ISDS 577 Group 1

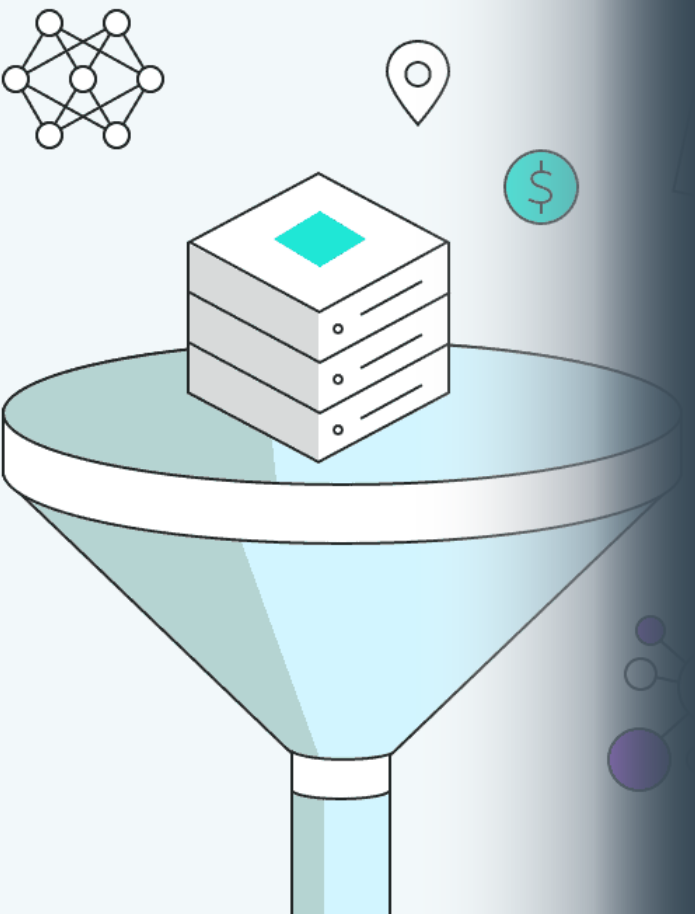
AIRLINE FLIGHT PERFORMANCE

Sairaj Prakash Jadhav, Ray Lien, Keti Lin, Hy Luong, Rucha Nilangekar

TABLE OF CONTENTS

DATASET

- *Kaggle.com*
- United States Bureau of Transportation Statistics
- 2 million flight from 1987 – 2020
- 92 attributes



DATA PREPARATION

Clean the data using R:

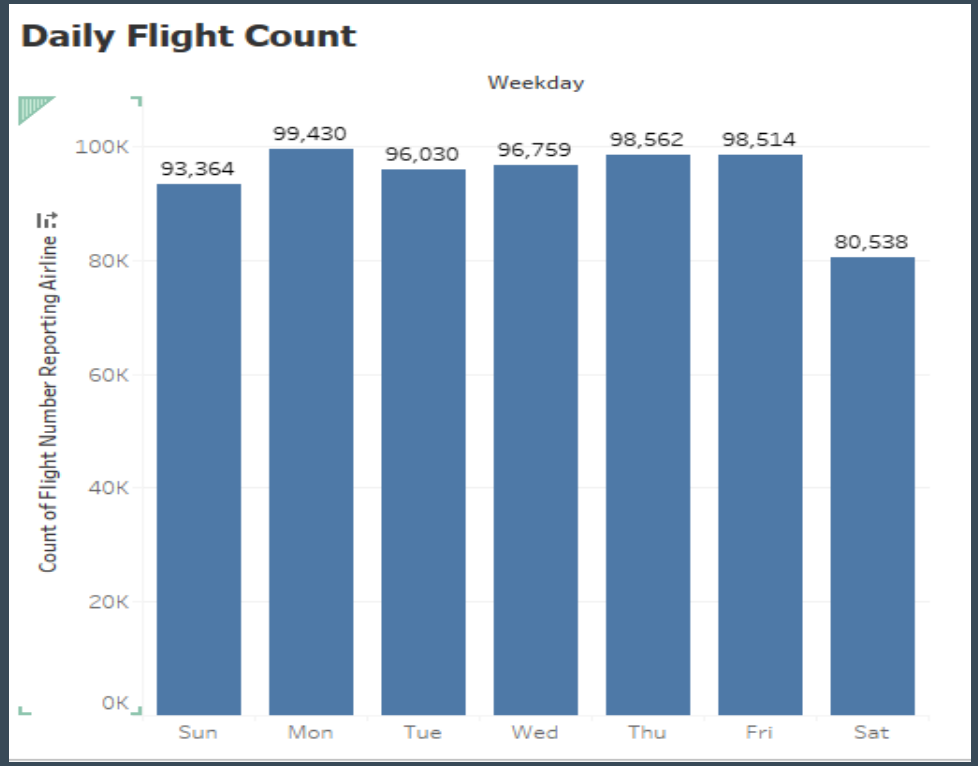
- Data from 2010 – 2020 for relevant.
- Remove empty columns
- Remove multicollinearity variables (|correlation| > 70%)
- Remove missing values
- Reduced from 2 mil rows to 650,000 rows
- 44 variables

Question 1

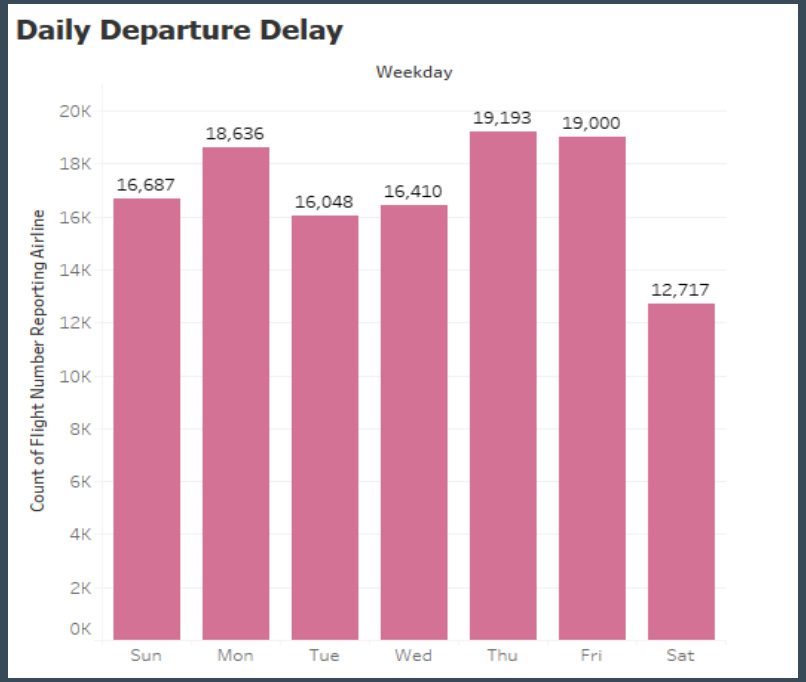
Which day of the week tends to have more delays or cancellations? What are the reasons ?

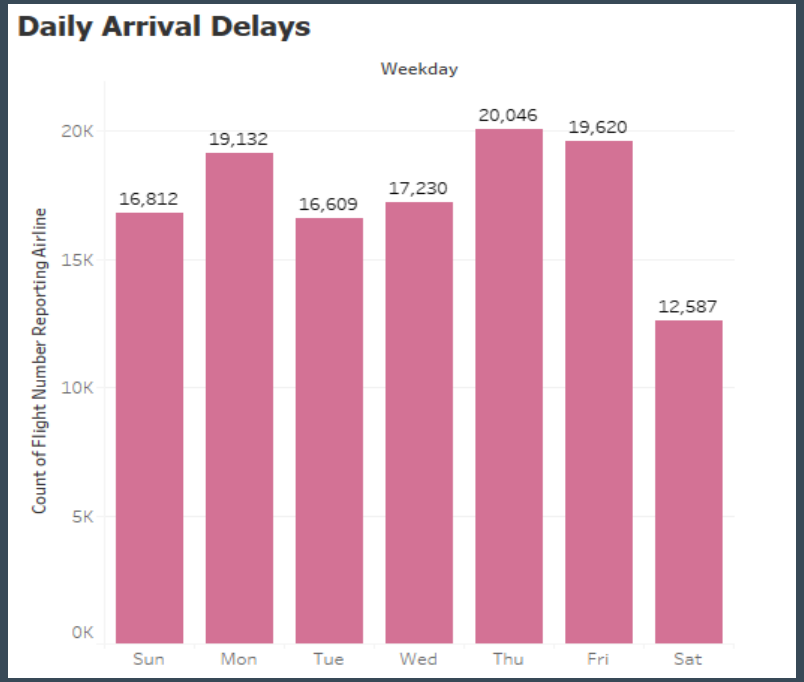
How does the on-time performance change across different months or seasons?

Daily Flight Count

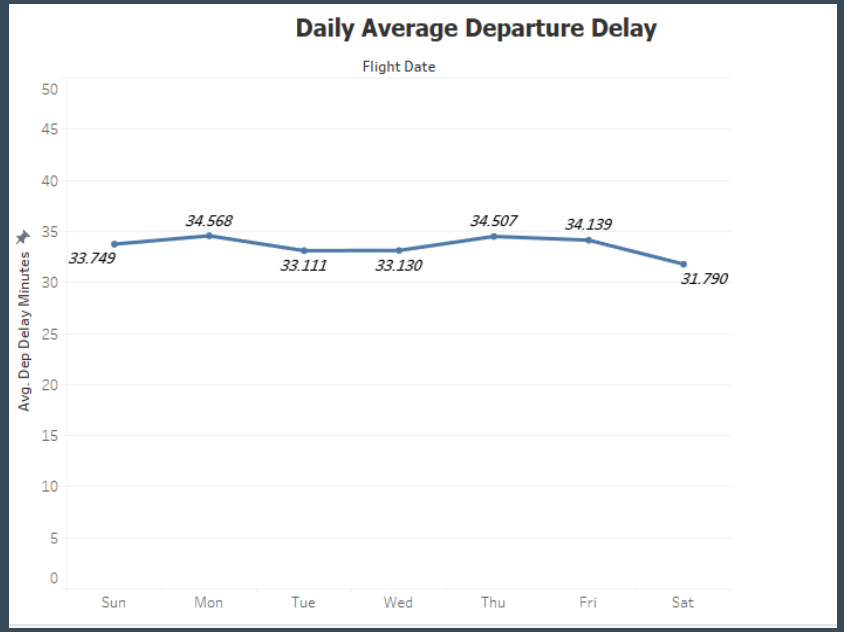


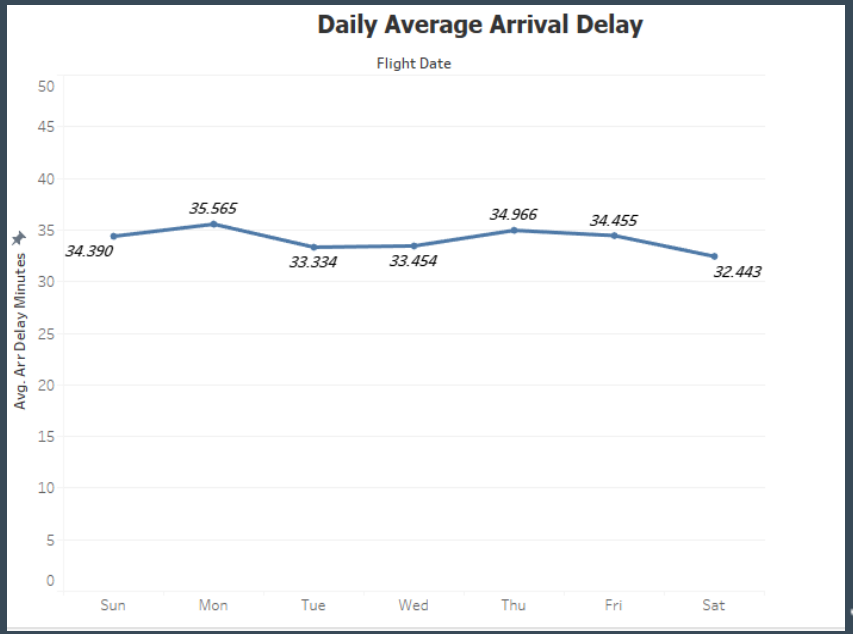
Daily Airline Delays



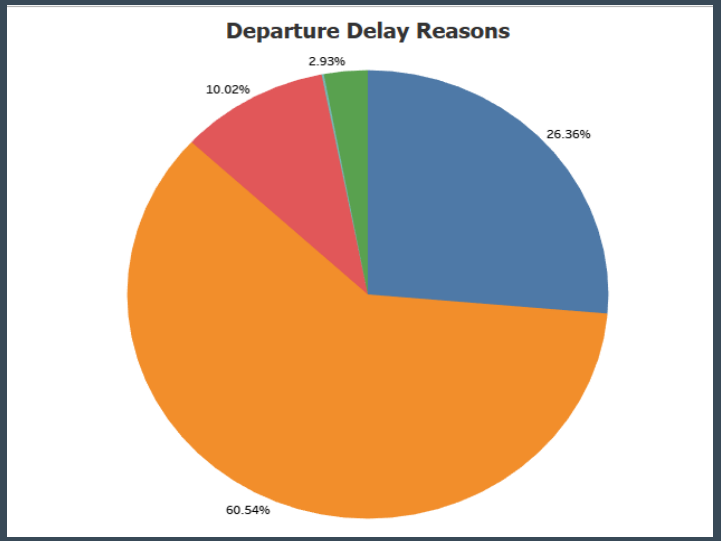


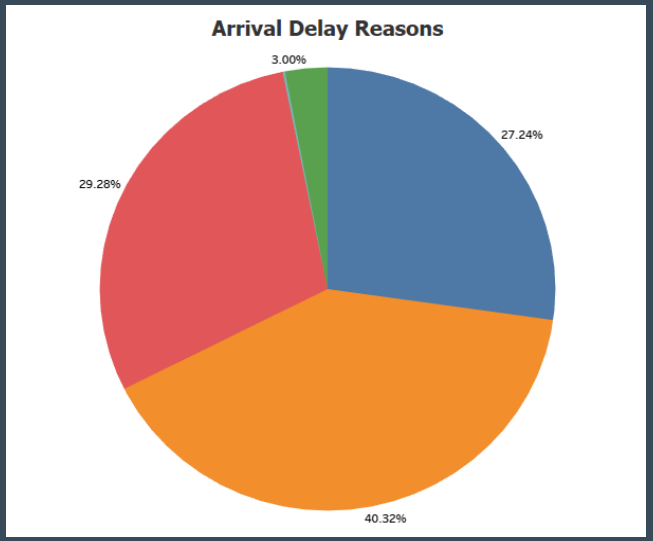
Daily Average Delay Minutes



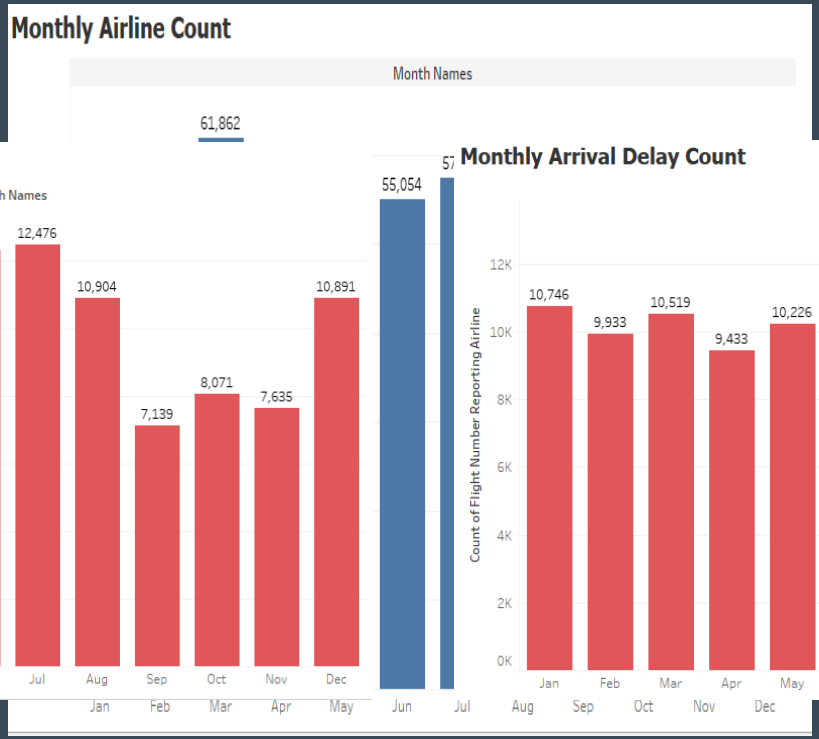


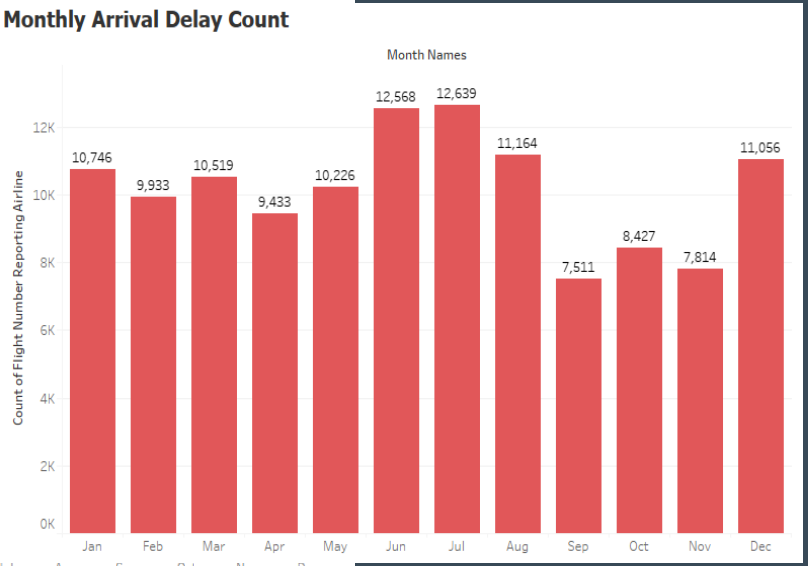
Delay Reasons



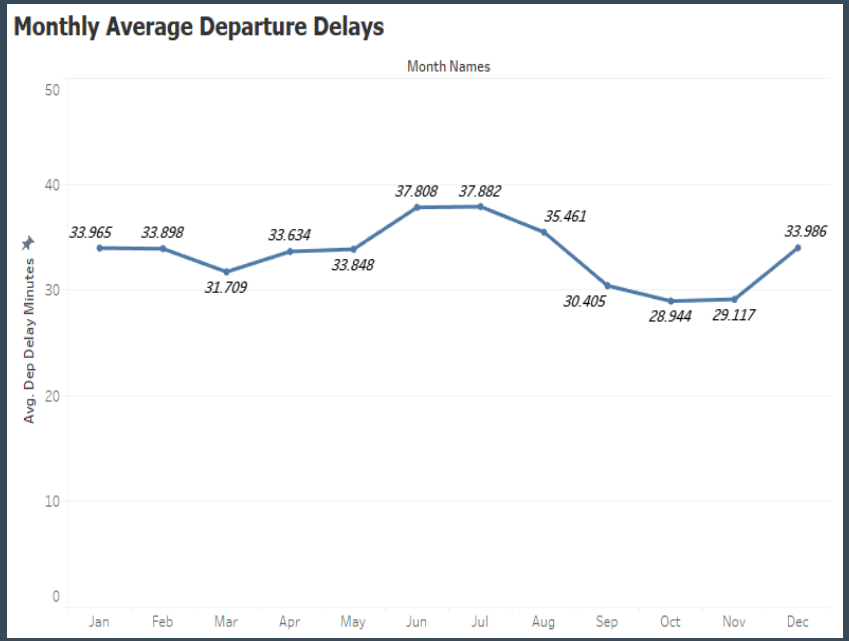


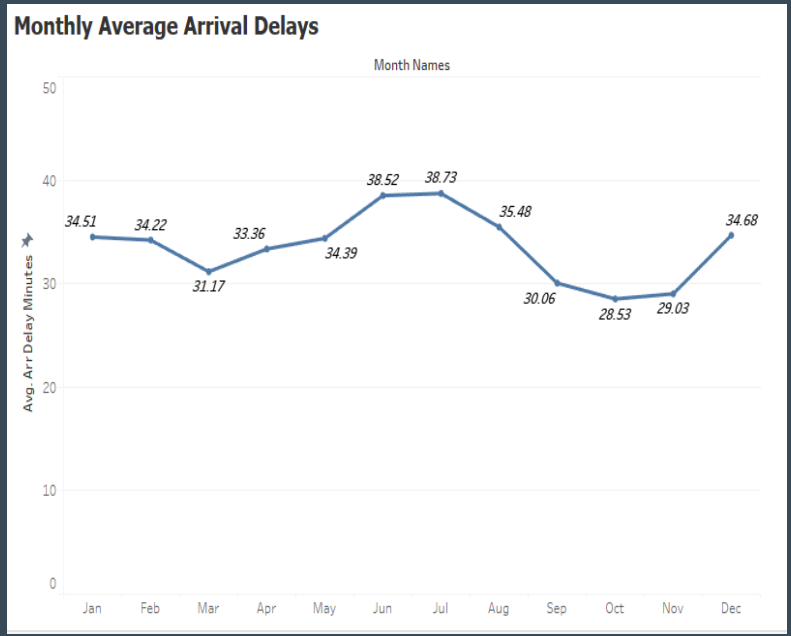
Monthly Airline Delays





Monthly Average Delays





Question 2

Are there certain types of flights (e.g., red-eye flights, early morning departures) that are more likely to experience delays?

Departure Delays for Each Airline

Count of Airline Depature Delays

	Reporting Airline																				
Dep Time Blk	9E	AA	AS	B6	CO	DL	EV	F9	FL	G4	HA	MQ	NK	OH	OO	UA	US	VX	WN	XE	YV	YX
0001-0559	174	1,784	431	710	110	1,646	882	295	63		268	405	221	276	1,538	1,203	506	6	2,208	98	199	186
0600-0659	811	4,808	1,366	1,845	368	5,909	3,263	867	620	159	296	2,405	627	395	5,418	4,176	1,305	83	9,455	652	590	646
0700-0759	625	5,997	1,555	1,845	384	6,448	2,271	619	671	215	496	2,040	734	568	3,735	3,808	1,962	455	8,305	397	635	389
0800-0859	690	5,133	1,279	1,672	328	6,526	2,776	599	719	124	652	1,946	538	295	4,185	4,086	1,600	258	8,997	491	802	480
0900-0959	729	4,358	1,125	1,606	306	5,381	2,826	420	353	140	611	1,861	488	610	4,010	3,312	1,508	339	7,711	449	727	434
1000-1059	847	4,531	1,144	1,564	259	4,921	3,496	786	872	149	611	2,185	400	361	4,229	3,110	1,208	221	8,673	599	674	385
1100-1159	712	4,518	1,115	1,663	295	6,119	2,901	491	655	104	463	2,059	487	663	4,890	3,154	1,733	241	7,615	478	643	503
1200-1259	843	4,585	987	1,311	331	5,674	3,504	418	681	143	492	2,297	413	450	4,147	3,307	1,070	210	7,676	548	1,015	455
1300-1359	767	4,495	1,089	1,484	300	5,458	2,981	513	539	152	612	1,824	372	578	4,964	3,165	1,457	252	7,908	568	508	413
1400-1459	691	4,702	824	1,563	327	4,864	3,219	631	625	184	619	2,096	357	519	3,583	2,873	1,368	237	7,505	621	872	427
1500-1559	798	4,267	1,084	1,424	321	5,812	2,902	614	588	165	600	2,023	490	456	4,902	2,918	1,119	175	7,535	445	454	437
1600-1659	722	4,162	819	1,414	229	5,203	3,337	567	636	123	454	2,067	476	497	4,083	2,848	1,380	215	7,885	604	903	417
1700-1759	905	4,455	1,351	1,565	379	6,400	3,148	506	635	176	468	1,861	409	582	4,927	3,552	1,607	386	8,429	556	672	526
1800-1859	452	4,742	1,358	1,798	284	3,697	2,288	718	705	136	381	2,110	483	403	3,763	2,800	1,452	275	8,189	452	750	445
1900-1959	700	3,236	1,037	1,601	276	5,573	2,594	646	468	99	325	1,627	498	453	3,461	3,130	1,062	204	8,047	503	702	415
2000-2059	398	3,562	752	1,331	178	3,107	1,637	481	440	93	303	1,541	617	355	3,238	1,803	1,192	148	6,610	241	372	357
2100-2159	301	1,849	681	1,125	176	2,836	1,203	408	370	91	207	762	451	89	1,921	1,607	332	92	4,447	202	224	320
2200-2259	140	1,827	283	665	26	2,429	467	208	312	3	147	392	214	283	1,326	1,280	995	60	1,690	58	309	119
2300-2359	9	551	405	639	41	758	59	136	87	5	47	2	191		201	835	136	104	76	21	14	2

- Most departure delays during the day for each airline are during early morning to late afternoon
- Southwest (WN) has the highest count of departure delay starting from the morning

Arrival Delays for Each Airline

Count of Airline Arrival Delays

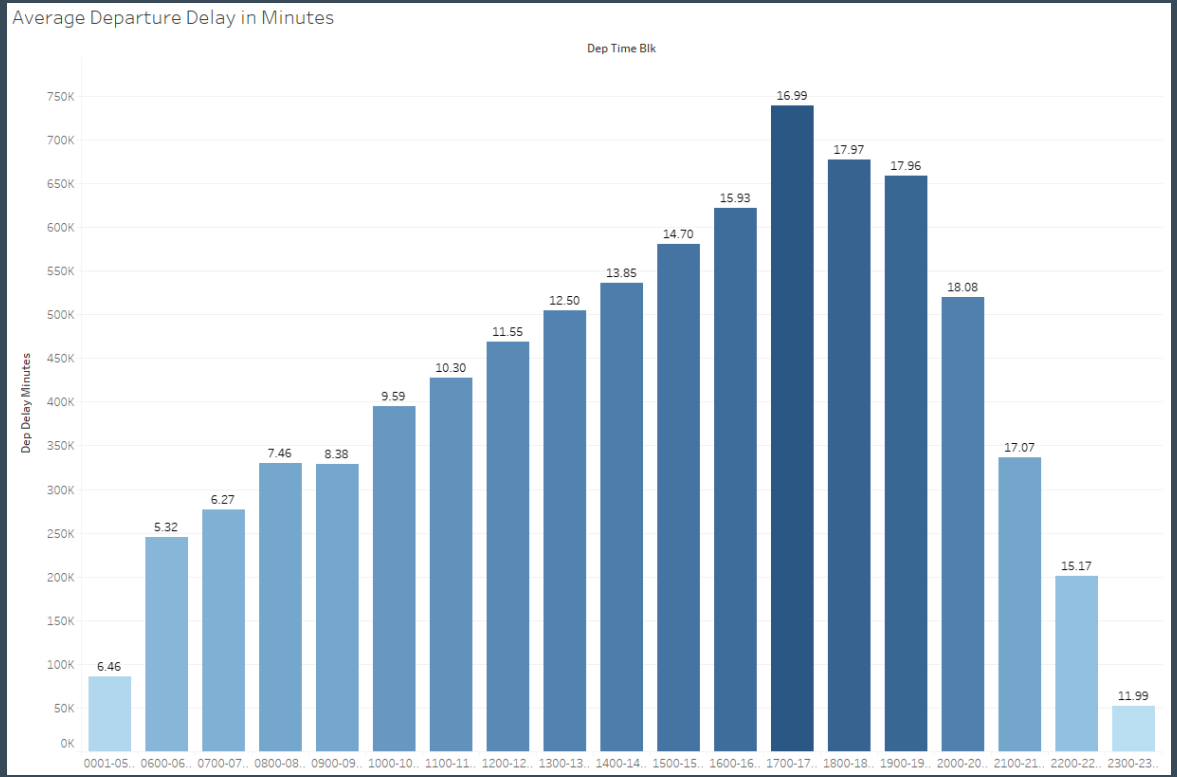
	Reporting Airline																				
Arr Time Blk	9E	AA	AS	B6	CO	DL	EV	F9	FL	G4	HA	MQ	NK	OH	OO	UA	US	VX	WN	XE	YV	YX
0001-0559	17	1,961	829	1,644	141	1,666	119	423	132		186	238	486	18	574	2,362	363	72	2,401	13	56	71
0600-0659	104	1,041	320	468	81	1,252	597	166	30	5	325	446	137	194	1,096	926	556	45	802	115	160	73
0700-0759	502	1,645	466	815	100	2,353	1,652	348	284	19	330	1,338	309	183	3,487	1,703	467	110	5,237	305	312	339
0800-0859	515	3,207	922	1,391	168	4,315	2,520	297	311	108	378	1,625	420	518	3,173	2,282	1,331	158	6,302	430	588	327
0900-0959	743	4,251	1,088	1,402	194	4,900	2,594	771	742	157	441	2,018	522	367	4,046	2,848	1,078	171	8,187	401	628	415
1000-1059	588	4,466	1,070	1,753	315	5,776	2,954	488	660	107	509	1,840	514	581	4,126	3,018	1,777	220	7,734	502	701	478
1100-1159	870	4,570	858	1,322	329	5,483	3,138	368	660	133	602	2,367	458	353	3,907	3,137	1,102	164	7,759	539	882	500
1200-1259	793	4,297	1,099	1,629	226	5,264	2,962	468	575	174	603	1,988	395	546	4,792	2,828	1,382	224	7,732	475	594	398
1300-1359	790	4,569	850	1,540	298	5,320	3,267	603	636	182	544	1,973	384	594	4,206	2,679	1,507	204	7,691	561	800	409
1400-1459	785	4,125	1,065	1,475	314	5,767	3,122	644	587	169	468	2,050	450	451	4,825	3,242	1,097	196	7,484	569	558	493
1500-1559	706	4,750	822	1,246	227	4,944	2,839	582	570	130	458	1,991	442	590	3,870	2,799	1,385	260	7,725	424	904	391
1600-1659	877	4,379	1,330	1,415	404	6,418	3,657	535	598	148	484	1,896	371	420	5,083	3,658	1,596	344	8,147	626	588	460
1700-1759	602	4,786	1,107	1,498	321	4,612	2,747	532	663	148	426	2,460	471	577	4,201	2,961	1,218	229	8,197	538	776	440
1800-1859	871	4,242	1,055	1,512	312	6,019	3,421	773	629	127	407	1,851	444	439	4,129	3,527	1,357	243	8,405	584	741	410
1900-1959	510	4,903	937	1,347	414	5,391	2,324	550	574	153	425	1,975	536	586	4,377	2,805	1,715	255	8,164	575	730	457
2000-2059	692	4,531	1,294	1,705	302	5,503	2,827	608	724	169	411	1,635	414	315	3,749	3,926	1,467	269	7,641	427	545	494
2100-2159	627	4,857	1,252	1,741	214	5,646	2,057	449	576	116	485	1,786	444	595	3,688	3,358	1,729	306	7,585	314	700	414
2200-2259	392	3,477	1,302	1,576	325	4,120	1,779	744	630	136	317	1,104	528	190	2,597	2,342	833	363	6,407	368	282	364
2300-2359	315	3,250	970	1,269	220	3,810	1,013	557	432	71	244	786	726	291	2,315	2,405	975	121	5,062	190	480	399

- Southwest (WN) also has the highest count of arrival delay starting from late morning to late evening
- The effect of departure delays causes arrival delays

Average Departure Delay in Minutes

Average Departure Delay

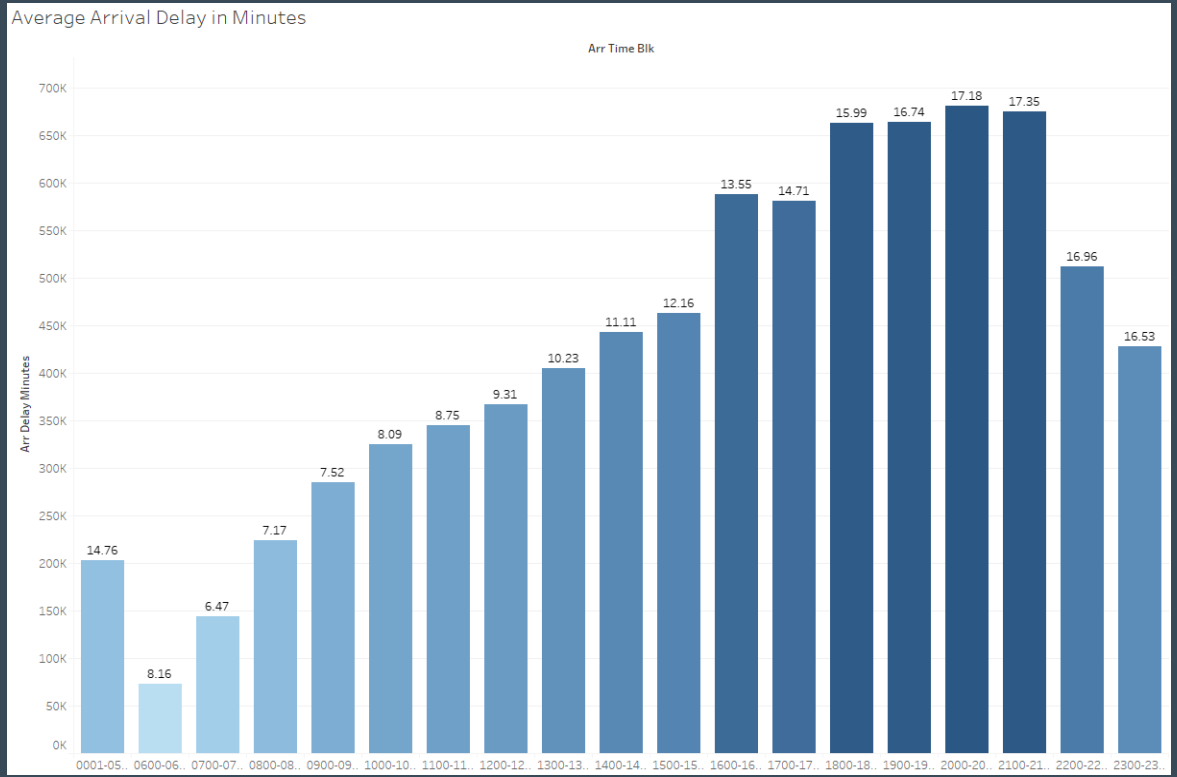
- Least amount of delay time is from 1:00am – 5:59am with an average of 6.46 minutes
- Highest amount of delay time is from 5:00pm – 5:59pm with an average of 16.99 minutes



Average Arrival Delay in Minutes

Average Arrival Delay

- Least amount of delay time is from 6:00am – 6:59am with an average of 8.16 minutes
- Highest amount of delay time is from 9:00pm – 9:59pm with an average of 17.35 minutes

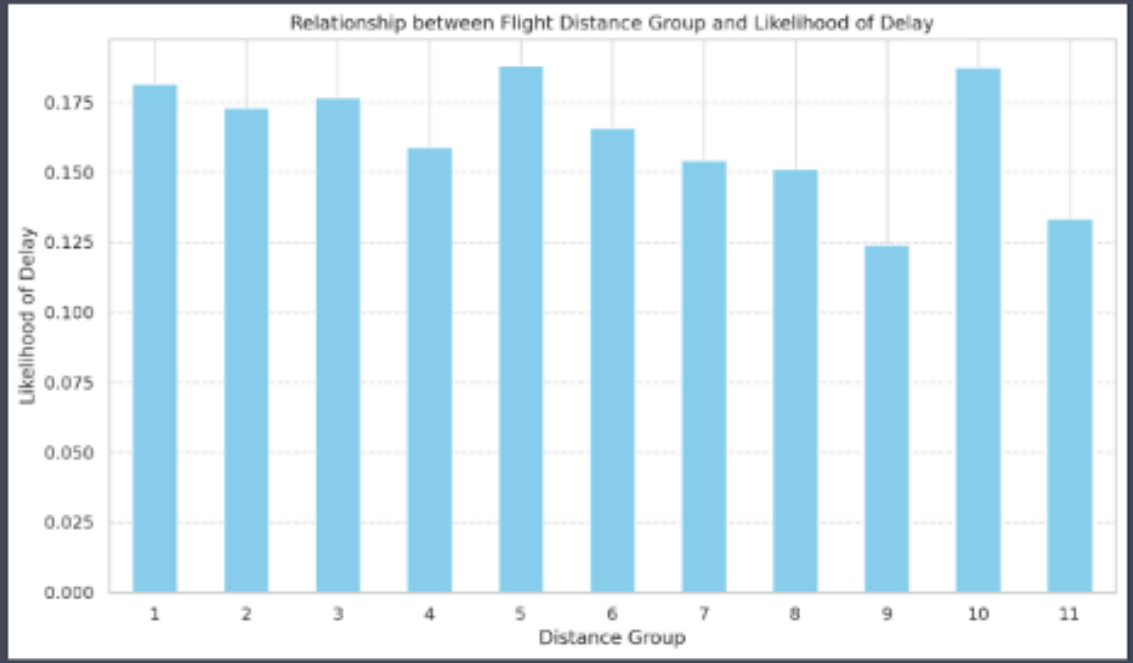


Question 3

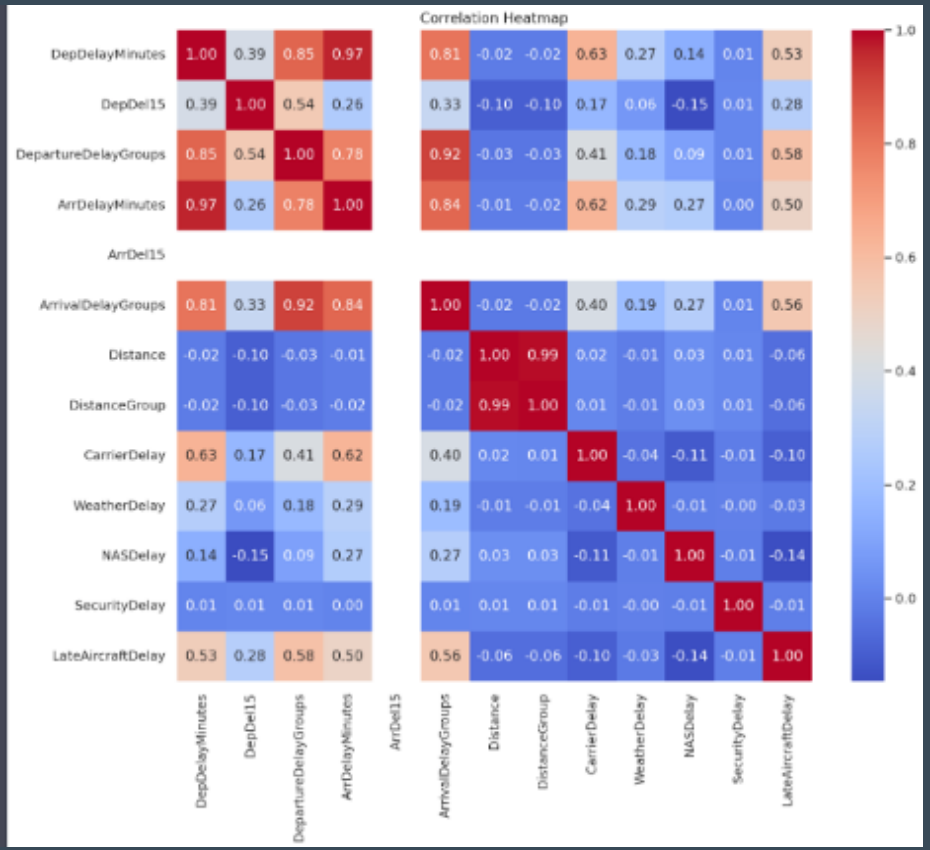
What is the relationship between flight distance and the likelihood of delay?

Are there specific routes or flight numbers that are consistently delayed?

Here's the bar chart visualizing the relationship between flight distance groups and the likelihood of delay:

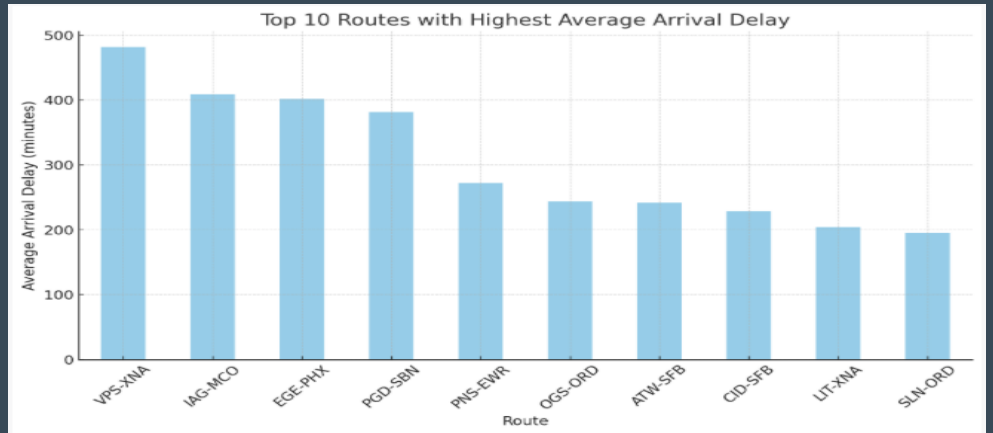


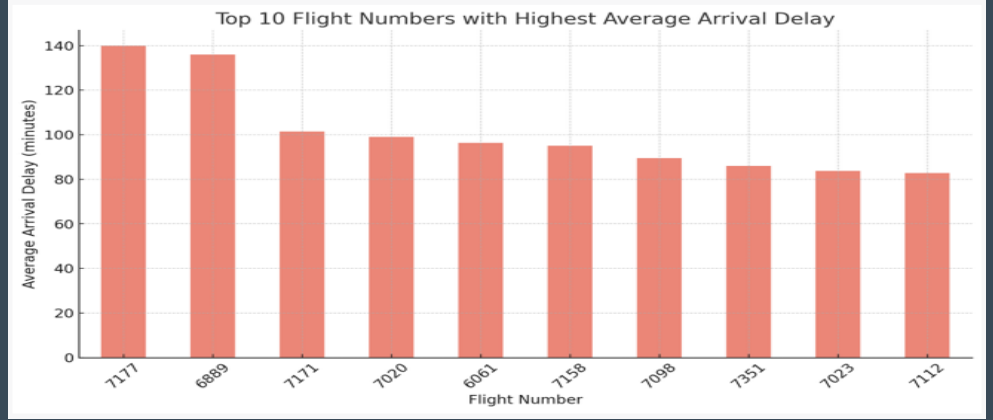
Heat Map



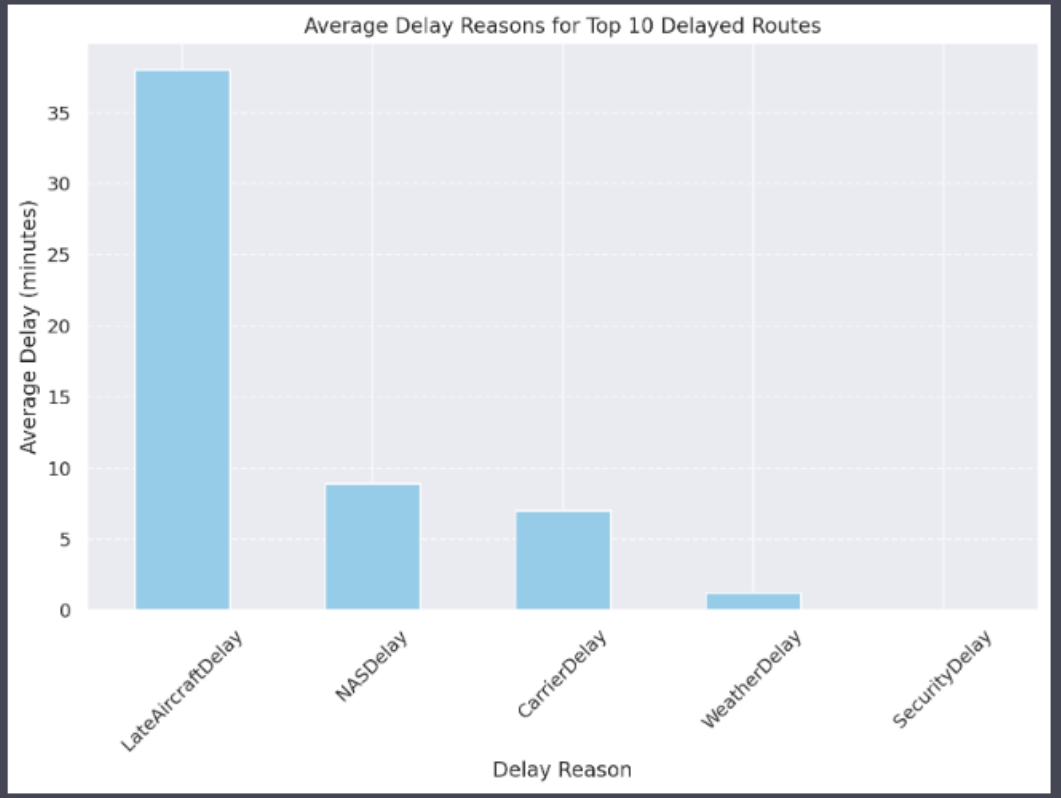
The heatmap displays the correlation between different variables. A correlation of 1 indicates a perfect positive relationship, while a correlation of -1 indicates a perfect negative relationship. A correlation close to 0 suggests no linear relationship between the variables.

The bar charts above show the top 10 routes and flight numbers with the highest average arrival delay in minutes





VPS: Destin-Fort Walton Beach Airport
XNA: Northwest Arkansas Regional Airport
IAG: Niagara Falls International Airport
MCO: Orlando International Airport
EGE: Eagle County Regional Airport
PHX: Phoenix Sky Harbor International Airport
PGD: Punta Gorda Airport
SBN: South Bend International Airport
OGS: Ogdensburg International Airport
ORD: O'Hare International Airport
ATW: Appleton International Airport
SFB: Orlando Sanford International Airport
PNS: Pensacola International Airport
EWR: Newark Liberty International Airport



The bar plot above shows the average delay (in minutes) for each delay reason (CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay) on the top 10 most delayed routes.

Question 4

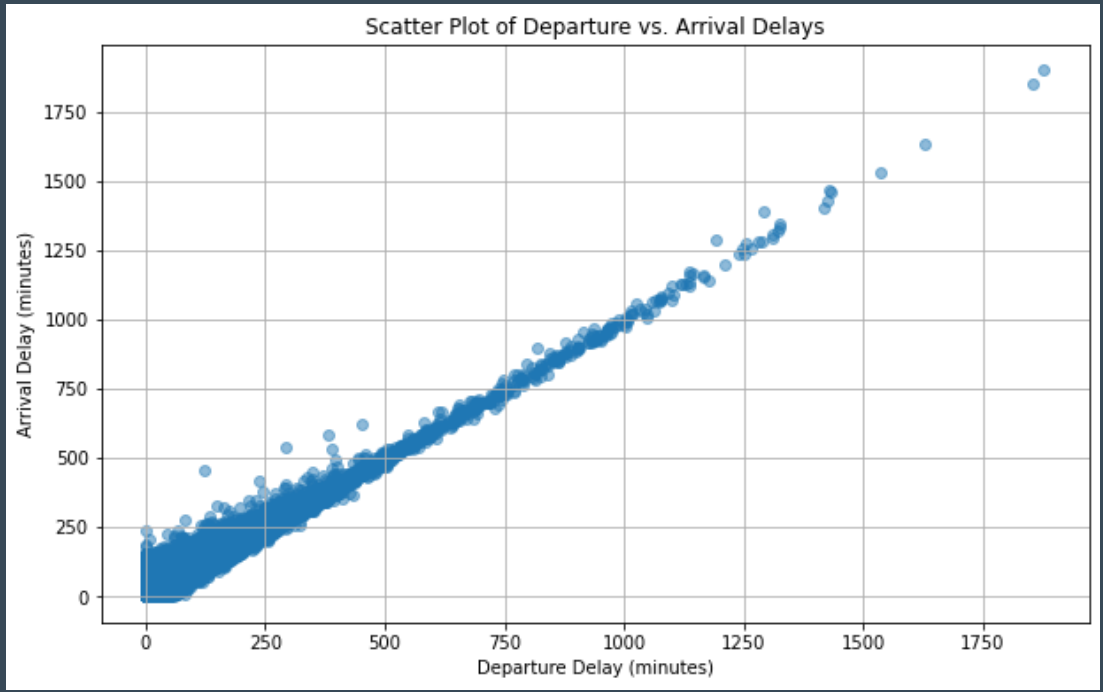
Is there a correlation between the departure delay and arrival delay for flights?

Which airline(s) consistently have the best and worst on-time performance?

Correlation between Departure and Arrival Delays

Direct Relationship:

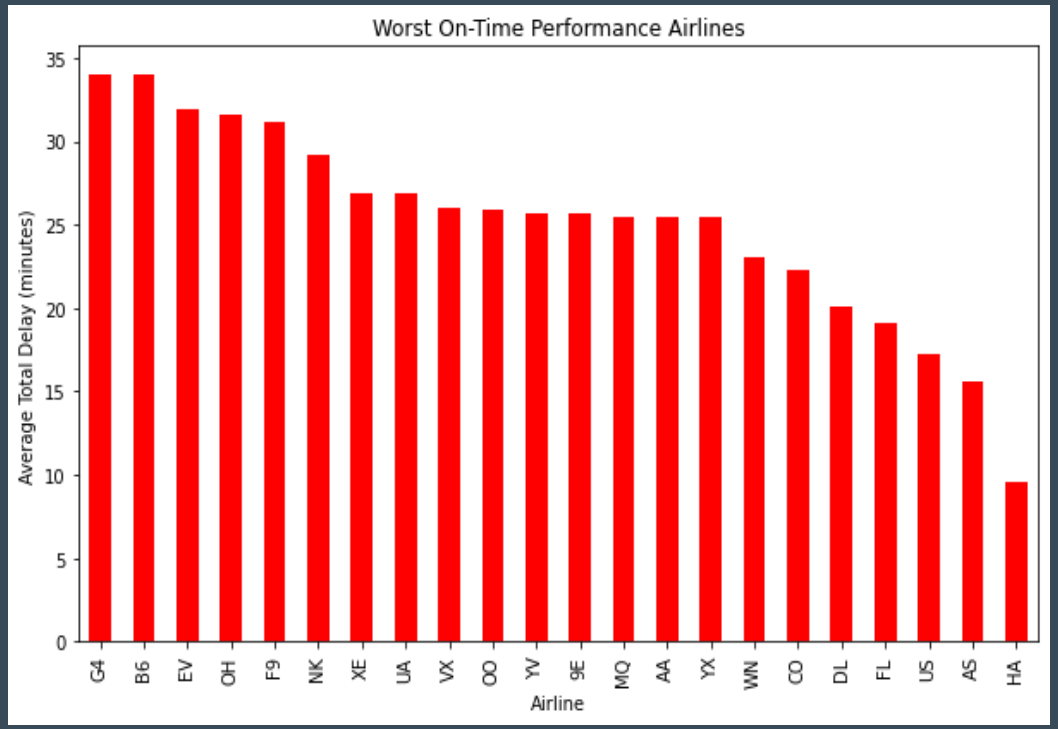
A clear positive trend indicates that flights that depart late tend to also arrive late.



Worst Performing Flights

Worst Performing Flights:

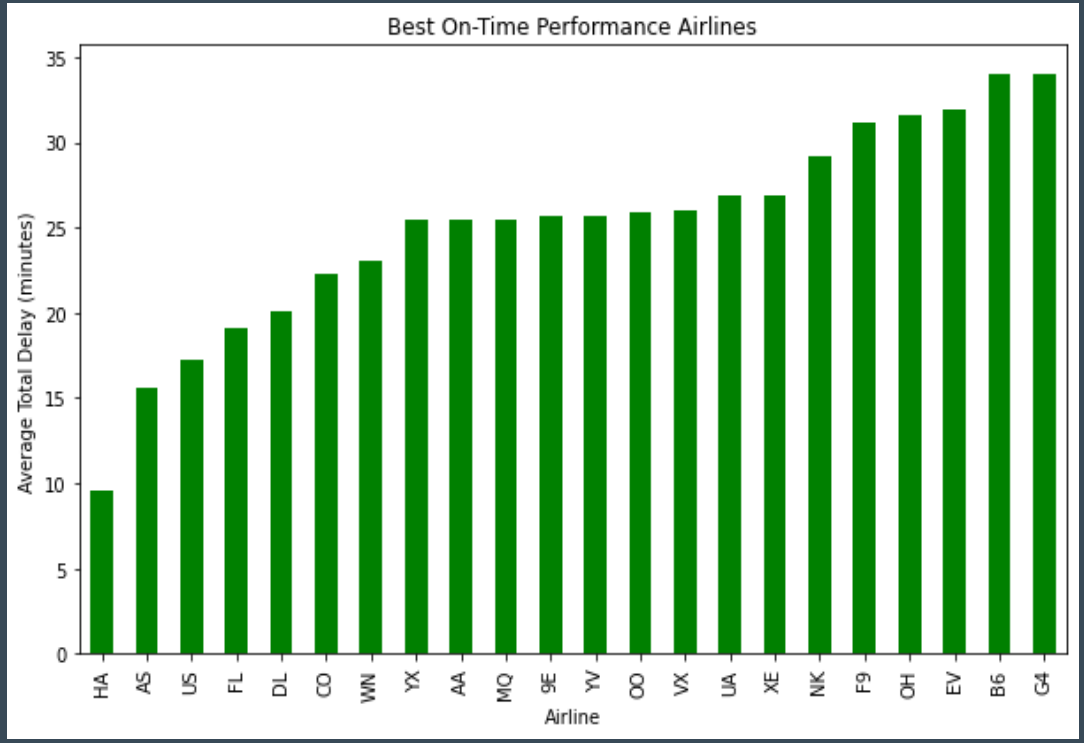
1. Allegiant Air LLC
2. JetBlue Airways
3. ExpressJet Airlines
4. PSA Airlines
5. Frontier Airlines



Best Performing Flights

Best Performing Flights:

1. Hawaiian Airlines, Inc.
2. Alaska Airlines
3. US Airways
4. AirTran Airways
5. Delta Air Lines



Question 5

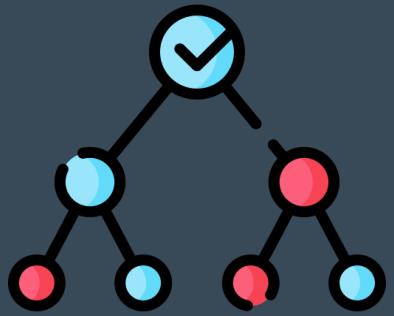
Can machine learning models accurately predict flight delays based on historical data from this dataset?

Data Preparation

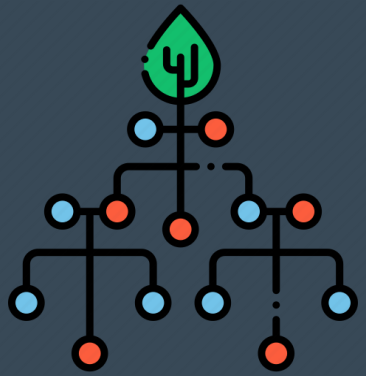
- Only use 5 years data due to computing power limitation (2015 – 2020)
- Data frame size (with dummy variables): 339132 observations, 2780 variables
- Split data into 70% training, 30% testing
- Removed any departure/arrival delay information (Delay minutes, Reason for delays,...)
- Removed any cancelations information
- Removed any high multicollinearity variables (|correlation| > 0.7)

Methods

Regression



Decision Tree



Random Forest

Regression Analysis

Accuracy		val Delay
MAE		797.694628 mins
MSE		9338376612e+22

Decision Tree

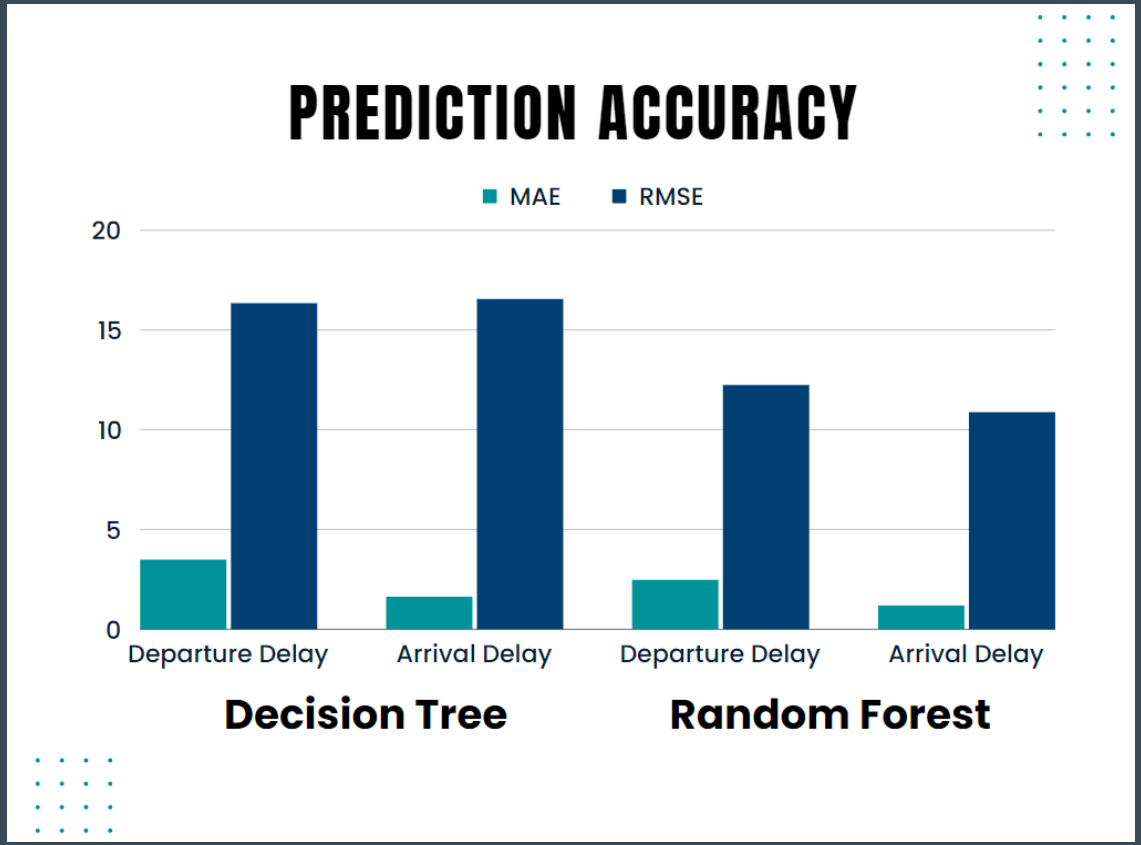
Accuracy	Departure Delay	Arrival Delay
MAE	3.491 minutes	1.6313 minutes
MSE	267.0044	273.9907
RMSE	16.34	16.55

Random Forest

(100 Random Trees)

Accuracy	Departure Delay	Arrival Delay
MAE	2.4776 minutes	1.1896 minute
MSE	149.8350	118.4285
RMSE	12.24	10.88

Conclusion


PREDICTION ACCURACY
MAE
RMSE
20
15
10
5
0
Departure Delay
Arrival Delay
Departure Delay
Arrival Delay
Decision Tree
Random Forest

RECOMMENDATIONS

RECOMMEND 1

Flights and routes

RECOMMEND 2

Days and timings

RECOMMEND 3

Increasing Customer satisfaction

Recommendation

Best airlines to book: Hawaiian, Alaska

Do not book: Allegiant, Jet Blue

Avoid route: Destin Fort Walton Beach Airport - Northwest Arkansas Regional Airport

Recommendation

Best day to book: Saturday

Days to avoid: Monday, Thursday, and Friday

Best Time: Early mornings or late nights

Recommendation

Incorporate flight delay prediction for better booking experience

Incorporate flight delay prediction for better scheduling for the airport

THANK YOU

ANY QUESTIONS OR COMMENTS?