

Jiwan Hwang
AIT 580
Final Project
July 25, 2018

2015 Flight Delays and Cancellations Analysis

[Introduction]

In the U.S., Air transportation is one of the most important transportation means. It is the best transportation especially for long distance travel, and unlike Europe and other large nations people in the states prefer air plane over train. Because of huge amounts of air traffics there have always been delays in major airports and sometimes cancellations of flights.

This project is about analysis of flight delays and cancellations happened in 2015. The scope of the project is analysis of delays and cancellations by airlines and by airports, relationships between air time and distance, and relationship between departure delays and arrival delays. Data cleansing for this dataset is not effectively and thoroughly applied because it is out of the scope of the AIT 580 course, but I tried best to avoid missing values altering the dataset with some conditions.

The dataset was downloaded from Kaggle.com but they are originally from Bureau of Transportation Statistics within Department of Transportation. It consists of three csv files: airlines.csv(1KB), airport.csv(24KB), and flights.csv(578MB). The airlines.csv and airport.csv are metadata which support flights.csv. Therefore, the flights.csv file is the most used one for this project by providing information about delays and cancellation of the flights in 2015.

[Who collected the data]

Bureau of Transportation Statistics is the independent statistical agency as a part of the U.S. Department of Transportation. They collect all data related nation's transportations in the U.S.

and provide database and analysis system online. “The Bureau of Transportation Statistics (BTS) is a politically objective supplier of trusted and statistically sound baseline, contextual, and trend information used to shape transportation policy, investments, and research across the U.S. and abroad (Bureau of Transportation Statistics).”

The purpose of the Bureau of Transportation is to improve transportation systems in the U.S. by analyzing and providing “timely, accurate, and reliable information (2018).”

[Need for data]

Bureau of Transportation Statics collected this data to analyze the flight status, such as delays and cancellations in terms of airlines, airports, seasons, and so on. Not only the Department of Transportation uses this data to improve conditions of airlines and make better transportation policy, but also it gives the data interest parties, such as airlines, airports, and even customers. Nowadays it is getting more important for busy and huge airports like JFK or ATL to understand and analyze all kinds of information about flights because they need to optimize operations of airport using the data. In the U.S., customers transfer to another airplane very often to go to their final destinations. In order to reduce their risks of missing another transferring plane, it is better for them to understand which airlines or airports have more delays.

Here are a few questions to answer by studying this data.

Q1. Which airlines had most arrival delays?

I wonder which airlines had most arrival delays in 2015.

Q2. Which airport has most arrival delays?

I wonder which airports has most arrival delays in 2015.

Q3. Is there a relationship between distance and airtime?

It looks pretty sure that the longer distance, the more airtime. However, it might not be always true because it depends on the speed of an air plane or a variety of sky routes.

Q4. Is there a relationship between departure delay time and arrival delay time?

I guess that if there is a longer departure delay time, a following arrival delay time would be longer, meaning a positive relationship between them.

Q5. How many (or what percentage of) flights were on time (or early) when they had departure delays?

Most people care of only arrival delays not departure delays, so it would be fine not to have arrival delays even if they had departure delays.

I want to analyze how many flights were on time even if they had departure delays.

Q6. How many (or what percentage of) flights had arrival delays when they did not have departure delays?

If the flights with arrival delays have no departure delays, the only reason of delays would be arrival delays (pure arrival delays).

These datasets (three CSV files) were found and downloaded in kaggles.com but they are originally from Bureau of Transportation Statics within Department of Transportation. Users can get a customized dataset using TranStats. Therefore, in terms of data privacy, the datasets were published publicly under the federal Open Data Policy and the OPEN (Open, public, Electronic and Necessary) Government Data Act (S. 760, H.R. 1770). Everyone can read and use these data.

[Requirements and Resources needed]

Flights.csv file's capacity is 578MB. It was slow to process the data using a laptop computer which is intel core i5 and 8GB because of the large size of the file. For that reason, two laptops were used to code. A laptop was used to code R and the other laptop was used to code Postgres SQL scripts during waiting for a result from the first laptop computer.

To explore the dataset overall, Postgres SQL 3.0 was used. To alter data formation, perform some analysis, and graph outcomes, R studio was mainly used.

[Dataset Description]

The flights.csv file was imported to Postgres SQL 3.0 manually. To check and validate whether it was successfully imported, 'View Top 100 rows' was performed in Object Browser of SQL. Data with 31 columns were successfully imported like below.

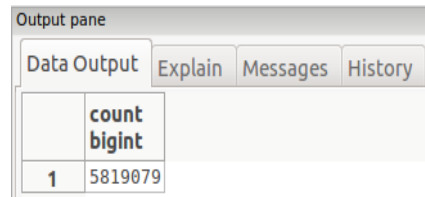
	year Integer	month Integer	day Integer	day_of_week Integer	airline character varying(255)	flight_number Integer	tail_number character varying(255)	origin_airport character varying(255)	destination_airport character varying(255)	scheduled_departure Integer	departure_time Integer	departure_delay Integer	taxi_out Integer	wheels_off Integer	schedule_time Integer	elapsed_time Integer
1	2015	1	1	4	AS	98	N407AS	ANC	SEA	5	2354	-11	21	15	205	194
2	2015	1	1	4	AA	2336	N3KJAA	LAX	PBI	10	2	-8	12	14	200	279
3	2015	1	1	4	US	848	N171US	SFO	CLT	20	18	-2	16	34	206	293
4	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20	15	-5	15	30	205	201
5	2015	1	1	4	AS	135	N527AS	SEA	ANC	25	24	-1	11	35	235	215
6	2015	1	1	4	DL	806	N3730B	SFO	MSP	25	20	-5	18	38	217	230
7	2015	1	1	4	NK	612	N635NK	LAS	MSP	25	19	-6	11	30	181	170
8	2015	1	1	4	US	2013	N584UW	LAX	CLT	30	44	14	13	57	273	249
9	2015	1	1	4	AA	1112	N3LAA	SFO	DFW	30	19	-11	17	36	195	193
10	2015	1	1	4	DL	1173	N826DN	LAS	ATL	30	33	3	12	45	221	203
11	2015	1	1	4	DL	2336	N958DN	DEN	ATL	30	24	-6	12	36	173	149
12	2015	1	1	4	AA	1674	N853AA	LAS	MIA	35	27	-8	21	48	268	266
13	2015	1	1	4	DL	1434	N547US	LAX	MSP	35	35	0	18	53	214	210
14	2015	1	1	4	DL	2324	N3751B	SLC	ATL	40	34	-6	18	52	215	199
15	2015	1	1	4	DL	2440	N651DL	SEA	MSP	40	39	-1	28	107	189	198
16	2015	1	1	4	AS	108	N309AS	ANC	SEA	45	41	-4	17	58	204	194
17	2015	1	1	4	DL	1560	N3743H	ANC	SEA	45	31	-14	25	56	210	200
18	2015	1	1	4	UA	1197	N78448	SFO	IAH	48	42	-6	11	53	218	217
19	2015	1	1	4	AS	122	N413AS	ANC	PDX	50	46	-4	11	57	215	201
20	2015	1	1	4	DL	1670	N806DN	PDX	MSP	50	45	-5	9	54	193	186
21	2015	1	1	4	NK	520	N525NK	LAS	MCI	55	120	25	11	131	162	143
22	2015	1	1	4	AA	371	N36XAA	SEA	MIA	100	52	-8	30	122	338	347
23	2015	1	1	4	NK	214	N632NK	LAS	DFW	103	102	-1	13	115	147	147
24	2015	1	1	4	AA	115	N3CTAA	LAX	MIA	105	103	-2	14	117	206	276
25	2015	1	1	4	DL	1450	N671DN	LAS	MSP	105	102	-3	11	113	183	163
26	2015	1	1	4	UA	1545	N76517	LAX	IAH	115	112	-3	11	123	183	175
27	2015	1	1	4	AS	130	N457AS	FAI	SEA	115	107	-8	25	132	213	218
28	2015	1	1	4	NK	597	N528NK	MSP	FLL	115	127	12	14	141	207	220
29	2015	1	1	4	US	413	N571UW	LAS	CLT	120	110	-10	12	122	245	224
30	2015	1	1	4	AA	2392	N3HRAA	DEN	MIA	120	141	21	12	153	227	208
31	2015	1	1	4	NK	168	N629NK	PHX	ORD	125	237	72	9	246	204	175
32	2015	1	1	4	AA	2211	N3CGAA	PHX	MIA	127	116	-11	10	126	239	234
33	2015	1	1	4	AS	136	N431AS	ANC	SEA	135					205	
34	2015	1	1	4	DL	95	N320US	SLC	ATL	140	134	-6	43	217	215	231
35	2015	1	1	4	NK	298	N514NK	LAS	IAH	144	140	-4	10	150	170	170
36	2015	1	1	4	HA	17	N389HA	LAS	HNL	145	145	0	16	201	370	385
37	2015	1	1	4	US	617	N804AW	ANC	PHX	152	143	-9	21	204	323	322
38	2015	1	1	4	UA	1528	N76519	SJU	ENR	154	157	3	12	209	255	241
39	2015	1	1	4	AS	134	N464AS	ANC	SEA	155	140	-15	17	157	218	198
40	2015	1	1	4	B6	304	N607JB	SJU	JFK	155	153	-2	12	205	235	248

100 rows.

<Table 1: Checking and Validating Import of Dataset>

After checking that the dataset was successfully imported, the total number of rows (records) were counted by following scripts and the results was 58019079 records.

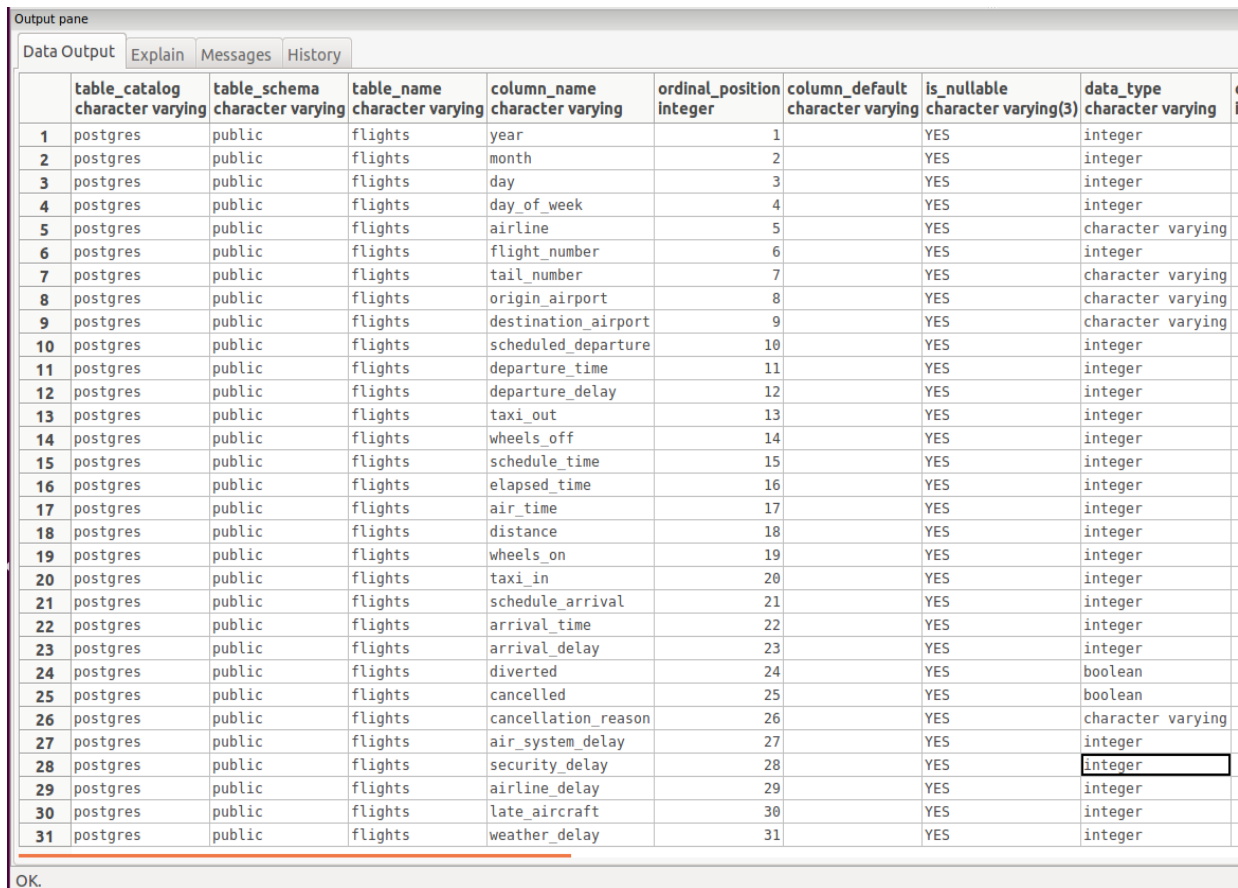
```
SELECT COUNT (*) FROM flights;
```



	count bigint
1	5819079

<Table 2: Result of Counting>

As a result, the dataset consists of 31 columns and 58019079 rows. The below is the schema of the flights dataset. In schema, the attributes of dataset are made of 24 integer, 5 character, and 2 boolean columns. For the boolean columns, True and Falls appeared instead of 1 and 0 in Postgres SQL.



	table_catalog	table_schema	table_name	column_name	ordinal_position	column_default	is_nullable	data_type
1	postgres	public	flights	year	1		YES	integer
2	postgres	public	flights	month	2		YES	integer
3	postgres	public	flights	day	3		YES	integer
4	postgres	public	flights	day_of_week	4		YES	integer
5	postgres	public	flights	airline	5		YES	character varying
6	postgres	public	flights	flight_number	6		YES	integer
7	postgres	public	flights	tail_number	7		YES	character varying
8	postgres	public	flights	origin_airport	8		YES	character varying
9	postgres	public	flights	destination_airport	9		YES	character varying
10	postgres	public	flights	scheduled_departure	10		YES	integer
11	postgres	public	flights	departure_time	11		YES	integer
12	postgres	public	flights	departure_delay	12		YES	integer
13	postgres	public	flights	taxi_out	13		YES	integer
14	postgres	public	flights	wheels_off	14		YES	integer
15	postgres	public	flights	schedule_time	15		YES	integer
16	postgres	public	flights	elapsed_time	16		YES	integer
17	postgres	public	flights	air_time	17		YES	integer
18	postgres	public	flights	distance	18		YES	integer
19	postgres	public	flights	wheels_on	19		YES	integer
20	postgres	public	flights	taxi_in	20		YES	integer
21	postgres	public	flights	schedule_arrival	21		YES	integer
22	postgres	public	flights	arrival_time	22		YES	integer
23	postgres	public	flights	arrival_delay	23		YES	integer
24	postgres	public	flights	diverted	24		YES	boolean
25	postgres	public	flights	cancelled	25		YES	boolean
26	postgres	public	flights	cancellation_reason	26		YES	character varying
27	postgres	public	flights	air_system_delay	27		YES	integer
28	postgres	public	flights	security_delay	28		YES	integer
29	postgres	public	flights	airline_delay	29		YES	integer
30	postgres	public	flights	late_aircraft	30		YES	integer
31	postgres	public	flights	weather_delay	31		YES	integer

<table 3: Dataset Schema>

There are already two metadata sets in the whole dataset. First metadata set is airlines.csv. In this metadata set, two columns which consist of names of 14 airlines and their IATA (International Air Transport Association) codes support the main dataset, flights.csv. Second one is airports.csv. Airports.csv consists of 7 columns and 322 rows of IATA code for airport, name of airport, city of airport, country of airport, latitude of airport and longitude of airport to explain more about flights.csv that is our main dataset. If more information is needed, one can download related metadata at TranStats of Bureau of Transportation Statistics.

[Result/Findings]

Six relevant columns were used to analyze for this project. These are air time, distance, departure delay time, arrival delay time, cancelled, and cancellation reason. First four columns are integer variables which are mainly used for statistical analysis. Cancellation is a boolean variable and cancellation reason is a categorized variable. There were missing values in this dataset and the technique of omitting the missing values was selected since the data is large enough. Other techniques might be applied but these are out of the scope of this project.

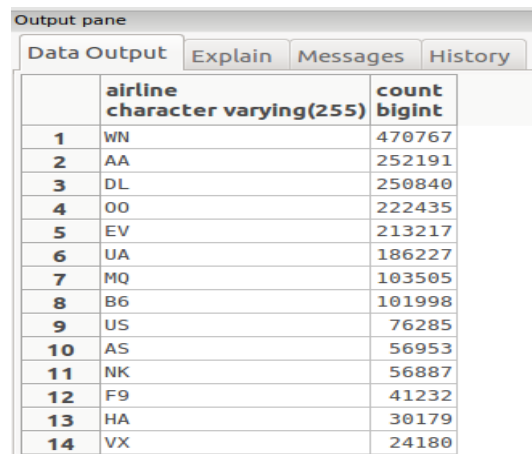
Summary of first four columns is below (without 'NA' values).

DEPARTURE_DELAY	AIR_TIME	DISTANCE	ARRIVAL_DELAY
Min. : -82.000	Min. : 7.0	Min. : 31.0	Min. : -87.000
1st Qu.: -5.000	1st Qu.: 60.0	1st Qu.: 373.0	1st Qu.: -13.000
Median : -2.000	Median : 94.0	Median : 650.0	Median : -5.000
Mean : 9.295	Mean : 113.5	Mean : 824.5	Mean : 4.407
3rd Qu.: 7.000	3rd Qu.: 144.0	3rd Qu.: 1065.0	3rd Qu.: 8.000
Max. : 1988.000	Max. : 690.0	Max. : 4983.0	Max. : 1971.000

<Result 1: Summary of Selected Data>

The number of arrival delays for each airline were analyzed using following SQL scripts in Postgres SQL. WN (Southwest Airlines) had most arrival delays in 2015 (*Q1*).

```
SELECT airline, COUNT(airline)
FROM flights WHERE arrival_delay > 0
GROUP BY airline ORDER BY COUNT(airline) DESC;
```

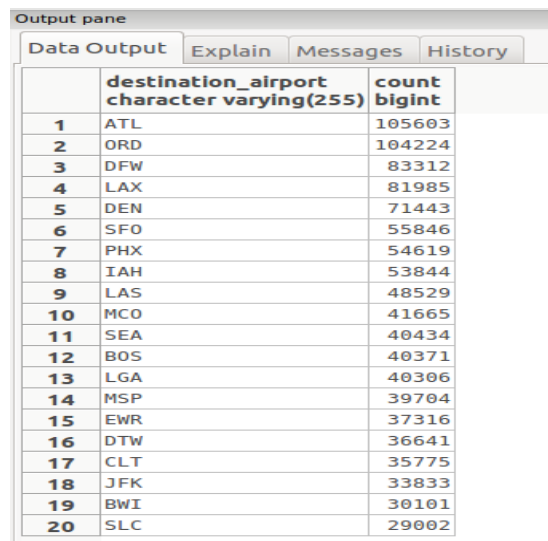


	airline character varying(255)	count bigint
1	WN	470767
2	AA	252191
3	DL	250840
4	OO	222435
5	EV	213217
6	UA	186227
7	MQ	103505
8	B6	101998
9	US	76285
10	AS	56953
11	NK	56887
12	F9	41232
13	HA	30179
14	VX	24180

<Result 2: Number of Arrival Delays by Airlines>

In the U.S., Atlanta Hartsfield Airport is usually known as the most notorious airport for arrival delays. This analysis confirmed it was true in 2015 (*Q2*).

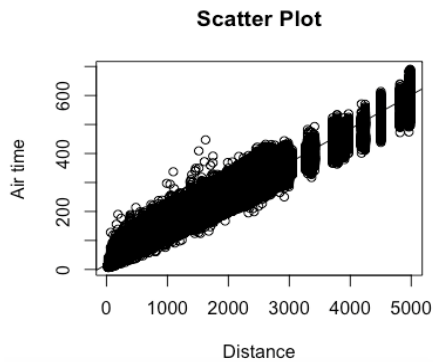
```
SELECT destination_airport, COUNT(destination_airport)
FROM flights WHERE arrival_delay > 0
GROUP BY destination_airport ORDER BY COUNT(destination_airport) DESC
LIMIT 20;
```



	destination_airport character varying(255)	count bigint
1	ATL	105603
2	ORD	104224
3	DFW	83312
4	LAX	81985
5	DEN	71443
6	SFO	55846
7	PHX	54619
8	IAH	53844
9	LAS	48529
10	MCO	41665
11	SEA	40434
12	BOS	40371
13	LGA	40306
14	MSP	39704
15	EWB	37316
16	DTW	36641
17	CLT	35775
18	JFK	33833
19	BWI	30101
20	SLC	29002

<Result 3: Number of Arrival Delays by Airports>

Generally speaking, the longer distance between airports, the more airtime it takes, but I thought it would not be always true because of variety of air courses (air routes). However, the former was true seeing the result of analysis. The correlation between distance and airtime is 0.9856435 meaning almost perfect positive relationship and R-squared (fit) is 0.9715.



```
Residuals:
    Min       1Q   Median       3Q      Max
-123.471  -6.103   -1.278    5.176   239.734

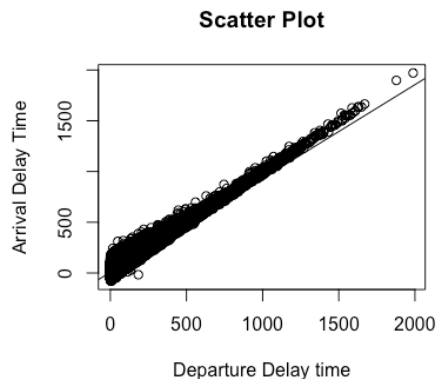
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.708e+01  8.590e-03   1988  <2e-16 ***
nonnadf$DISTANCE 1.170e-01  8.382e-06  13955  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.2 on 5714006 degrees of freedom
Multiple R-squared:  0.9715,    Adjusted R-squared:  0.9715
F-statistic: 1.947e+08 on 1 and 5714006 DF,  p-value: < 2.2e-16
```

<Result 4: Relationship between Distance and Airtime>

For the linear regression analysis, it has $\text{Airtime} = 17.077 + 0.117 \times \text{Distance}$. P-values for both coefficients are significant ($2e-16$), so the null hypothesis (the coefficients = 0) would be rejected at any level of significant level. The relationship is statistically significant (**Q3**).

I want to confirm that if longer departure delay time, the longer arrival delay time. For the conclusion, the relationship between departure delay time and arrival delay time was also almost perfectly positive. Their correlation is 0.9654397 and R-squared (fit) is 0.9321.



```
Residuals:
    Min       1Q   Median       3Q      Max
-246.308  -6.050    0.737    7.556   192.751

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9599066  0.0106068   656.2  <2e-16 ***
deptdelaydf$ARRIVAL_DELAY 0.9241937  0.0001716  5387.2  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.79 on 2115047 degrees of freedom
Multiple R-squared:  0.9321,    Adjusted R-squared:  0.9321
F-statistic: 2.902e+07 on 1 and 2115047 DF,  p-value: < 2.2e-16
```

<Result 5: Relationship between Departure Delay Time and Arrival Departure Time>

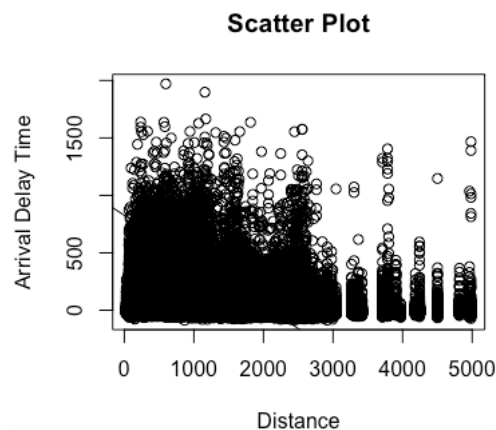
P-values of both coefficients are significant, so the null hypothesis (the coefficients = 0) is rejected at any level of significant level. The relationship is statistically significant (**Q4**).

Sometimes, people experience that an airplane arrives on time (or early) despite having a departure delay. From calculations using R, 28.69% of number of departure delays had arrivals on time (or early), ($=606902 / 2115049$). 71.31% of departure delays experienced arrival delays ($=1508147 / 2115049$). This means that if there is a departure delay, the airplane will be on time with about 30% probability (**Q5**).

How many flights had arrival delays when they did not have departure delays? It is a question about pure arrival delays. Total number of arrival delays is 2086896 with omitting NA values. Among these arrival delays, 606902 of arrival delays had no departure delays. Therefore 29.08% ($=606902 / 2086896$) of number of arrival delays had no departure delays. This means 29.08% of number of arrival delays are pure arrival delays(**Q6**).

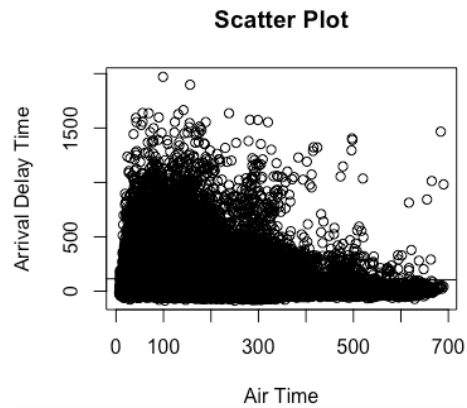
Additionally, more analyses were performed, and the results are below.

- There is a very little negative relationship between distance and arrival delay time but not strong. Its correlation is -0.025 which is near 0.



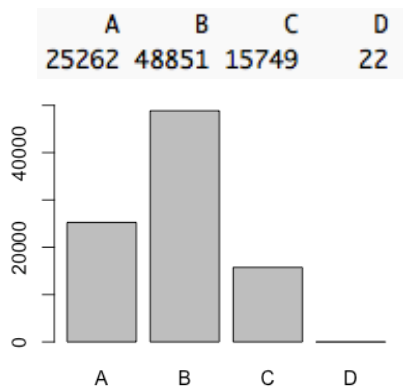
<Result 6: Relationship between Distance and Arrival Delay Time>

- There is no specific relationship between airtime and arrival delay time. The correlation coefficient is -0.0071 which is almost 0. They do not have any trend.



<Result 7: Relationship between Air Time and Arrival Delay Time>

- The biggest cancellation reason was weather(b). Codes of cancellation are found in TranStats in Bureau of Transportation Statistics Homepage.



		On-Time : On-Time Performance	
Search this site: <input type="text"/> <input type="button" value="Go"/>		Field: Specifies The Reason For Cancellation	
Advanced Search		Format results for printing Download Lookup Table	
Resources		Code	Description
Database Directory		A	Carrier
Glossary		B	Weather
Upcoming Releases		C	National Air System
Data Release History		D	Security

<Result 8: Cancellation Reason>

<Source: Bureau of Transportation Statistics (n.d.)>

Reference

Department of Transportation. n.d. "2015 Flight Delays and Cancellations: Which airline should you fly on to avoid significant delays? Version 1." Retrieved July 25, 2018.
<https://www.kaggle.com/usdot/flight-delays/version/1>.

Bureau of Transportation Statistics. July 25, 2018. "About BTS" Retrieved July 25, 2018.
<https://www.bts.gov/about-BTS>.

Bureau of Transportation Statistics. n.d. Airline On-Time Performance Data.
Retrieved July 25, 2018.
https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time.