Flight Delay and Cancellation Prediction

Team 4

Dmitry Veretennikov

Gauri Tiwari

Mostafa Omidi

Poojitha Anuvala Setty

Sai Siva Sri Harsha Vasinenku

George Mason University

AIT614-005 Big Data Essentials
(Spring 2024)

Dr. Lindi Liao

April 23, 2024

INTRODUCTION

- Every day in the US, 2.9 million passengers fly from nearly 20,000 airports on 45,000 flights (FAA, 2023).
- In 2023, over 200 million US passengers faced flight delays and cancellations, costing $30-34 billion (Junginger, 2023).
- Our project aims to analyze flight data in the Washington DC area to understand patterns and causes of delays and cancellations using Machine Learning techniques.

OBJECTIVES

- Determine the timing of high-risk periods for flight delays and cancellations departing from Washington DC.
- Develop a predictive model to estimate the probability of flight delay and cancellation.

PROJECT TIMELINE

The project is divided into four main tasks, each with designated team members and timelines.

Data Acquisition and Preprocessing:

Team Members:
Mostafa, Dmitry
Start Date: 3/13/2024
End Date: 3/19/2024

Exploratory Analysis:

Team Members:
Dmitry
Start Date: 3/20/2024
End Date: 3/26/2024

Building and Comparing ML Models:

Team Members:
Dmitry, Poojitha, Harsha
Start Date: 3/27/2024
End Date: 4/9/2024

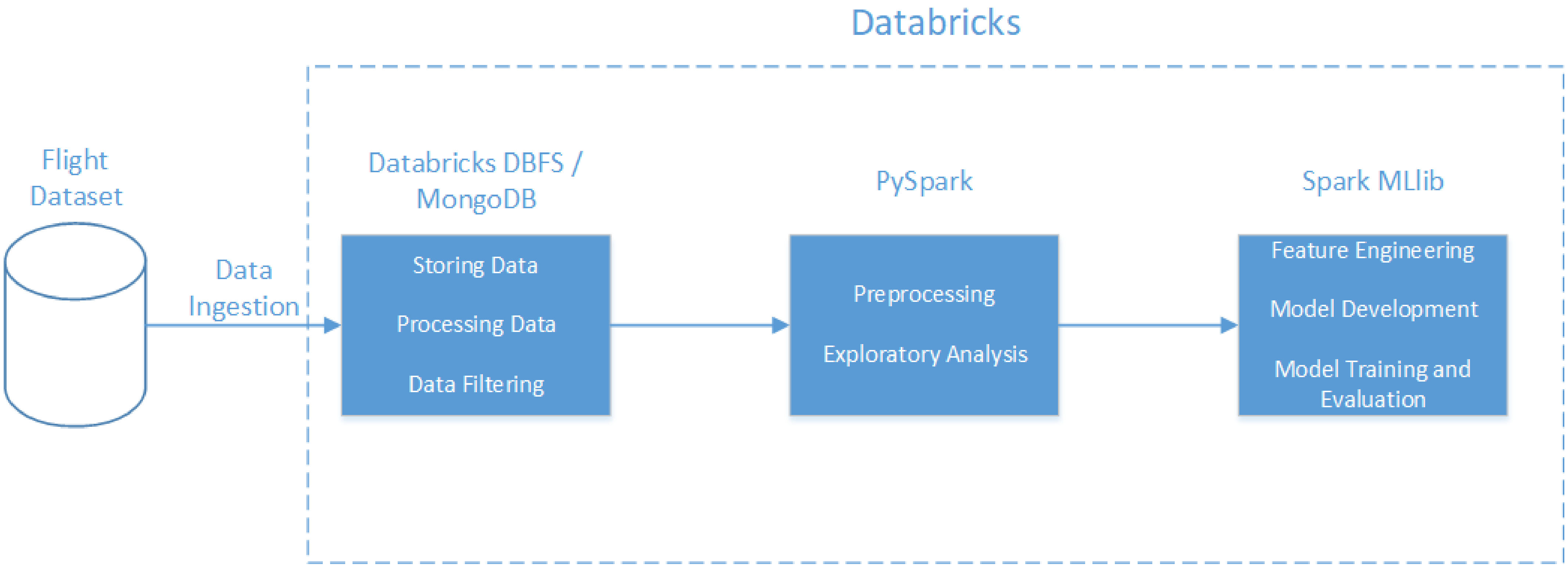
Report Writing and Presentation Preparation:

Team Members:
Poojitha, Gauri
Start Date: 4/10/2024
End Date: 4/16/2024

DATASET

- The proposed dataset (Zelazko, 2023) consists of 3 million flight records. Among these, over 79,000 flights originated from Washington, DC.
- The dataset comprises 32 original features (Appendix 1) categorized into 19 decimals, 8 strings, 4 integers, 1 date

ARCHITECTURE



Databricks

Flight Dataset

Data Ingestion

Databricks DBFS / MongoDB

Storing Data

Processing Data

Data Filtering

PySpark

Preprocessing

Exploratory Analysis

Spark MLlib

Feature Engineering

Model Development

Model Training and Evaluation

DATA PROCESSING

- **Data Handling:**
  - Uploading dataset to Databricks DBFS.
  - Filtering and selecting flights from Washington DC.
  - Cleaning data.
  - Converting data to appropriate data types.
  - Introducing new variables.
  - Balancing dataset (oversampling).

ANALYTIC APPROACHES

Exploratory Analysis with PySpark:

- Utilize PySpark for data exploration.
- Analyze data distribution, trends, and correlations.
- Visualize insights for understanding flight patterns.

Machine Learning Model Development using Spark MLlib:

- Employ Spark MLlib for predictive modeling.
- Train models to forecast flight delays and cancellations.
- Experiment with diverse algorithms for optimal performance.

HARDWARE AND SOFTWARE DEVELOPMENT PLATFORMS

Databricks:

Cloud-based platform providing both hardware infrastructure and software tools.

Primary platform for data processing, analysis, and model development.

PySpark:

Apache Spark for distributed data processing.

Used for data manipulation, analysis, and visualization.

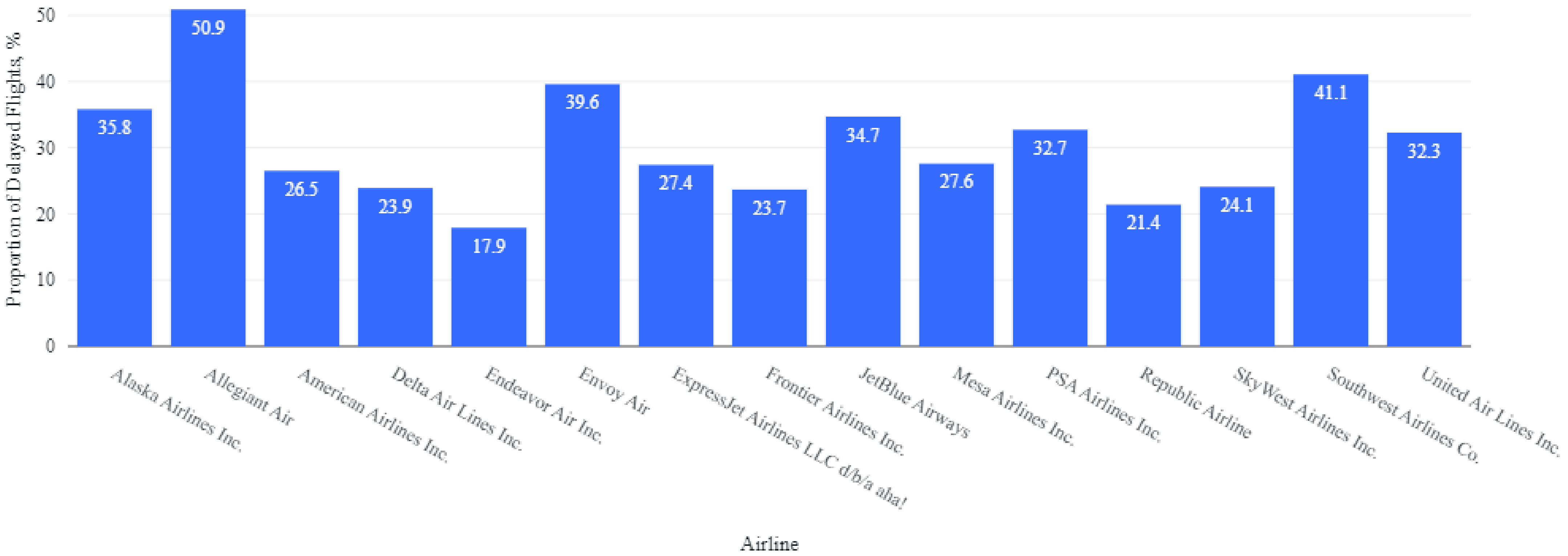
Spark MLlib:

Machine learning library for building predictive models within Spark.

Integrates seamlessly with PySpark for streamlined workflow.

EXPLORATORY ANALYSIS RESULTS: DELAYS BY AIRLINES

- Allegiant Air has the biggest proportion of delayed flights (50.9%).
- The most punctual airline is Endeavor Air (17.9% of delays).

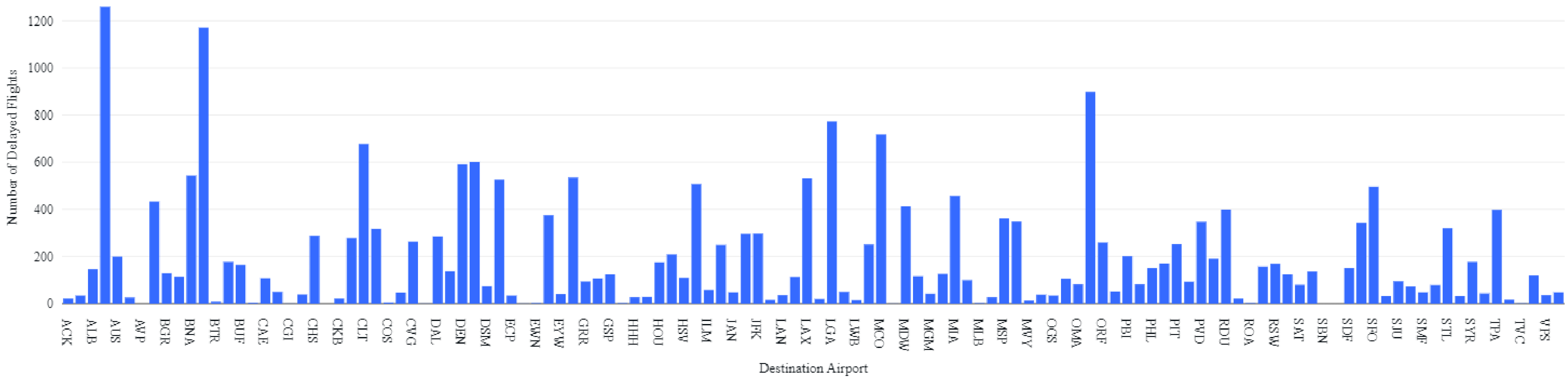


Proportion of Delayed Flights by Airline (%)

EXPLORATORY ANALYSIS RESULTS: DELAYS BY DESTINATION AIRPORTS

The airports with maximum number of delays are:

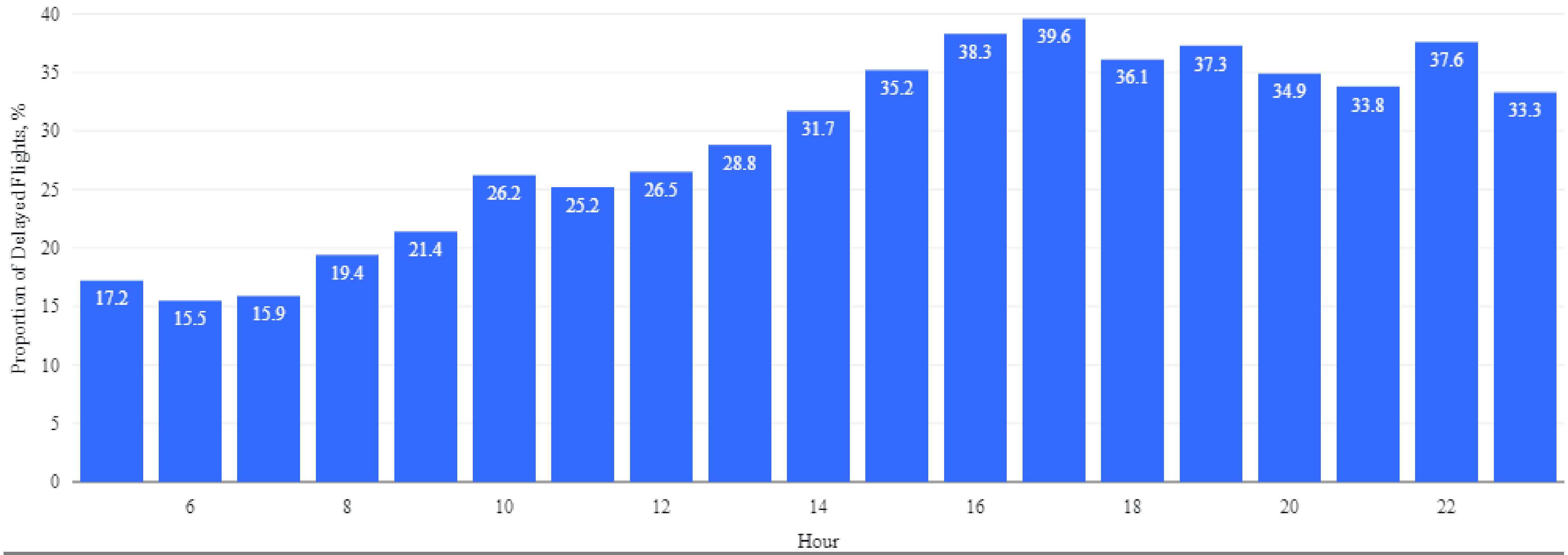
- Hartsfield-Jackson Atlanta International Airport (1,259 delays),
- Boston Logan International Airport (1,171 delays),
- Chicago O'Hare International Airport (898 delays).



Number of Delayed Flights by Destination Airport

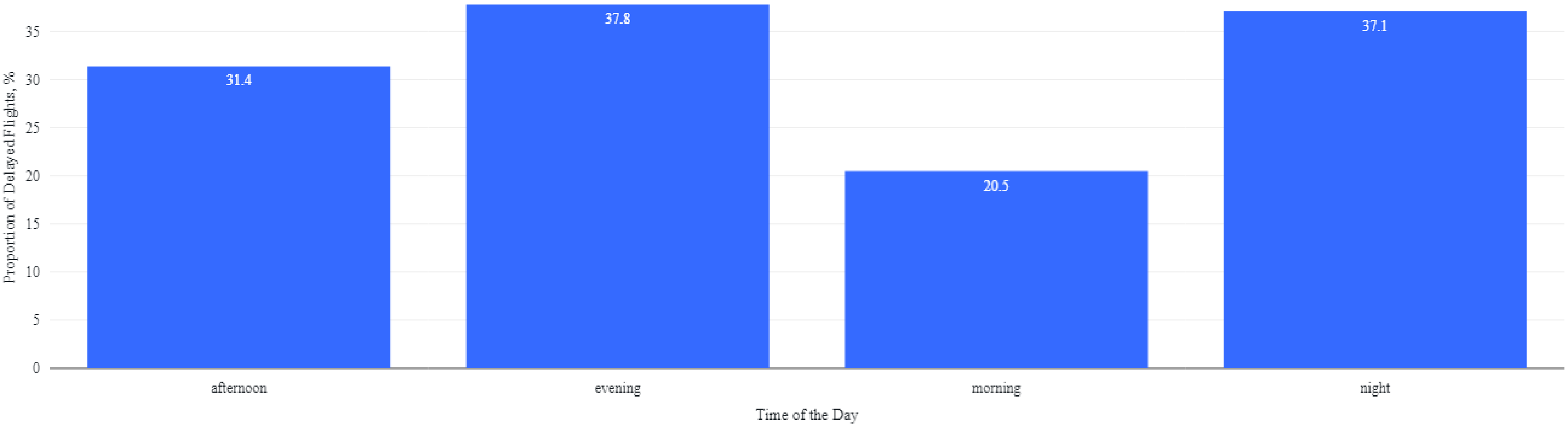
EXPLORATORY ANALYSIS: DELAYS BY HOUR

- Flights from 2 pm to midnight have the biggest proportion of delays (33.3-39.6%).



Proportion of Delayed Flights by Hour

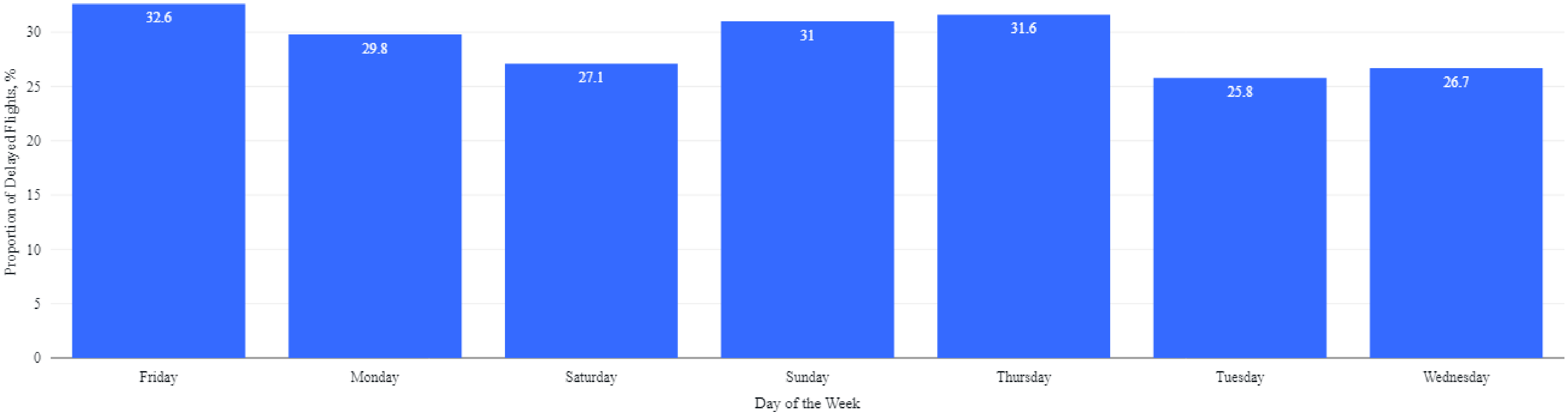
EXPLORATORY ANALYSIS RESULTS: DELAYS BY THE TIME OF THE DAY



Proportion of Delayed Flights by the Time of the Day (%)

Evenings (37.8%) and nights (37.1%) have more flight delays compared to the other times.

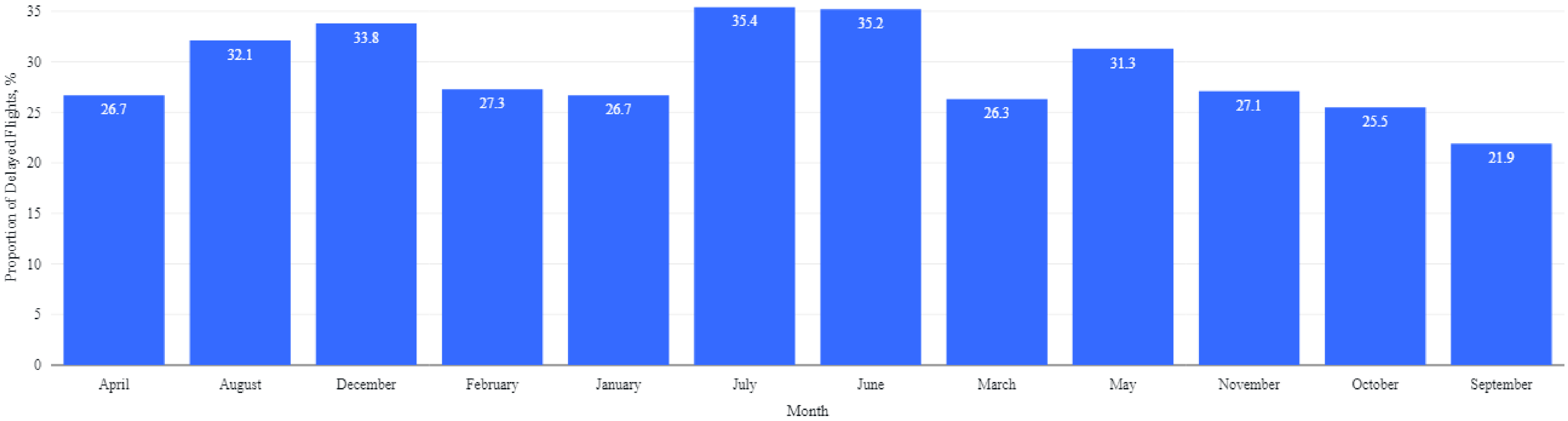
EXPLORATORY ANALYSIS RESULTS: DELAYS BY THE DAY OF THE WEEK



Proportion of Delayed Flights by Day of the Week (%)

Most of the delayed flights are on Friday(32.6%), followed by Thursday(31.6%).

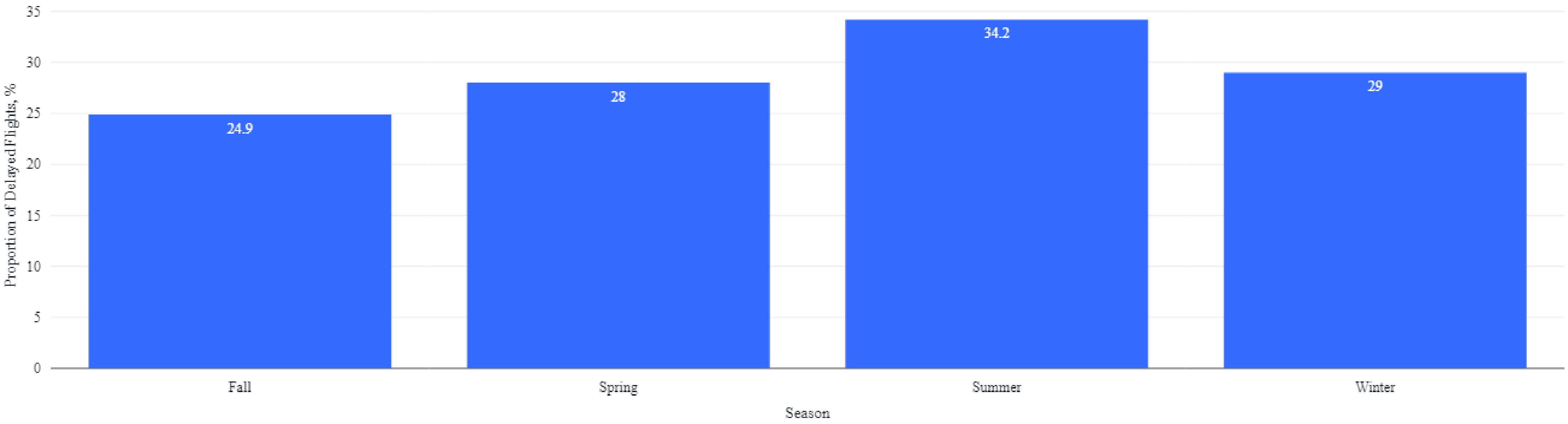
EXPLORATORY ANALYSIS RESULTS: DELAYS BY MONTH



Proportion of Delayed Flights by Month (%)

July(35.4%) has the most delayed flights followed by June(35.2%).

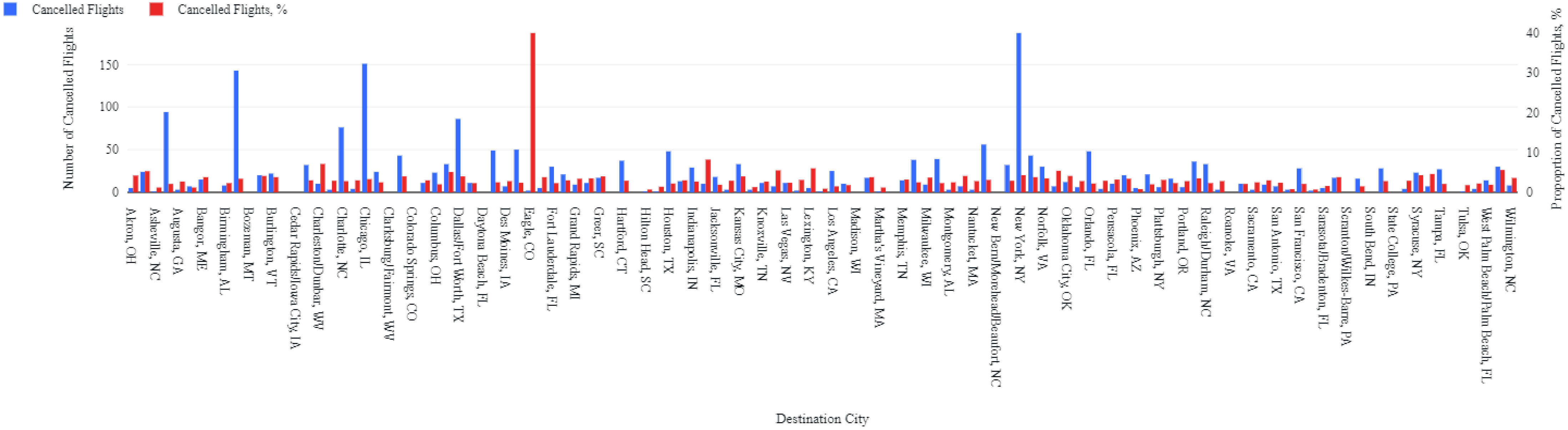
EXPLORATORY ANALYSIS RESULTS: DELAYS BY THE SEASON



Proportion of Delayed Flights by Season(%)

Summer is the season with most delays (34.2%).

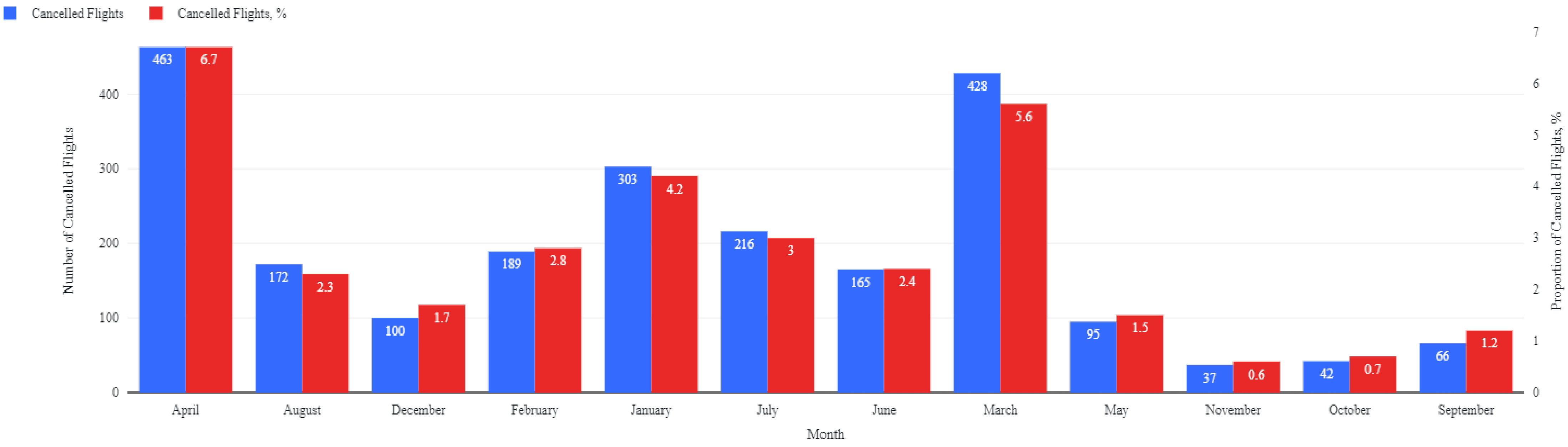
EXPLORATORY ANALYSIS RESULTS: CANCELLATIONS BY THE DESTINATION CITY



Cancelled Flights by the Destination City

The destination city with the maximum number of cancellations is New York (187 flights). However, the biggest proportion of the cancelled flights (40%) have Eagle, CO as a destination city.

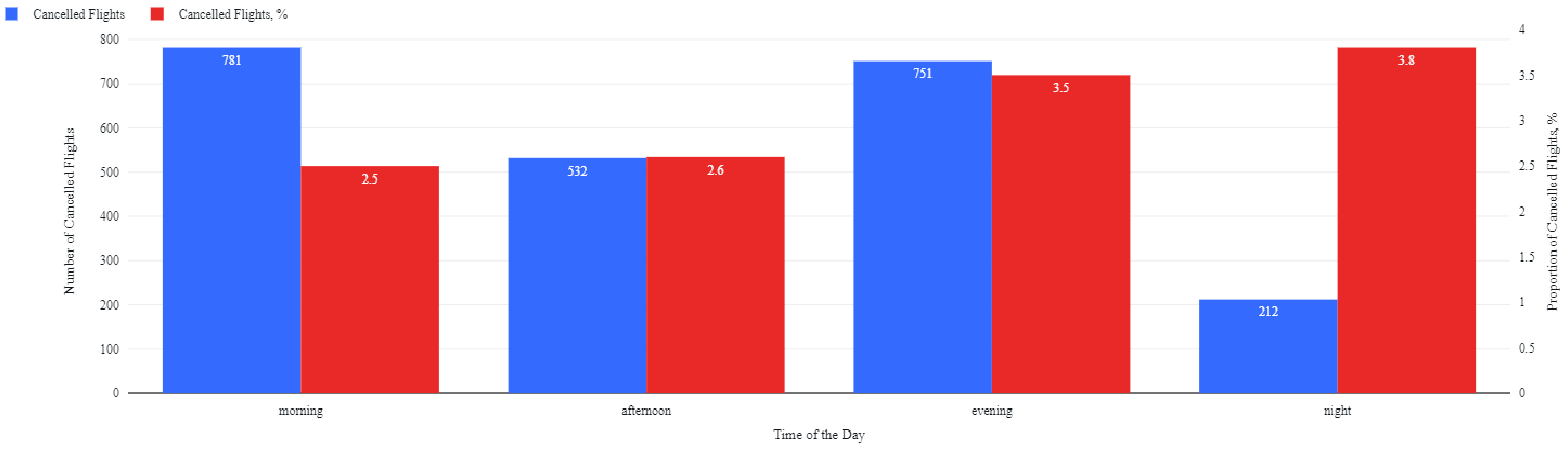
EXPLORATORY ANALYSIS RESULTS: CANCELLATIONS BY MONTH



Cancelled Flights by Month

April (463 flights, or 6.7%) has the most cancelled flights followed by March (428 flights, or 5.6%).

EXPLORATORY ANALYSIS RESULTS: CANCELLATIONS BY THE TIME OF THE DAY



Cancelled Flights by the Time of the Day

Mornings has the most cancelled flights (781 flights) followed by Evening (751 flights). However, the biggest proportion of flights was cancelled at night (3.8%).

SOME OTHER EXPLORATORY ANALISYS RESULTS

- Reagan Washington Airport and Dulles International Airport have the same proportion of delayed flights (29.3%) and average delay time (45.2-46.2 min).
- Destination airport with the max proportion of delayed flights – CGI and DAB (100%)
- Destination airport with the minimum proportion of delayed flights – AVP, CID, SBN, SCE (0%)

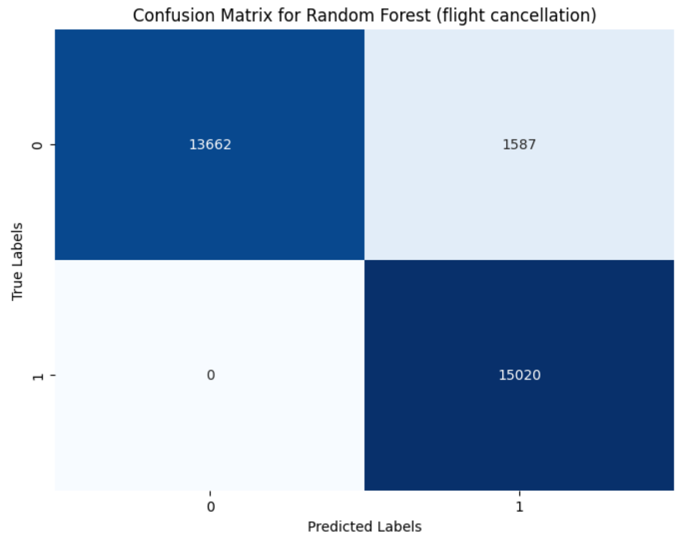
MACHINE LEARNING RESULTS

Model	Cancellation Prediction Accuracy	Delay Prediction Accuracy
Random Forest	0.95	0.69
Logistic Regression	0.73	0.64
XG Boost	0.94	0.68
Gradient Boosting	0.92	0.66
Artificial Neural Network	0.62	0.54

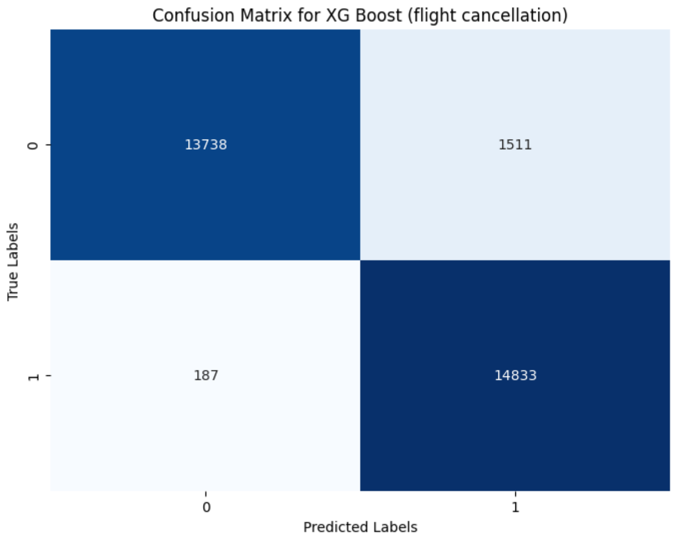
CANCELLATION PREDICTION: BEST MODELS

Accuracy:

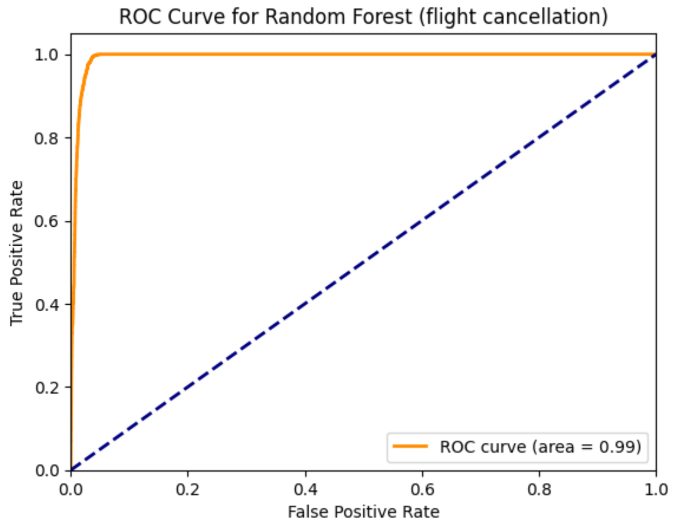
- Random Forest – 0.95
- XG Boost – 0.94



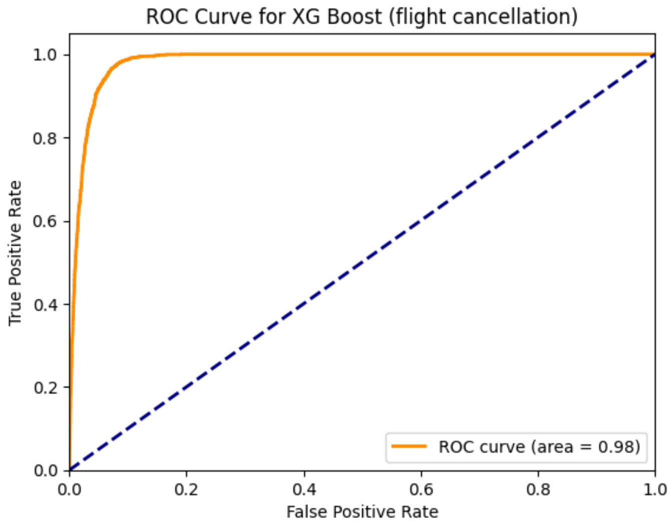
Confusion Matrix for Random Forest (flight cancellation)



Confusion Matrix for XG Boost (flight cancellation)



ROC Curve for Random Forest (flight cancellation)

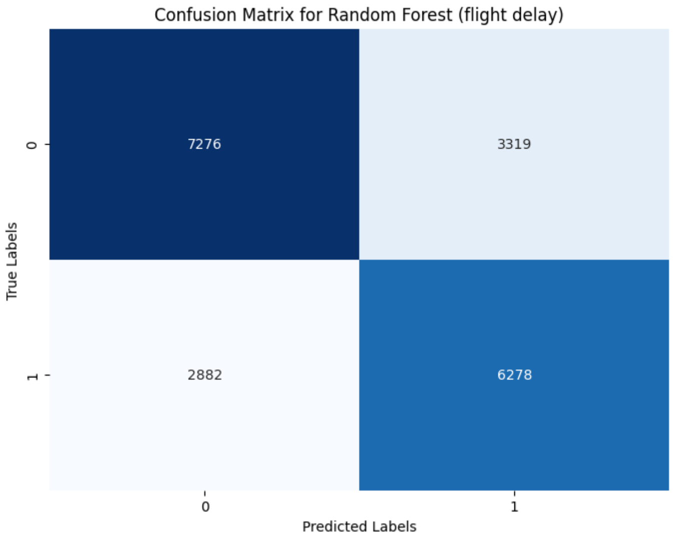


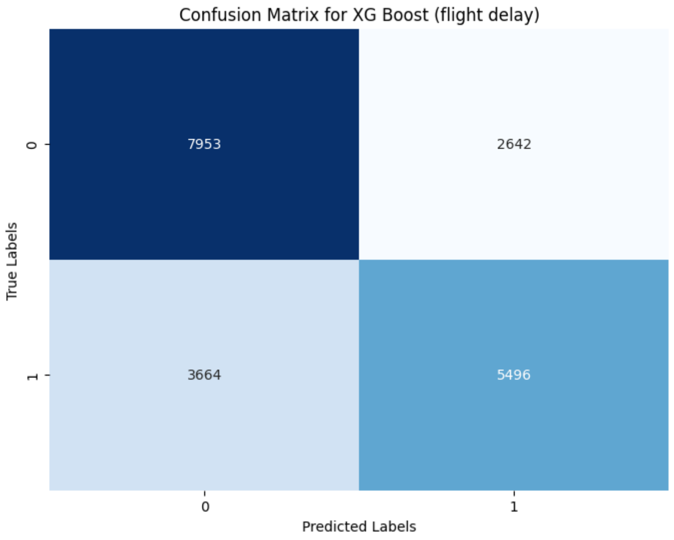
ROC Curve for XG Boost (flight cancellation)

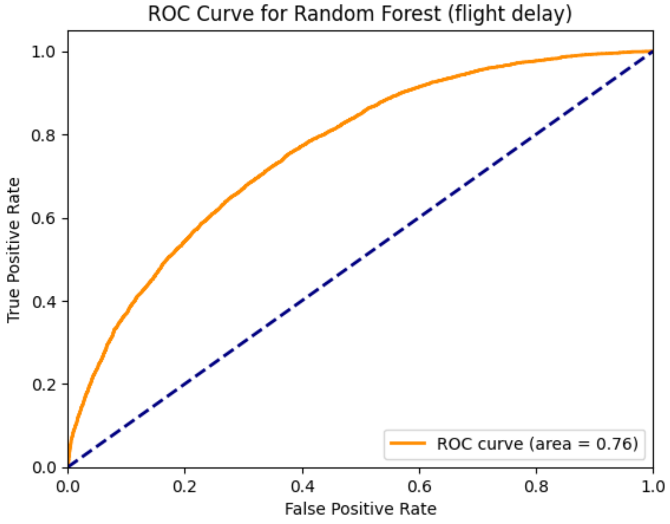
DELAY PREDICTION: BEST MODELS

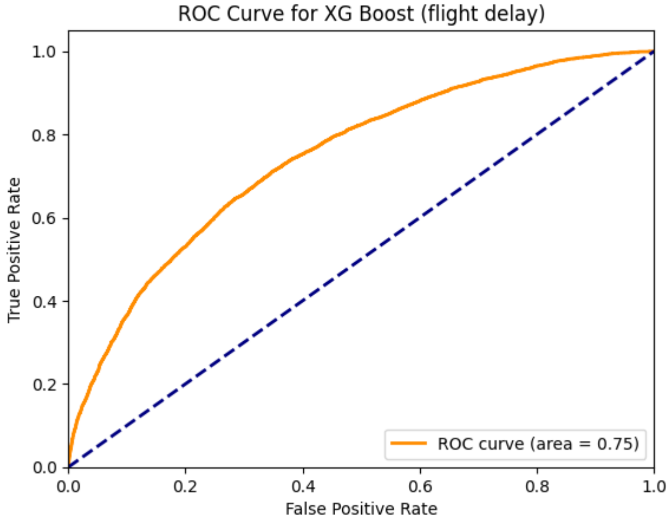
Accuracy:

- Random Forest – 0.69
- XG Boost – 0.68



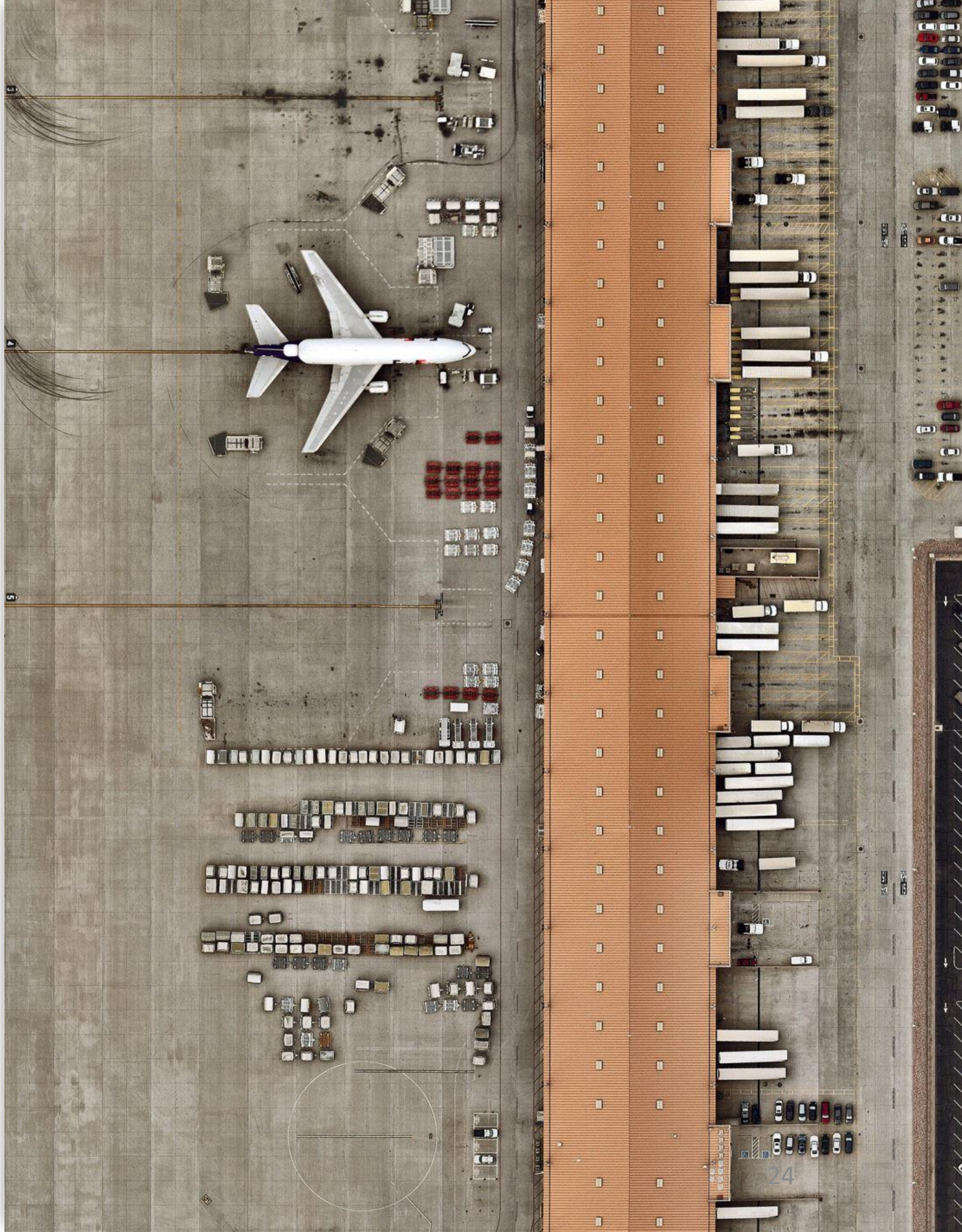






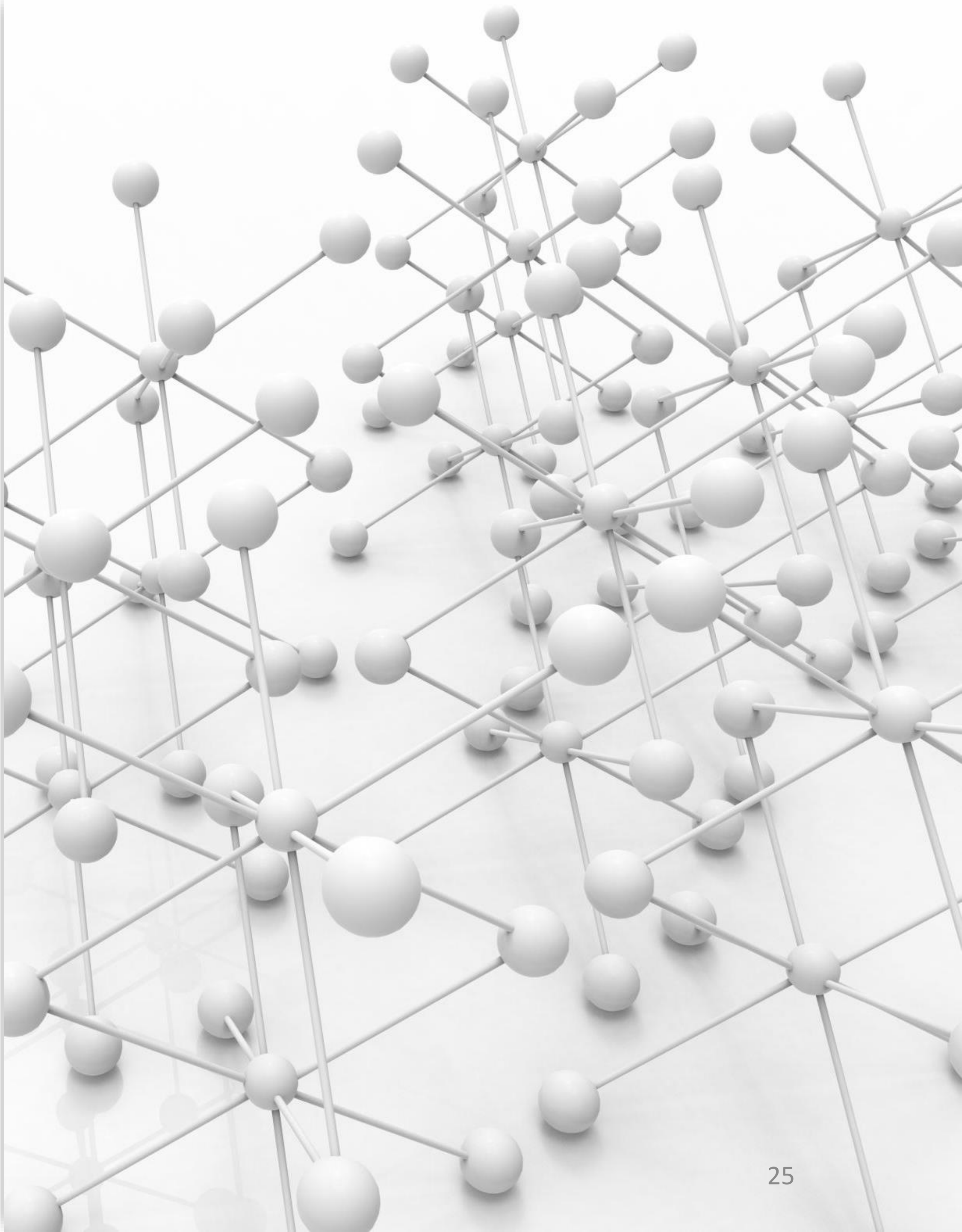
CONCLUSION

- Through extensive data processing and analysis, we gained valuable insights into flight patterns and disruptions in the Washington DC area.
- Random Forest and XG Boost demonstrated high accuracy in predicting flight cancellations and moderate performance in predicting delays.
- The outcomes of our analysis and modeling have significant implications for travelers, airlines, and airport authorities, enabling informed decision-making and proactive measures to mitigate flight disruptions.



FUTURE WORK

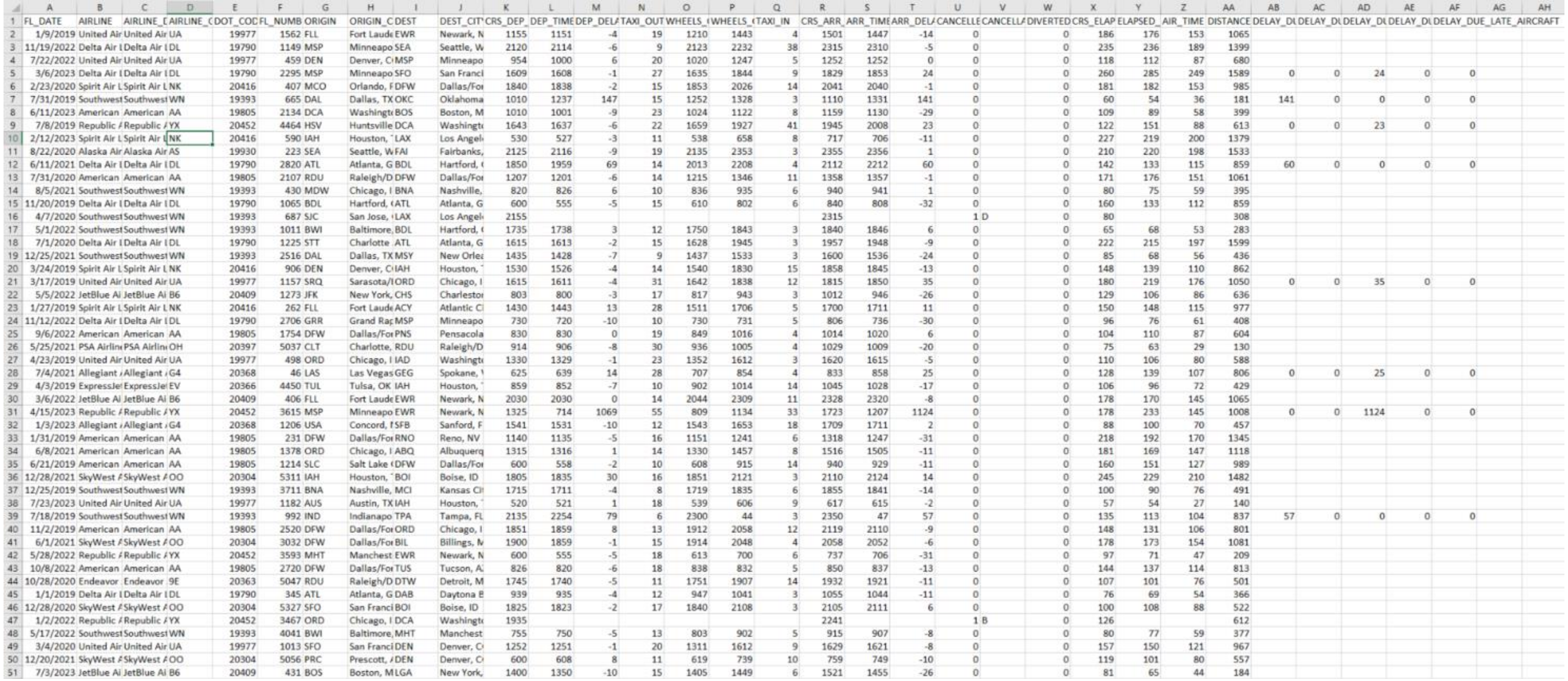
- Optimize parameters and adjust model architecture to improve performance.
- Identify new predictors to capture additional insights from the data.
- Test the model on other locations to find common and specific factors of flight delays and cancellations
- Include factors like weather conditions, holidays, and airport congestion for improved predictive accuracy.
- Systematically experiment with hyperparameters to find optimal configurations.



REFERENCES

- Balamurugan, R., Maria, A. V., Baranidaran, G., MaryGladence, L., & Revathy, S. (2022, April 1). *Error Calculation for Prediction of Flight Delays using Machine Learning Classifiers*. IEEE Xplore. https://doi.org/10.1109/ICOEI53556.2022.9776709. Retrieved March 12, 2024.
- Federal Aviation Administration. (2023, April 10). *Air Traffic By The Numbers*. https://www.faa.gov/air_traffic/by_the_numbers. Retrieved March 12, 2024.
- Junginger, J. (2023, October 6). *In numbers: The economic impact of flight disruptions*. AirHelp. https://www.airhelp.com/en-int/blog/in-numbers-the-economic-impact-of-flight-disruptions/. Retrieved March 12, 2024.
- Metropolitan Washington Airports Authority. (2024, February 22). *Washington's airports set new passenger record*. https://www.flydulles.com/news/washingtons-airports-set-new-passenger-record. Retrieved March 12, 2024.
- Publisher Research and Innovative Technology Administration. (2023, December 7). *Airline on-time performance and causes of flight delays*. Catalog. https://catalog.data.gov/dataset/airline-on-time- performance-and-causes-of-flight-delays. Retrieved March 12, 2024.
- Zelazko, P. (2023, December 8). *Flight Delay and Cancellation Dataset (2019-2023)*. Kaggle. https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023. Retrieved March 11, 2024.

APPENDIX 1. DATASET SAMPLE



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	FL_DATE	AIRLINE	AIRLINE_D	AIRLINE_C	DOT_COD	FL_NUMB	ORIGIN	ORIGIN_C	DEST	DEST_CIT	CRS_DEP_	DEP_TIME	DEP_DEL	TAXI_OUT	WHEELS_(	WHEELS_(	TAXI_IN	CRS_ARR_	ARR_TIME	ARR_DEL	CANCELLE	CANCELL	DIVERTED	CRS_ELAP	ELAPSED_	AIR_TIME	DISTANCE	DELAY_DU	DELAY_DU	DELAY_DU	DELAY_DU	DELAY_DUE_LATE_AIRCRAFT		
2	1/9/2019	United Air	United Air	UA	19977	1562	FLL	Fort Laude	EWR	Newark, N	1155	1151	-4	19	1210	1443	4	1501	1447	-14	0		0	186	176	153	1065							
3	11/19/2022	Delta Air I	Delta Air I	DL	19790	1149	MSP	Minneapo	SEA	Seattle, W	2120	2114	-6	9	2123	2232	38	2315	2310	-5	0		0	235	236	189	1399							
4	7/22/2022	United Air	United Air	UA	19977	459	DEN	Denver, C	MSP	Minneapo	954	1000	6	20	1020	1247	5	1252	1252	0	0		0	118	112	87	680							
5	3/6/2023	Delta Air I	Delta Air I	DL	19790	2295	MSP	Minneapo	SFO	San Franci	1609	1608	-1	27	1635	1844	9	1829	1853	24	0		0	260	285	249	1589	0	0	24	0	0		
6	2/23/2020	Spirit Air L	Spirit Air L	NK	20416	407	MCO	Orlando, F	DFW	Dallas/For	1840	1838	-2	15	1853	2026	14	2041	2040	-1	0		0	181	182	153	985							
7	7/31/2019	Southwest	Southwest	WN	19393	665	DAL	Dallas, TX	OKC	Oklahoma	1010	1237	147	15	1252	1328	3	1110	1331	141	0		0	60	54	36	181	141	0	0	0	0		
8	6/11/2023	American	American	AA	19805	2134	DCA	Washingt	BOS	Boston, M	1010	1001	-9	23	1024	1122	8	1159	1130	-29	0		0	109	89	58	399							
9	7/8/2019	Republic /	Republic /	YX	20452	4464	HSV	Huntsville	DCA	Washingt	1643	1637	-6	22	1659	1927	41	1945	2008	23	0		0	122	151	88	613	0	0	23	0	0		
10	2/12/2023	Spirit Air L	Spirit Air L	NK	20416	590	IAH	Houston,	LAX	Los Angel	530	527	-3	11	538	658	8	717	706	-11	0		0	227	219	200	1379							
11	8/22/2020	Alaska Air	Alaska Air	AS	19930	223	SEA	Seattle, W	FAI	Fairbanks,	2125	2116	-9	19	2135	2353	3	2355	2356	1	0		0	210	220	198	1533							
12	6/11/2021	Delta Air I	Delta Air I	DL	19790	2820	ATL	Atlanta, G	BDL	Hartford, (	1850	1959	69	14	2013	2208	4	2112	2212	60	0		0	142	133	115	859	60	0	0	0	0		
13	7/31/2020	American	American	AA	19805	2107	RDU	Raleigh/D	DFW	Dallas/For	1207	1201	-6	14	1215	1346	11	1358	1357	-1	0		0	171	176	151	1061							
14	8/5/2021	Southwest	Southwest	WN	19393	430	MDW	Chicago, I	BNA	Nashville,	820	826	6	10	836	935	6	940	941	1	0		0	80	75	59	395							
15	11/20/2019	Delta Air I	Delta Air I	DL	19790	1065	BDL	Hartford, (	ATL	Atlanta, G	600	555	-5	15	610	802	6	840	808	-32	0		0	160	133	112	859							
16	4/7/2020	Southwest	Southwest	WN	19393	687	SJC	San Jose, (	LAX	Los Angel	2155							2315			1	D	0	80			308							
17	5/1/2022	Southwest	Southwest	WN	19393	1011	BWI	Baltimore,	BDL	Hartford, (	1735	1738	3	12	1750	1843	3	1840	1846	6	0		0	65	68	53	283							
18	7/1/2020	Delta Air I	Delta Air I	DL	19790	1225	STT	Charlotte	ATL	Atlanta, G	1615	1613	-2	15	1628	1945	3	1957	1948	-9	0		0	222	215	197	1599							
19	12/25/2021	Southwest	Southwest	WN	19393	2516	DAL	Dallas, TX	MSY	New Orlea	1435	1428	-7	9	1437	1533	3	1600	1536	-24	0		0	85	68	56	436							
20	3/24/2019	Spirit Air L	Spirit Air L	NK	20416	906	DEN	Denver, C	IAH	Houston,	1530	1526	-4	14	1540	1830	15	1858	1845	-13	0		0	148	139	110	862							
21	3/17/2019	United Air	United Air	UA	19977	1157	SRQ	Sarasota/I	ORD	Chicago, I	1615	1611	-4	31	1642	1838	12	1815	1850	35	0		0	180	219	176	1050	0	0	35	0	0		
22	5/5/2022	JetBlue Ai	JetBlue Ai	B6	20409	1273	JFK	New York,	CHS	Charlestor	803	800	-3	17	817	943	3	1012	946	-26	0		0	129	106	86	636							
23	1/27/2019	Spirit Air L	Spirit Air L	NK	20416	262	FLL	Fort Laude	ACY	Atlantic C	1430	1443	13	28	1511	1706	5	1700	1711	11	0		0	150	148	115	977							
24	11/12/2022	Delta Air I	Delta Air I	DL	19790	2706	GRR	Grand Rap	MSP	Minneapo	730	720	-10	10	730	731	5	806	736	-30	0		0	96	76	61	408							
25	9/6/2022	American	American	AA	19805	1754	DFW	Dallas/For	PNS	Pensacola	830	830	0	19	849	1016	4	1014	1020	6	0		0	104	110	87	604							
26	5/25/2021	PSA Airlin	PSA Airlin	OH	20397	5037	CLT	Charlotte,	RDU	Raleigh/D	914	906	-8	30	936	1005	4	1029	1009	-20	0		0	75	63	29	130							
27	4/23/2019	United Air	United Air	UA	19977	498	ORD	Chicago, I	IAD	Washingt	1330	1329	-1	23	1352	1612	3	1620	1615	-5	0		0	110	106	80	588							
28	7/4/2021	Allegiant /	Allegiant /	G4	20368	46	LAS	Las Vegas	GEG	Spokane, '	625	639	14	28	707	854	4	833	858	25	0		0	128	139	107	806	0	0	25	0	0		
29	4/3/2019	ExpressJe	ExpressJe	EV	20366	4450	TUL	Tulsa, OK	IAH	Houston,	859	852	-7	10	902	1014	14	1045	1028	-17	0		0	106	96	72	429							
30	3/6/2022	JetBlue Ai	JetBlue Ai	B6	20409	406	FLL	Fort Laude	EWR	Newark, N	2030	2030	0	14	2044	2309	11	2328	2320	-8	0		0	178	170	145	1065							
31	4/15/2023	Republic /	Republic /	YX	20452	3615	MSP	Minneapo	EWR	Newark, N	1325	714	1069	55	809	1134	33	1723	1207	1124	0		0	178	233	145	1008	0	0	1124	0	0		
32	1/3/2023	Allegiant /	Allegiant /	G4	20368	1206	USA	Concord, I	SFB	Sanford, F	1541	1531	-10	12	1543	1653	18	1709	1711	2	0		0	88	100	70	457							
33	1/31/2019	American	American	AA	19805	231	DFW	Dallas/For	RNO	Reno, NV	1140	1135	-5	16	1151	1241	6	1318	1247	-31	0		0	218	192	170	1345							
34	6/8/2021	American	American	AA	19805	1378	ORD	Chicago, I	ABQ	Albuquerq	1315	1316	1	14	1330	1457	8	1516	1505	-11	0		0	181	169	147	1118							
35	6/21/2019	American	American	AA	19805	1214	SLC	Salt Lake (	DFW	Dallas/For	600	558	-2	10	608	915	14	940	929	-11	0		0	160	151	127	989							
36	12/28/2021	SkyWest /	SkyWest /	OO	20304	5311	IAH	Houston,	BOI	Boise, ID	1805	1835	30	16	1851	2121	3	2110	2124	14	0		0	245	229	210	1482							
37	12/25/2019	Southwest	Southwest	WN	19393	3711	BNA	Nashville,	MCI	Kansas Ci	1715	1711	-4	8	1719	1835	6	1855	1841	-14	0		0	100	90	76	491							
38	7/23/2023	United Air	United Air	UA	19977	1182	AUS	Austin, TX	IAH	Houston,	520	521	1	18	539	606	9	617	615	-2	0		0	57	54	27	140							
39	7/18/2019	Southwest	Southwest	WN	19393	992	IND	Indianapo	TPA	Tampa, FL	2135	2254	79	6	2300	44	3	2350	47	57	0		0	135	113	104	837	57	0	0	0	0		
40	11/2/2019	American	American	AA	19805	2520	DFW	Dallas/For	ORD	Chicago, I	1851	1859	8	13	1912	2058	12	2119	2110	-9	0		0	148	131	106	801							
41	6/1/2021	SkyWest /	SkyWest /	OO	20304	3032	DFW	Dallas/For	BIL	Billings, M	1900	1859	-1	15	1914	2048	4	2058	2052	-6	0		0	178	173	154	1081							
42	5/28/2022	Republic /	Republic /	YX	20452	3593	MHT	Manchest	EWR	Newark, N	600	555	-5	18	613	700	6	737	706	-31	0		0	97	71	47	209							
43	10/8/2022	American	American	AA	19805	2720	DFW	Dallas/For	TUS	Tucson, A	826	820	-6	18	838	832	5	850	837	-13	0		0	144	137	114	813							
44	10/28/2020	Endeavor	Endeavor	9E	20363	5047	RDU	Raleigh/D	DTW	Detroit, M	1745	1740	-5	11	1751	1907	14	1932	1921	-11	0		0	107	101	76	501							
45	1/1/2019	Delta Air I	Delta Air I	DL	19790	345	ATL	Atlanta, G	DAB	Daytona B	939	935	-4	12	947	1041	3	1055	1044	-11	0		0	76	69	54	366							
46	12/28/2020	SkyWest /	SkyWest /	OO	20304	5327	SFO	San Franci	BOI	Boise, ID	1825	1823	-2	17	1840	2108	3	2105	2111	6	0		0	100	108	88	522							
47	1/2/2022	Republic /	Republic /	YX	20452	3467	ORD	Chicago, I	DCA	Washingt	1935							2241			1	B	0	126			612							
48	5/17/2022	Southwest	Southwest	WN	19393	4041	BWI	Baltimore,	MHT	Manchest	755	750	-5	13	803	902	5	915	907	-8	0		0	80	77	59	377							
49	3/4/2020	United Air	United Air	UA	19977	1013	SFO	San Franci	DEN	Denver, C	1252	1251	-1	20	1311	1612	9	1629	1621	-8	0		0	157	150	121	967							
50	12/20/2021	SkyWest /	SkyWest /	OO	20304	5056	PRC	Prescott, /	DEN	Denver, C	600	608	8	11	619	739	10	759	749	-10	0		0	119	101	80	557							
51	7/3/2023	JetBlue Ai	JetBlue Ai	B6	20409	431	BOS	Boston, M	LGA	New York,	1400	1350	-10	15	1405	1449	6	1521	1455	-26	0		0	81	65	44	184							