INFO 7250 Final Project Flight On-time Performance Analysis

YUNAN SHAO

Introduction

- Air transport is now very common for travelling. Although, it is the fastest method of transport, passengers waste both time and money when their flights are delayed.

- This project use the flight dataset to analyze the flight on-time performance and try to implement a machine learning model to predict if a flight will be delayed.

- Possible Questions:
  - During which time period does flight delay the most?
  - Relations between Airports' air traffic and flight delay.
  - Airliner on-time performance.

Objective & Techniques

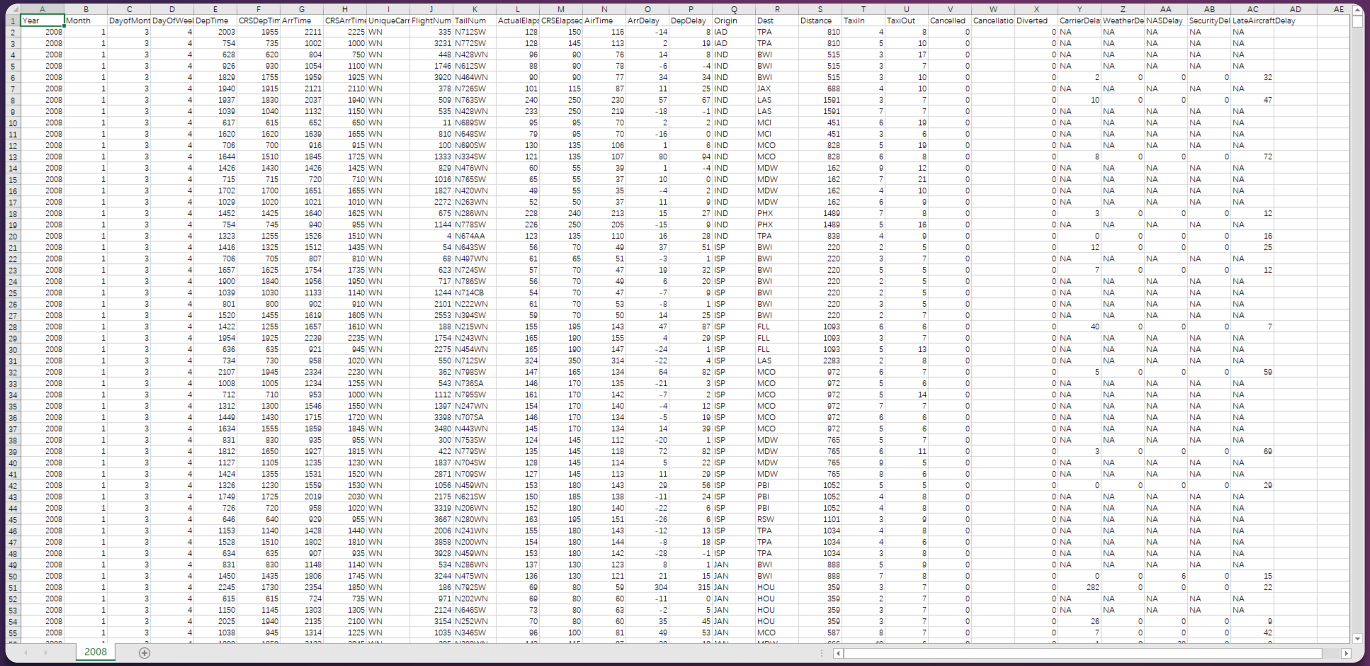
1. Analyze the flight on-time performance from different aspects (time, airport and carriers) and provide visualized results

- MultipleOutput Binning
- MapReduce, Job Chaining
- MultipleInput Reducer side Join
- TopK Pattern
- Combiner, Memory-Conscious Implementation
- Custom Writable for processing selected attributes

2. Try to implement machine learning model for prediction flight delay

- Apache Mahout
- AWS Machine Learning

Dataset Flight Data From 1987 to 2008



Year	Month	DayofMont	DayOfWeel	DepTime	CRSDepTim	ArrTime	CRSArrTime	UniqueCarr	FlightNum	TailNum	ActualElaps	CRSElapsec	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	Cancellatio	Diverted	CarrierDela	WeatherDe	NASDelay	SecurityDel	LateAircraftDelay
2008	1	3	4	2003	1955	2211	2225	WN	335	N712SW	128	150	116	-14	8	IAD	TPA	810	4	8	0		0	NA	NA	NA	NA	NA
2008	1	3	4	754	735	1002	1000	WN	3231	N772SW	128	145	113	2	19	IAD	TPA	810	5	10	0		0	NA	NA	NA	NA	NA
2008	1	3	4	628	620	804	750	WN	448	N428WN	96	90	76	14	8	IND	BWI	515	3	17	0		0	NA	NA	NA	NA	NA
2008	1	3	4	926	930	1054	1100	WN	1746	N612SW	88	90	78	-6	-4	IND	BWI	515	3	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1829	1755	1959	1925	WN	3920	N464WN	90	90	77	34	34	IND	BWI	515	3	10	0		0	2	0	0	0	32
2008	1	3	4	1940	1915	2121	2110	WN	378	N726SW	101	115	87	11	25	IND	JAX	688	4	10	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1937	1830	2037	1940	WN	509	N763SW	240	250	230	57	67	IND	LAS	1591	3	7	0		0	10	0	0	0	47
2008	1	3	4	1039	1040	1132	1150	WN	535	N428WN	233	250	219	-18	-1	IND	LAS	1591	7	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	617	615	652	650	WN	11	N689SW	95	95	70	2	2	IND	MCI	451	6	19	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1620	1620	1639	1655	WN	810	N648SW	79	95	70	-16	0	IND	MCI	451	3	6	0		0	NA	NA	NA	NA	NA
2008	1	3	4	706	700	916	915	WN	100	N690SW	130	135	106	1	6	IND	MCO	828	5	19	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1644	1510	1845	1725	WN	1333	N334SW	121	135	107	80	94	IND	MCO	828	6	8	0		0	8	0	0	0	72
2008	1	3	4	1426	1430	1426	1425	WN	829	N476WN	60	55	39	1	-4	IND	MDW	162	9	12	0		0	NA	NA	NA	NA	NA
2008	1	3	4	715	715	720	710	WN	1016	N765SW	65	55	37	10	0	IND	MDW	162	7	21	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1702	1700	1651	1655	WN	1827	N420WN	49	55	35	-4	2	IND	MDW	162	4	10	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1029	1020	1021	1010	WN	2272	N263WN	52	50	37	11	9	IND	MDW	162	6	9	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1452	1425	1640	1625	WN	675	N286WN	228	240	213	15	27	IND	PHX	1489	7	8	0		0	3	0	0	0	12
2008	1	3	4	754	745	940	955	WN	1144	N778SW	226	250	205	-15	9	IND	PHX	1489	5	16	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1323	1255	1526	1510	WN	4	N674AA	123	135	110	16	28	IND	TPA	838	4	9	0		0	0	0	0	0	16
2008	1	3	4	1416	1325	1512	1435	WN	54	N643SW	56	70	49	37	51	ISP	BWI	220	2	5	0		0	12	0	0	0	25
2008	1	3	4	706	705	807	810	WN	68	N497WN	61	65	51	-3	1	ISP	BWI	220	3	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1657	1625	1754	1735	WN	623	N724SW	57	70	47	19	32	ISP	BWI	220	5	5	0		0	7	0	0	0	12
2008	1	3	4	1900	1840	1956	1950	WN	717	N786SW	56	70	49	6	20	ISP	BWI	220	2	5	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1039	1030	1133	1140	WN	1244	N714CB	54	70	47	-7	9	ISP	BWI	220	2	5	0		0	NA	NA	NA	NA	NA
2008	1	3	4	801	800	902	910	WN	2101	N222WN	61	70	53	-8	1	ISP	BWI	220	3	5	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1520	1455	1619	1605	WN	2553	N394SW	59	70	50	14	25	ISP	BWI	220	2	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1422	1255	1657	1610	WN	188	N215WN	155	195	143	47	87	ISP	FLL	1093	6	6	0		0	40	0	0	0	7
2008	1	3	4	1954	1925	2239	2235	WN	1754	N243WN	165	190	155	4	29	ISP	FLL	1093	3	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	636	635	921	945	WN	2275	N454WN	165	190	147	-24	1	ISP	FLL	1093	5	13	0		0	NA	NA	NA	NA	NA
2008	1	3	4	734	730	958	1020	WN	550	N712SW	324	350	314	-22	4	ISP	LAS	2283	2	8	0		0	NA	NA	NA	NA	NA
2008	1	3	4	2107	1945	2334	2230	WN	362	N798SW	147	165	134	64	82	ISP	MCO	972	6	7	0		0	5	0	0	0	59
2008	1	3	4	1008	1005	1234	1255	WN	543	N736SA	146	170	135	-21	3	ISP	MCO	972	5	6	0		0	NA	NA	NA	NA	NA
2008	1	3	4	712	710	953	1000	WN	1112	N795SW	161	170	142	-7	2	ISP	MCO	972	5	14	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1312	1300	1546	1550	WN	1397	N247WN	154	170	140	-4	12	ISP	MCO	972	7	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1449	1430	1715	1720	WN	3398	N707SA	146	170	134	-5	19	ISP	MCO	972	6	6	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1634	1555	1859	1845	WN	3480	N443WN	145	170	134	14	39	ISP	MCO	972	5	6	0		0	NA	NA	NA	NA	NA
2008	1	3	4	831	830	935	955	WN	300	N753SW	124	145	112	-20	1	ISP	MDW	765	5	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1812	1650	1927	1815	WN	422	N779SW	135	145	118	72	82	ISP	MDW	765	6	11	0		0	3	0	0	0	69
2008	1	3	4	1127	1105	1235	1230	WN	1837	N704SW	128	145	114	5	22	ISP	MDW	765	9	5	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1424	1355	1531	1520	WN	2871	N709SW	127	145	113	11	29	ISP	MDW	765	8	6	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1326	1230	1559	1530	WN	1056	N459WN	153	180	143	29	56	ISP	PBI	1052	5	5	0		0	0	0	0	0	29
2008	1	3	4	1749	1725	2019	2030	WN	2175	N621SW	150	185	138	-11	24	ISP	PBI	1052	4	8	0		0	NA	NA	NA	NA	NA
2008	1	3	4	726	720	958	1020	WN	3319	N206WN	152	180	140	-22	6	ISP	PBI	1052	4	8	0		0	NA	NA	NA	NA	NA
2008	1	3	4	646	640	929	955	WN	3667	N280WN	163	195	151	-26	6	ISP	RSW	1101	3	9	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1153	1140	1428	1440	WN	2006	N241WN	155	180	143	-12	13	ISP	TPA	1034	4	8	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1528	1510	1802	1810	WN	3858	N200WN	154	180	144	-8	18	ISP	TPA	1034	4	6	0		0	NA	NA	NA	NA	NA
2008	1	3	4	634	635	907	935	WN	3928	N459WN	153	180	142	-28	-1	ISP	TPA	1034	3	8	0		0	NA	NA	NA	NA	NA
2008	1	3	4	831	830	1148	1140	WN	534	N286WN	137	130	123	8	1	JAN	BWI	888	5	9	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1450	1435	1806	1745	WN	3244	N475WN	136	130	121	21	15	JAN	BWI	888	7	8	0		0	0	0	6	0	15
2008	1	3	4	2245	1730	2354	1850	WN	186	N792SW	69	80	59	304	315	JAN	HOU	359	3	7	0		0	282	0	0	0	22
2008	1	3	4	615	615	724	735	WN	971	N202WN	69	80	60	-11	0	JAN	HOU	359	2	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	1150	1145	1303	1305	WN	2124	N646SW	73	80	63	-2	5	JAN	HOU	359	3	7	0		0	NA	NA	NA	NA	NA
2008	1	3	4	2025	1940	2135	2100	WN	3154	N252WN	70	80	60	35	45	JAN	HOU	359	3	7	0		0	26	0	0	0	9
2008	1	3	4	1038	945	1314	1225	WN	1035	N346SW	96	100	81	49	53	JAN	MCO	587	8	7	0		0	7	0	0	0	42

http://stat-computing.org/dataexpo/2009/the-data.html

Field Attributes

Variable descriptions

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Most of analyses are using files from 2003-2008 because these fields are ‘NA’ before 2003

Type of Delay

- **Carrier Delay**

Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.

- **Late Arrival Delay**

Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

- **NAS Delay**

Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc. Delays that occur after Actual Gate Out are usually attributed to the NAS and are also reported through OPSNET.

- **Security Delay**

Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

- **Weather Delay**

Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.

Supplemental Data

- Airport and Carrier tables for JOIN in MapReduce Jobs
- * The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time.

Airports

airports.csv describes the locations of US airports, with the fields:

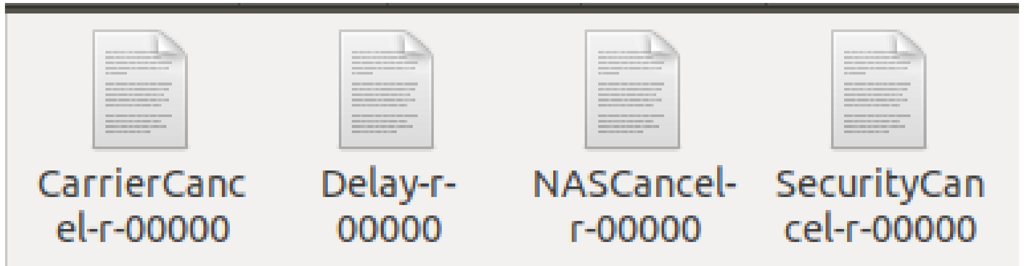
- `iata`: the international airport abbreviation code
- name of the `airport`
- `city` and `country` in which airport is located.
- `lat` and `long`: the latitude and longitude of the airport

This majority of this data comes from the FAA, but a few extra airports (mainly military bases and US protectorates) were collected from other web sources by Ryan Hafen and Hadley Wickham.

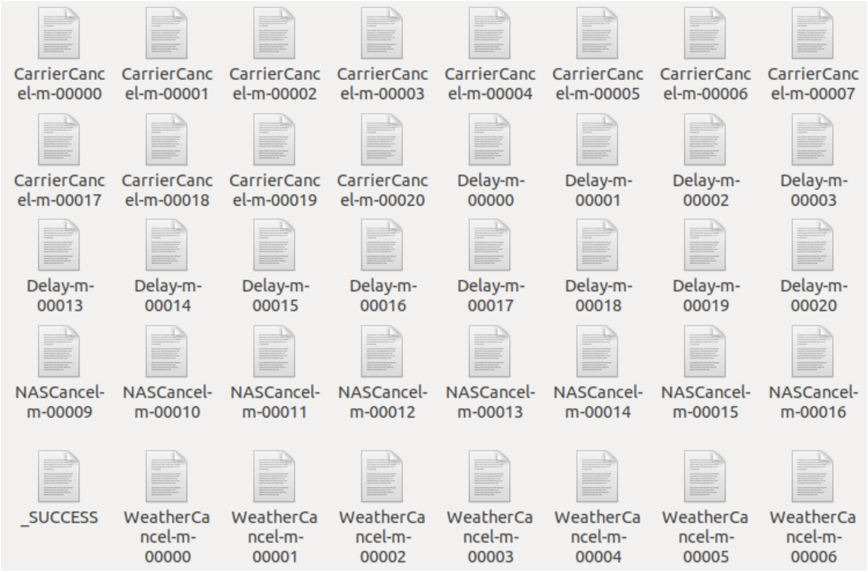
Carrier codes

Listing of carrier codes with full names: carriers.csv

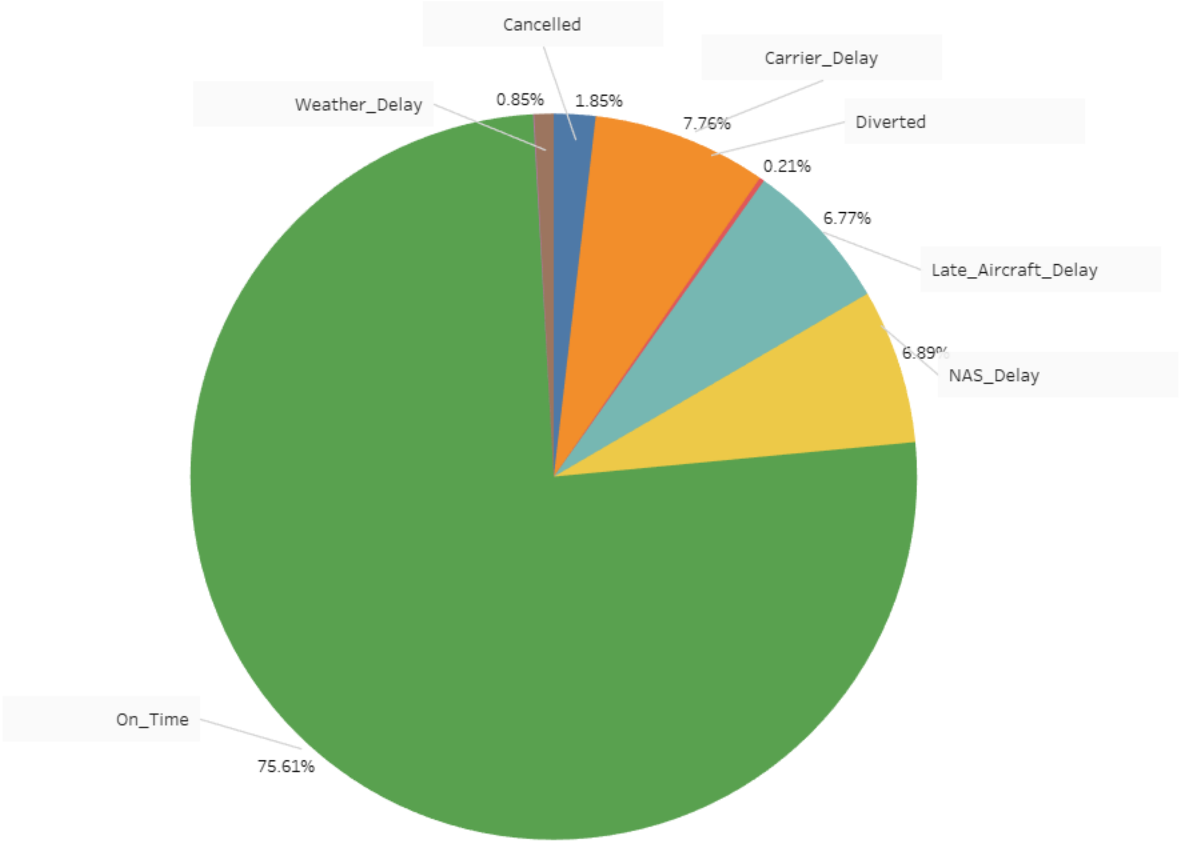
Inspection on the Dataset



- Because the dataset is too large, even one csv file takes too long to load. In order to inspect the data, I use MultipleOutput for binning. (mapper and reducer side)



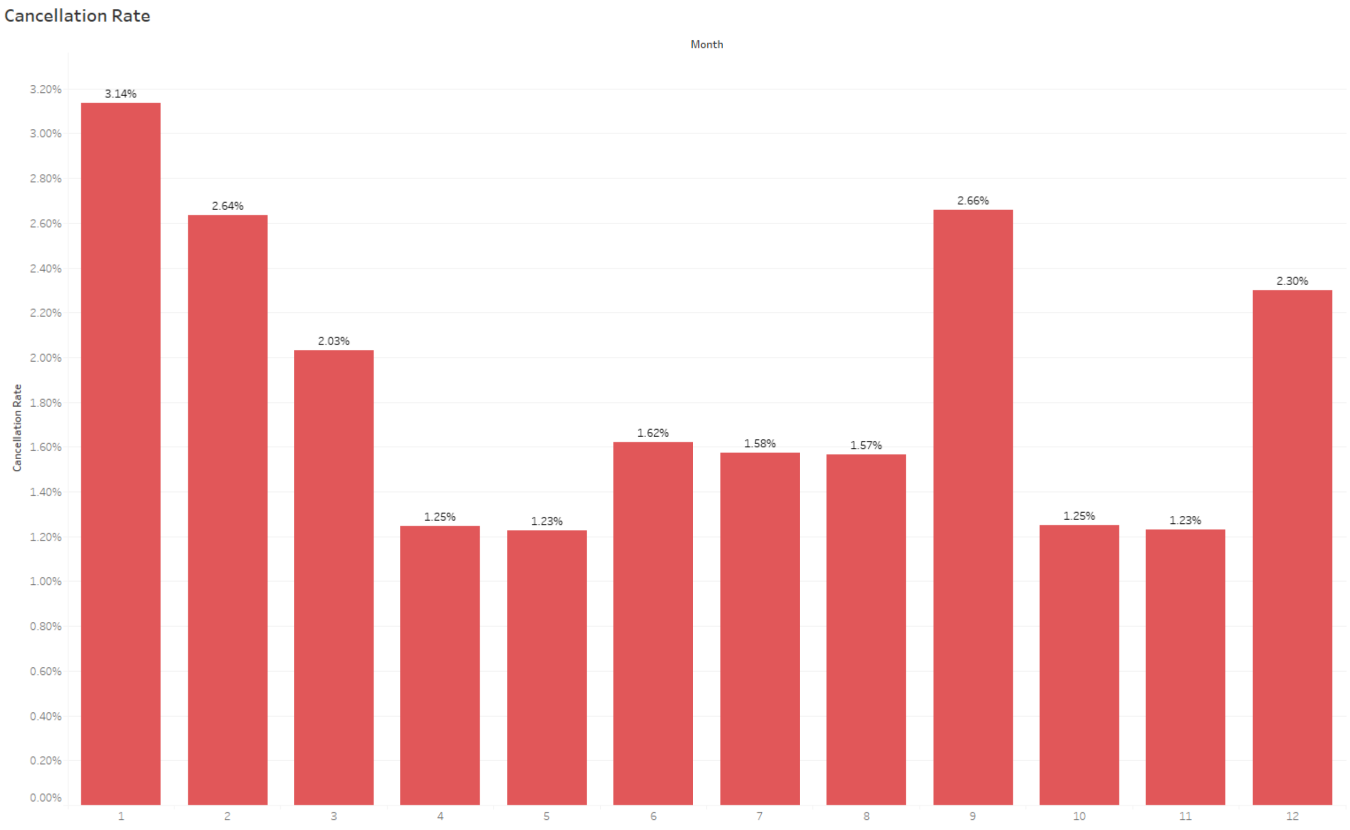
Causes of Delay 2003-2008



Main Causes

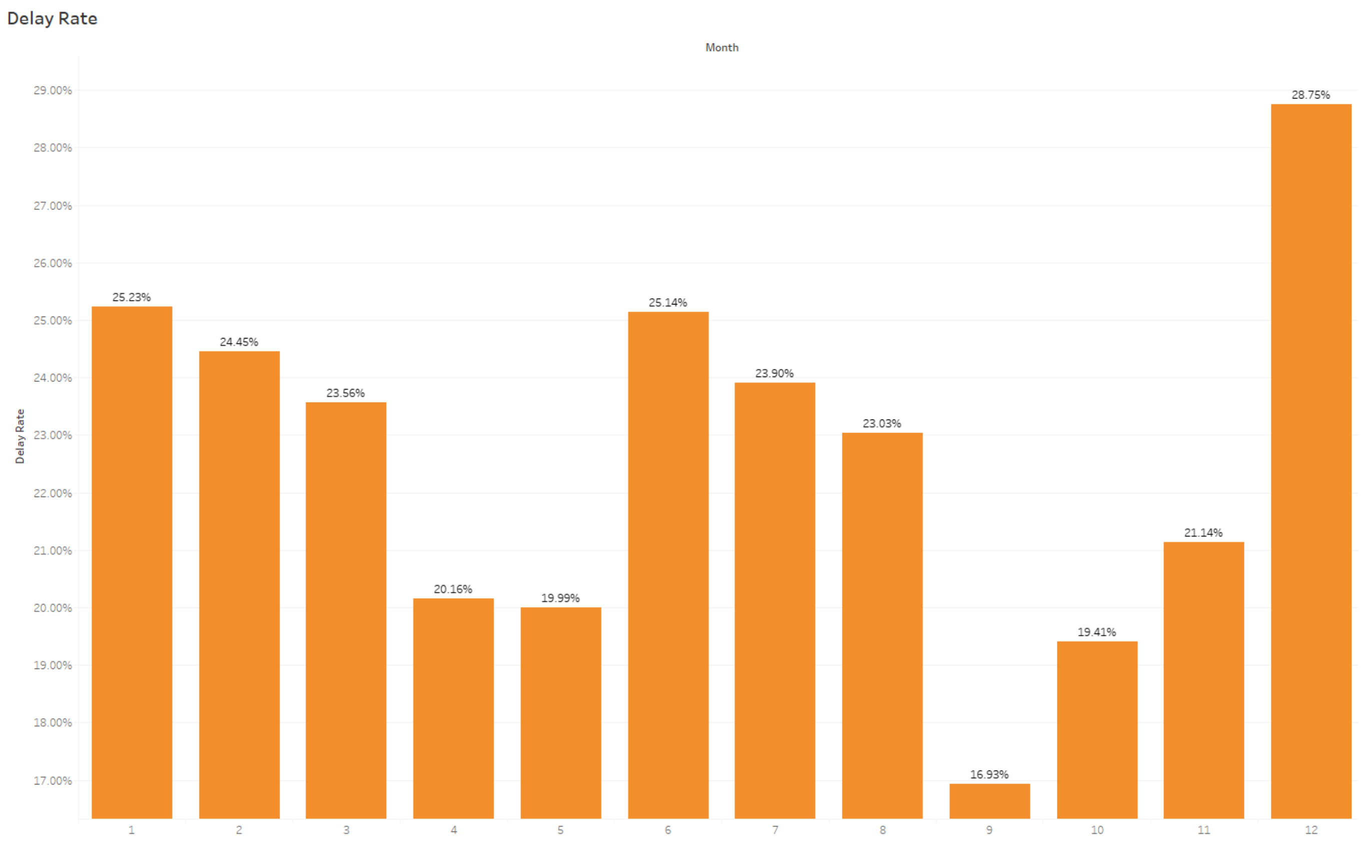
1. Carrier Delay
2. NAS Delay (National Airspace System)
3. Late Aircraft Delay

*Detailed results for each year can be accessed in Tableau file.

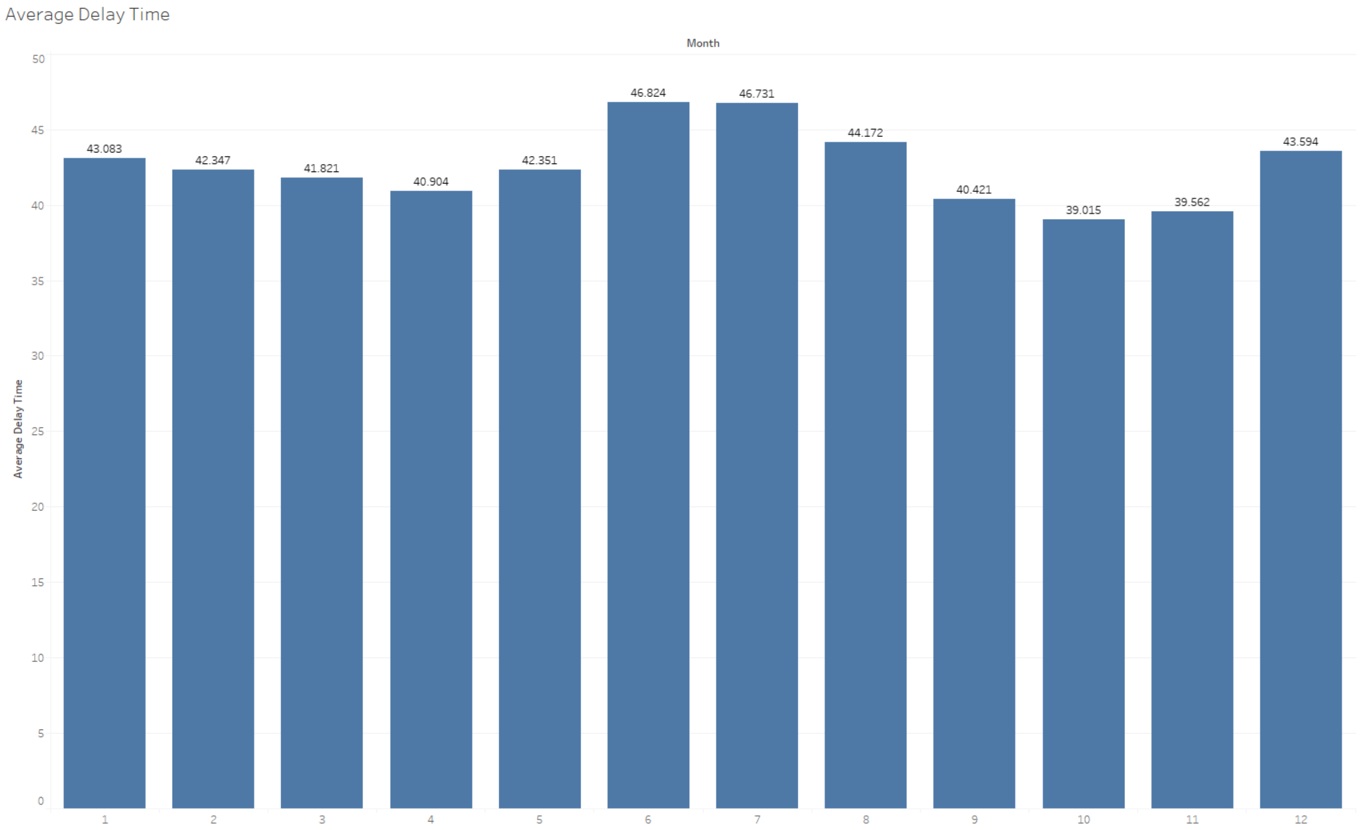


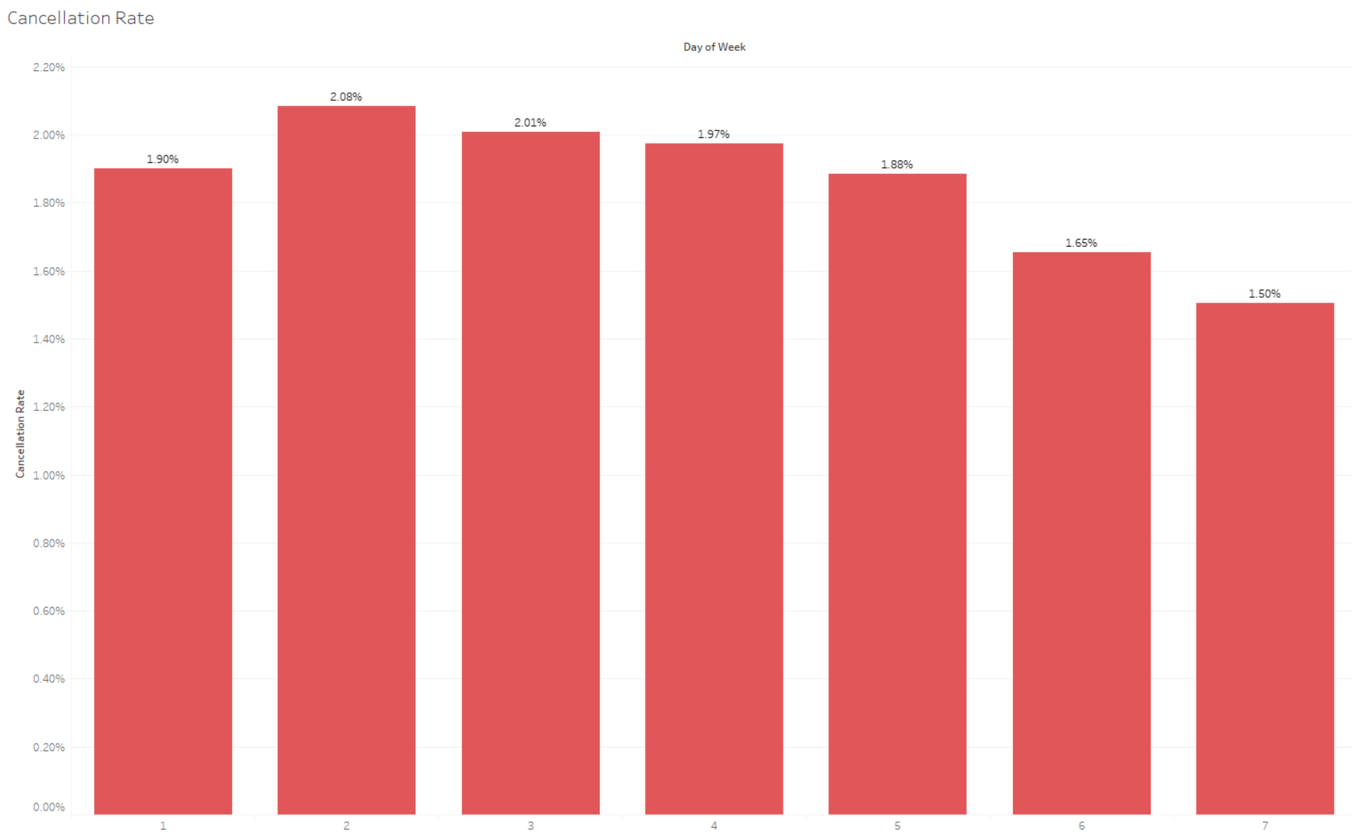
Cancellation Rate By Month

Delay Rate By Month



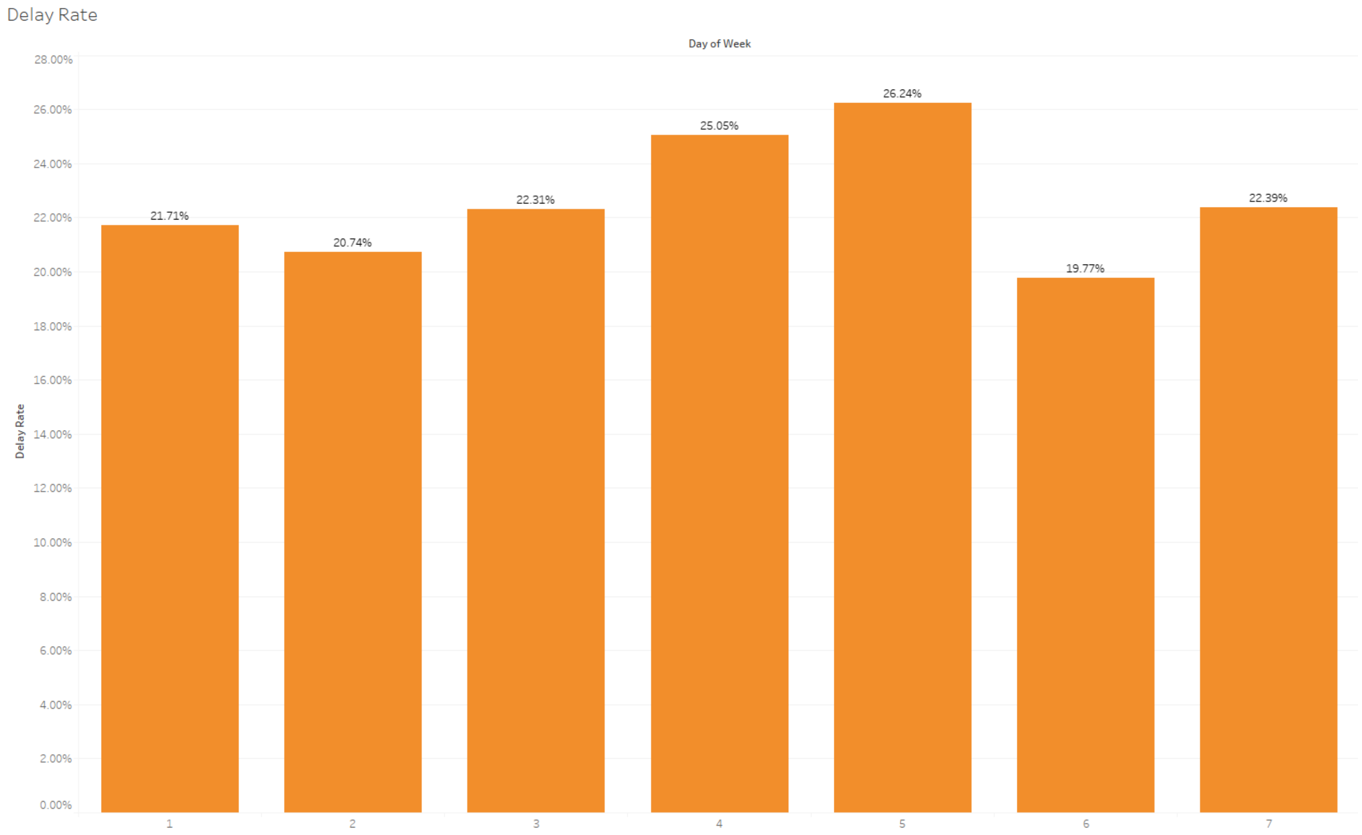
Average Delay Time By Month

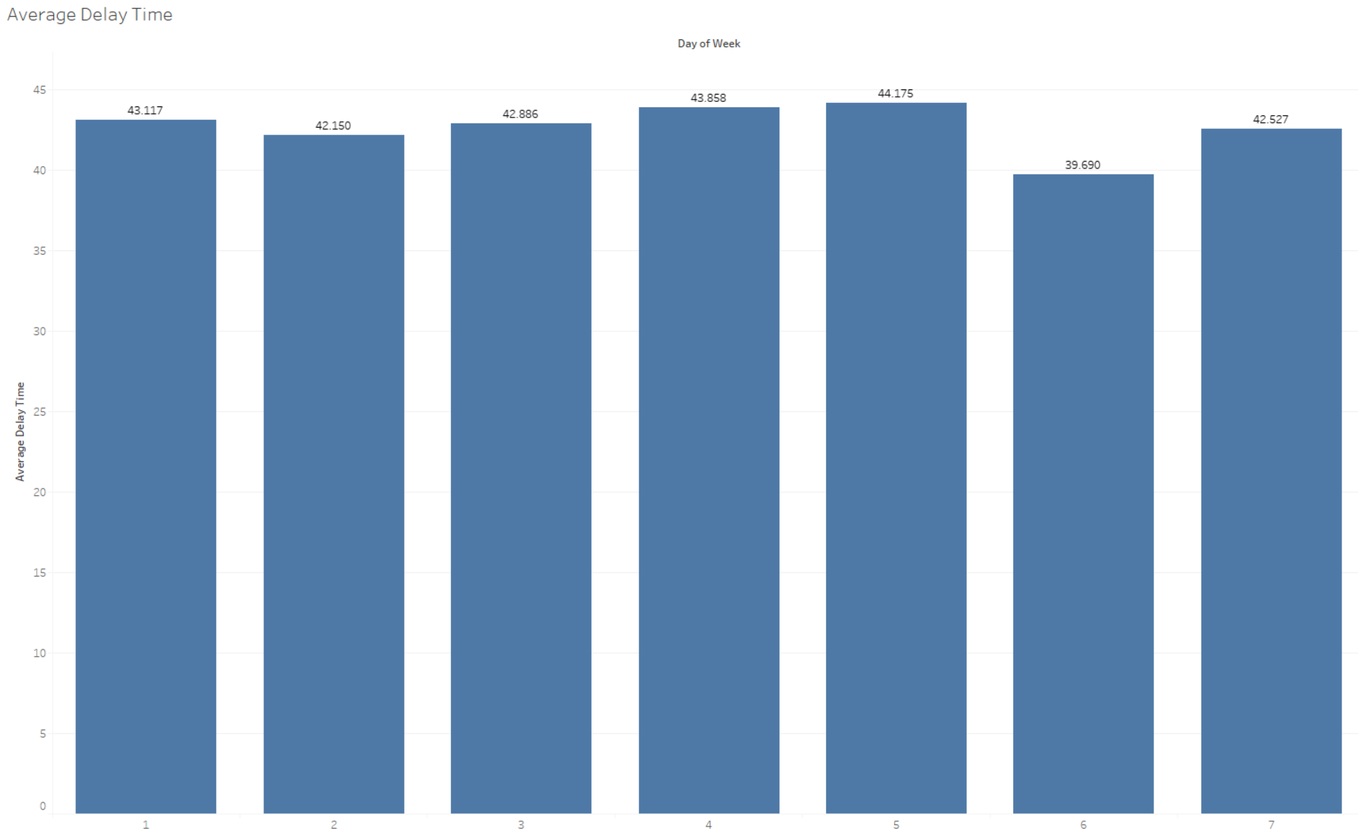




Cancellation Rate By Day of Week

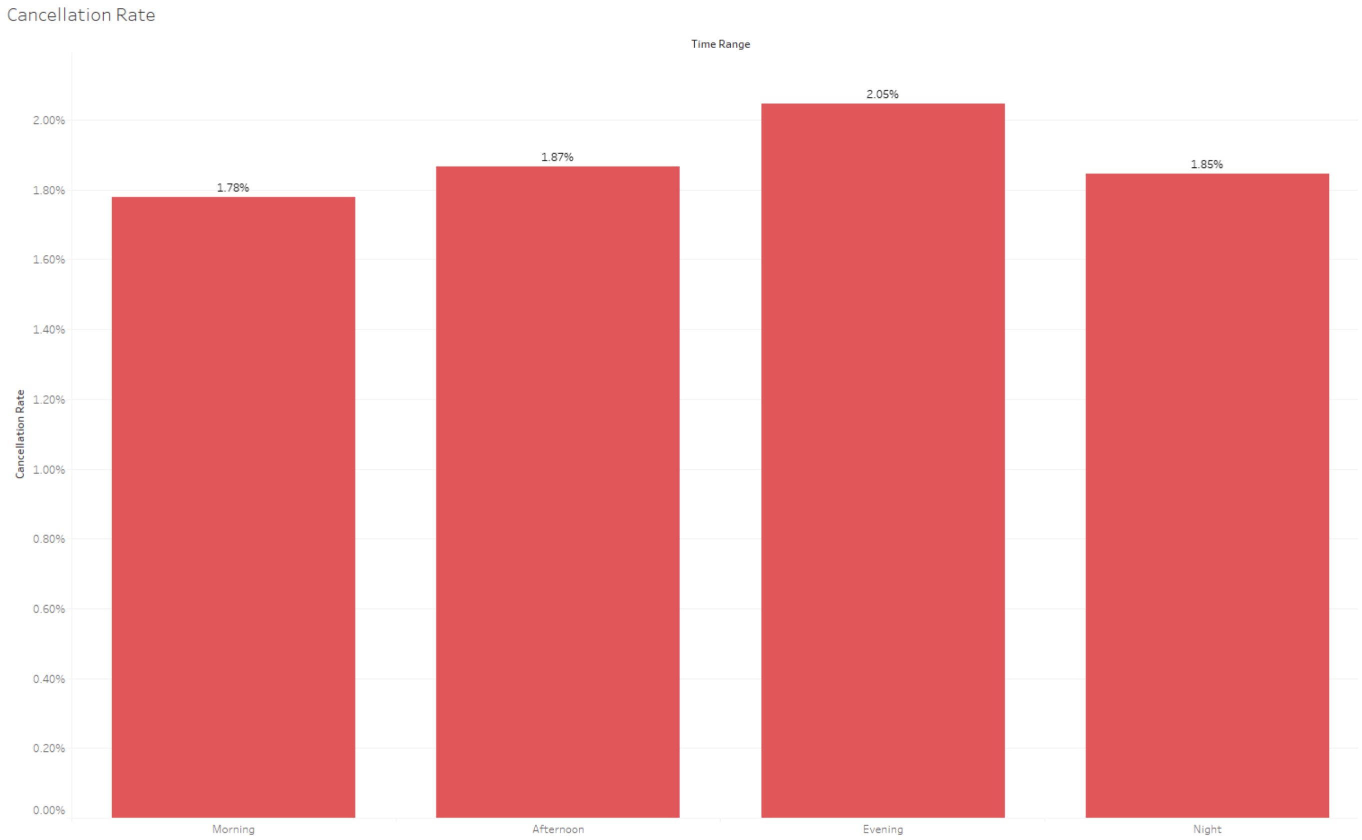
Delay Rate By Day of Week



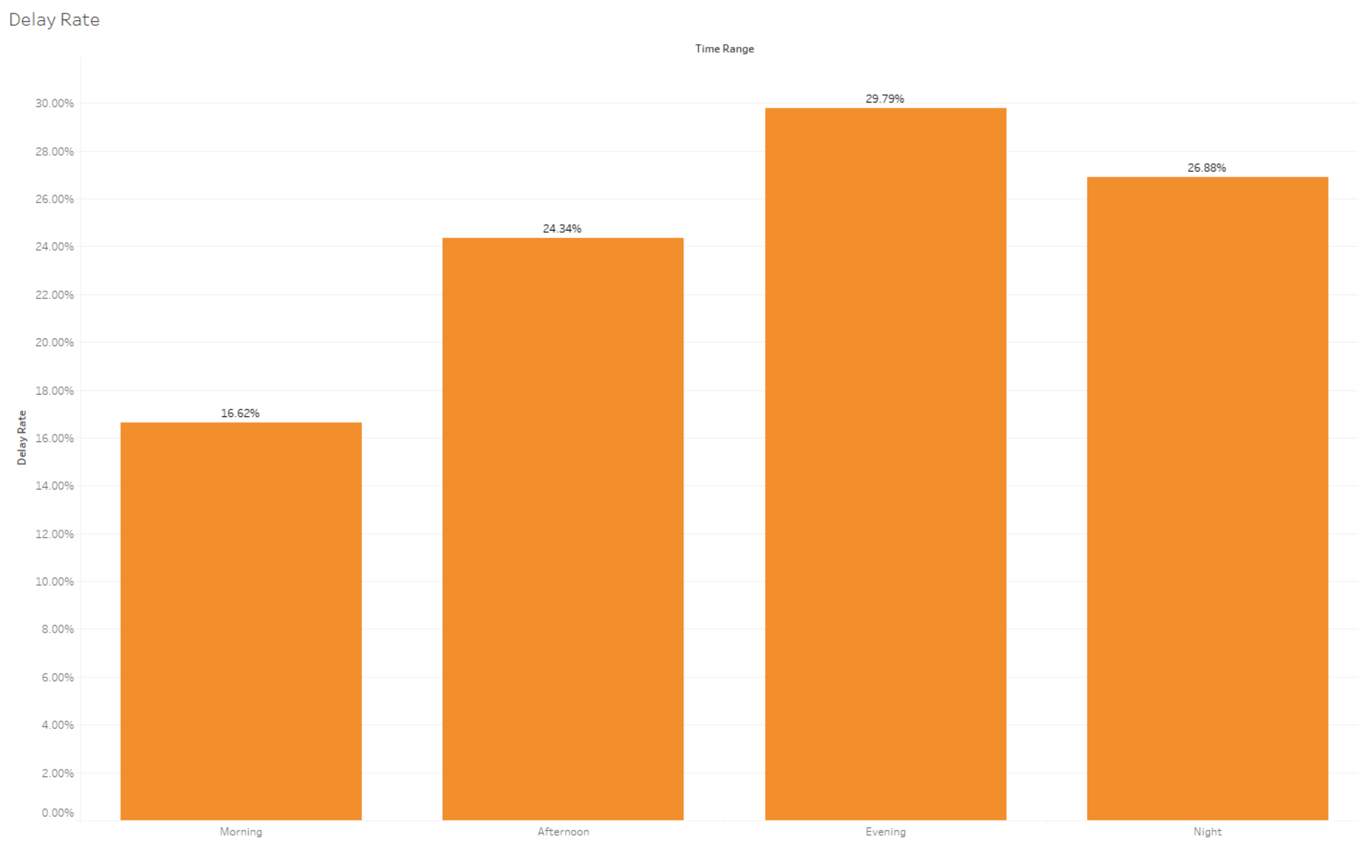


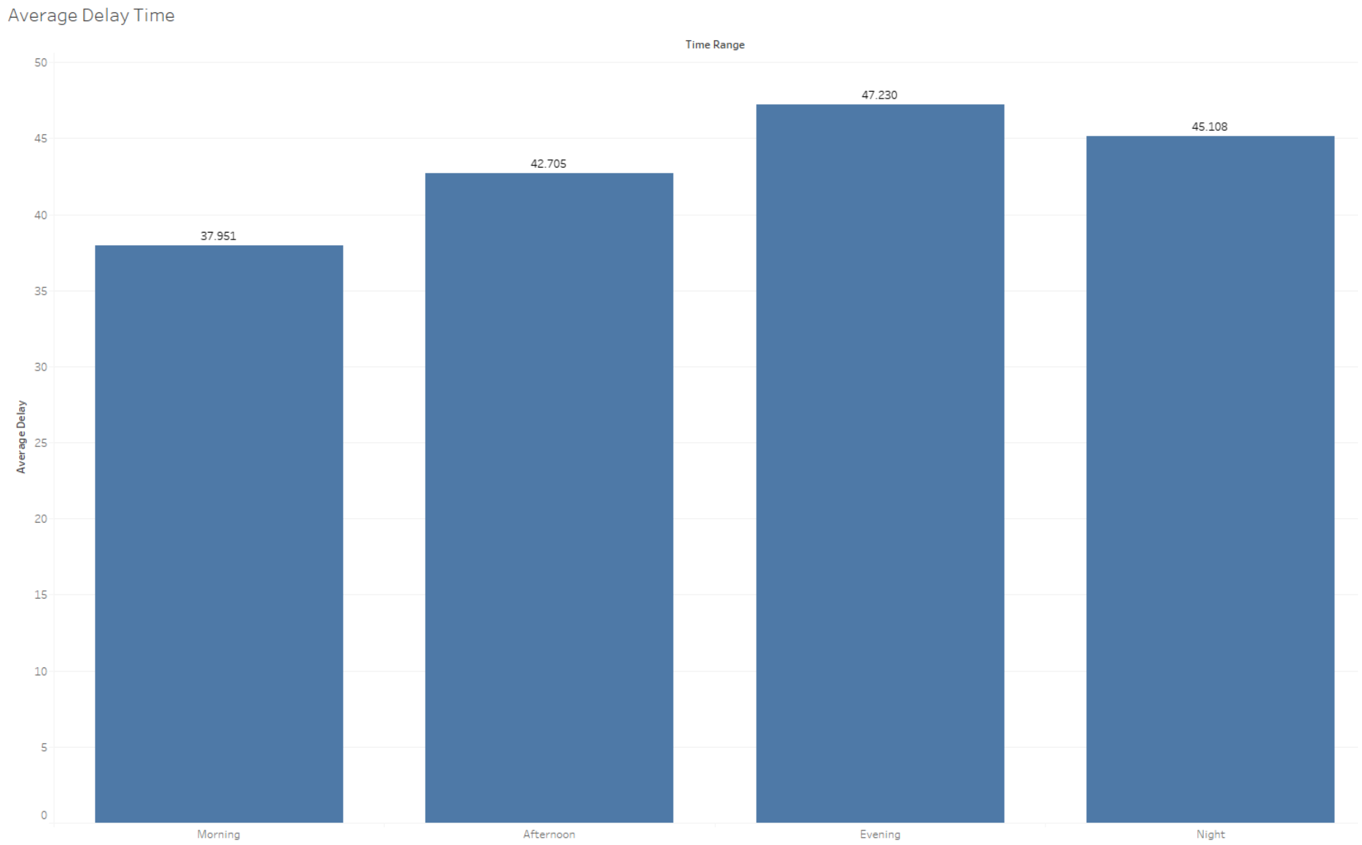
Average Delay Time By Day of Week

Cancellation Rate By Time Range



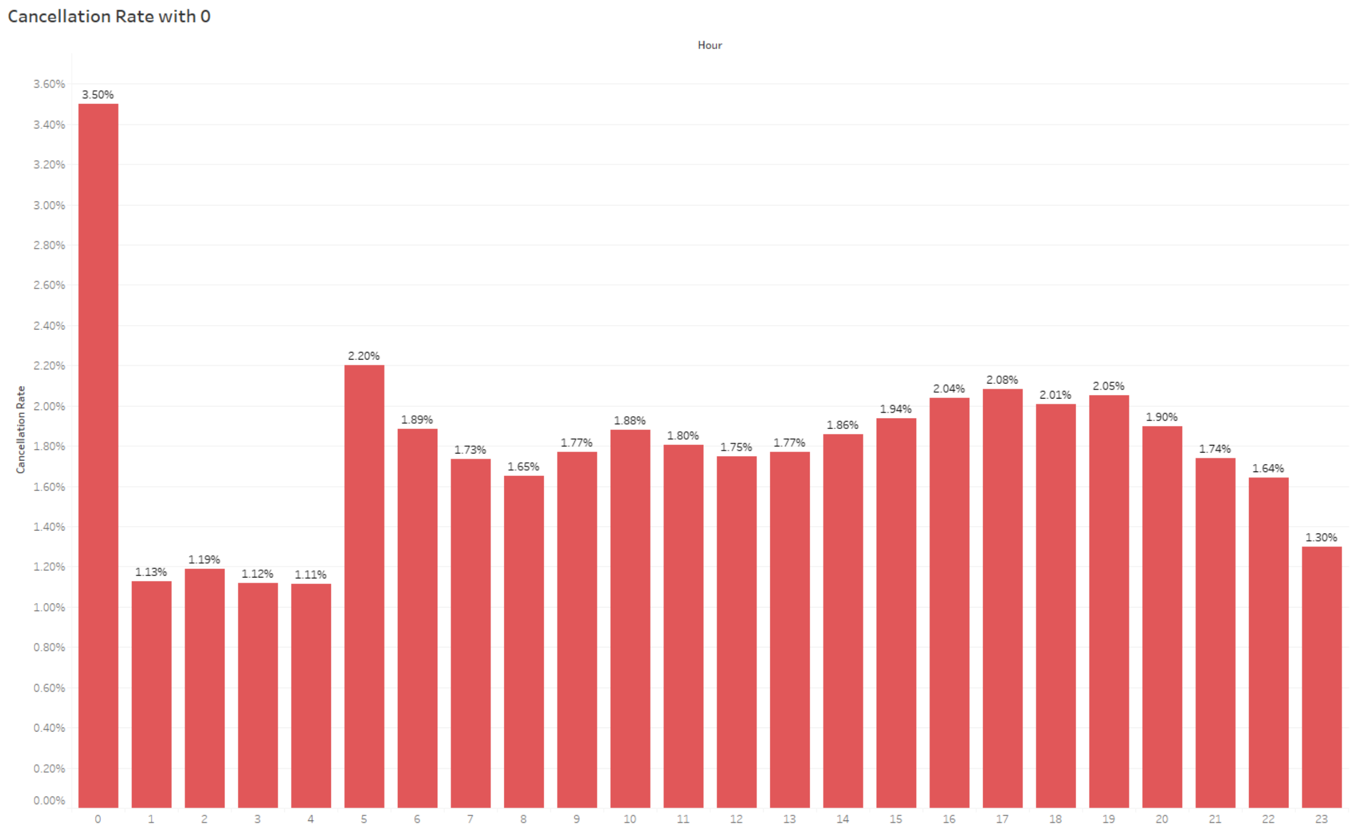
Delay Rate By Time Range

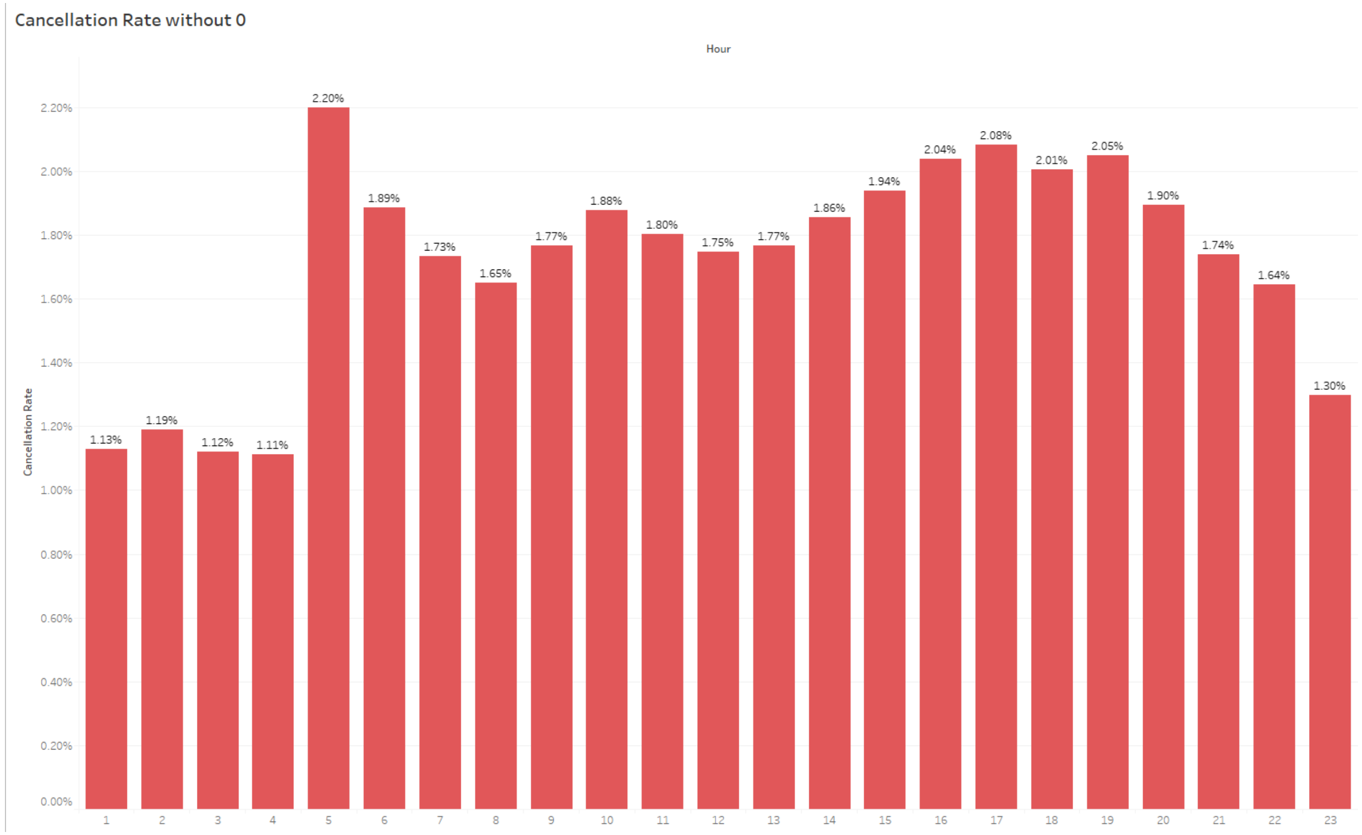




Average Delay Time By Time Range

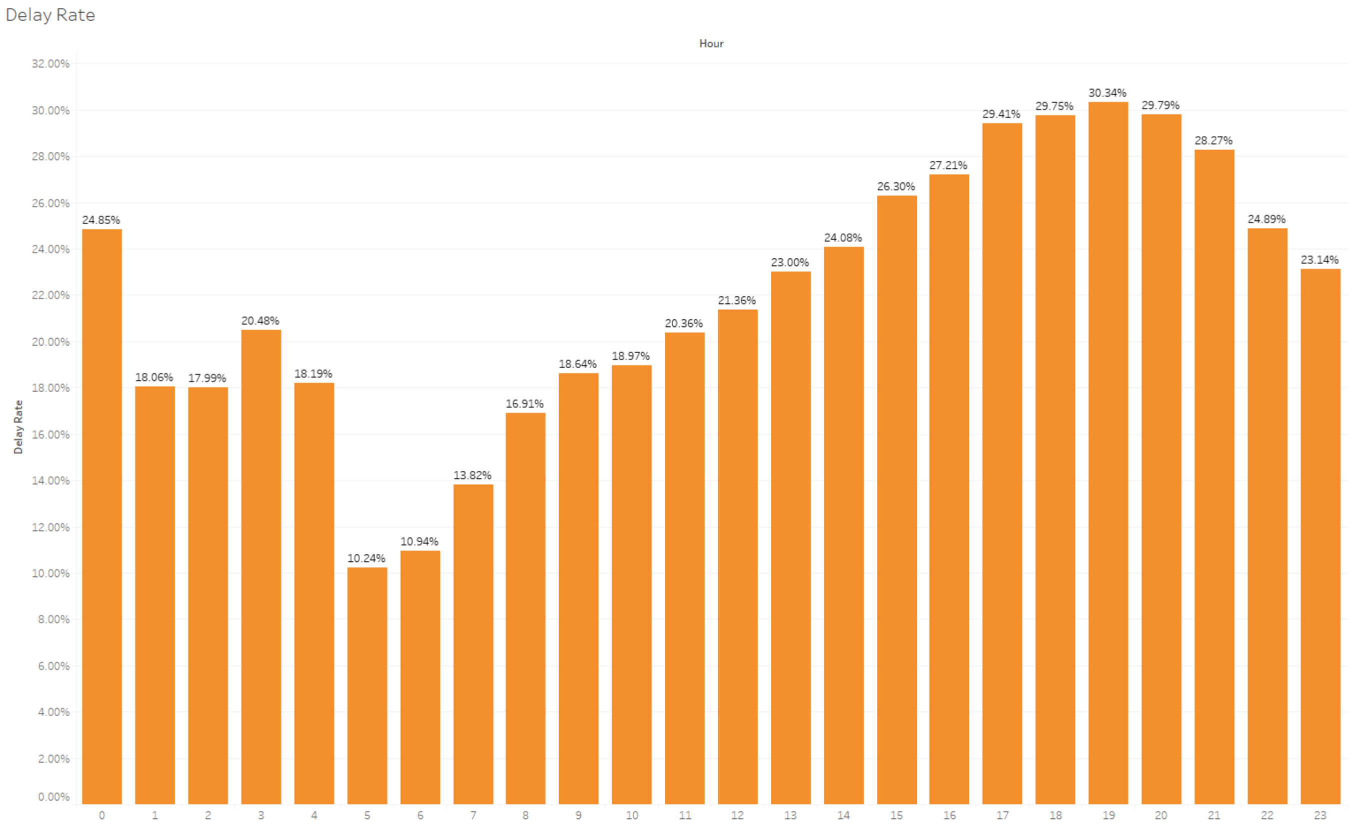
Cancellation Rate By Hour



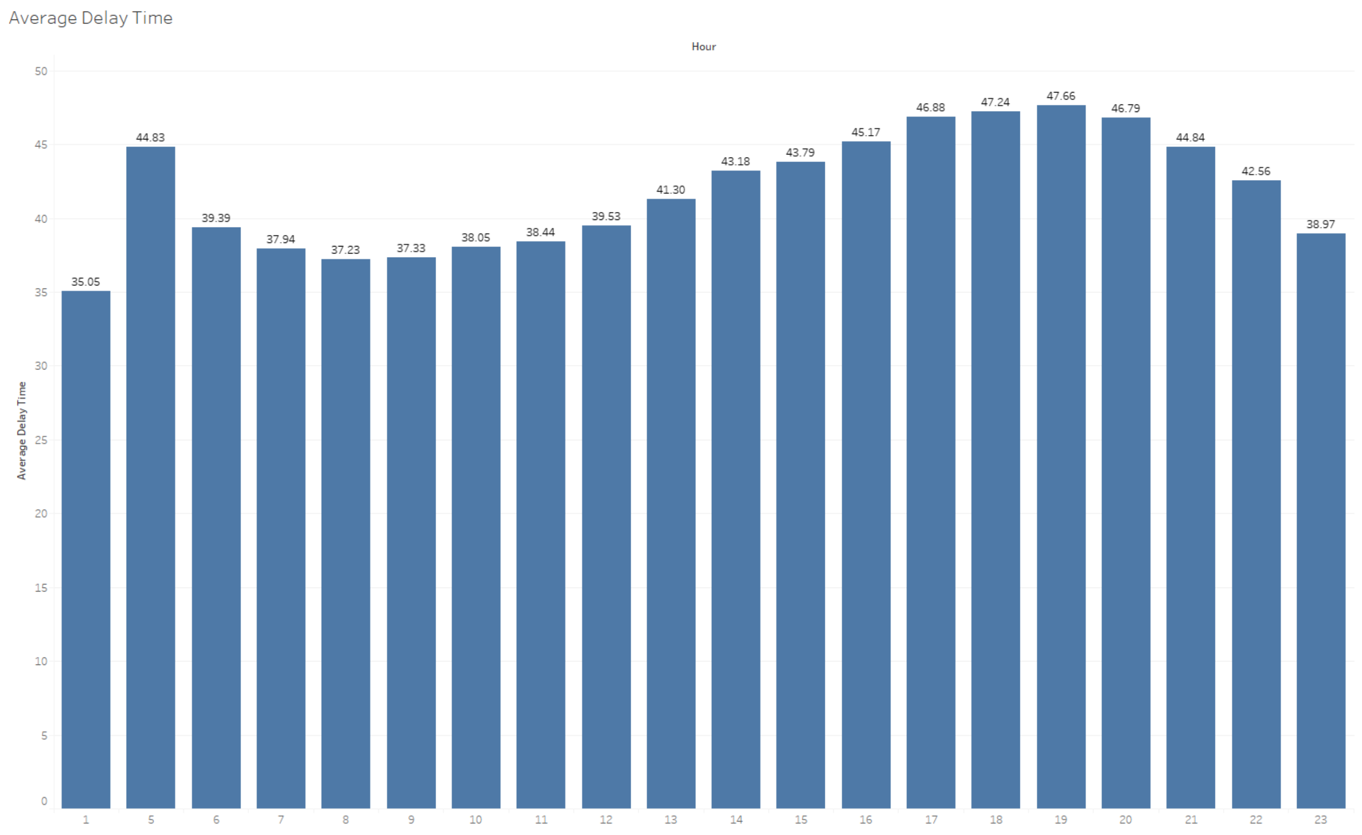


Cancellation Rate By Hour

Delay Rate By Hour



Average Delay Time By Hour

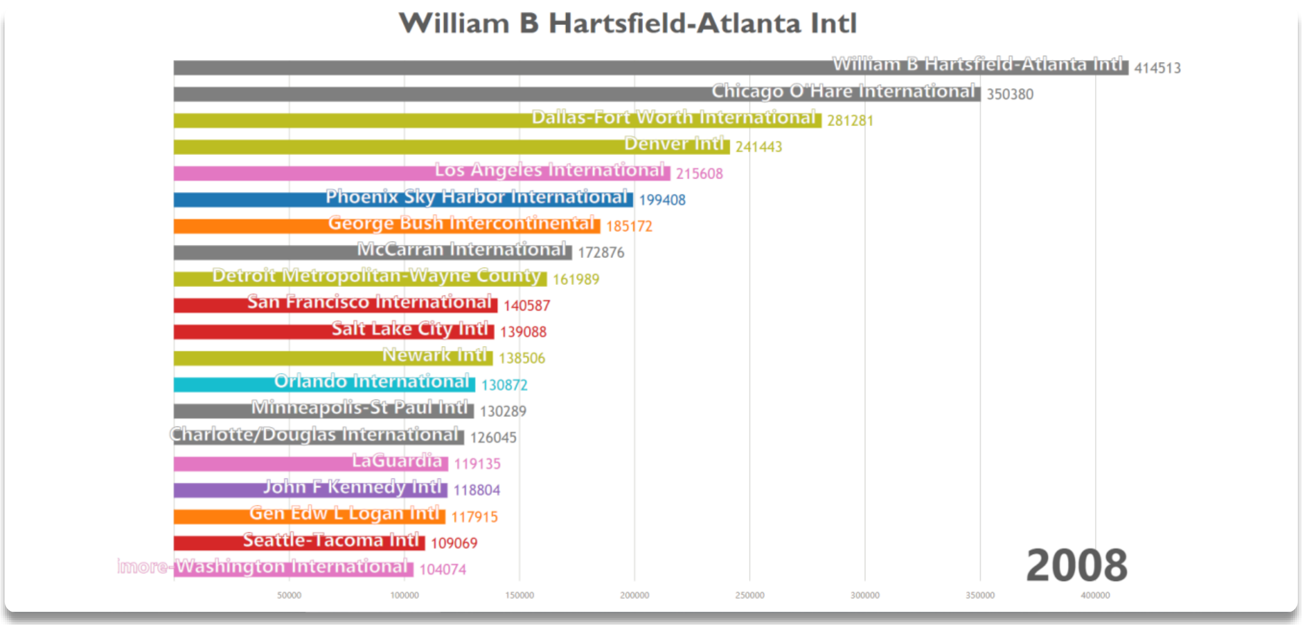


Time Analysis Summary & Conclusion

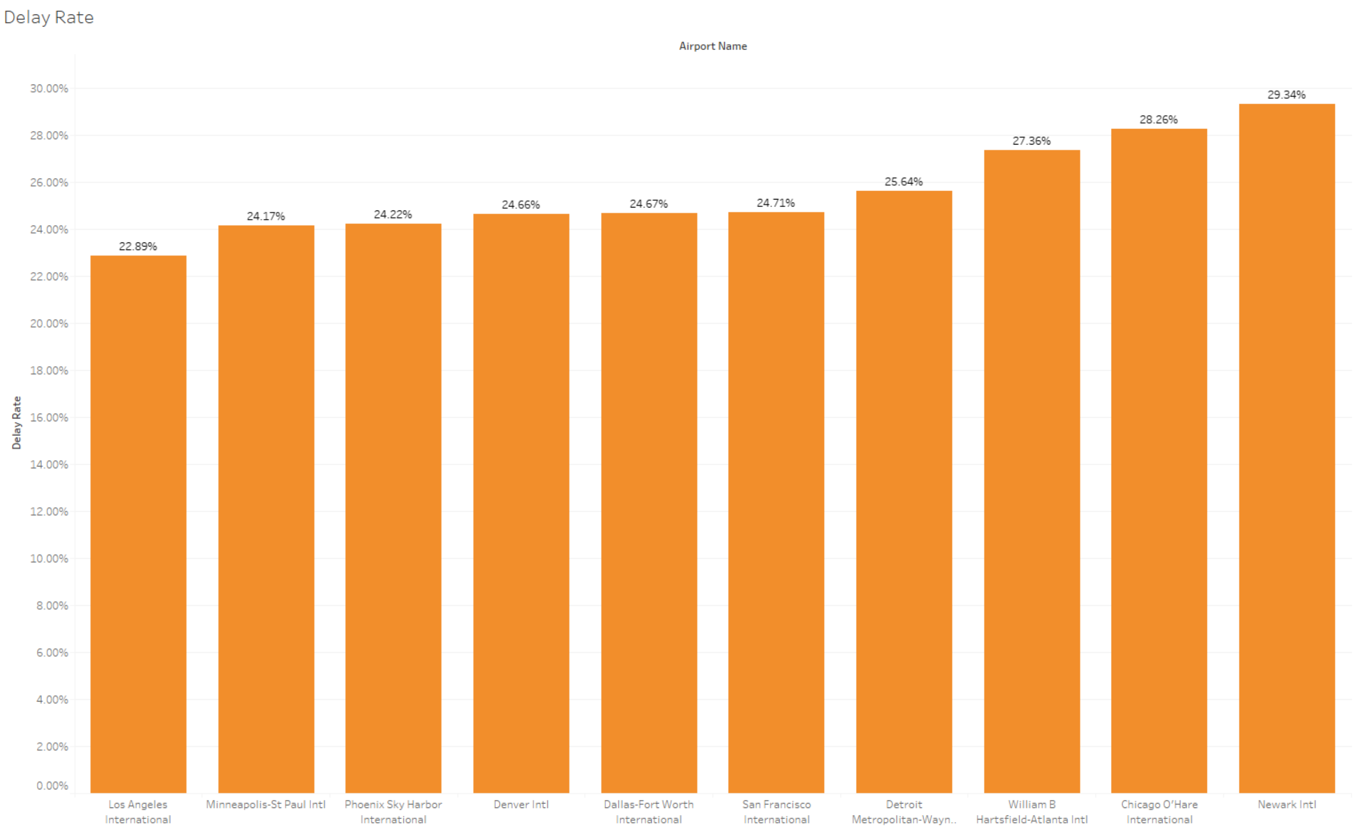
- January, September and December have the highest cancellation rates December, January, and June have the highest delay rates. Possible causes are high traffic volume during winter holidays and Christmas from Dec to Jan. Summer holidays starts in June and ends in Aug.
- Friday has the highest delay rate.
- Evening period (especially 6pm – 8pm) has the highest delay rate and relatively high cancellation rate.
- Not enough information to know why night especially midnight has higher cancellation rate. The cancellation rates are 'NA' even in most of the records in 03-08 files.
- *Detailed results for each year can be accessed in Tableau file.

Top 20 Airports 1987-2008
Number of Flights per Year

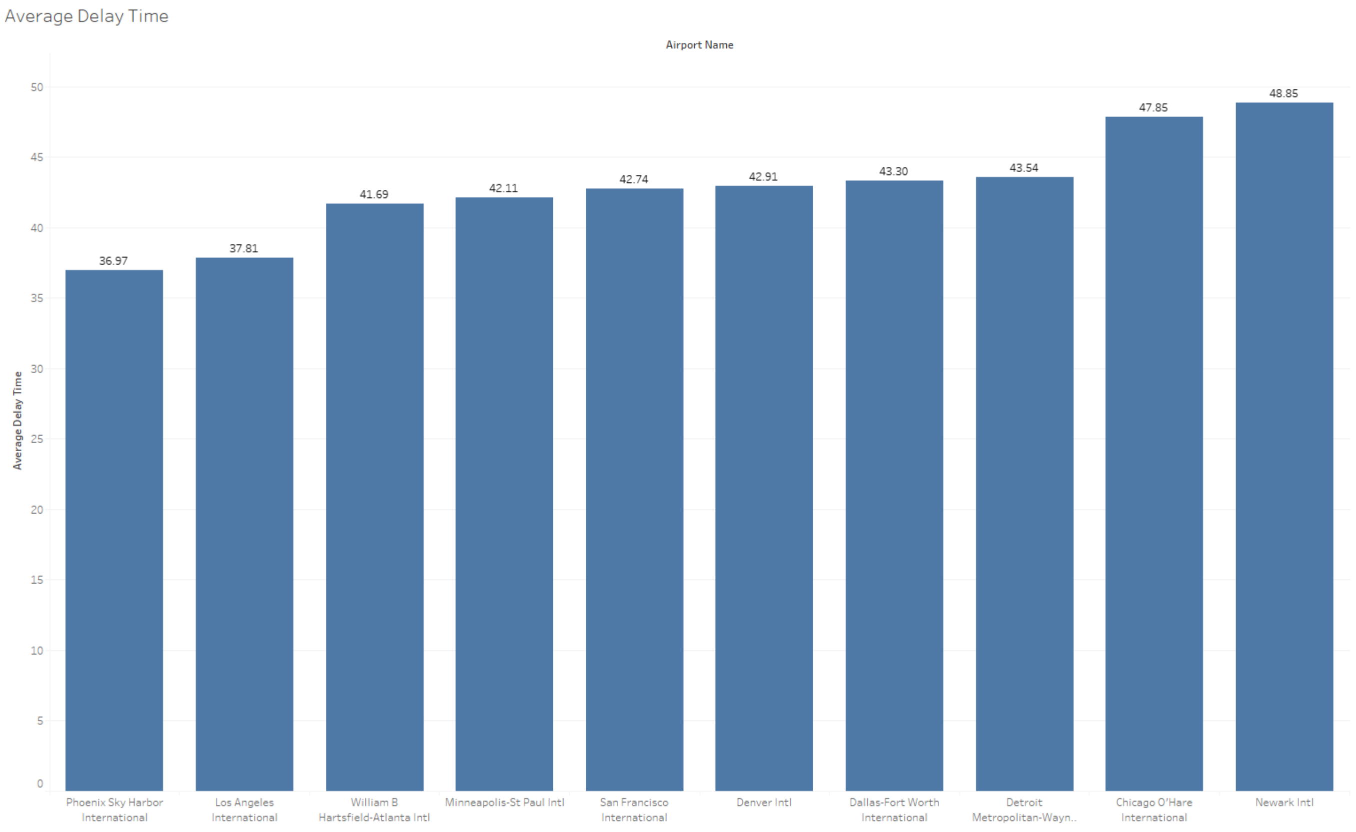
- Dynamic chart. See d3js folder in codes

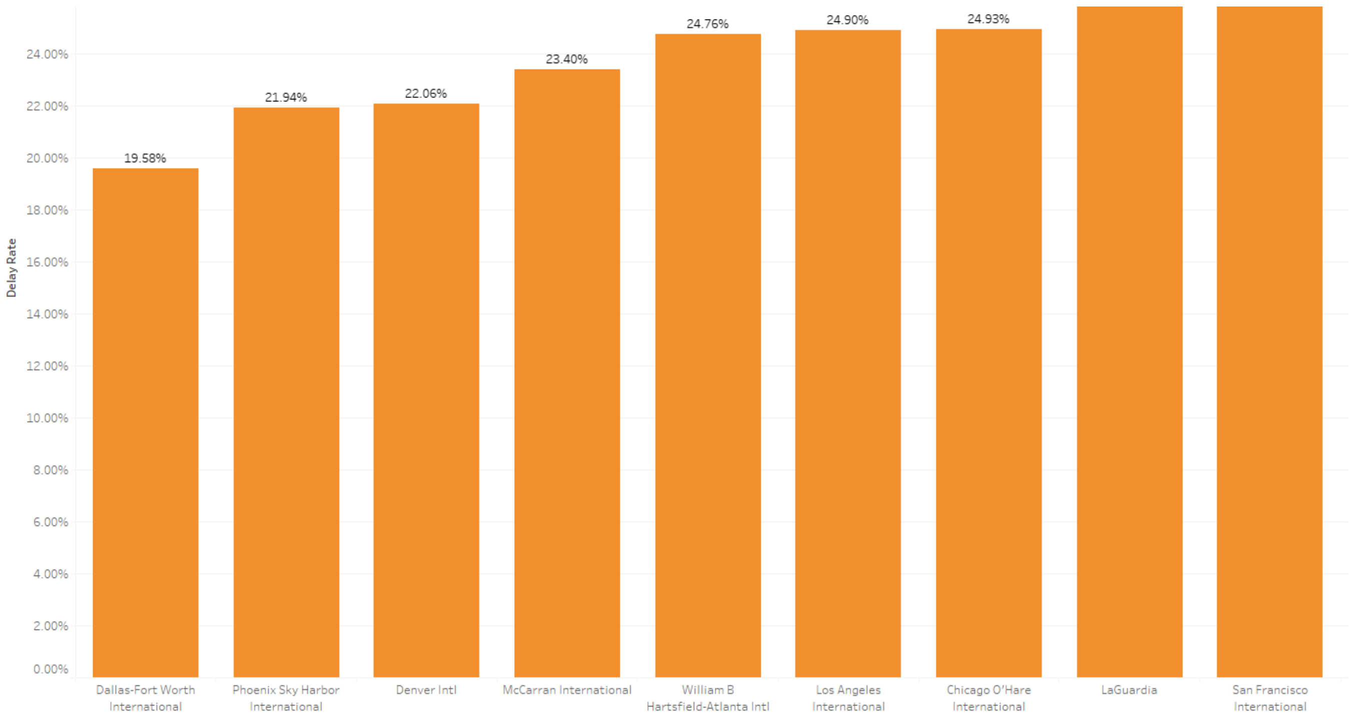


Top 10 Origin Airports Delay Rate

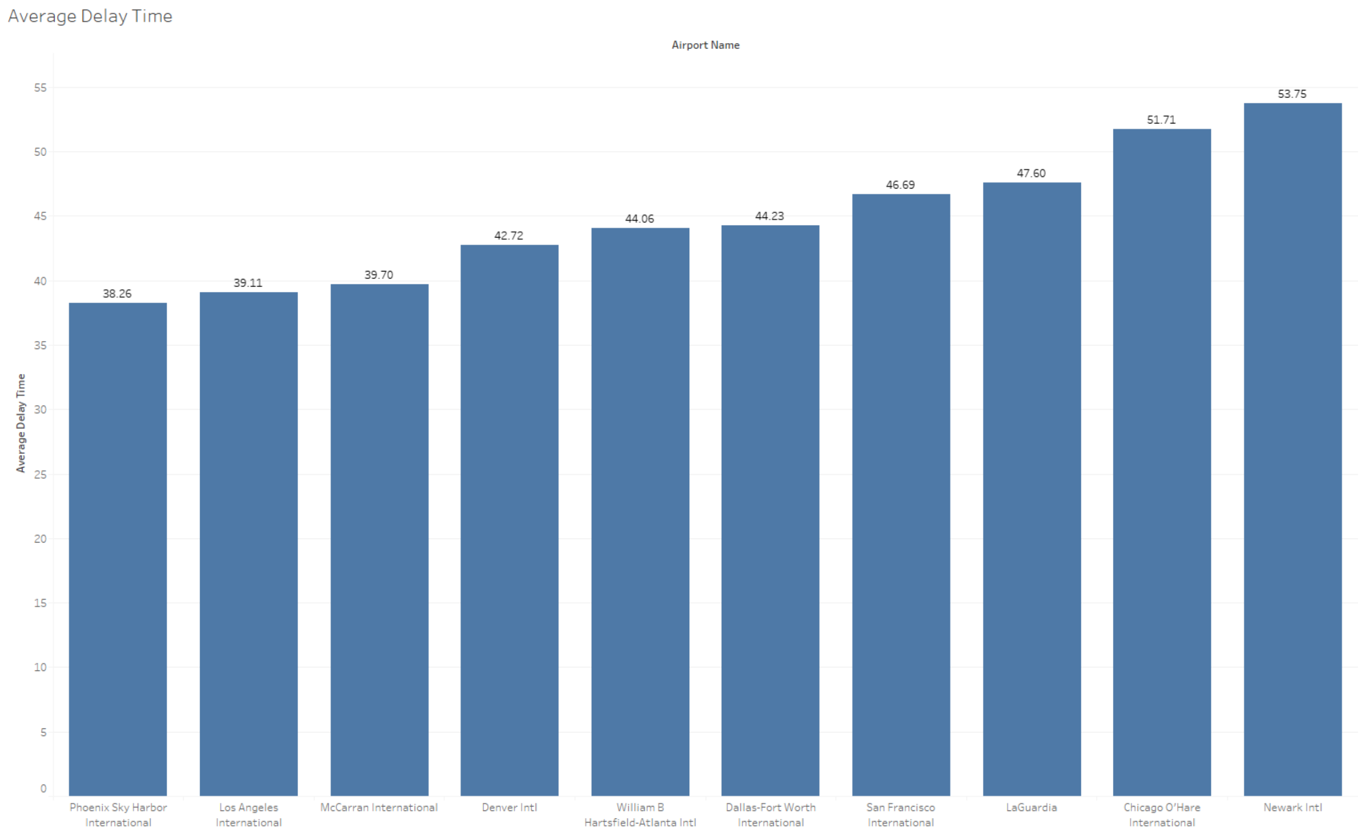


Average Delay Time

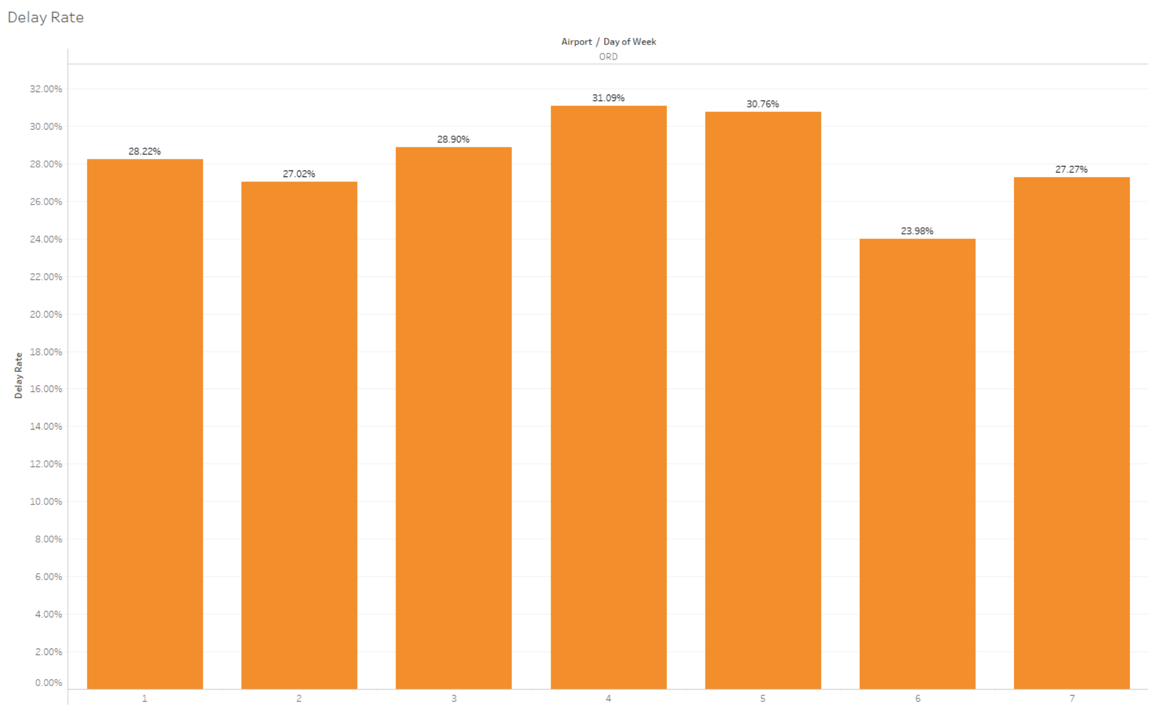


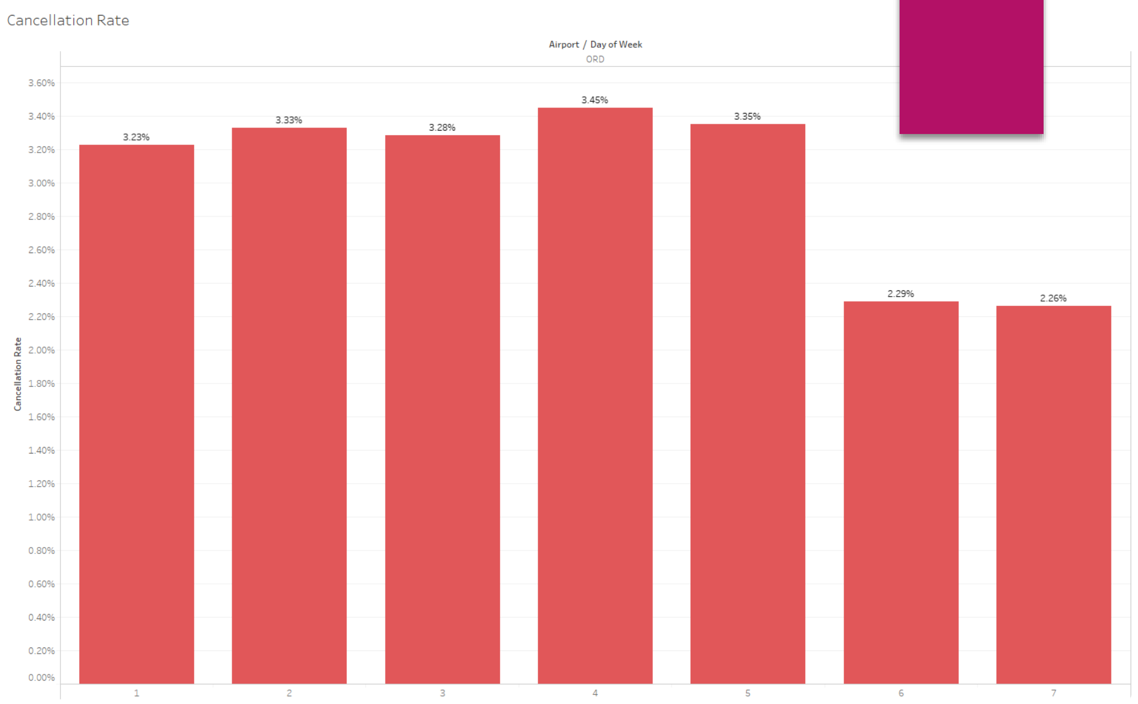


Top 10 Destination Airports Delay Rate



Average Delay Time





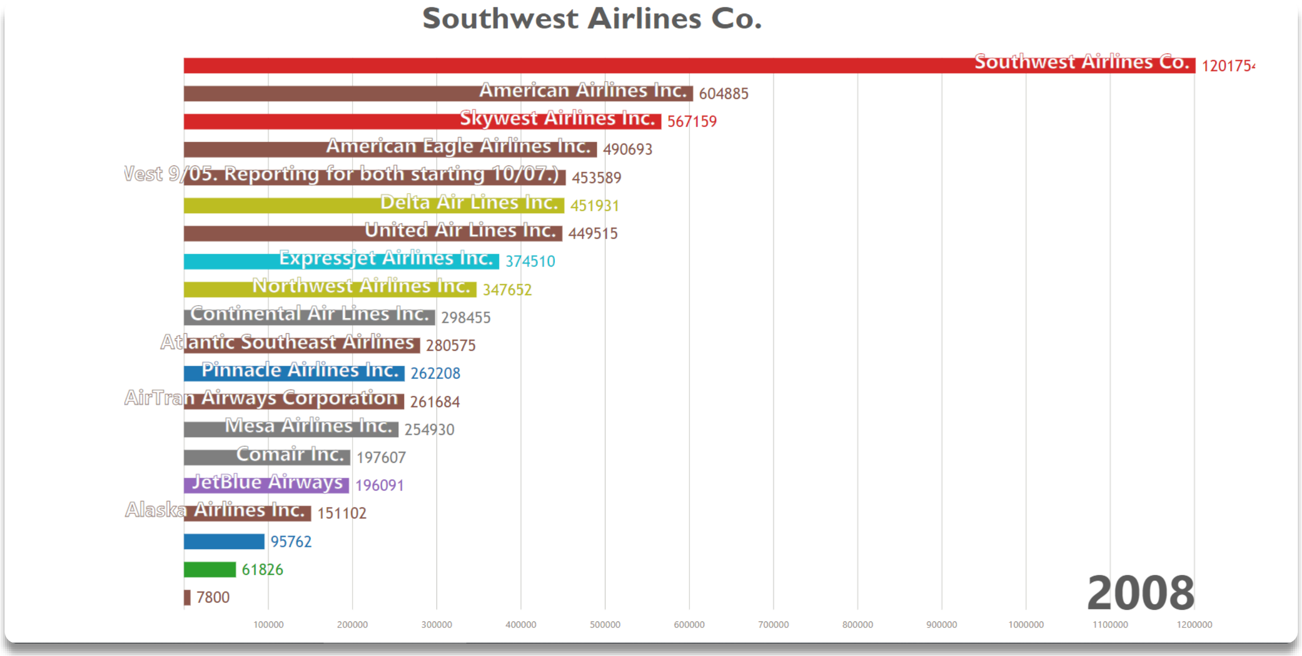
Airport Time Example Chicago O'Hare

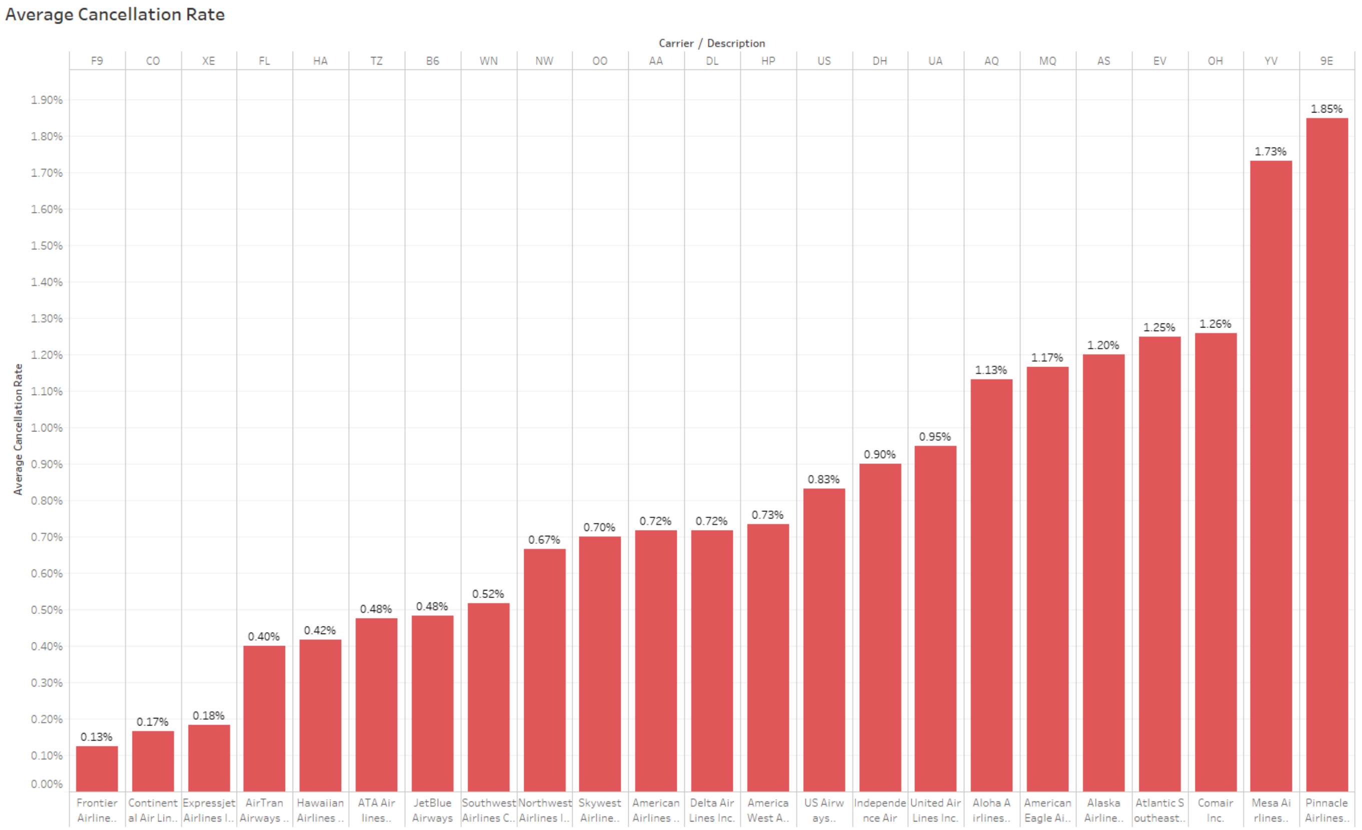
Airport Analysis Conclusion

- Atlanta, Chicago and Dallas have top total number of flights through all time.
- Newark Intl is the worst airport to travel as both origin and destination, it has the highest delay rate and average delay time.
- Chicago, Atlanta, Detroit have relatively high delay rate as origin
- As a destination, San Francisco has almost the same delay rate as Newark
- If consider the airports with high number of delayed flights and delayed rate both as busy, most of them have also higher total number of flights
- *Detailed results for each year can be accessed in Tableau file.

Total Number of Flights By Carriers 1987-2008

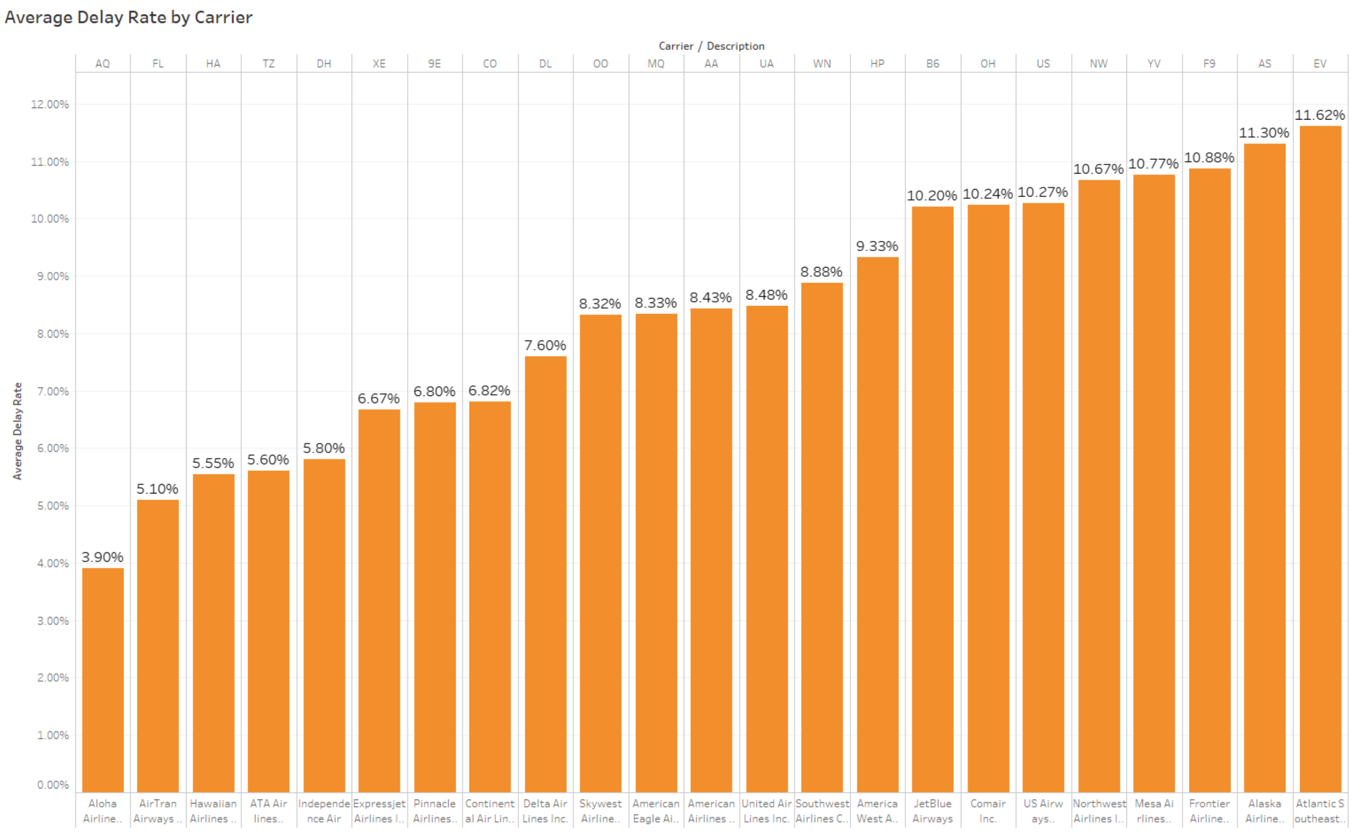
- Dynamic chart. See d3js folder in codes



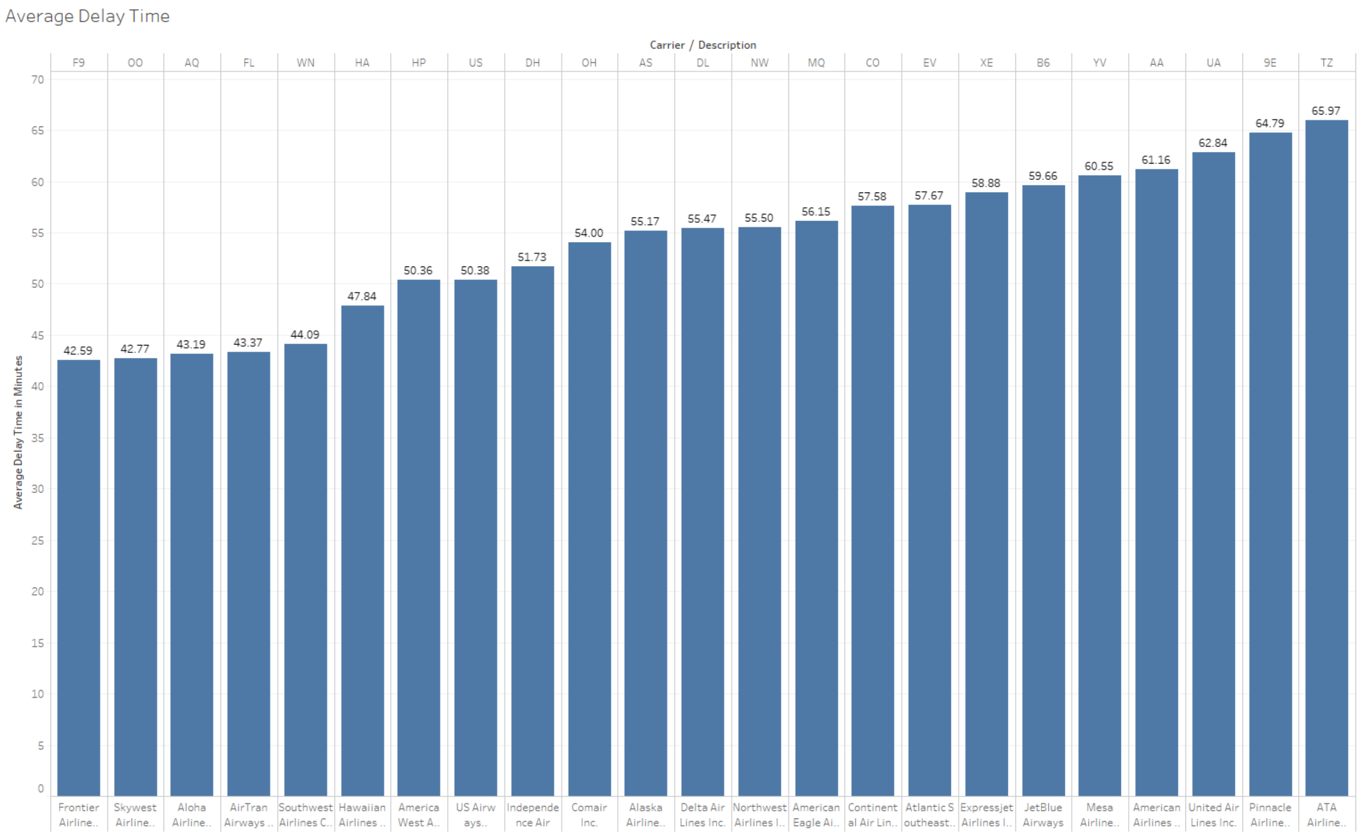


Cancellation Rate

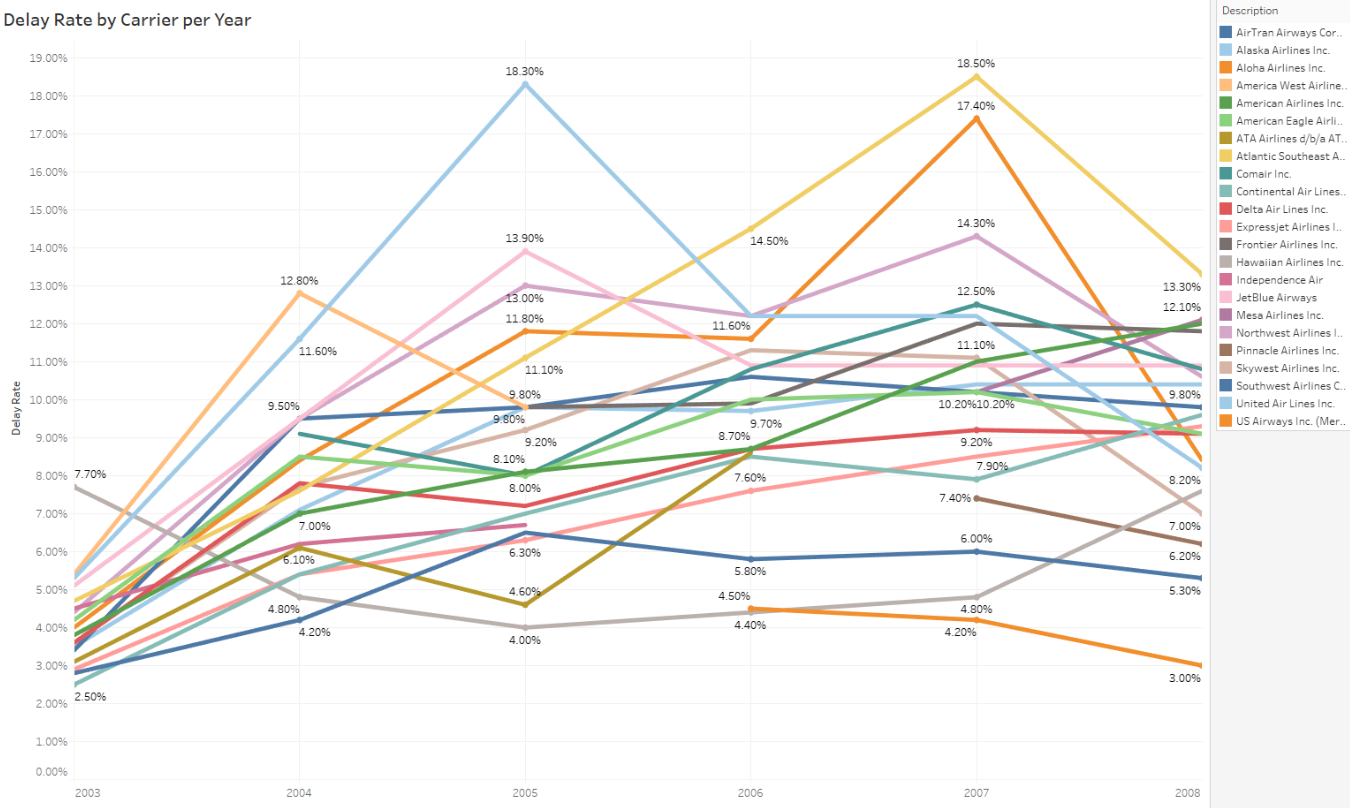
Delay Rate


Average Delay Rate by Carrier
Carrier / Description
AQ
FL
HA
TZ
DH
XE
9E
CO
DL
OO
MQ
AA
UA
WN
HP
B6
OH
US
NW
YV
F9
AS
EV
Average Delay Rate
12.00%
11.00%
10.00%
9.00%
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
3.90%
5.10%
5.55%
5.60%
5.80%
6.67%
6.80%
6.82%
7.60%
8.32%
8.33%
8.43%
8.48%
8.88%
9.33%
10.20%
10.24%
10.27%
10.67%
10.77%
10.88%
11.30%
11.62%
Aloha Airline..
AirTran Airways ..
Hawaiian Airlines ..
ATA Air lines..
Independence Air
Expressjet Airlines I..
Pinnacle Airlines..
Continental Air Lin..
Delta Air Lines Inc.
Skywest Airline..
American Eagle Ai..
American Airlines ..
United Air Lines Inc.
Southwest Airlines C..
America West A..
JetBlue Airways
Comair Inc.
US Airways..
Northwest Airlines I..
Mesa Airlines..
Frontier Airline..
Alaska Airline..
Atlantic Southeast..

Average Delay Time



Delay Rate Line Chart 2003-2008



Carrier Analysis Conclusion

- The carrier performance analysis only counts the records with carrier delay as the main reason of the delay
- Pinnacle and Mesa have highest average cancellation rates
- Atlantic Southeast and Alaska have highest average delay rates
- Airlines that delay rates keep raising: American Airlines, Continental, ExpressJet
- Alaska, Us Airways, Atlantic Southeast, US Airways, SkyWest and Northwest have big improvements in recent one year
- Aloha Airline is a small carrier and started service at 2006 but has lowest average delay rate and still improving
- *Detailed results for each year can be accessed in Tableau file.

Travel Advice

- Passengers should avoid peak time period such as the evening. Avoid travelling in busy month is not really possible if they have a travel plan but they can still try to book flights in the morning, afternoon or early night.
- Another possible way is to rearrange the day you fly, for example flying on Saturday will be better than Friday.
- If your origin or destination airport is a big one, consider again to avoid peak time. (ORD example shows Saturday is much more better even for a busy airport)
- Booking flights operated by carriers have better on-time performance and avoid bad ones such as Atlantic SouthWest.

Machine Learning (Attempt)
Data Preparation

Use 2008.csv as sample data. Similar patterns for both implementations. MapReduce to get training data then train the model with selected fields.

Fields Selected

- 1. Month
- 2. DayofWeek
- 3. CRSDepTime (Scheduled Departure Time) - Hour
- 4. Carrier
- 5. Origin
- 6. Dest
- *7. Delayed(0-not delayed, 1-delayed) arrdelay>=15 minutes

Apache Mahout OnlineLogisticRegression

LogisticRegression.java

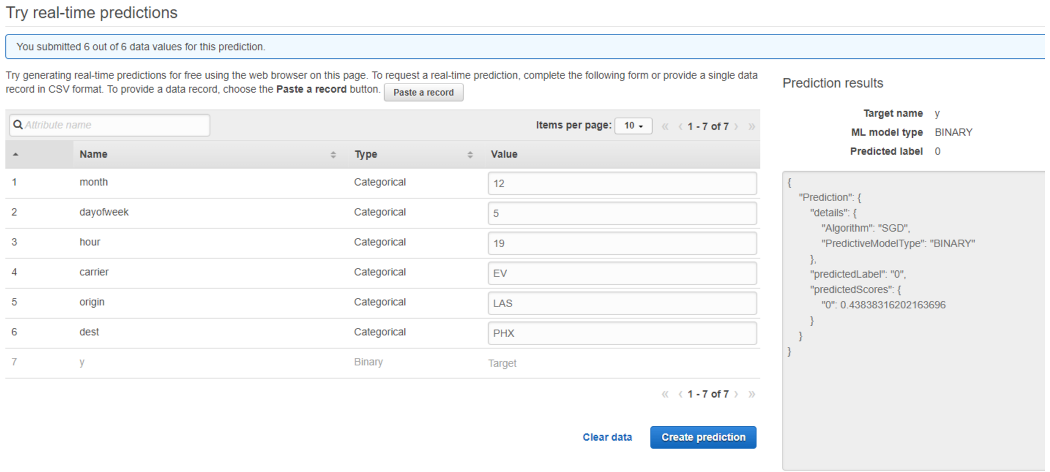
```java
        }
        return result;
    }

    public OnlineLogisticRegression train(List<Observation> trainData) {
        // System.out.println(trainData.size());
        System.out.println("Start Training");
        OnlineLogisticRegression olr = new OnlineLogisticRegression(2, 7, new L1());
        // Train the model
        for (int pass = 0; pass < 5; pass++) {
            for (Observation observation : trainData) {
                olr.train(observation.getActual(), observation.getVector());
            }

            if (pass % 1 == 0) {
                Auc eval = new Auc(0.5);
                for (Observation observation : trainData) {
                    eval.add(observation.getActual(), olr.classifyScalar(observation.getVector()));
                }

                System.out.format("Pass: %2d, Accuracy: %2.4f\n", pass + 1, eval.auc());
            }
        }
        return olr;
    }

    void testModel(OnlineLogisticRegression olr) {
        Observation newObservation = new Observation(new String[] { "12", "5", "19", "EV", "LAS", "PHX", "0" });
        Vector result = olr.classifyFull(newObservation.getVector());

        System.out.println("------------ Testing ------------");
        System.out.format("Probability of not Delay (0) = %.3f\n", result.get(0));
        System.out.format("Probability of Delay (1)     = %.3f\n", result.get(1));
    }

    class Observation {
        private DenseVector vector = new DenseVector(7);
        private int actual;

        public Observation(String[] values) {
```

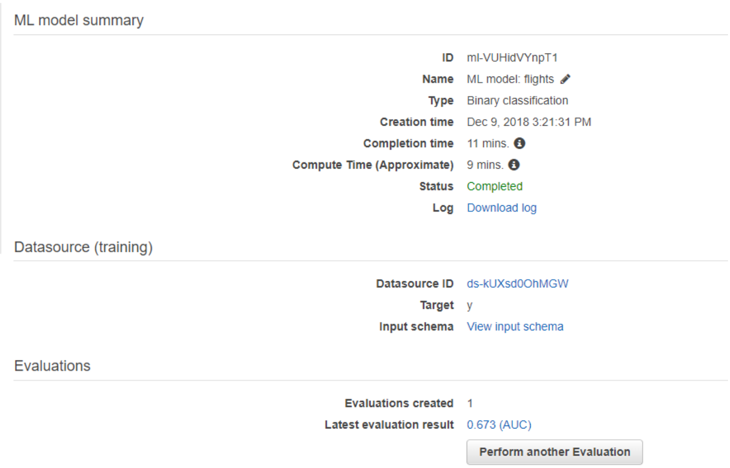
Problems | Javadoc | Declaration | Console

<terminated> LogisticRegression [Java Application] /usr/local/lib/jdk1.8.0_192/bin/java (Dec 12, 2018, 12:30:47 AM)

```
Start Importing
Start Training
Pass:  1, Accuracy: 0.5495
Pass:  2, Accuracy: 0.5481
Pass:  3, Accuracy: 0.5456
Pass:  4, Accuracy: 0.5477
Pass:  5, Accuracy: 0.5511
------------ Testing ------------
Probability of not Delay (0) = 0.717
Probability of Delay (1)     = 0.283
```

AWS Machine Learning





Machine Learning Conclusion

- The accuracy of my prediction model is not good enough (best 60% and average 55%). One of the reason could be that Apache Mahout is normally used for recommendation and classification model. It's very hard to find a detailed guide showing how to implement the logistic regression and tune the model.

- The AWS Machine Learning Model has a better score (67%) but still not good enough. I have tried to used MapReduce to get training data with different parameters but this is the only one data that I have successfully imported to the model due to the permission problem of S3 bucket.

Additional Ideas

- Selecting datasets with such big time range is good for showing results through years such as performance trending. However, making recommendations with average/overall result may not be a good idea since airports and carriers operate differently because of the increased air traffic and more advanced technologies they use.
- For prediction using machine learning, the idea is similar to the first one. My opinion is to use most recent data (2-3 years) to increase the performance of the prediction model. Another possible reason that my models don't get ideal scores is I don't have the weather data. There are weather cancellations and delays in the data and the weather data could be related to the month results in some seasons such as Summer and Winter (Storm, Snow).

References

- Apache Mahout OnlineLogisticRegression example

  http://technobium.com/logistic-regression-using-apache-mahout/

- D3.js

  https://github.com/d3/d3

  https://github.com/Jannchie/Historical-ranking-data-visualization-based-on-d3.js