aligning with the return of weekend travelers.
It is also interesting to notice that in the day where we have less flights, we also notice less delays.

1.2 Task 2 : Describe the average delays per airport/airline

On this task we will discuss different ways and approaches in order to describe the average delays.

UniqueCarrier

We will start by representing the number of flights per airline, but before we show our data it is important to explain a noise which we fixed in order to have a efficient dataset. We've noticed an odd airline type labeled 'PA (1)', which seems to be linked to Airblue's IATA code. However, this code originally belonged to the no longer operational Pan American World Airways. To correct this, we need to simplify it to just 'PA'.

```
Number of flights per airline:
UniqueCarrier
UA          63706
CO           9219
AA           8620
NW           5521
DL           4793
US           3513
TW           3056
EA           1420
'PA (1)'      313
Name: count, dtype: int64
```

(a) Airline type 'PA (1)'

```
Number of flights per airline:
UniqueCarrier
UA    63706
CO     9219
AA     8620
NW     5521
DL     4793
US     3513
TW     3056
EA     1420
PA      313
```

(b) Fixed Airline PA

Figure 8: Number of Flights per airline

The following Flight Statistics will help us to gain a better picture of the data:

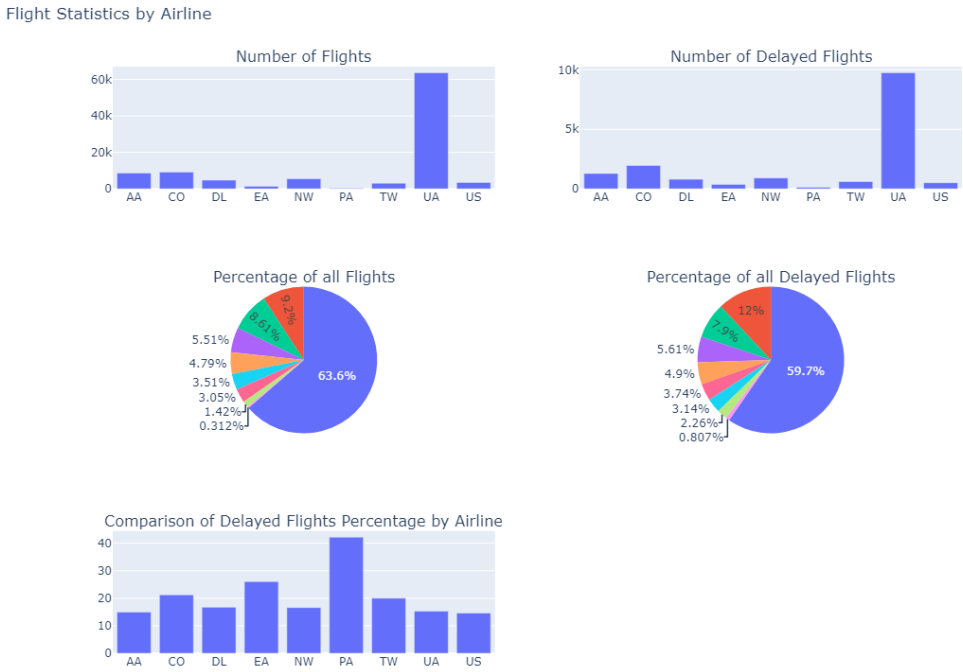


Figure 9: Flight Statistics

From figure above we see that the airlines with the most domestic flights in the US are:

- UA (United Airlines) by far (63%)
- CO (Continental Airlines)
- AA (American Airlines)

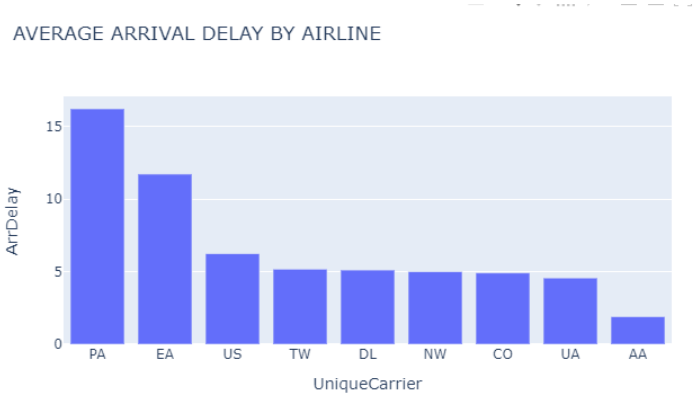
Also, it is interesting to notice that while UA airline experiences the highest number of delayed flights in absolute terms (refer to the first two figures), these delays constitute only 15% of their total flights. Conversely, PA, despite having the lowest overall and delayed flight counts, exhibits a higher proportion of delays, accounting for 42% of its total flights, as evident in the last figure.This suggests that nearly half of PA's flights encounter delays. A similar observation applies to other less popular airlines such as EA and TW , despite their lower popularity, they exhibit a notable percentage of delayed flights.

Since, the best airline would be the one with the least average delays, the average minute delays will be sorted in ascending order below.

At first we are going to take into consideration all kinds of delays, even the negative ones, indicating that a flight arrived earlier than expected

	UniqueCarrier	ArrDelay
0	PA	16.223642
1	EA	11.728169
2	US	6.246798
3	TW	5.179647
4	DL	5.118715
5	NW	5.004890
6	CO	4.919948
7	UA	4.569664
8	AA	1.908237

(a) Arrival Delay

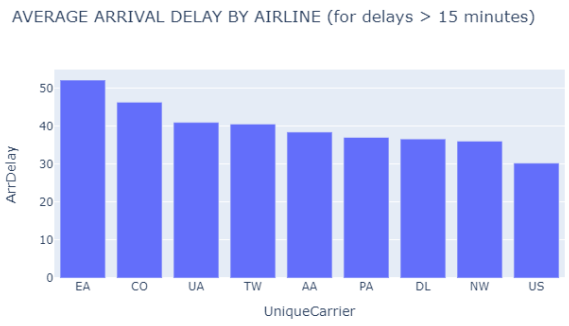


(b) Distribution of Arrival delay per Airline

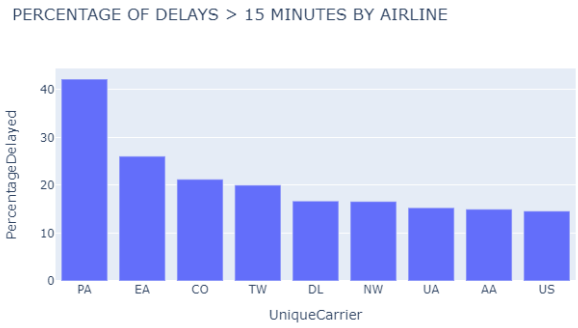
Figure 10: Arrival delay per Airline

It is interesting to note here again, that less popular airlines and especially PA which is the least popular, seem to have the biggest average total delays, while the contrary can be said for more popular airlines.

Now, let's take a deeper look and examine instances where delays exceed 15 minutes, categorizing a flight as delayed.



(a) Average arrival delay by airline (delays > 15 minutes)



(b) Percentage of delays > 15 minutes by airline

Figure 11: Delays by Airline > 15 minutes

Airport (Origin-Destination) : Now we will see the top 30 destination cites

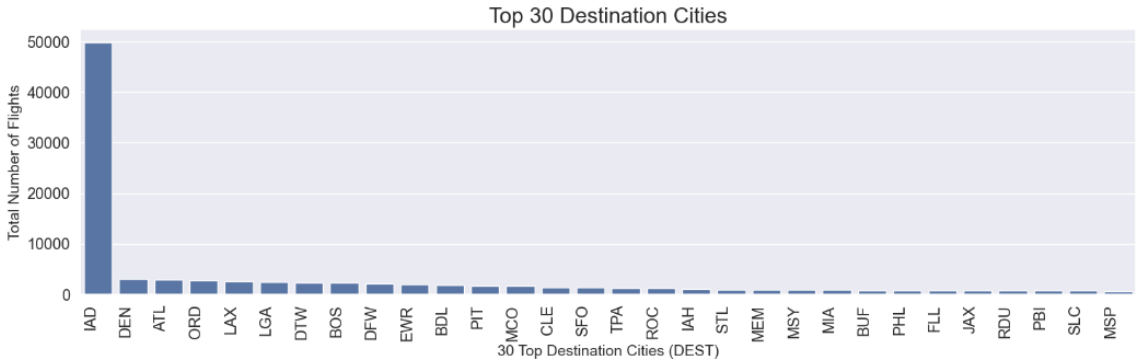
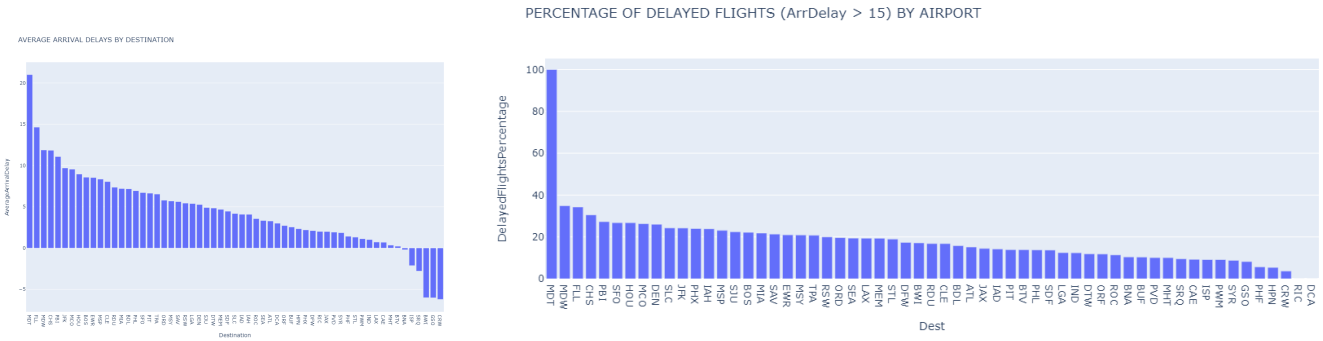


Figure 12: 30 Top destination Cities

It appears that Washington Dulles International Airport (IAD) is the most frequently traveled to location.

After determining the average arrival delays for each destination, we will illustrate the data in a bar plot, which you can view in the following images. The left-side graph illustrates the average arrival by destination, considering both on-time arrivals and those that arrived earlier. On the right-side graph, the focus shifts to the percentage of arrival delays, accounting solely for the flights that experienced delays.



(a) Average arrival delays by destination

(b) Percentage of delayed flights (ArrDelay > 15 minutes) by airport

Figure 13: Delaed Flights by airport

We see that all flights arriving in MDT airport are delayed. That is because though in this airport only one flight has arrived. It is more interesting to look at the airports with a lot of flights like IAD , DEN, ATL that have a 15-25 % percent of delayed flights. It would be interesting to see which airlines fly to each airport or which travels are the most popular and check if certain route by certain airlines tend to be delayed more often.

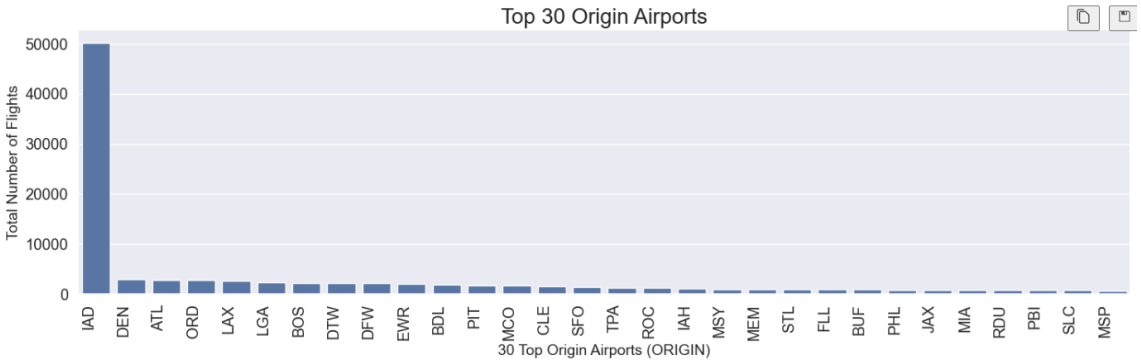
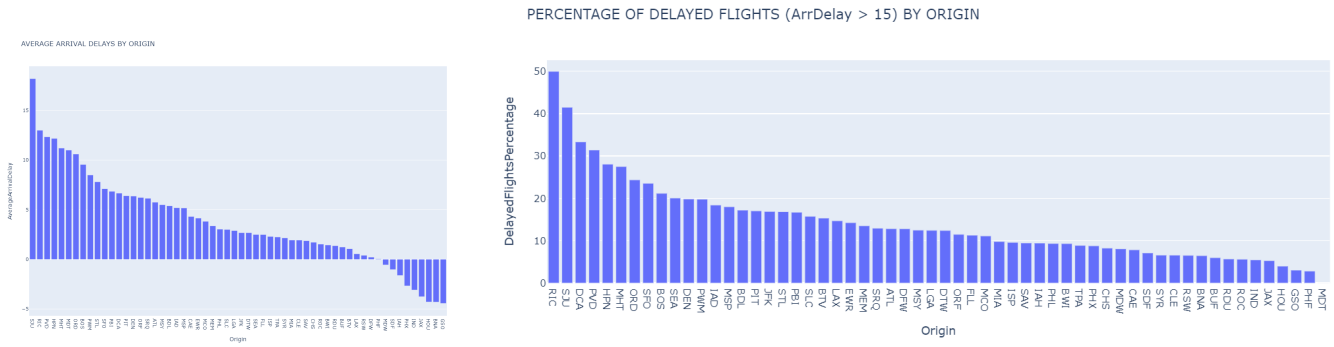


Figure 14: Top 30 Origin Airports



(a) Average arrival delays by origin

(b) Percentage of delayed flights (ArrDelay > 15) by origin

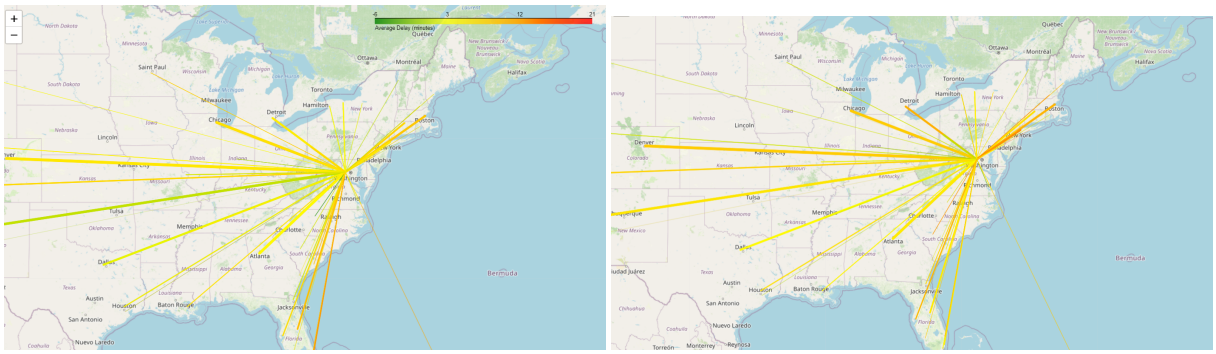
Figure 15: Arrival Delayed Flights

As we see from the first figure the most popular origin airports are the same as the most popular destination airports. That must be because these are popular routes to take or are in-between flights. It is also interesting that certain airports like GSO, BNA and HOU in general seem to not have as many delays. On the contrary a lot of flights seem to arrive earlier as well.

The map below illustrates the geographical positions of the airports.

Figure 16: Airports on the map

We can clearly see that the most airports are located in the eastern side of America.



(a) Average delay per route (taking account all the routes)

(b) Average delay per route (taking account only the delayed routes)

Figure 17: Average delay per route

These maps display the flight routes, with colors indicating delays. Orange-Red colors signify longer delays, while yellow-green colors represent routes with shorter or without earlier arrivals. Additionally, the thickness of the lines reflects the frequency of each route. The left map takes account all the routes, but the right map takes account only the delayed routes. It appears that flights between Washington and Florida, Miami and Boston show some delay in contrast with for example Los Angeles.

Analyzing the average flight delays using a bar plot that is sorted in descending order is an interesting topic to investigate. This graphic can give important insights about the amount of

delays experienced by various aircraft by presenting a thorough picture of the general pattern and emphasizing the flights with the biggest delays.

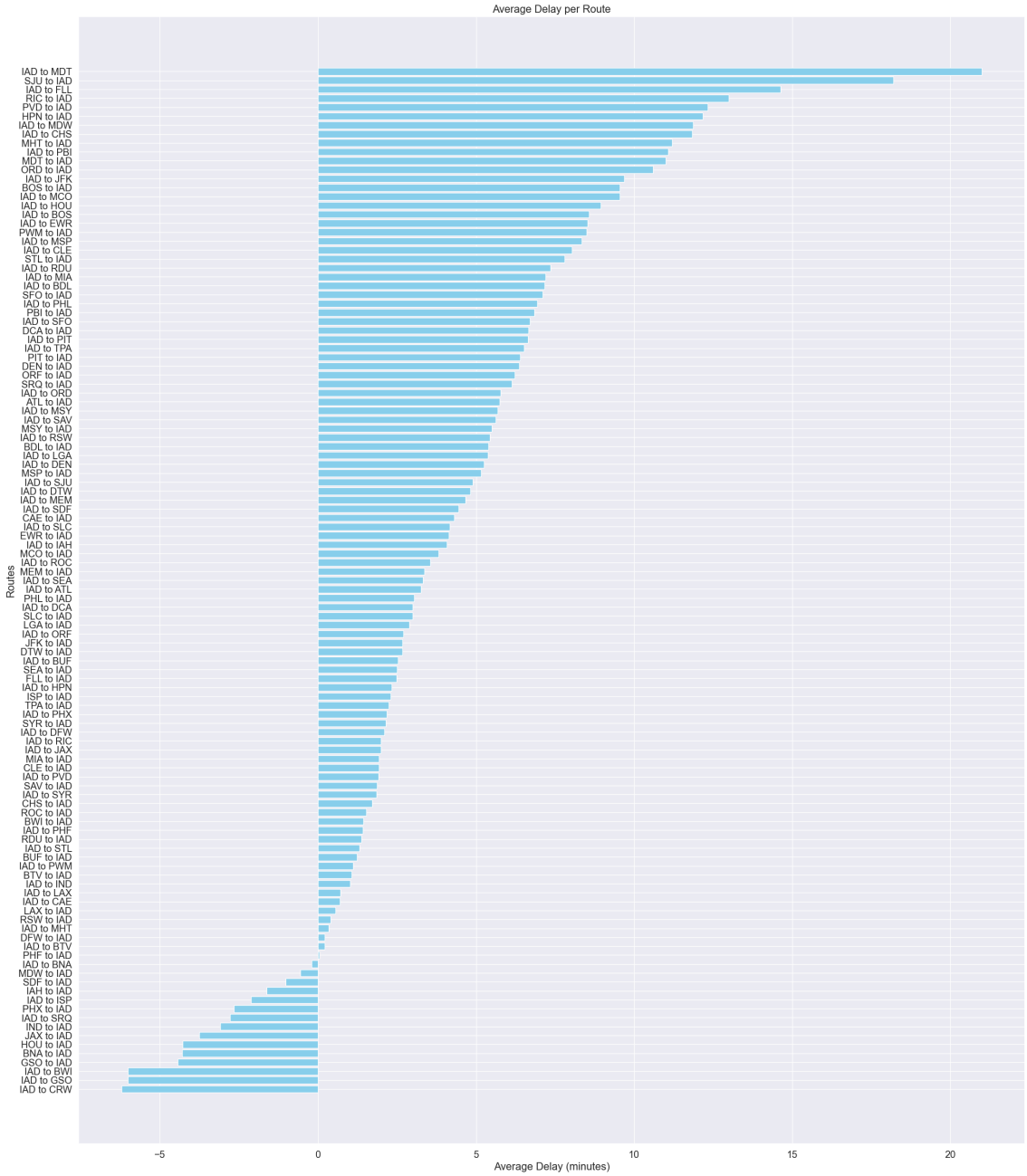


Figure 18: Average Delay in descending order

With a large number of routes, the above bar plot might has become crowded and difficult to interpret. One alternative is to plot only the top N routes with the highest average delays.

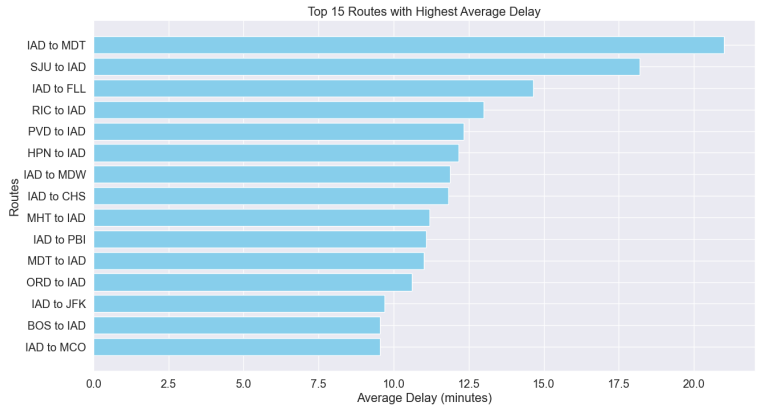


Figure 19: Top 15 Routes with Highest Average Delay

So we observe that for certain routes like IAD to CRW it is likely for the flight to arrive early, but for other types of routes like SJU to IAD there is a lot of delay.
Let's check for these cases which airlines had these routes.

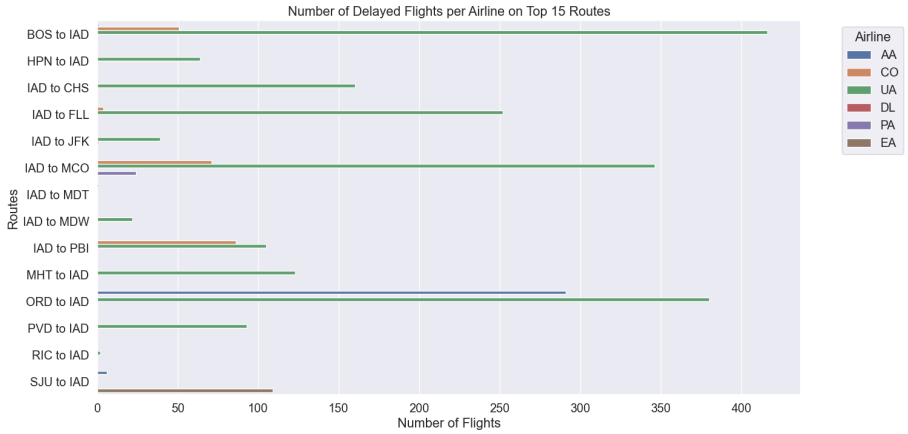


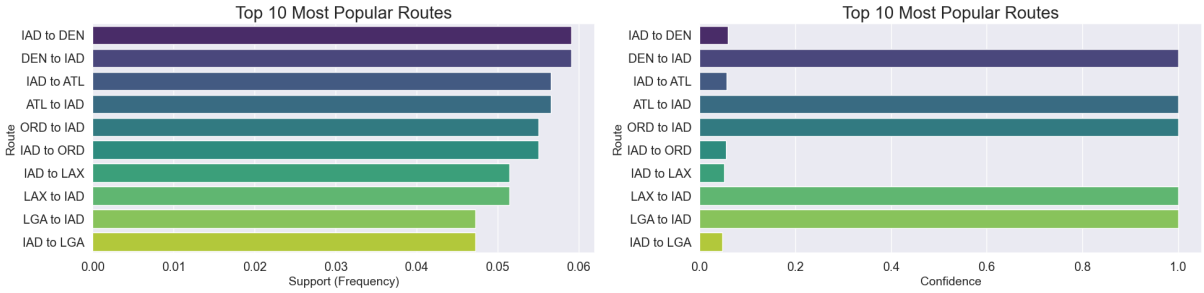
Figure 20: Number of Delayed Flights per Airline on Top 15 Routes

1.3 Task 3 : Identify and report the most prominent rules of association between [delays] and [point of origin AND/OR point of arrival]

Frequent patterns are patterns that appear frequently in a data set. We want to indicate association between items.
A rule is measured not only by its support and confidence but also by the correlation between item-sets A and B (Lift).
The Apriori algorithm is a key method that is essential for identifying popular item sets.
The method finds item-sets that meet a minimal support criterion via an iterative technique,

creating strong correlations between qualities.
Its main function in associative categorization is to produce a set of frequent item sets from which association rules may be derived.

First let's check most popular routes based on their support value and confidence value.We're specifically interested in **arrival** delays, so we're narrowing our focus to routes originating from specific destinations.



(a) Support **(b)** Confidence

Figure 21: Top 10 Most Popular Routes

Nothing noteworthy can be observed from here, other than the fact that all routes departing from IAD airport have extremely low confidence, while all routes arriving at IAD airport have significantly higher confidence. So it's easier to predict where flights are coming from when they arrive at IAD compared to where they are going when they depart from IAD.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(ATL)	(IAD)	0.048569	1.0	0.048569	1.0	1.0	0.0	inf	0.0
1	(BDL)	(IAD)	0.036151	1.0	0.036151	1.0	1.0	0.0	inf	0.0
2	(BOS)	(IAD)	0.058600	1.0	0.058600	1.0	1.0	0.0	inf	0.0
3	(CHS)	(IAD)	0.012479	1.0	0.012479	1.0	1.0	0.0	inf	0.0
4	(CLE)	(IAD)	0.020431	1.0	0.020431	1.0	1.0	0.0	inf	0.0
5	(DEN)	(IAD)	0.083252	1.0	0.083252	1.0	1.0	0.0	inf	0.0
6	(DFW)	(IAD)	0.040066	1.0	0.040066	1.0	1.0	0.0	inf	0.0
7	(DTW)	(IAD)	0.032726	1.0	0.032726	1.0	1.0	0.0	inf	0.0
8	(EWR)	(IAD)	0.042023	1.0	0.042023	1.0	1.0	0.0	inf	0.0
9	(FLL)	(IAD)	0.021593	1.0	0.021593	1.0	1.0	0.0	inf	0.0
10	(IAH)	(IAD)	0.022877	1.0	0.022877	1.0	1.0	0.0	inf	0.0
11	(LAX)	(IAD)	0.053707	1.0	0.053707	1.0	1.0	0.0	inf	0.0
12	(LGA)	(IAD)	0.036090	1.0	0.036090	1.0	1.0	0.0	inf	0.0
13	(MCO)	(IAD)	0.038414	1.0	0.038414	1.0	1.0	0.0	inf	0.0
14	(MEM)	(IAD)	0.018534	1.0	0.018534	1.0	1.0	0.0	inf	0.0
15	(MHT)	(IAD)	0.010276	1.0	0.010276	1.0	1.0	0.0	inf	0.0

Figure 22: Association rules(First 18 rows)

We will only concentrate on the top routes that are **delayed** in order to gain a better understanding of the rules based on the delays. The results will be as follows.:

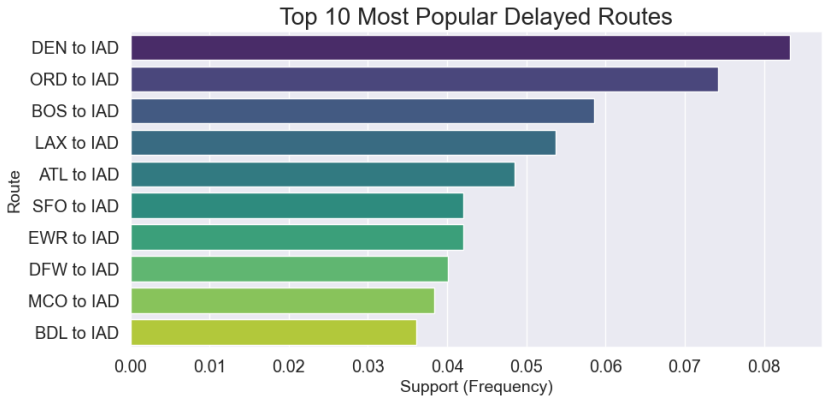
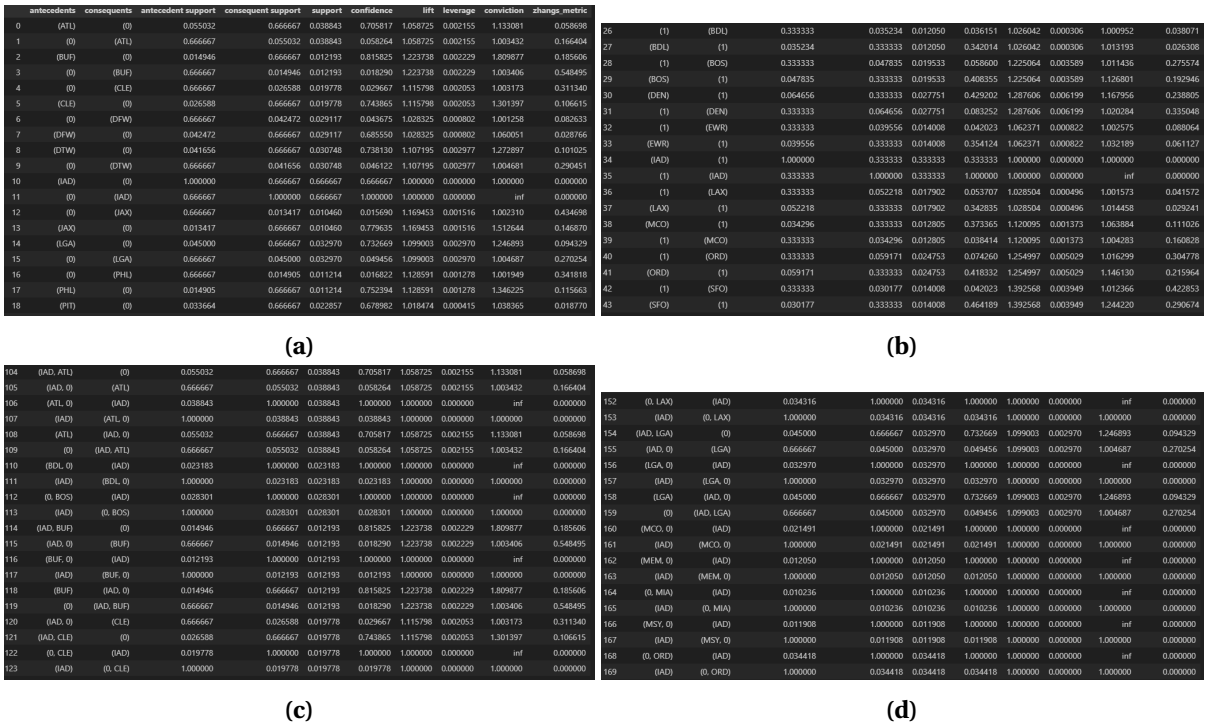


Figure 23: Support for delayed routes

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
6	(IAD)	(0)	0.498517	0.836783	0.427681	0.857907	1.025244	0.010531	1.148663	0.049100
7	(0)	(IAD)	0.836783	0.498517	0.427681	0.511102	1.025244	0.010531	1.025741	0.150858
0	(ATL)	(0)	0.028424	0.836783	0.024111	0.848261	1.013717	0.000326	1.075647	0.013928
1	(0)	(ATL)	0.836783	0.028424	0.024111	0.028814	1.013717	0.000326	1.000401	0.082907
8	(LGA)	(0)	0.023692	0.836783	0.020737	0.875263	1.045986	0.000912	1.308495	0.045031
9	(0)	(LGA)	0.836783	0.023692	0.020737	0.024781	1.045986	0.000912	1.001117	0.269362
4	(DTW)	(0)	0.022224	0.836783	0.019578	0.880952	1.052785	0.000982	1.371025	0.051278
5	(0)	(DTW)	0.836783	0.022224	0.019578	0.023397	1.052785	0.000982	1.001201	0.307189
2	(BDL)	(0)	0.017801	0.836783	0.014986	0.841840	1.006043	0.000090	1.031973	0.006116
3	(0)	(BDL)	0.836783	0.017801	0.014986	0.017909	1.006043	0.000090	1.000110	0.036803

Figure 24: Association rules for delayed routes

Nothing important seems to occur from this case.Some association rules are visible, but they only cover "not delayed cases," which makes sense given the **imbalanced** data.We use under-sampling to minimize bias in our data in order to help deal with the imbalanced data.Using the lift metric > 1 and the Appriori algorithm, we come at the following outcome(for space management we put in the report some samples of the output).



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(ATL)	(0)	0.055032	0.666667	0.038843	0.705817	1.058725	0.002155	1.133081	0.058698
1	(0)	(ATL)	0.666667	0.055032	0.038843	0.058264	1.058725	0.002155	1.003432	0.166404
2	(BUF)	(0)	0.014946	0.666667	0.012193	0.815825	1.223738	0.002229	1.809877	0.185606
3	(0)	(BUF)	0.666667	0.014946	0.012193	0.018290	1.223738	0.002229	1.003406	0.548495
4	(0)	(CLE)	0.666667	0.026588	0.019778	0.029667	1.115798	0.002053	1.003173	0.311340
5	(CLE)	(0)	0.026588	0.666667	0.019778	0.743865	1.115798	0.002053	1.301397	0.106615
6	(0)	(DFW)	0.666667	0.042472	0.029117	0.043675	1.028325	0.000802	1.001258	0.082633
7	(DFW)	(0)	0.042472	0.666667	0.029117	0.685550	1.028325	0.000802	1.060051	0.028766
8	(DTW)	(0)	0.041656	0.666667	0.030748	0.738130	1.107195	0.002977	1.272897	0.101025
9	(0)	(DTW)	0.666667	0.041656	0.030748	0.046122	1.107195	0.002977	1.004681	0.290451
10	(IAD)	(0)	1.000000	0.666667	0.666667	0.666667	1.000000	0.000000	1.000000	0.000000
11	(0)	(IAD)	0.666667	1.000000	0.666667	1.000000	1.000000	0.000000	inf	0.000000
12	(0)	(JAX)	0.666667	0.013417	0.010460	0.015690	1.169453	0.001516	1.002310	0.434698
13	(JAX)	(0)	0.013417	0.666667	0.010460	0.779635	1.169453	0.001516	1.512644	0.146870
14	(LGA)	(0)	0.045000	0.666667	0.032970	0.732669	1.099003	0.002970	1.246893	0.094329
15	(0)	(LGA)	0.666667	0.045000	0.032970	0.049456	1.099003	0.002970	1.004687	0.270254
16	(0)	(PHL)	0.666667	0.014905	0.011214	0.016822	1.128591	0.001278	1.001949	0.341818
17	(PHL)	(0)	0.014905	0.666667	0.011214	0.752394	1.128591	0.001278	1.346225	0.115663
18	(PIT)	(0)	0.033664	0.666667	0.022857	0.678962	1.018474	0.000415	1.038365	0.018770

(a)

26	(1)	(BDL)	0.333333	0.035234	0.012050	0.036151	1.026042	0.000306	1.000952	0.038071
27	(BDL)	(1)	0.035234	0.333333	0.012050	0.342014	1.026042	0.000306	1.013193	0.026308
28	(1)	(BOS)	0.333333	0.047835	0.019533	0.058600	1.225064	0.003589	1.011436	0.275574
29	(BOS)	(1)	0.047835	0.333333	0.019533	0.408355	1.225064	0.003589	1.126801	0.192946
30	(DEN)	(1)	0.064656	0.333333	0.027751	0.429202	1.287606	0.006199	1.167956	0.238805
31	(1)	(DEN)	0.333333	0.064656	0.027751	0.083252	1.287606	0.006199	1.020284	0.335048
32	(1)	(EWR)	0.333333	0.039556	0.014008	0.042023	1.062371	0.000822	1.002575	0.088064
33	(EWR)	(1)	0.039556	0.333333	0.014008	0.354124	1.062371	0.000822	1.032189	0.061127
34	(IAD)	(1)	1.000000	0.333333	0.333333	0.333333	1.000000	0.000000	1.000000	0.000000
35	(1)	(IAD)	0.333333	1.000000	0.333333	1.000000	1.000000	0.000000	inf	0.000000
36	(1)	(LAX)	0.333333	0.052218	0.017902	0.053707	1.028504	0.000496	1.001573	0.041572
37	(LAX)	(1)	0.052218	0.333333	0.017902	0.342835	1.028504	0.000496	1.014458	0.029241
38	(MCO)	(1)	0.034296	0.333333	0.012805	0.373365	1.120095	0.001373	1.063884	0.111026
39	(1)	(MCO)	0.333333	0.034296	0.012805	0.038414	1.120095	0.001373	1.004283	0.160828
40	(1)	(ORD)	0.333333	0.059171	0.024753	0.074260	1.254997	0.005029	1.016299	0.304778
41	(ORD)	(1)	0.059171	0.333333	0.024753	0.418332	1.254997	0.005029	1.146130	0.215964
42	(1)	(SFO)	0.333333	0.030177	0.014008	0.042023	1.392568	0.003949	1.012366	0.422853
43	(SFO)	(1)	0.030177	0.333333	0.014008	0.464189	1.392568	0.003949	1.244220	0.290674

(b)

104	(IAD, ATL)	(0)	0.055032	0.666667	0.038843	0.705817	1.058725	0.002155	1.133081	0.058698
105	(IAD, 0)	(ATL)	0.666667	0.055032	0.038843	0.058264	1.058725	0.002155	1.003432	0.166404
106	(ATL, 0)	(IAD)	0.038843	1.000000	0.038843	1.000000	1.000000	0.000000	inf	0.000000
107	(IAD)	(ATL, 0)	1.000000	0.038843	0.038843	0.038843	1.000000	0.000000	1.000000	0.000000
108	(ATL)	(IAD, 0)	0.055032	0.666667	0.038843	0.705817	1.058725	0.002155	1.133081	0.058698
109	(0)	(IAD, ATL)	0.666667	0.055032	0.038843	0.058264	1.058725	0.002155	1.003432	0.166404
110	(BDL, 0)	(IAD)	0.023183	1.000000	0.023183	1.000000	1.000000	0.000000	inf	0.000000
111	(IAD)	(BDL, 0)	1.000000	0.023183	0.023183	0.023183	1.000000	0.000000	1.000000	0.000000
112	(0, BOS)	(IAD)	0.028301	1.000000	0.028301	1.000000	1.000000	0.000000	inf	0.000000
113	(IAD)	(0, BOS)	1.000000	0.028301	0.028301	0.028301	1.000000	0.000000	1.000000	0.000000
114	(IAD, BUF)	(0)	0.014946	0.666667	0.012193	0.815825	1.223738	0.002229	1.809877	0.185606
115	(IAD, 0)	(BUF)	0.666667	0.014946	0.012193	0.018290	1.223738	0.002229	1.003406	0.548495
116	(BUF, 0)	(IAD)	0.012193	1.000000	0.012193	1.000000	1.000000	0.000000	inf	0.000000
117	(IAD)	(BUF, 0)	1.000000	0.012193	0.012193	0.012193	1.000000	0.000000	1.000000	0.000000
118	(BUF)	(IAD, 0)	0.014946	0.666667	0.012193	0.815825	1.223738	0.002229	1.809877	0.185606
119	(0)	(IAD, BUF)	0.666667	0.014946	0.012193	0.018290	1.223738	0.002229	1.003406	0.548495
120	(IAD, 0)	(CLE)	0.666667	0.026588	0.019778	0.029667	1.115798	0.002053	1.003173	0.311340
121	(IAD, CLE)	(0)	0.026588	0.666667	0.019778	0.743865	1.115798	0.002053	1.301397	0.106615
122	(0, CLE)	(IAD)	0.019778	1.000000	0.019778	1.000000	1.000000	0.000000	inf	0.000000
123	(IAD)	(0, CLE)	1.000000	0.019778	0.019778	0.019778	1.000000	0.000000	1.000000	0.000000

(c)

152	(0, LAX)	(IAD)	0.034316	1.000000	0.034316	1.000000	1.000000	0.000000	inf	0.000000
153	(IAD)	(0, LAX)	1.000000	0.034316	0.034316	0.034316	1.000000	0.000000	1.000000	0.000000
154	(IAD, LGA)	(0)	0.045000	0.666667	0.032970	0.732669	1.099003	0.002970	1.246893	0.094329
155	(IAD, 0)	(LGA)	0.666667	0.045000	0.032970	0.049456	1.099003	0.002970	1.004687	0.270254
156	(LGA, 0)	(IAD)	0.032970	1.000000	0.032970	1.000000	1.000000	0.000000	inf	0.000000
157	(IAD)	(LGA, 0)	1.000000	0.032970	0.032970	0.032970	1.000000	0.000000	1.000000	0.000000
158	(LGA)	(IAD, 0)	0.045000	0.666667	0.032970	0.732669	1.099003	0.002970	1.246893	0.094329
159	(0)	(IAD, LGA)	0.666667	0.045000	0.032970	0.049456	1.099003	0.002970	1.004687	0.270254
160	(MCO, 0)	(IAD)	0.021491	1.000000	0.021491	1.000000	1.000000	0.000000	inf	0.000000
161	(IAD)	(MCO, 0)	1.000000	0.021491	0.021491	0.021491	1.000000	0.000000	1.000000	0.000000
162	(MEM, 0)	(IAD)	0.012050	1.000000	0.012050	1.000000	1.000000	0.000000	inf	0.000000
163	(IAD)	(MEM, 0)	1.000000	0.012050	0.012050	0.012050	1.000000	0.000000	1.000000	0.000000
164	(0, MIA)	(IAD)	0.010236	1.000000	0.010236	1.000000	1.000000	0.000000	inf	0.000000
165	(IAD)	(0, MIA)	1.000000	0.010236	0.010236	0.010236	1.000000	0.000000	1.000000	0.000000
166	(MSY, 0)	(IAD)	0.011908	1.000000	0.011908	1.000000	1.000000	0.000000	inf	0.000000
167	(IAD)	(MSY, 0)	1.000000	0.011908	0.011908	0.011908	1.000000	0.000000	1.000000	0.000000
168	(0, ORD)	(IAD)	0.034418	1.000000	0.034418	1.000000	1.000000	0.000000	inf	0.000000
169	(IAD)	(0, ORD)	1.000000	0.034418	0.034418	0.034418	1.000000	0.000000	1.000000	0.000000

(d)

Figure 25: Association rules after under-sampling

We notice that for many destinations, there are two-way rules. For instance, if the destination is JAX, there are no delays both when departing from and arriving at JAX. Conversely, if the destination is BDL, there are delays both when departing from and arriving at BDL.

1.4 Task 4 : Try to predict the delay given all other features and report the appropriate performance on cross-validation

Although the task is presented as predicting delays, it is advisable to treat it as a binary classification problem. The model is structured to differentiate between two classes: 0 for flights that are not delayed and 1 for delayed flights. In this context, a flight is categorized as delayed if its arrival delays surpass 15 minutes, as specified earlier, thereby introducing an imbalanced data challenge to be addressed.

1.4.1 Data Preparation

To begin, our initial step involves preparing the data. We will pursue two distinct dataset approaches, tailoring one specifically for tree-based algorithms and the other for parametric models such as Logistic Regression.
The initial dataframe that we will use in our datasets is :

	DayofWeek	DepTimeBucket	UniqueCarrier	Origin	Dest	is_delayed
0	5	06:00-12:00	UA	ORD	IAD	0
1	5	06:00-12:00	DL	IAD	ATL	0
2	5	06:00-12:00	UA	IAD	DEN	0

Figure 26: Initial Dataset

Dataset 1 (for tree-based algorithms) :

Our first task is to transform the existing dataframe into a usable format. We will employ label encoding for the features 'DepTimeBucket,' 'UniqueCarrier,' 'Origin,' and 'Dest.' This encoding process will provide a structured dataset, as illustrated in the subsequent image.

	DayofWeek	DepTimeBucket	UniqueCarrier	Origin	Dest	is_delayed
0	4	1	7	35	20	0
1	4	1	2	19	0	0
2	4	1	7	19	12	0

Figure 27: Dataset after label encoding

Next, we'll partition the dataset into training and testing sets, allocating 70% to the train set and 30% to the test set. It's important to maintain the distribution of the target variable, as indicated by the graphs below. This ensures that both the training and testing datasets accurately represent the patterns and characteristics present in the overall dataset.

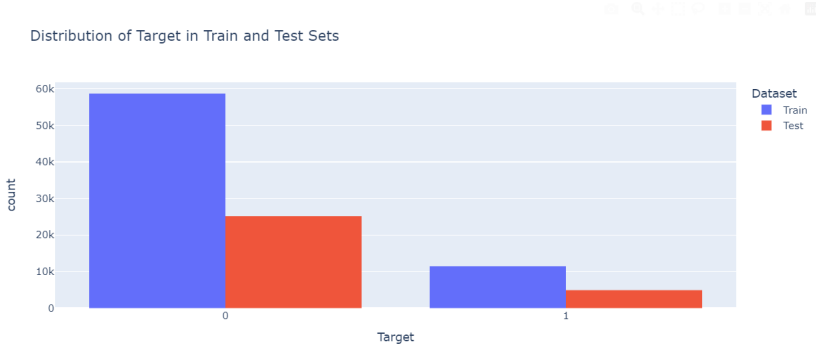


Figure 28: Distribution of the target of train-test sets

Finally, in dataset 1, we implemented an undersampling method to achieve a more balanced distribution of data. It is crucial to note that despite choosing undersampling, we will still maintain an unbalanced dataset. The decision to go for undersampling instead of oversampling was driven by the substantial size of the dataset, exceeding 100,000 rows. So, as a result after the undersampling method our dataset has 22,886 for the class 0 and 11,443 for

the class 1.

Dataset2 (for parametric model) : In this dataset we are going to use one-hot encoding on the features 'DayofWeek', 'UniqueCarrier', 'Origin', 'Dest', and 'DepTimeBucket' in order to have an efficient dataset. So, a sample of our dataset would be like this:

is_delayed	DayofWeek0	DayofWeek1	DayofWeek2	DayofWeek3	DayofWeek4	DayofWeek5	DayofWeek6	UniqueCarrier0	UniqueCarrier1	UniqueCarrier2	UniqueCarrier3
0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0
0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 29: Dataset after one-hot encoding

Subsequently, mirroring the procedures applied to dataset 1, we divided our dataset into 70% for training and 30% for testing. Employing an undersampling technique, we created a new dataset that, while still imbalanced, exhibited a less pronounced imbalance compared to the initial dataset. As a result having again 22,886 for the class 0 and 11,443 for the class 1.

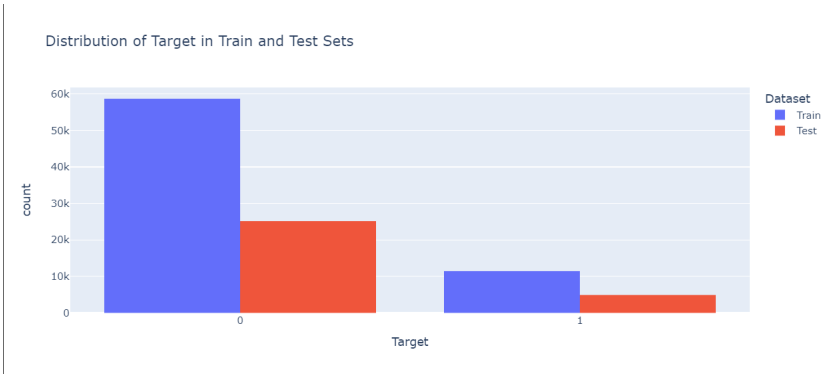


Figure 30: Distribution of the target of train-test sets

RFECV using Stratified k-fold and f1 scoring

After conducting Recursive Feature Elimination with Cross-Validation (RFECV) utilizing Stratified k-fold and employing the F1 statistic as the scoring metric, it was observed that all features were consistently deemed important. Given this result, indicating the significance of each feature in the model, we proceeded with retaining all of them in our analysis.

1.4.2 Model Selection

- Macro F1 Score: This metric treats all classes equally, making it suitable for scenarios with a balanced class distribution.
- Micro F1 Score: It aggregates the contributions of all classes, proving valuable in the presence of class imbalance. Notably, in a binary class dataset, the micro F1 score simplifies to the accuracy score.
- Weighted F1 Score: This metric considers the number of true instances when averaging the F1 scores. It strikes a balance between treating all classes equally and addressing class imbalance.

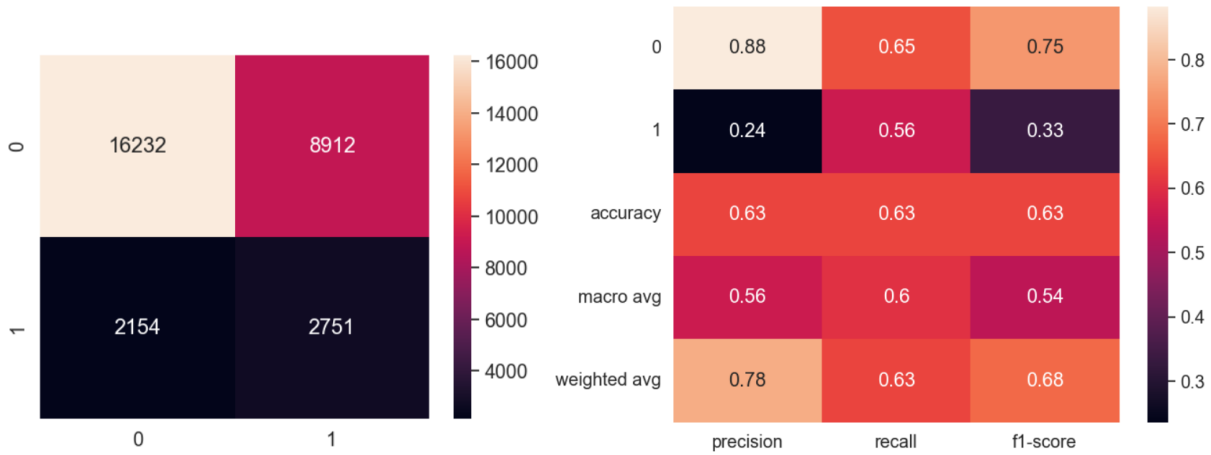
The F1 score, ranging from 0 to 1, serves as a valuable metric in cases of imbalanced class problems, where one class may dominate the dataset. An ideal model achieves a F1 score of 1, signifying all predictions are correct. When dealing with scenarios like credit card fraud detection, the focus is often on maximizing the recall (catching fraudulent transactions) while minimizing false positives (high precision). Therefore, precision-recall curves and F1 score become informative tools for such specific use cases.
After executing the algorithms, let's examine the outcomes of each model.

Table 2: F1 scores for each model across the datasets.

Dataset	Baseline	With undersampling	Baseline: only Tree - based algorithms with extra criteria instead of undersampling the dataset
Dataset 1	LR : 0.000000 RF : 0.073700 SVC : 0.000000 NB : 0.000000 XGB : 0.047096 LGBM : 0.022959	LR : 0.000175 RF : 0.368289 SVC : 0.000000 NB: 0.038738 XGB : 0.344374 LGBM : 0.280582	RF : 0.331541 XGB : 0.050694
Dataset 2	LR : 0.018761 RF : 0.071134 SVC : 0.026453 NB : 0.302389 XGB : 0.039791 LGBM : 0.025122	LR : 0.238791 RF : 0.371167 SVC : 0.321932 NB : 0.516740 XGB : 0.315106 LGBM : 0.312559	RF : 0.330963 XGB : 0.050826

1.4.3 Finalize model (perform predictions and finalize model)

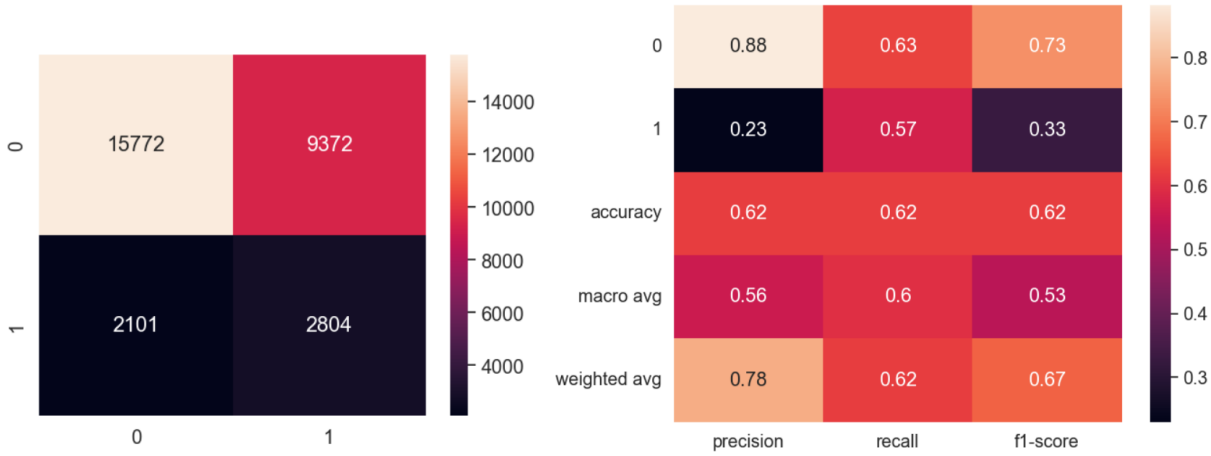
After taking a look at our models, and performing some optimizations, it is observed that 2 methods seem to stand out overall, NB (Dataset 2- with and without undersampling) and Random Forest (Dataset 1-with and without undersampling) . For these 2 winning algorithms, for each case now, we are going to make predictions based on the test set.
For the **Random Forest**, we have :



(a) Confusion Matrix **(b)** Precision, Recall, f1 score

Figure 31: Random Forest

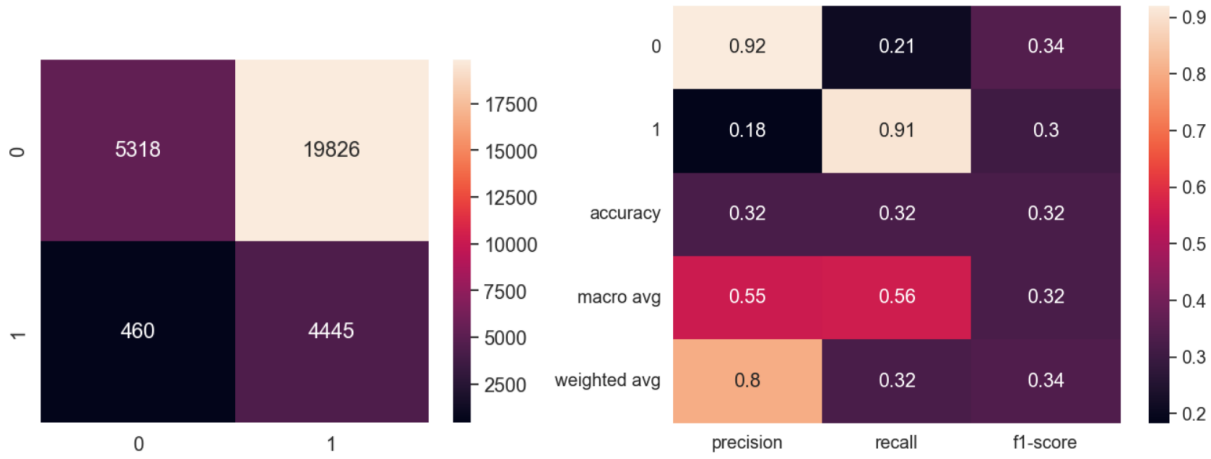
For the **Random Forest** of undersampled dataset, we have :



(a) Confusion Matrix **(b)** Precision, Recall, f1 score

Figure 32: Random Forest with undersampling

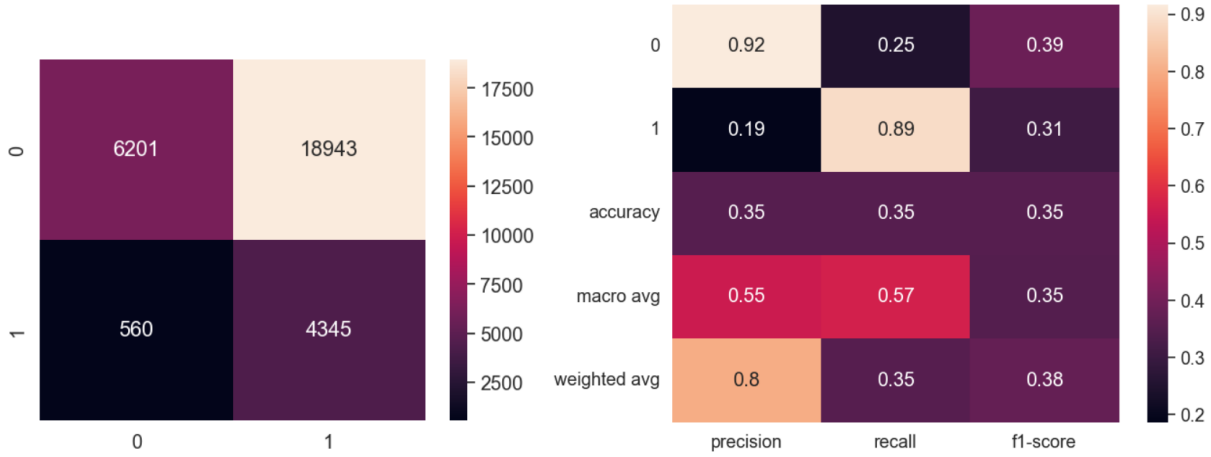
For the **Gaussian Naive Bayes**, we have :



(a) Confusion Matrix **(b)** Precision, Recall, f1 score

Figure 33: Gaussian Naive Bayes

For the **Gaussian Naive Bayes** of undersampled dataset, we have :

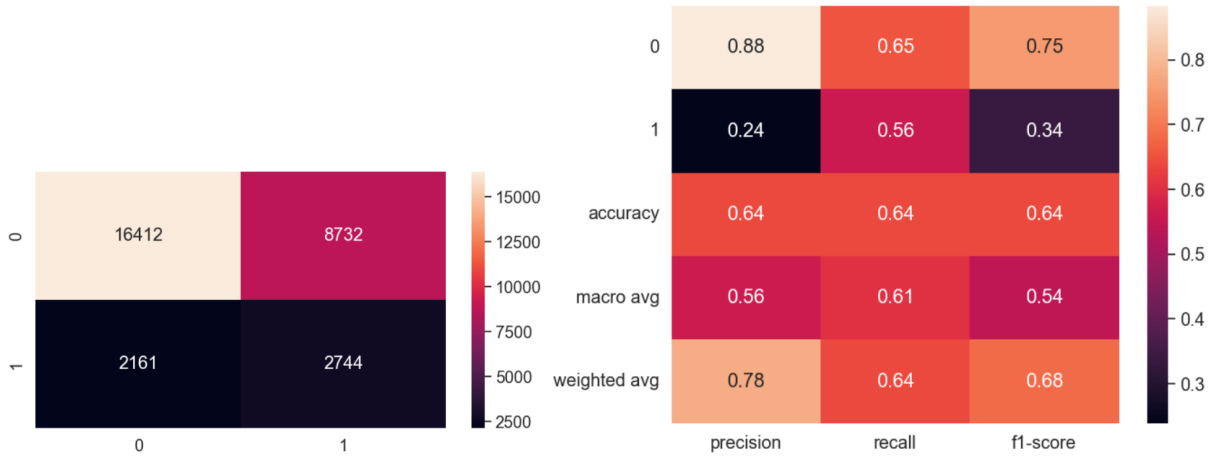


(a) Confusion Matrix **(b)** Precision, Recall, f1 score

Figure 34: Gaussian Naive Bayes with undersampling

1.4.4 Optimize Models

In our scenario, the model available for optimization is the Random Forest. Consequently, we engaged in hyperparameter tuning for a RandomForestClassifier utilizing RandomizedSearchCV. The optimized hyperparameter values obtained were 'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 20, 'class_weight': 'balanced_subsample', and 'bootstrap': True. Armed with this information, let's proceed to examine our confusion matrix and delve into the metric values.



(a) Confusion Matrix **(b)** Precision, Recall, f1 score

Figure 35: Random Forest with hyperparameter tuning

Having considered all the information presented above, it is time to draw some conclusions. Throughout our model selection process, where we assessed the performance of Logistic Regression, Random Forest, SVC, Gaussian Naive Bayes, XGBoost, and LGBM, we identified the Random Forest and Gaussian Naive Bayes models as the most effective across both the baseline and undersampled datasets. Subsequently, we applied these finalized models to make predictions. Upon comparing the two, it became apparent that Random Forest outperformed Gaussian Naive Bayes in terms of F1 score across both datasets. However, a closer look at precision and recall revealed that Random Forest exhibited superior precision, while Gaussian Naive Bayes had higher recall. Notably, following hyperparameter tuning in Random Forest, we observed a slight improvement in results compared to the initial performance.

1.5 Task 5 : Identify patterns/rules regarding delays and try to explain when delays should be expected based on these patterns

We will tackle this task using two methods. Firstly, we'll employ association rules utilizing the Apriori algorithm. Secondly, we'll utilize exploratory data analysis, similar to the analysis we conducted in Task 2 for average delays.

Association rules : We aim to establish rules exclusively for delayed flights, focusing on the relationships between days of the week and airlines, as well as between departure time buckets and airlines. The following pictures represent our results

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(06:00-12:00)	(UA)	0.286763	0.596587	0.178737	0.623294	1.044766	0.007659	1.070895	0.060075
1	(18:00-24:00)	(UA)	0.277649	0.596587	0.167788	0.604318	1.012959	0.002147	1.019539	0.017711

Figure 36: Departure time - Airline

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2	(3)	(UA)	0.160692	0.596587	0.098483	0.612866	1.027288	0.002616	1.042052	0.031649
3	(4)	(UA)	0.153964	0.596587	0.092733	0.602304	1.009584	0.000880	1.014377	0.011220
1	(2)	(UA)	0.147113	0.596587	0.089124	0.605821	1.015479	0.001359	1.023427	0.017872
0	(1)	(UA)	0.143014	0.596587	0.086738	0.606501	1.016619	0.001418	1.025196	0.019075

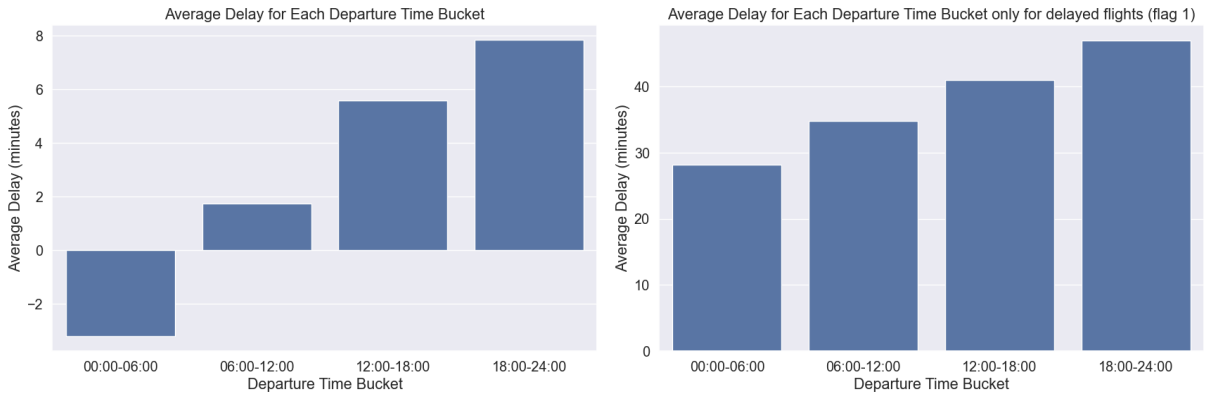
Figure 37: Days of week - Airline

Observations:

- We expect to have a delay if the flight departs around 06:00-12:00 or 18:00-24:00 and the airline is UA
- We expect to have a delay if the flight departs around Monday to Thursday and the airline is UA

Exploratory Data Analysis: From our previous EDA we can detect certain patterns

- **Departure time:** As the day progresses from the early morning hours (00:00-06:00), there is a noticeable trend where flights tend to arrive on time or even early. However, as we transition into the afternoon, delays show an increasing pattern. This observation could be attributed to various factors such as air traffic congestion, increased airport activities, and the cumulative effects of delays that accumulate throughout the day. Additionally, during the morning hours, operational efficiency may be higher with fewer disruptions, contributing to the punctuality of flights. In contrast, as the day unfolds, the likelihood of encountering challenges like weather changes, peak travel times, and other operational complexities may lead to a higher incidence of delays in the afternoon.



(a) Average Delay for Each Departure Time Bucket **(b)** Av. Delay for Each Dep.Time Bucket (delayed flights)

Figure 38: Average Delay

- **Day of week:** We also notice, that if a flight is happening during the weekend, we may have less delays. The opposite can be said for days like Wednesday and Friday. That may be as mentioned before due to the fact that a lot of people may fly on Friday for the weekend or in the middle of the week, when it is cheaper.

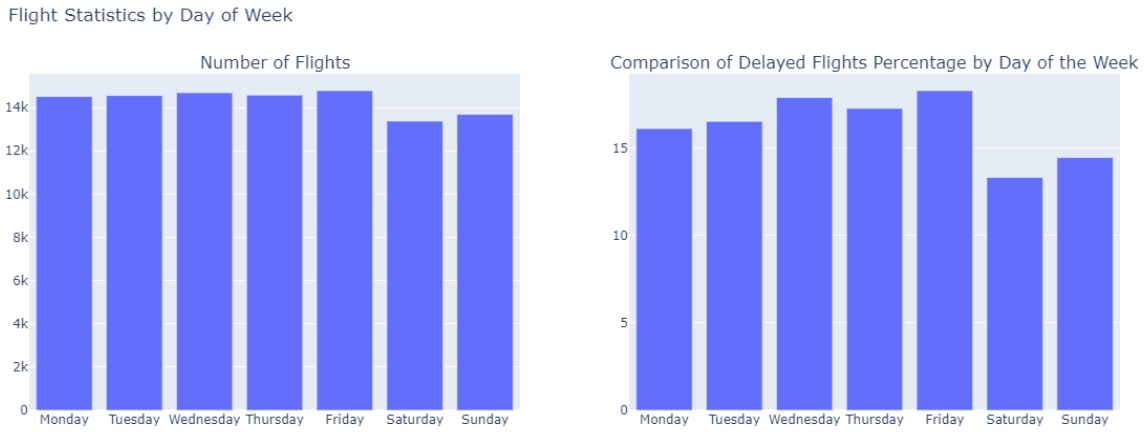


Figure 39: Days of Week

- **Airline:** It seems that less popular airlines like PA and EA, have the biggest percentage of delayed flights, so if a flight is done by them it is likely that they will be delayed.

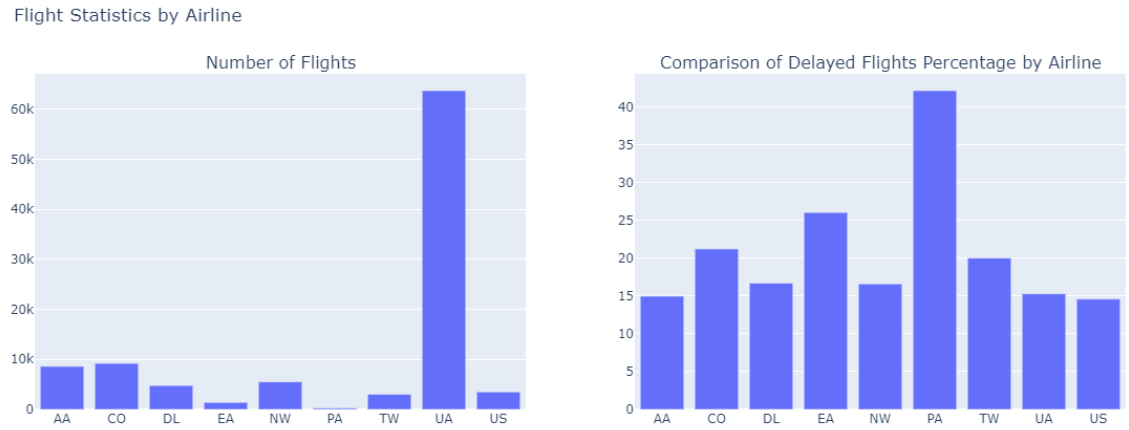


Figure 40: Airline

- **Airport:** When we look at the most common routes for delayed flights, we notice that some routes have more delays than others. What's interesting is that these delays often happen at airports that have a lot of flights going in and out. This suggests that the number of flights at an airport might affect how often delays happen. It shows us that there's a connection between how busy an airport is and the likelihood of delays happening there. This highlights the need to understand how airport traffic and other factors can lead to delays at busy airports.

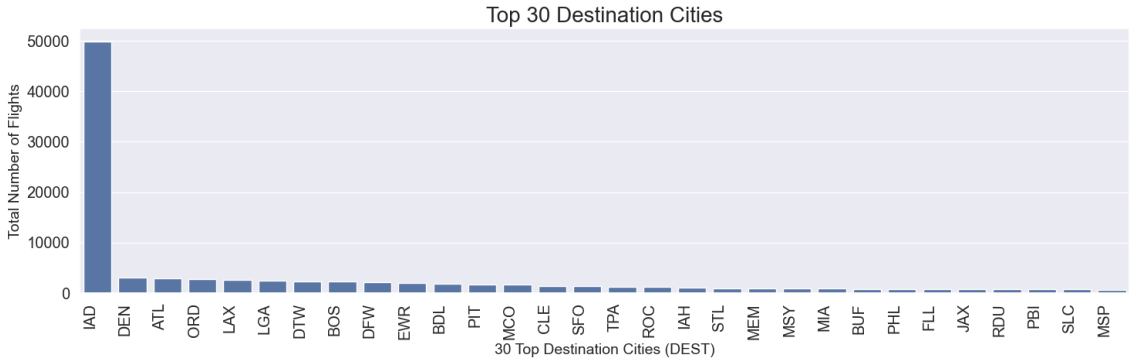


Figure 41: 30 Top Destination Cities (DEST)

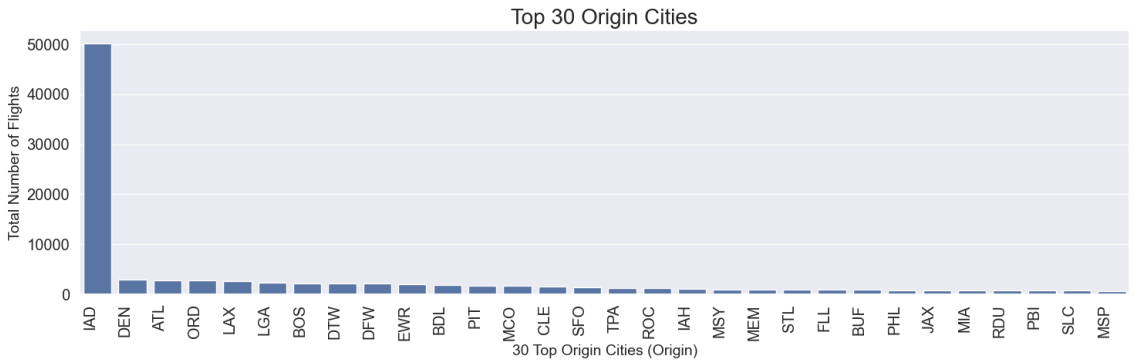


Figure 42: 30 Top Origin Cities (Origin)

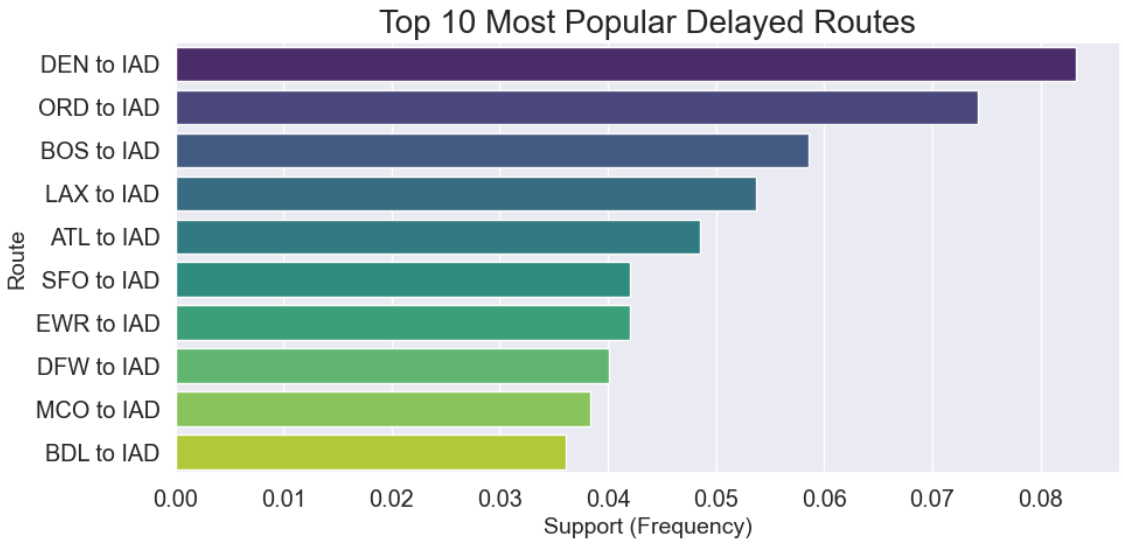


Figure 43: Top 10 Most Popular Delayed Routes

2 Dataset 2 : Religions in America

2.1 Task 1: Data Information

In the following dataset we have a list of attributes related to religious denominations and their membership and church counts in 1952, organized by county and state. Each attribute represents the number of members and churches for a specific denomination.On the following image we can see a representation of our dataset:

Attributes explanation: Based on the tasks below we will use the first 6 attribute , which are demographic information about the religions

- CNAME = County Name

	CaseID$	CNAME	STCODE	CCODE	TOTPOP	TOTMEMB	TOTCHUR	SDA_M	SDA_C	AOG_M	...	UCHRC_M	UCHRC_C	UCA_M	UCA_C
0	33	Hale, AL	1	65	20832	5145	52	0	0	0	...	0	0	0	0
1	34	Henry, AL	1	67	18674	4773	31	0	0	23	...	0	0	0	0
2	35	Houston, AL	1	69	46522	19420	88	87	2	379	...	0	0	0	0
3	36	Jackson, AL	1	71	38998	9030	82	110	1	0	...	0	0	0	0
4	37	Jefferson, AL	1	73	558928	212326	595	991	4	648	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3070	3071	Teton, WY	56	39	2593	968	3	0	0	10	...	0	0	0	0
3071	3072	Uinta, WY	56	41	7331	3817	8	0	0	5	...	0	0	0	0
3072	3073	Washakie, WY	56	43	7252	2378	15	70	3	90	...	0	0	0	0
3073	3074	Weston, WY	56	45	6733	2025	13	92	2	69	...	0	0	0	0
3074	3075	Yellowstone National Park, WY	56	47	353	0	0	0	0	0	...	0	0	0	0

Figure 44: Religions Dataset

- STCODE = State Census Code
- CCODE = County Census Code
- TOTPOP = Total Population–1952
- TOTMEMB = Total number of members–1952
- TOTCHUR = Total Number of churches–1952

The remaining 228 attributes contain data on the number of churches and members per religion.

Structure information: For better understating of our dataset use the following findings

- Dataset size : 3075 rows and 235 columns
- Data types : We have 235 attributes that are int64 and one attribute ,the County Names (CNAME), which is object
- Missing Values: There are no missing values
- Descriptive Statistics about the totals :

2.2 Task 2: Summarize the data

To help fully understand the overall picture of religious groups in the US, we will summarize the data. First we will make some groupings to understand the importance of attributes .

	TOTPOP	TOTCHUR	TOTMEMB
count	3.075000e+03	3075.000000	3.075000e+03
mean	4.898718e+04	59.465366	2.410584e+04
std	1.698913e+05	80.284820	9.496974e+04
min	5.200000e+01	0.000000	0.000000e+00
25%	1.000850e+04	24.000000	4.006500e+03
50%	1.876000e+04	42.000000	7.905000e+03
75%	3.584500e+04	68.000000	1.589350e+04
max	4.508792e+06	1939.000000	2.685524e+06

Figure 45: Descriptive Statistics

- **Grouping by State (STCODE)**: We group the data by state to analyze the religious demographics and distribution of denominations across different states.We also are going to calculate the rates of total population , total members and total churches and the percentage of believers per state. The head of the final table of our calculations will look like the following figure:

	STCODE	TOTPOP	TOTMEMB	TOTCHUR	TOTPOP_RATE	TOTMEMB_RATE	TOTCHUR_RATE	%_OF_BELIEVERS
0	1	3061743	1046460	5620	2.032550	1.411742	3.073457	34.178571
1	4	749587	336938	767	0.497616	0.454551	0.419456	44.949819
2	5	1909511	598593	3715	1.267636	0.807540	2.031653	31.347973
3	6	10586223	4306690	6794	7.027704	5.810001	3.715492	40.682026
4	8	1325089	550993	1578	0.879665	0.743325	0.862974	41.581584

Figure 46: State Summary

- **Distribution of Believers and Total Members per State:**Using that table we create two bar plots, as seen below, to display the distribution of believers by state::

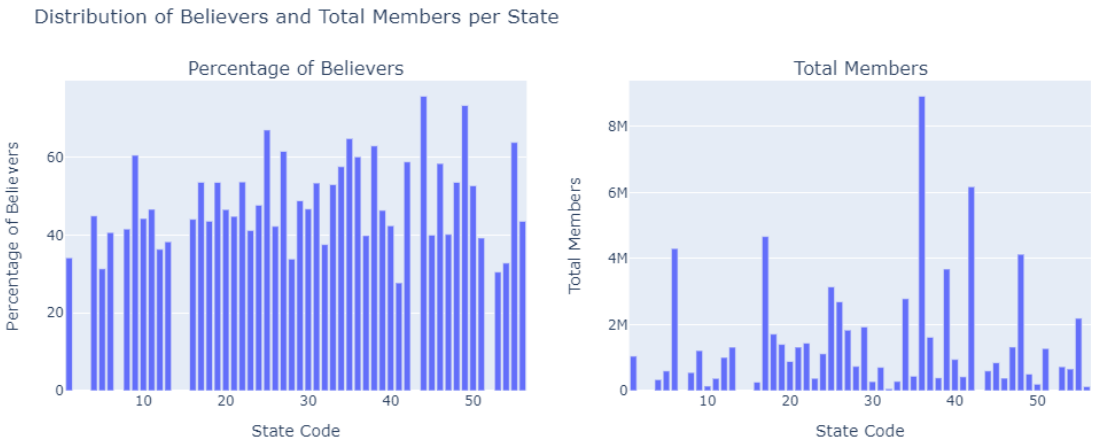


Figure 47: Distribution of Believers and Total Members per State

Observations:

1.State 36 (New York) has the largest population, indicating a high population density.State 36 has a slightly lower percentage of believers than State 44 (Rhode Island), but the sheer size of its population makes a significant difference to the total number of believers.

2.Although State 44 might have a higher percentage of believers than State 36, its total population might be lower. As is common in smaller societies, this shows that the concentration of believers in State 44 is higher than the population size.

3.In contrast, State 36 with its large population and relatively high percentage of believers implies a substantial number of believers overall. The combination of a large population and a reasonably high percentage of believers in State 36 makes it a significant contributor to the total number of believers in the area.

Conclusion : Despite State 44 having the highest percentage of believers, the larger population and reasonable percentage of believers in State 36 make it a more substantial area in terms of the total number of believers. This observation underscores the importance of considering both population size and percentage of believers when analyzing religious demographics across different regions.

- **Map representation of the summarized data :** With information about the total number of members, total number of churches or total rate of believers per county or state, we can show geographic patterns and relationships between the regions. We can see . We provide a summary using 3 maps

1. Representation of Total Believers

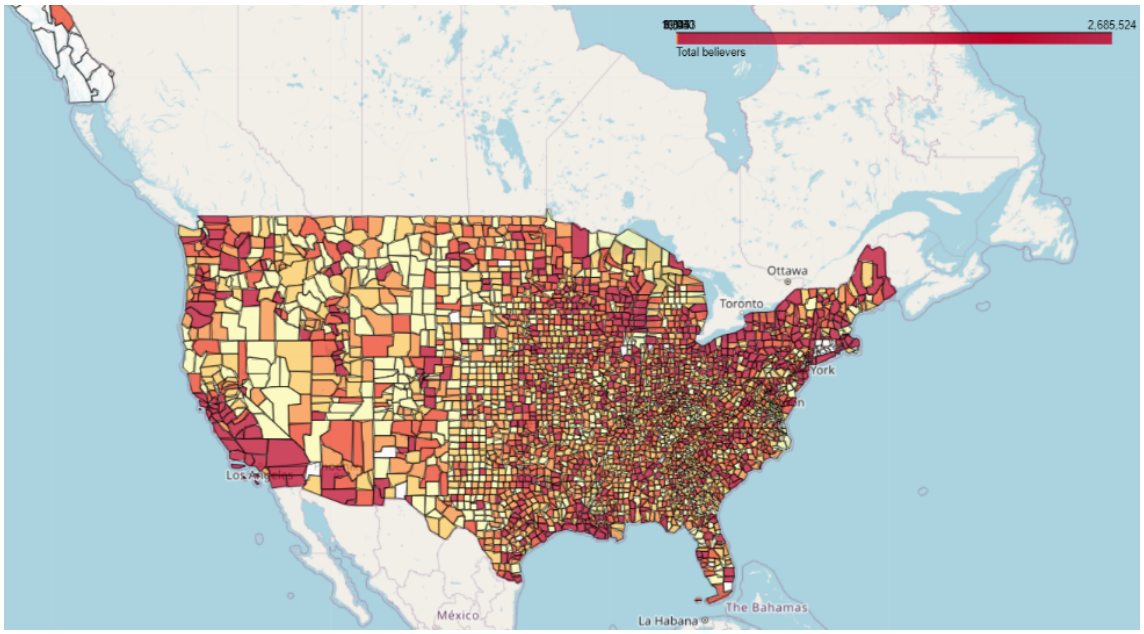


Figure 48: Total Believers

In the depicted map, the coloration of each county is determined by the total number of members or believers, creating a gradient that reflects varying population sizes. Counties with a larger population of believers appear darker, while those with a smaller population exhibit lighter colors. Upon closer inspection, the map reveals a notable concentration of believers in the eastern side of America, as evidenced by the darker hues in that region. With its spatial depiction of the distribution of all believers across several counties, this visualization offers important insights into the regional differences in religious demography.

2. Representation of Total Rate

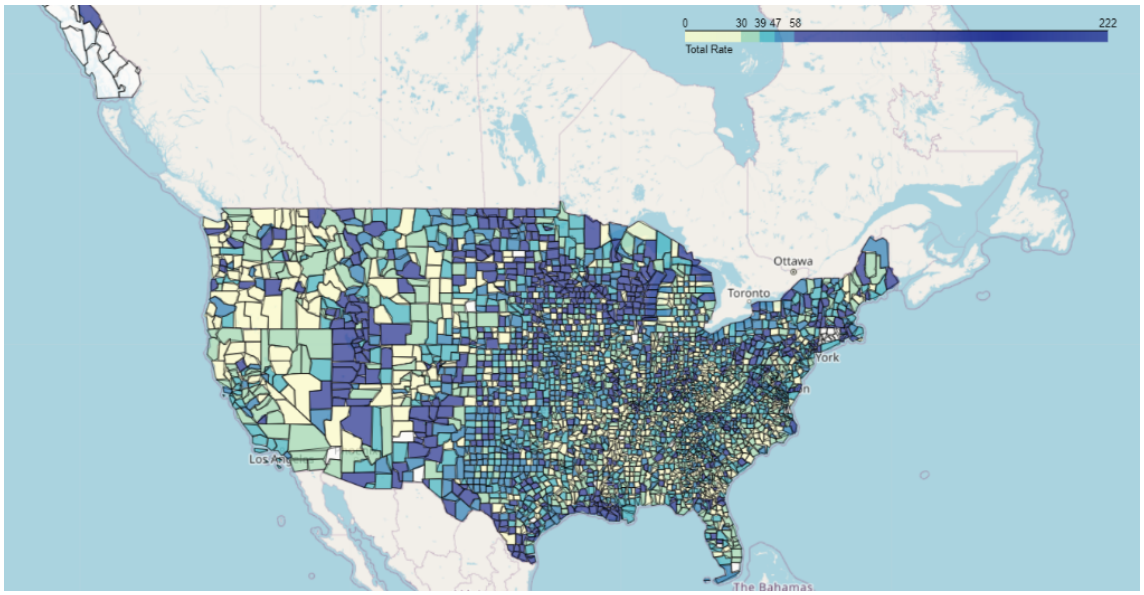


Figure 49: Total Rate

The above map illustrates the total rate across each county, and as the previous map, with darker colors indicating higher total rates and lighter colors, corresponding to lower rates. The visual representation highlights that the northeastern region of America exhibits a higher total rate, while an intriguing observation is the presence of a notable high-rate center on the western side.

3. Representation of Total Churches

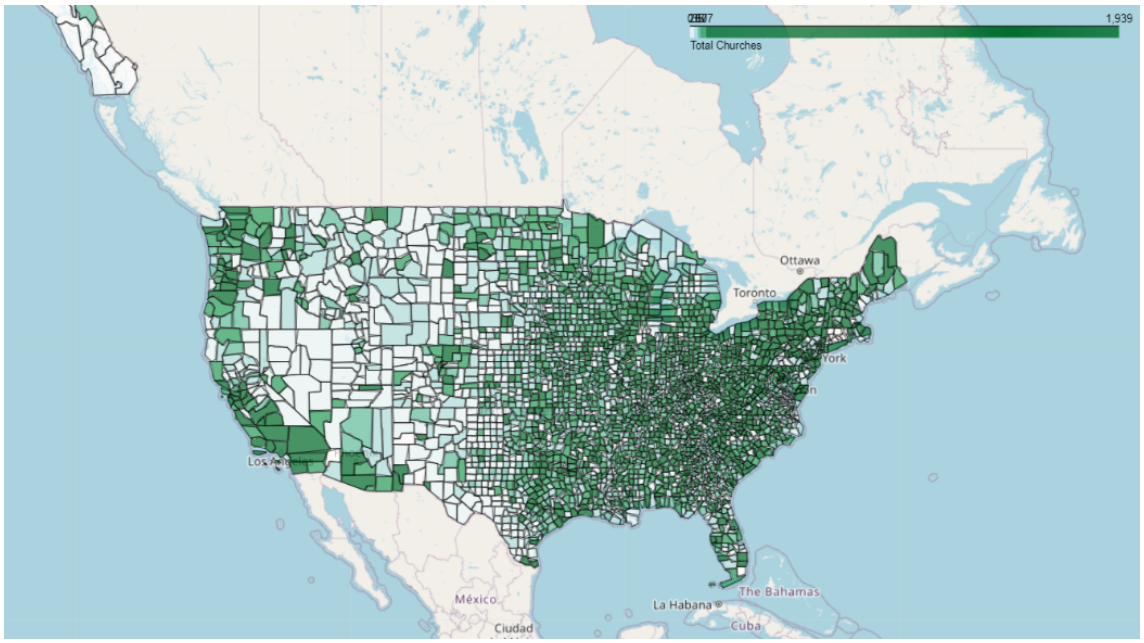


Figure 50: Total Churches

The map above illustrates the total number of churches in each county across America. Similar to the previous maps, darker colors signify a higher count of churches, while lighter colors indicate fewer churches in a county. Notably, the eastern region of America emerges as having a significant concentration of churches, as evidenced by the darker hues in this area.

2.3 Task 3 : Which are the counties with the highest per-person ratio of Orthodox Christian members?

For that task we had to identify the religions with Christian Orthodox members . After searching on the given description we found these four religions :

- Greek Orthodox Archdiocese of North and South America
- Armenian Apostolic Orthodox Church of America
- American Carpatho-Russian Orthodox Greek Catholic Church
- Bulgarian Eastern Orthodox Church

Using these four religions we calculate the per-person ratio of Orthodox Christian members (sorted in descending order) of each county and we had the following findings

	CNAME	Orthodox_Ratio
164	Cherokee, AL	0.011342
618	Madison, IL	0.002304
133	Fresno, CA	0.002275
3030	Racine, WI	0.001944
2278	Providence, RI	0.001734
...	...	...
1034	Marion, KY	0.000000
1035	Marshall, KY	0.000000
1036	Martin, KY	0.000000
1037	Mason, KY	0.000000
3074	Yellowstone National Park, WY	0.000000

Figure 51: Ratio of Orthodox Christian members

We can clearly observe that the county with the highest per-person ratio of Orthodox Christian members is **Cherokee,AL** with 'Orthodox Ratio'=0,011342.

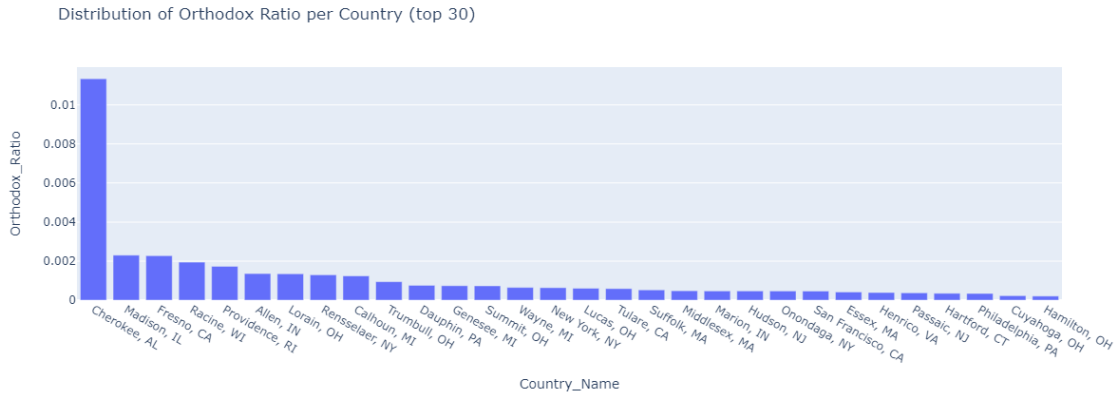
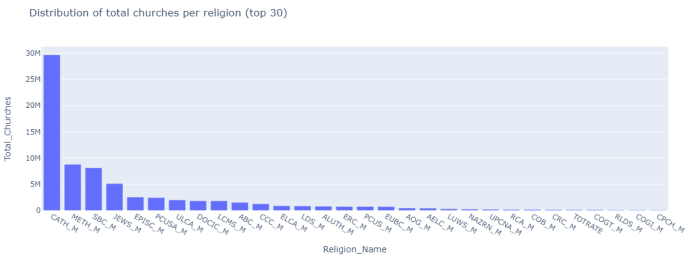


Figure 52: Distribution of Orthodox Ratio per county

On that point it would be interesting to see which is the most popular religion and compare their ratio of members and churches in Cherokee AL. In order to find the most popular religion we make the following dataset and with their distribution plots:

	index	total_churhes	total_members
0	METH_C	36813.0	NaN
1	SBC_C	29412.0	NaN
2	CATH_C	15728.0	NaN
3	PCUSA_C	8306.0	NaN
4	DOCIC_C	7860.0	NaN

(a) Religions with most Churches

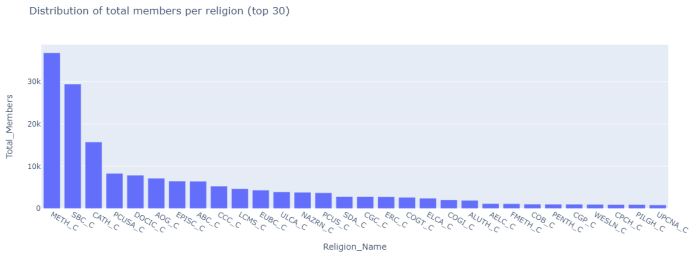


(b) Distribution of total churches per religion

Figure 53: Total Churches per religion

	index	total_churhes	total_members
0	CATH_M	NaN	29688058.0
1	METH_M	NaN	8790025.0
2	SBC_M	NaN	8121069.0
3	JEWS_M	NaN	5112024.0
4	EPISC_M	NaN	2544320.0

(a) Religions with most Member



(b) Distribution of total members per religion

Figure 54: Total members per religion

We see similar religions in the results regarding the population and the churches of each religion. It is logical, as when there are many members of a religion in an area, it is very likely that there will also be a church of that religion.

Now that we know that Roman Catholic is the most popular religion we can compare it with the Christian Orthodox in Cherokee,AL so that we can see the difference between religion's population

	CNAME	Roman_Catholic Ratio	Orthodox_Ratio
0	Cherokee, AL	0.0	0.011342

Figure 55: Ratio Comparison

The Roman Catholic ratio of 0.0 suggests that Cherokee, Alabama may not have a sizable Roman Catholic population. On the other hand, the Orthodox ratio of 0.011342 indicates the existence of a minor yet significant Orthodox community in Cherokee, AL.

2.4 Task 4: Find the 3 most extreme counties with respect to the distribution of their churches across religions

Firstly, let's create the dataset which will use in order to give answers to this task. We will need the names of every county and the total numbers of churches of every religion in each county. The dataset will look like that :

	CNAME	SDA_C	AOG_C	ABC_C	SBC_C	COB_C	COGT_C	COGI_C	CGC_C	NAZRN_C	...	RECH_C	SOCBR_C	NSAC_C	UNCHU_C	UBC_C
0	Hale, AL	0	0	0	14	0	0	4	0	2	...	0	0	0	0	0
1	Henry, AL	0	3	0	21	0	0	0	1	0	...	0	0	0	0	0
2	Houston, AL	2	10	0	42	0	1	0	3	1	...	0	0	0	0	0
3	Jackson, AL	1	0	0	58	0	5	0	0	2	...	0	0	0	0	0
4	Jefferson, AL	4	15	0	232	0	29	18	19	8	...	0	0	0	1	0

Figure 56: Total Churches - 5 first rows

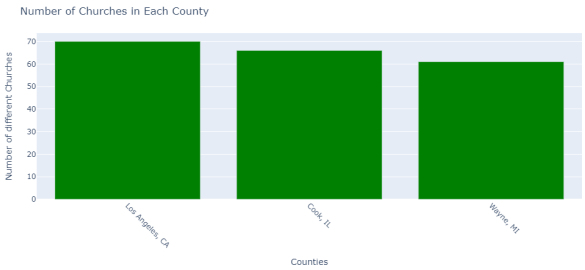
We will continue by using two different ways as a solution of this task :

- **First approach:** We'll count the number of churches corresponding to various religions in each county.

  Findings :

	CNAME	DIFCHURCHES
172	Los Angeles, CA	70
574	Cook, IL	66
1276	Wayne, MI	61

(a) Top 3 - Count Method



(b) Distribution of Top-3 Counties

Figure 57: Top 3 - First approach

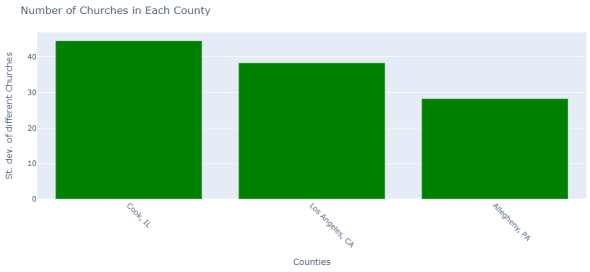
Looking at the provided Data-Frame, it is apparent that Los Angeles (CA) stands out as the county with the most diverse distribution of churches across various religions, boasting a total of 70 distinct churches. Following closely, Cook (IL) ranks second with 66 different churches, and Wayne (ML) holds the third position with 61 diverse churches representing different religions.

- **Second approach:** We'll compute the standard deviation for the number of churches of each religion within each county.This approach differs as it takes into account the variation in the number of churches for specific religions.

Findings :

	CNAME	STD_DEV
574	Cook, IL	44.532370
172	Los Angeles, CA	38.327505
2209	Allegheny, PA	28.252443

(a) Top 3 - Standard Divination Method



(b) Distribution of Top-3 Counties

Figure 58: Top 3 - Second approach

Upon reviewing both the Data-Frame and the bar plots depicting the standard deviation in the number of churches for each religion within every county, it becomes evident that Cook County (IL) holds the highest deviation at 44.53. Following closely is Los Angeles County (CA) with a standard deviation of 38.33, and Allegheny County (PA) secures the third position with a standard deviation of 28.25.

From the approaches above we can make some observations :

- Los Angeles (CA) stands out as the leading county in America, boasting the highest count of unique churches across various religions, specifically with 70 distinct churches. However, in terms of standard deviation, it takes the second position, considering both the variety and number of churches for each religion within the county.
- Cook (IL) County secures the second position in the count of unique churches with 66 distinct establishments. Interestingly, it claims the first position in standard deviation, indicating a notable diversity in the number of churches for each unique religion.
- Wayne (MI) claims the third position in terms of the count of unique churches. In the standard deviation of the number of churches, Allegheny (PA) rank third, respectively, showcasing a mix of unique churches and variability in religious institutions.

2.5 Task 5: Where would you create a cross-religion center of discussion between religions to maximize its impact? Support the proposal based on data analysis results

We will start by creating a similar dataset as the one from 'Task 4' which will use the names of every county and the total numbers of members of every religion in each county. The dataset

will look like that :

	CNAME	SDA_M	AOG_M	ABC_M	SBC_M	COB_M	COGT_M	COGI_M	CGC_M	NAZRN_M	...	RECH_M	SOCBR_M	NSAC_M	UNCHU_M	U
0	Hale, AL	0	0	0	1522	0	0	86	0	55	...	0	0	0	0	
1	Henry, AL	0	23	0	4051	0	0	0	0	0	...	0	0	0	0	
2	Houston, AL	87	379	0	11921	0	14	0	0	29	...	0	0	0	0	
3	Jackson, AL	110	0	0	7875	0	252	0	0	37	...	0	0	0	0	
4	Jefferson, AL	991	648	0	94219	0	2070	935	0	634	...	0	0	0	47	

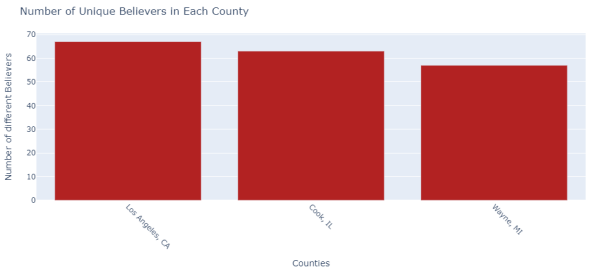
Figure 59: Total Members - 5 first rows

Using that dataset, we now apply the first approach from 'Task 4', the count method, to determine the diversity of religions in each county .

Findings :

	CNAME	DIFMEMBERS
172	Los Angeles, CA	67
574	Cook, IL	63
1276	Wayne, MI	57

(a) Top 3 - Count Method



(b) Distribution of Top-3 Counties

Figure 60: Top 3

Upon analyzing the provided Data-Frame and the related bar plot, it can be observed that Los Angeles (CA) is the county with the greatest diversity of believers from different religions, consisting of 67 different groups. Following closely behind, Cook (IL) claims the second place with 63 distinct parts of believers, and Wayne (ML) claims the third place with 57 varied groups representing various faiths.

Taking account the number of unique believers and unique churches let's check once again what we have.

	CNAME	DIFMEMBERS	CNAME	DIFCHURCHES
172	Los Angeles, CA	67	Los Angeles, CA	70
574	Cook, IL	63	Cook, IL	66
1276	Wayne, MI	57	Wayne, MI	61

Figure 61: Final Diversity Results

It is reasonable to observe diversity in religions and churches concentrated in the same geographic areas, and this outcome aligns with our expectations.

Considering the various datasets, bar plots, and maps, the selection of an optimal location for establishing a cross-religion center of discussion depends on multiple factors. While Los Angeles appears favorable based solely on the distribution of religion's members and churches, a more comprehensive analysis that considers both the geographic positions of counties and the distribution of religion's members and churches suggests a county Cook or Wayne, and specifically Wayne. This preference is influenced by the combined population of Cook and Wayne being higher than that of Los Angeles. Additionally, the surrounding counties near especially Wayne boast larger populations such New York, compared to the area around Los Angeles.

3 Conclusion

This assignment centered around the analysis of two diverse datasets – one pertaining to airline flights across America, with a specific focus on delays, and the other exploring the religious affiliations within each county. Both datasets provided a real-world perspective, offering intriguing challenges and necessitating distinct analytical approaches to address varied questions. Real-world datasets' richness adds a level of complexity to the analysis process, but it also makes them incredibly useful by offering possibilities for learning through a variety of analytical tools and rich insights.