What’s Up with Flight Delays?

Xueyan Liu, Yiran Liu, Anderson Monken, and Shiying Sun
December 5, 2019

Introduction

- Motivation
- Data from Bureau of Transportation Statistics
  - All flight departures from New York City in 2013
    - Airlines
    - Airports
    - Flights
    - Planes
    - Weather
- Factors affecting on-time performance of flights
  - Time periods during a day
  - Departing Airports
  - Airline Carriers
  - Weather Conditions
- Prediction

Motivation

- In 2018 there were over 1 billion passengers on domestic and international flights
- Oct, 2019 78.1 million passengers
- Average minutely costs to airlines for a single plane is estimated to be $74.20
- In 2018 the estimated direct cost to airlines and passengers was 28 billion dollars

Airline Dataset

carrier	name
9E	Endeavor Air Inc.
AA	American Airlines Inc.
AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines Inc.
F9	Frontier Airlines Inc.
FL	AirTran Airways Corporation
HA	Hawaiian Airlines Inc.

Airport Dataset

faa	name	lat	lon	alt	tz	dst	tzone
04G	Lansdowne Airport	41.13047	-80.61958	1044	-5	A	America/New_York
06A	Moton Field Municipal Airport	32.46057	-85.68003	264	-6	A	America/Chicago
06C	Schaumburg Regional	41.98934	-88.10124	801	-6	A	America/Chicago
06N	Randall Airport	41.43191	-74.39156	523	-5	A	America/New_York
09J	Jekyll Island Airport	31.07447	-81.42778	11	-5	A	America/New_York
0A9	Elizabethton Municipal Airport	36.37122	-82.17342	1593	-5	A	America/New_York
0G6	Williams County Airport	41.46731	-84.50678	730	-5	A	America/New_York
0G7	Finger Lakes Regional Airport	42.88356	-76.78123	492	-5	A	America/New_York
0P2	Shoestring Aviation Airfield	39.79482	-76.64719	1000	-5	U	America/New_York
0S9	Jefferson County Intl	48.05381	-122.81064	108	-8	A	America/Los_Angeles

Flight Dataset

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
2013	1	1	533	529	4	850	830	20	UA	1714	N24211	LGA	IAH	227	1416	5	29	2013-01-01 05:00:00
2013	1	1	542	540	2	923	850	33	AA	1141	N619AA	JFK	MIA	160	1089	5	40	2013-01-01 05:00:00
2013	1	1	544	545	-1	1004	1022	-18	B6	725	N804JB	JFK	BQN	183	1576	5	45	2013-01-01 05:00:00
2013	1	1	554	600	-6	812	837	-25	DL	461	N668DN	LGA	ATL	116	762	6	0	2013-01-01 06:00:00
2013	1	1	554	558	-4	740	728	12	UA	1696	N39463	EWR	ORD	150	719	5	58	2013-01-01 05:00:00
2013	1	1	555	600	-5	913	854	19	B6	507	N516JB	EWR	FLL	158	1065	6	0	2013-01-01 06:00:00
2013	1	1	557	600	-3	709	723	-14	EV	5708	N829AS	LGA	IAD	53	229	6	0	2013-01-01 06:00:00
2013	1	1	557	600	-3	838	846	-8	B6	79	N593JB	JFK	MCO	140	944	6	0	2013-01-01 06:00:00
2013	1	1	558	600	-2	753	745	8	AA	301	N3ALAA	LGA	ORD	138	733	6	0	2013-01-01 06:00:00

Weather Dataset

origin	year	month	day	hour	temp	dewp	humid	wind_dir	wind_speed	wind_gust	precip	pressure	visib	time_hour
EWR	2013	1	1	1	39.02	26.06	59.37	270	10.35702	NA	0.00	1012.0	10.00	2013-01-01 01:00:00
EWR	2013	1	1	2	39.02	26.96	61.63	250	8.05546	NA	0.00	1012.3	10.00	2013-01-01 02:00:00
EWR	2013	1	1	3	39.02	28.04	64.43	240	11.50780	NA	0.00	1012.5	10.00	2013-01-01 03:00:00
EWR	2013	1	1	4	39.92	28.04	62.21	250	12.65858	NA	0.00	1012.2	10.00	2013-01-01 04:00:00
EWR	2013	1	1	5	39.02	28.04	64.43	260	12.65858	NA	0.00	1011.9	10.00	2013-01-01 05:00:00
EWR	2013	1	1	6	37.94	28.04	67.21	240	11.50780	NA	0.00	1012.4	10.00	2013-01-01 06:00:00
EWR	2013	1	1	7	39.02	28.04	64.43	240	14.96014	NA	0.00	1012.2	10.00	2013-01-01 07:00:00
EWR	2013	1	1	8	39.92	28.04	62.21	250	10.35702	NA	0.00	1012.2	10.00	2013-01-01 08:00:00
EWR	2013	1	1	9	39.92	28.04	62.21	260	14.96014	NA	0.00	1012.7	10.00	2013-01-01 09:00:00
EWR	2013	1	1	10	41.00	28.04	59.65	260	13.80936	NA	0.00	1012.4	10.00	2013-01-01 10:00:00

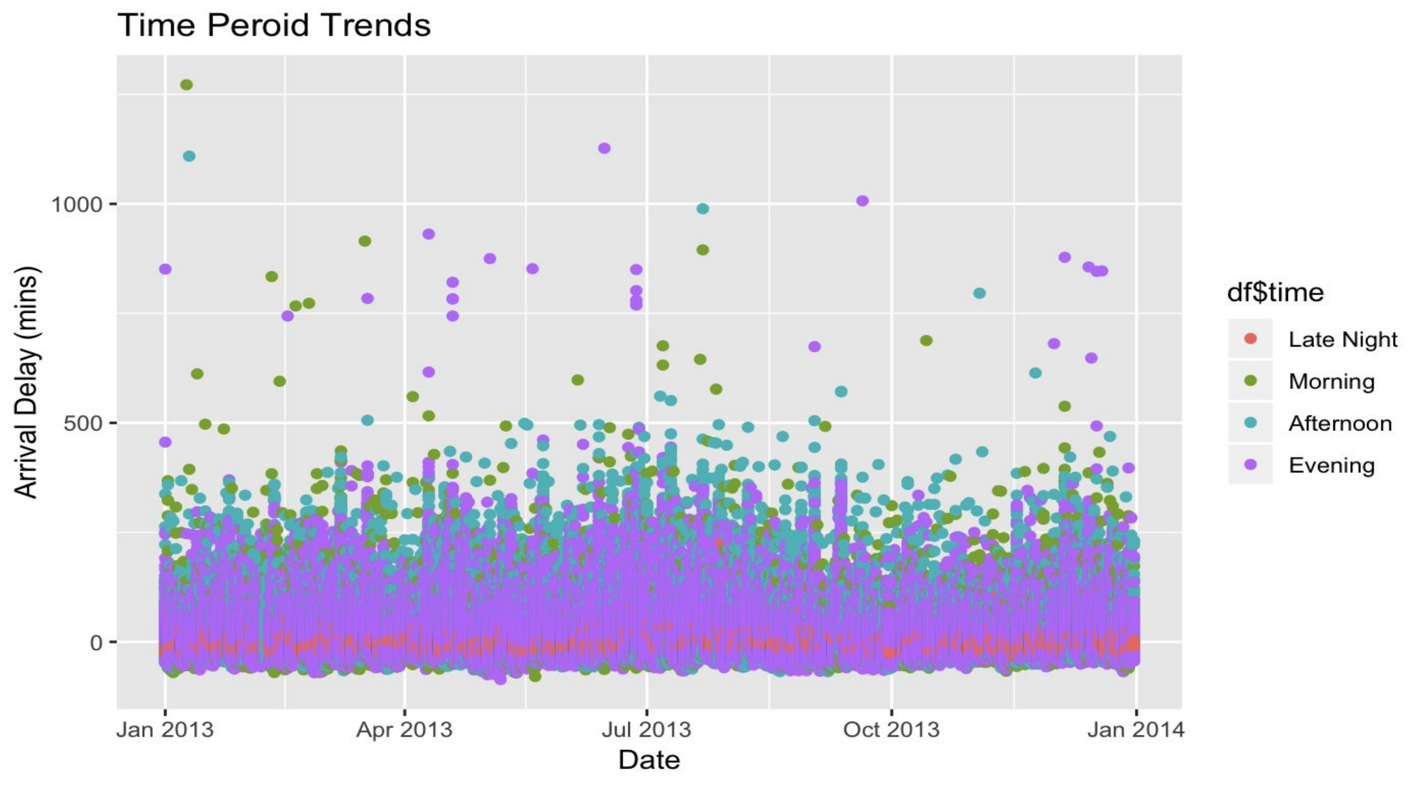
Different Time Periods

- Scheduled Departure Time
  - Morning: 5am-11am
  - Afternoon: 11am-5pm
  - Evening: 5pm-11pm
  - Late Night: 11pm-5am
- Arrival Delay
  - On Time: arr_delay<=0
  - Not On Time: arr_delay>0, arr_delay is NA

Different time periods

fl.sched_dep_time <int>	fl.arr_delay <dbl>	fl.carrier <fctr>	time <fctr>	ontime <chr>	date <date>
515	11	UA	Morning	Not On Time	2013-01-01
529	20	UA	Morning	Not On Time	2013-01-01
540	33	AA	Morning	Not On Time	2013-01-01
545	-18	B6	Morning	On Time	2013-01-01
600	-25	DL	Morning	On Time	2013-01-01
558	12	UA	Morning	Not On Time	2013-01-01
600	19	B6	Morning	Not On Time	2013-01-01
600	-14	EV	Morning	On Time	2013-01-01
600	-8	B6	Morning	On Time	2013-01-01
600	8	AA	Morning	Not On Time	2013-01-01

Different Time Periods



Chi-Squared Test for Independence

Null Hypothesis: On-time rate and day time periods are independent. On-time rate do not vary by day time periods.

Alternative Hypothesis: On-time rate and day time periods are dependent. On-time rate do vary by day time periods.

Chi-Squared Test for Independence

	Late Night	Morning	Afternoon	Evening
Not On Time	562	37586	54094	50192
On Time	500	77401	68672	47769

```
	Pearson's Chi-squared test

data:  tbl
X-squared = 7764.6, df = 3, p-value < 2.2e-16
```

Chi-Squared Test for Independence

Test the hypothesis that day time periods and on time rate are associated using a significance level of 0.05.

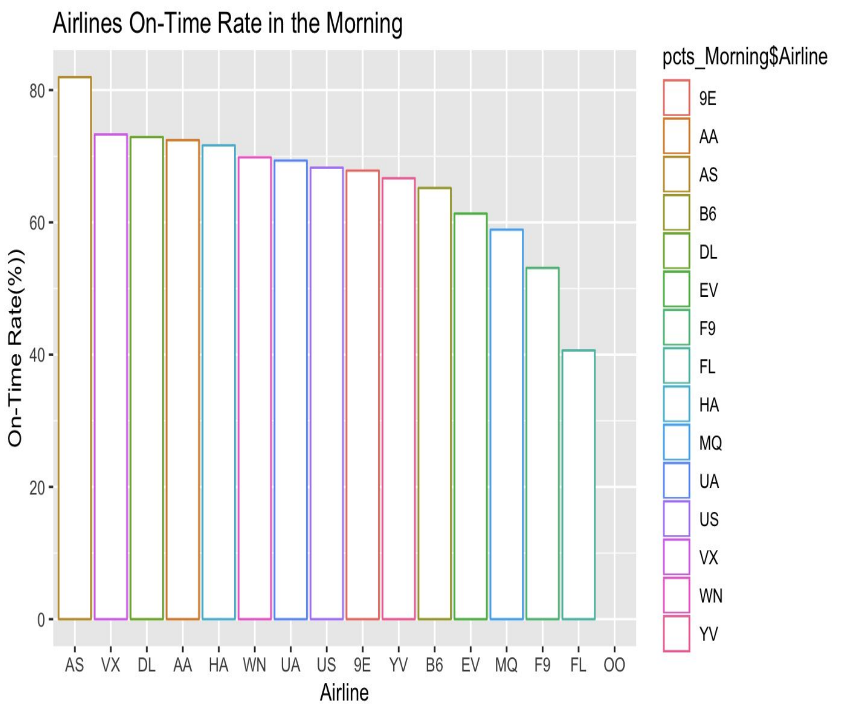
Since p-value is smaller than 0.05, we reject the null hypothesis.

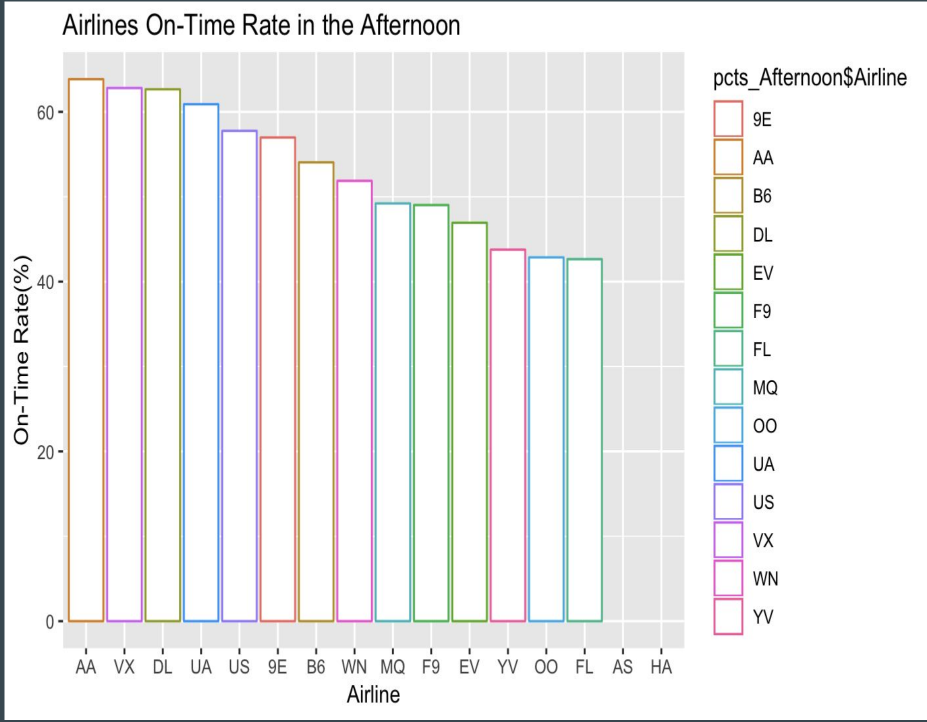
Thus, there is enough evidence to conclude that there is a significant relationship between on time rate and day time periods.

Conditional Probabilities

$$P(OnTime|Morning) = 77401/114987 = 0.6731$$
$$P(OnTime|Afternoon) = 68672/122766 = 0.5594$$
$$P(OnTime|Evening) = 47769/97961 = 0.4876$$
$$P(OnTime|LateNight) = 500/1062 = 0.4708$$

Conditional Probabilities

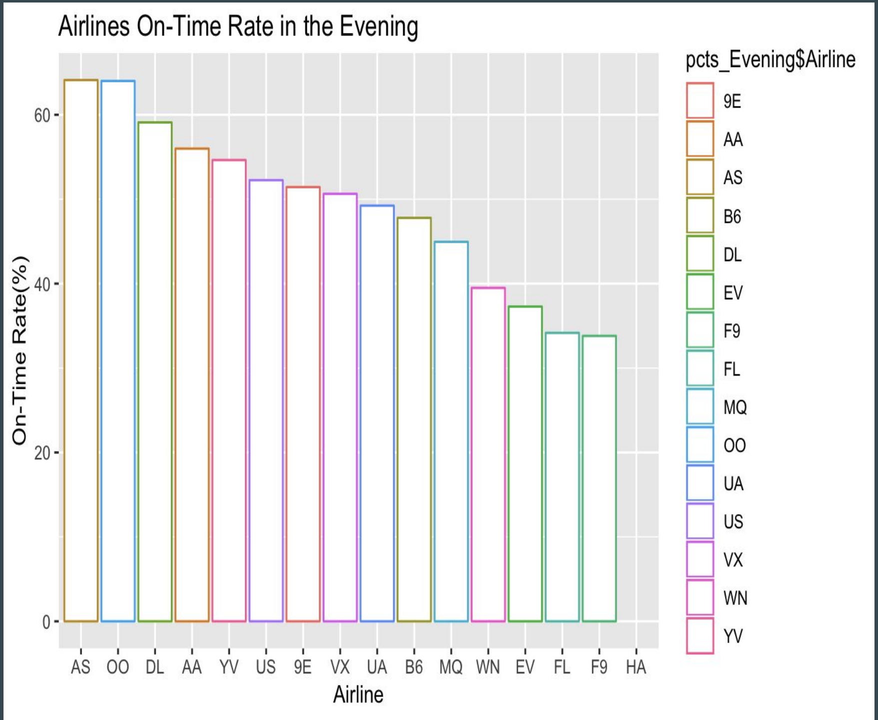


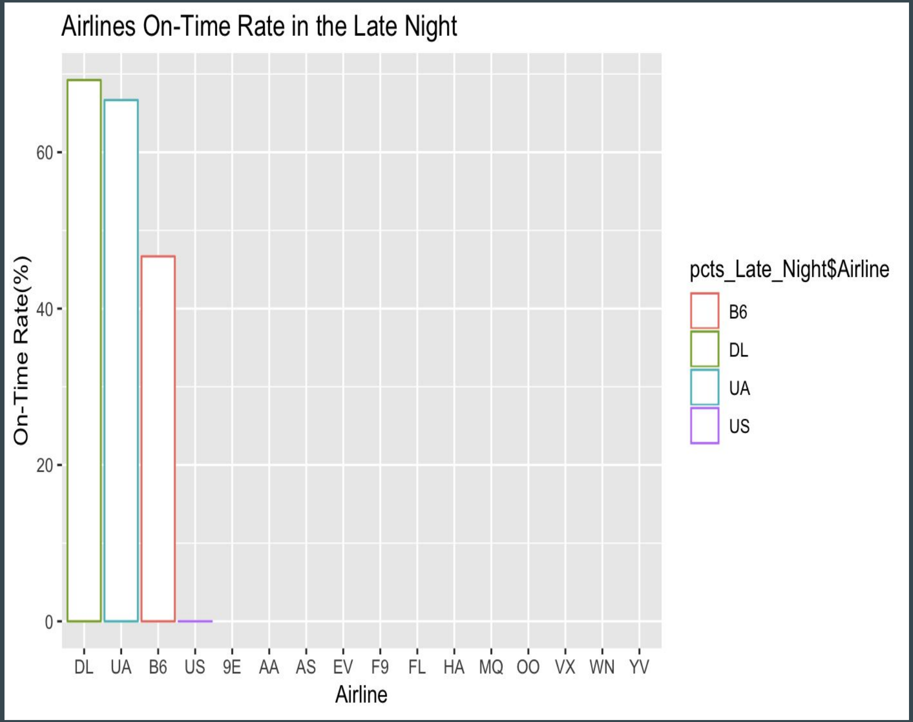


$P\ (\ OnTime \mid Morning)$

$P\ (\ OnTime \mid Afternoon)$

Conditional Probabilities



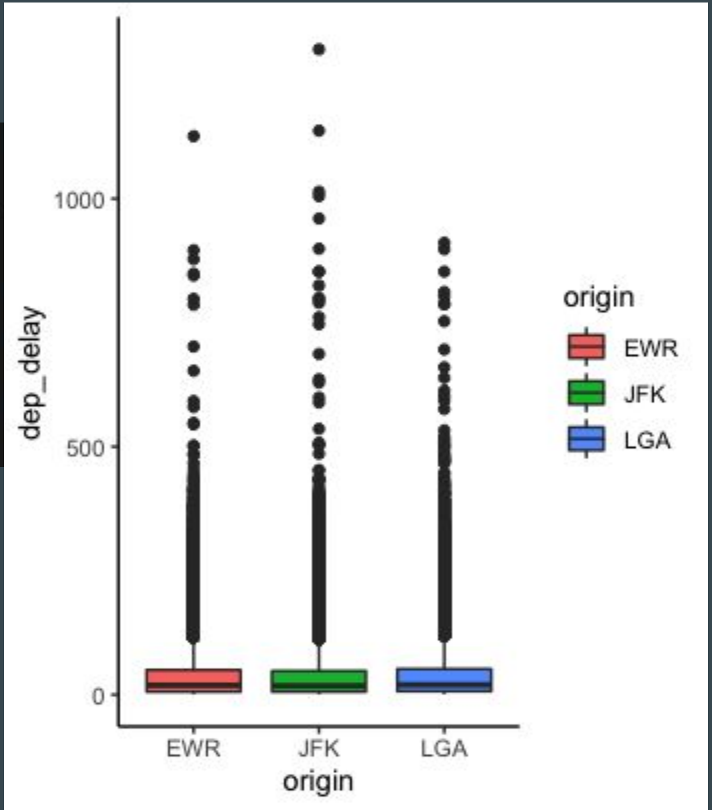


$p(OnTime \mid Evening)$

$p(OnTime \mid LateNight)$

Airports

	origin	count_origin	mean_time	sd_time
	<chr>	<int>	<dbl>	<dbl>
1	EWR	52414	38.8	52.5
2	JFK	41833	37.9	53.2
3	LGA	33498	41.5	57.7



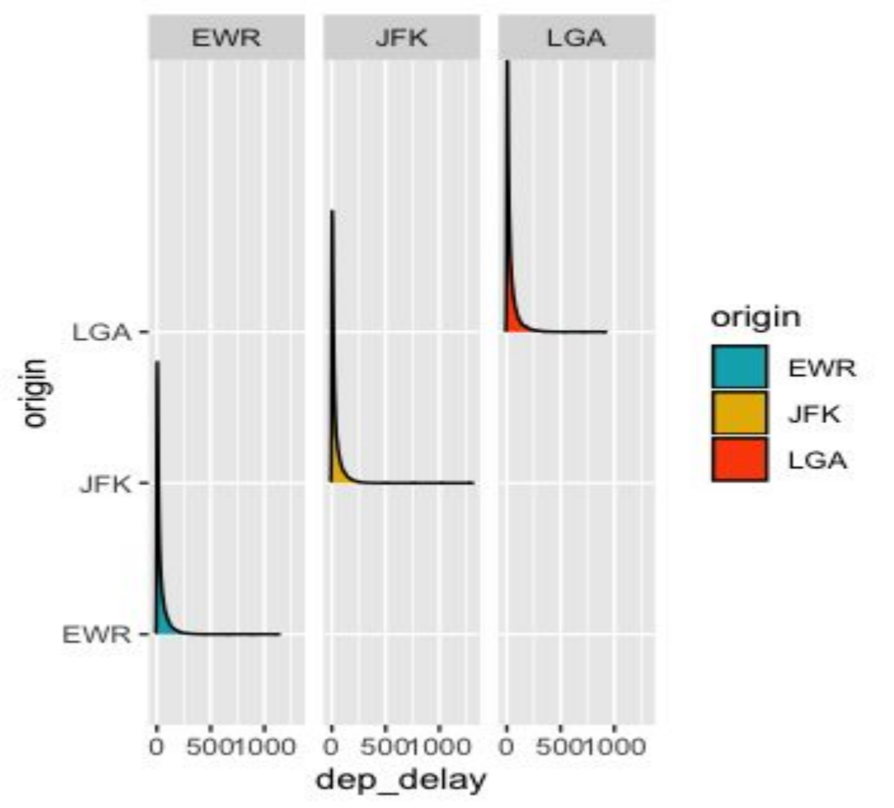
ANOVA on origin variable

```
              Df    Sum Sq Mean Sq F value Pr(>F)
origin         2    263912  131956   45.04 <2e-16 ***
Residuals 127742 374218364    2929
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

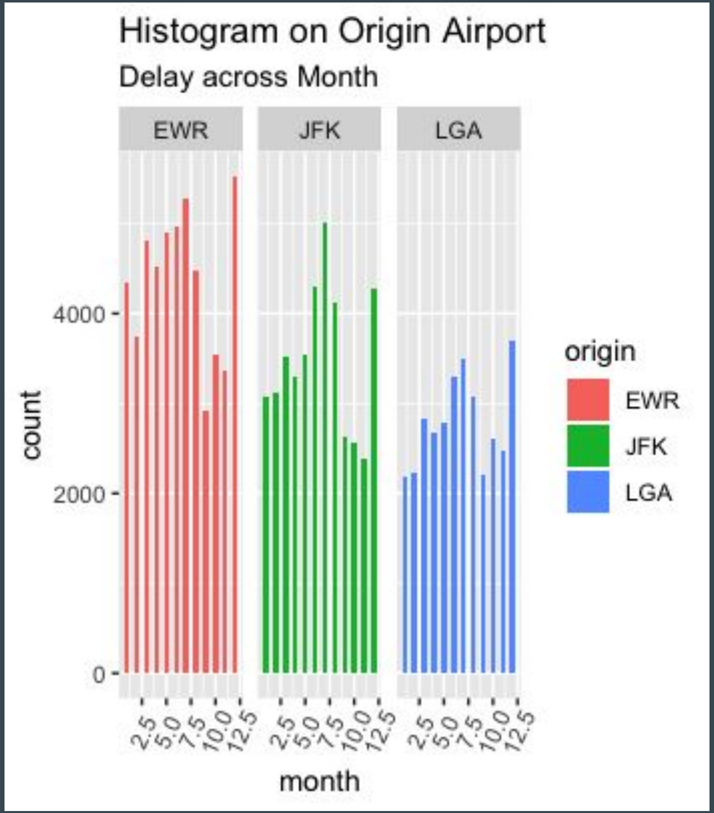
Pairwise ANOVA

```
              diff       lwr        upr     p adj
JFK-EWR -0.9540362 -1.785701 -0.1223712 0.0196443
LGA-EWR  2.7105178  1.823175  3.5978610 0.0000000
LGA-JFK  3.6645540  2.734484  4.5946243 0.0000000
```

Distribution of delay on origin airports



Delay across Month



Two way ANOVA on month and origin airport

```
> summary(anova_two_way)
                Df    Sum Sq Mean Sq F value Pr(>F)    
origin           2    263912  131956   45.08 <2e-16 ***
month            1    309451  309451  105.72 <2e-16 ***
Residuals   127741 373908913    2927                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Origin & Destination

Noticeable:

To ORD: Pick JFK

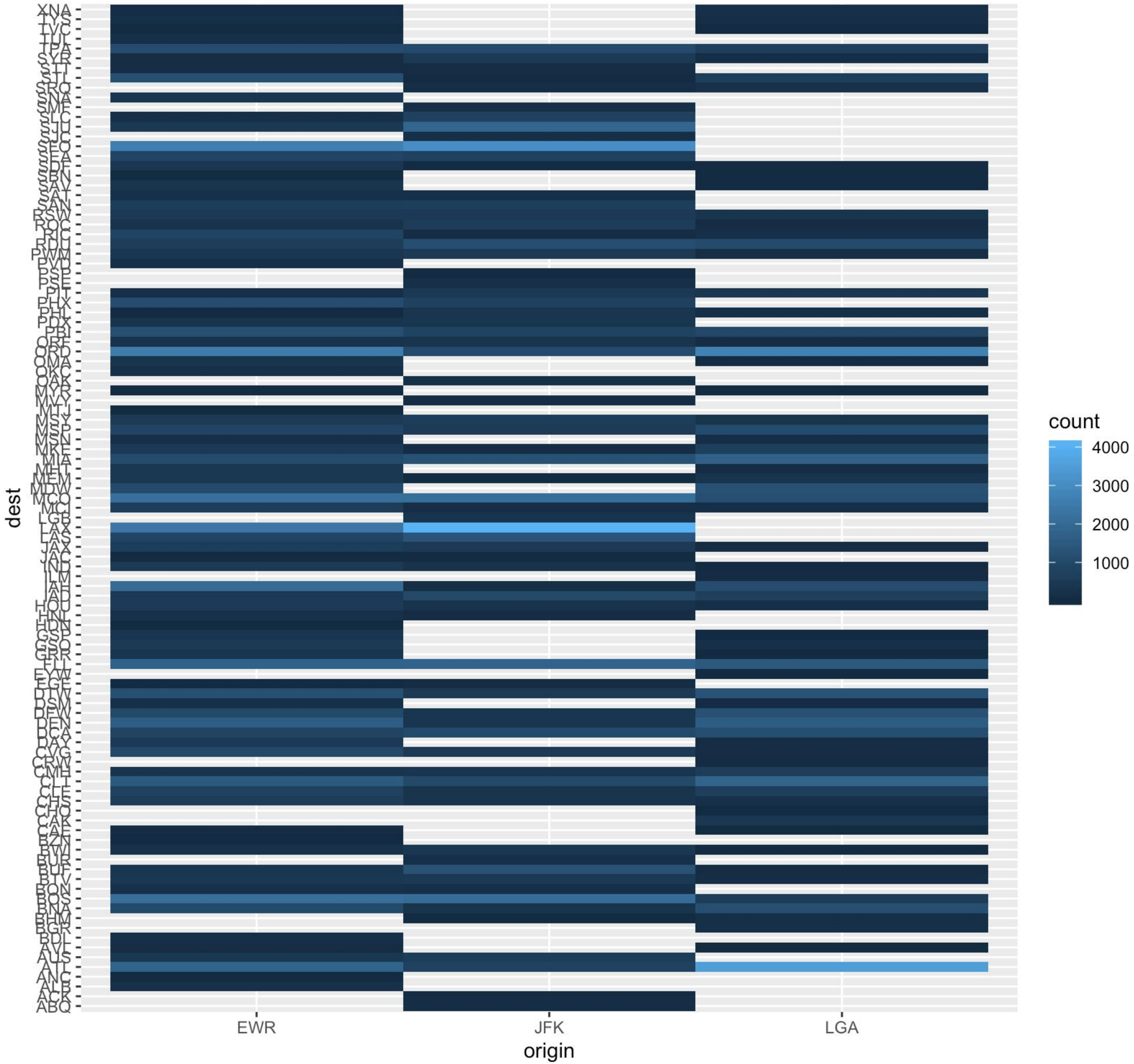
To LAX: Pick EWR

To LGB: Pick EWR

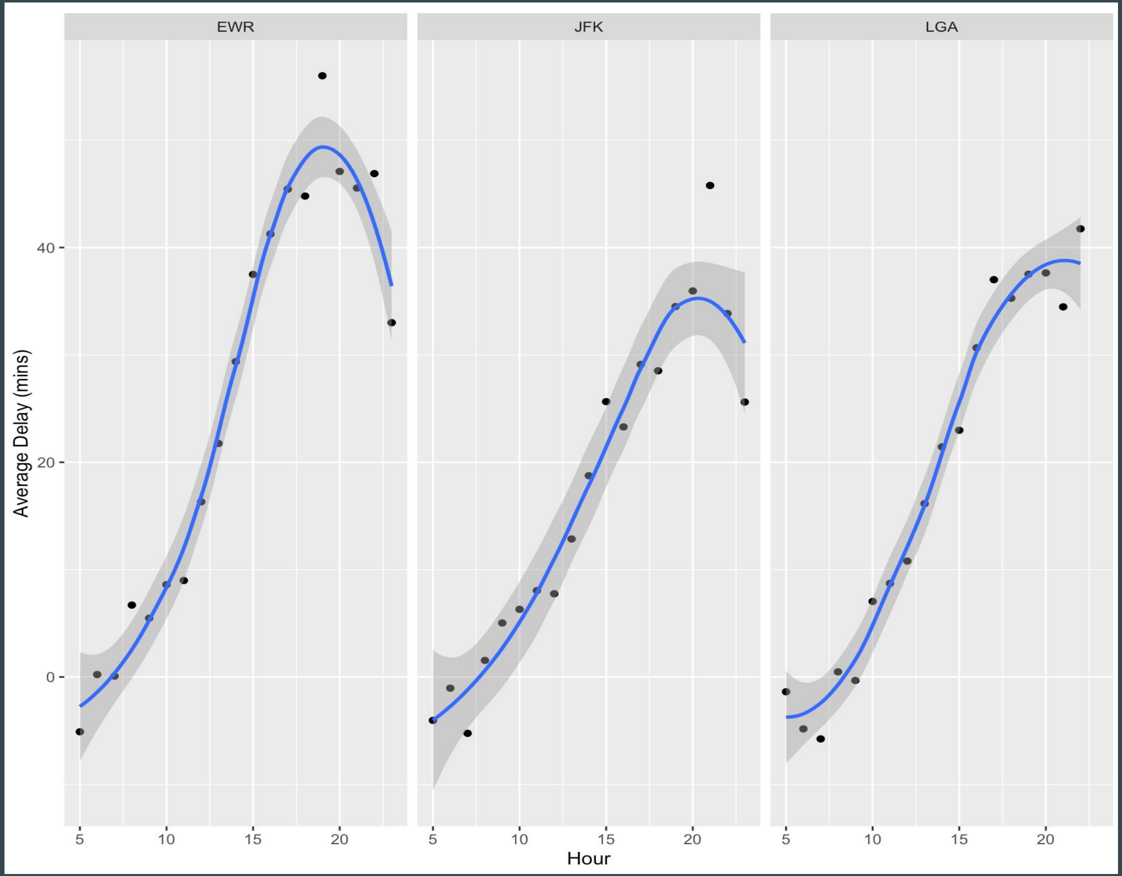
To ATL: Pick JFK

To SLC: Pick EWR

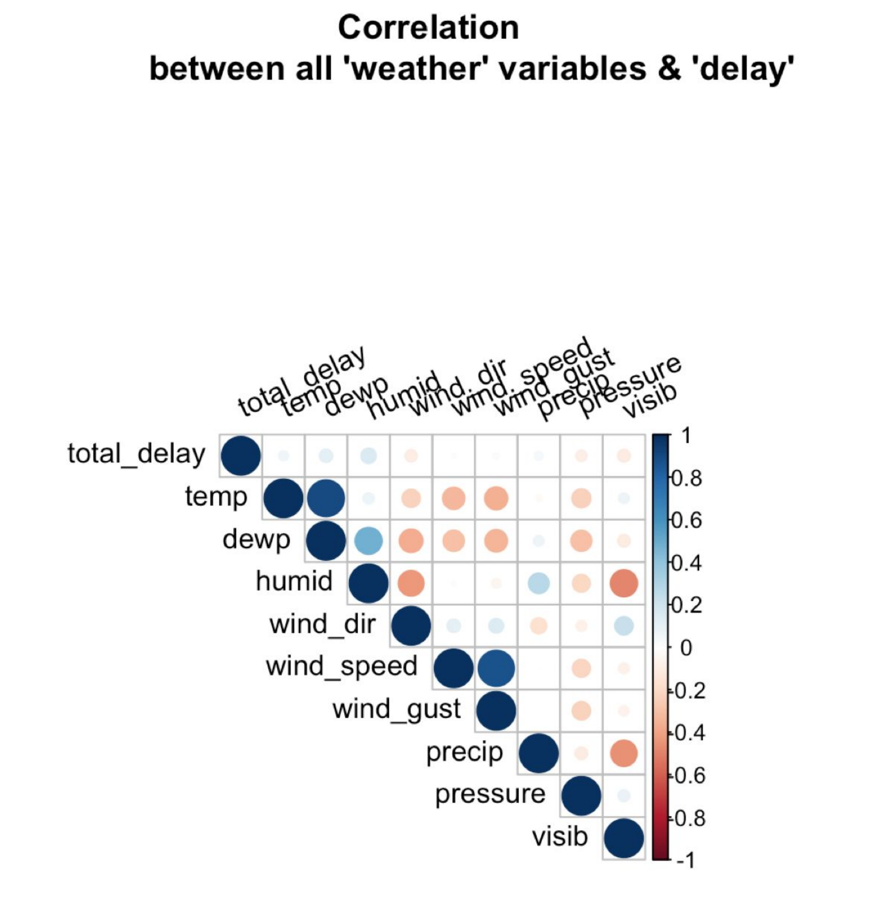
TO BOS: Pick LGA



Origin airport by hour



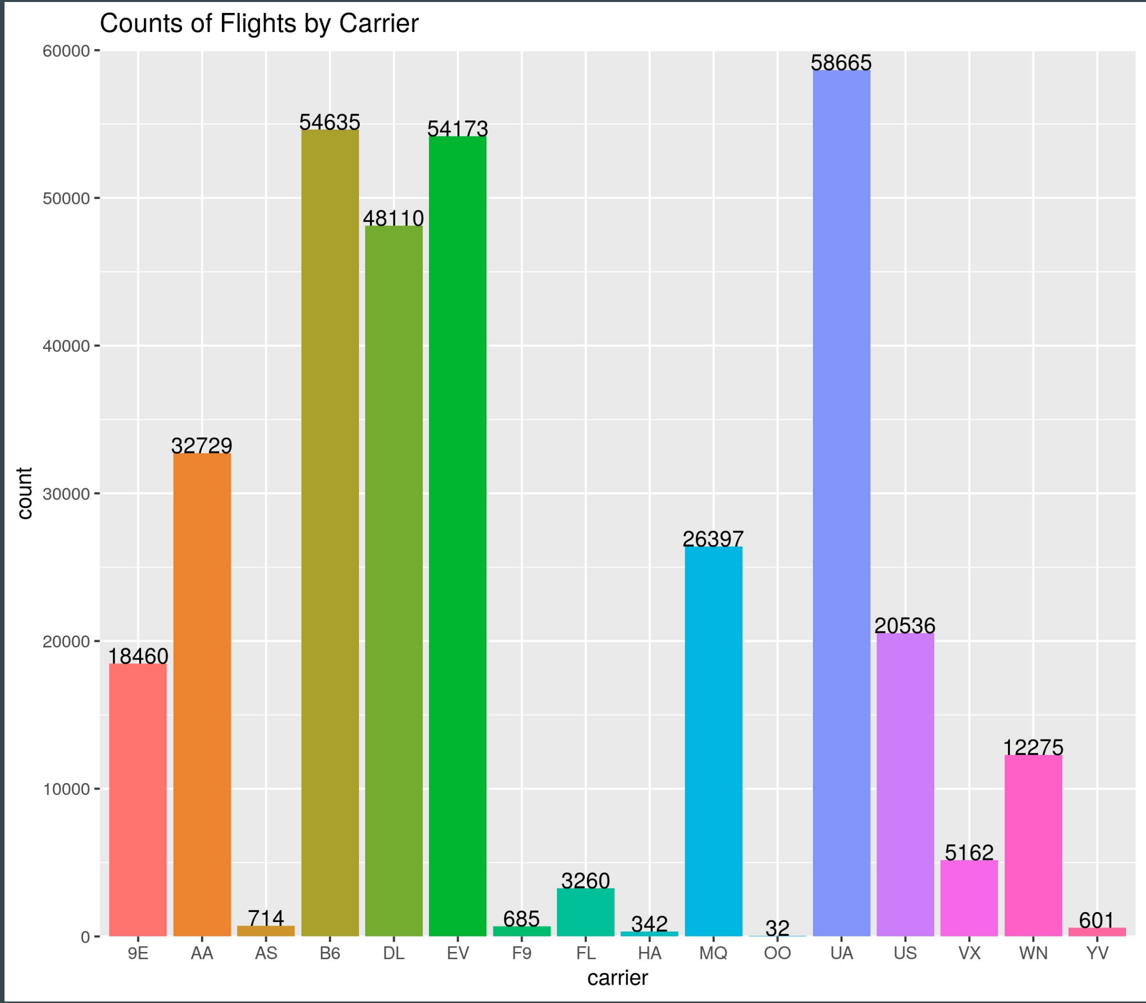
Correlation on Weather Var



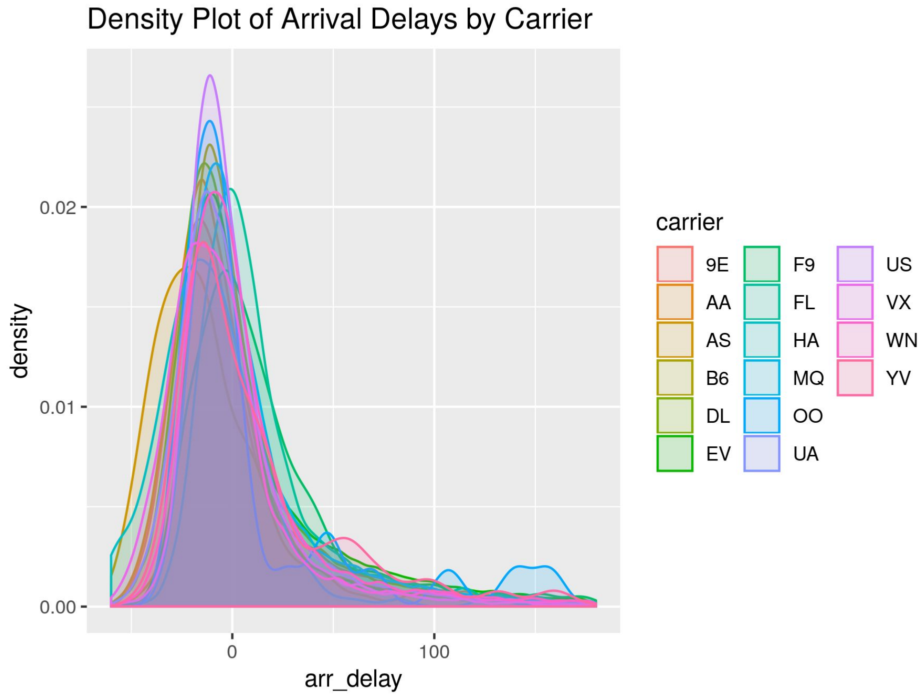
Carrier Analysis

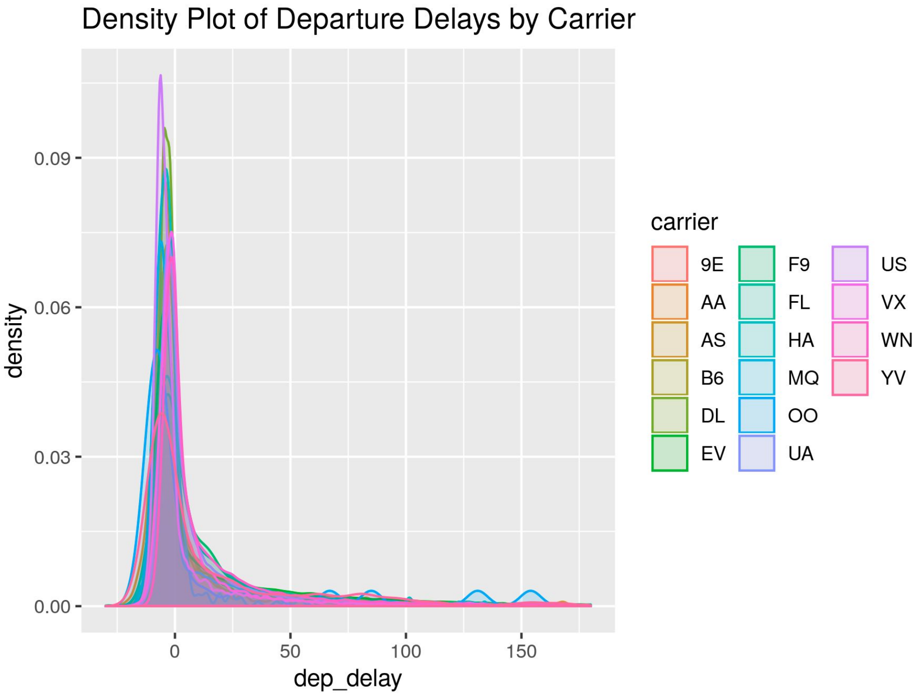
Drop carriers OO and F9

carrier <chr>	prop.delay <dbl>
FL	0.5079755
F9	0.5036496
YV	0.4592346
EV	0.4434866
MQ	0.4190628
WN	0.3724644
B6	0.3696898
9E	0.3693391
OO	0.3437500
UA	0.3257479



Distribution of Delays





Chi Squared Test for Independence Departing Delay by Carrier

- Null hypothesis: There is no relationship between airline carriers and departure delays
- Alternate hypothesis: There is a relationship between airline carriers and departure delays
- X-squared = 13680, df = 65, p-value < 2.2e-16
- We reject the null hypothesis
- Same p-value occurs for arrival delays

	early	ontime	5-30min	30-60min	1-2hr	>2hr
9E	10314	1178	2601	1275	1181	745
AA	21842	2844	3747	1523	1275	716
AS	484	69	93	24	22	17
B6	32677	4209	8789	3825	2944	1605
DL	32472	3653	6578	2317	1555	1083
EV	28132	3295	8133	4762	4370	2416
F9	341	63	147	57	39	34
FL	1528	415	689	233	161	149
MQ	17071	1106	3174	1715	1374	597
UA	30657	8272	11303	3770	2438	1342
US	15069	1204	1953	846	524	235
VX	2900	763	872	220	181	180
WN	5509	1754	2761	966	606	448
YV	312	35	83	36	55	23

Prediction - Linear regression

	Dependent variable:				
	arr_delay				
	(1)	(2)	(3)	(4)	(5)
temp	0.137***	0.038***	0.030***	0.030***	0.104***
	(0.004)	(0.007)	(0.007)	(0.007)	(0.004)
wind_speed	0.636***	0.433***	0.507***	0.509***	0.522***
	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)
visib	−2.811***	−3.087***	−3.129***	−3.146***	−3.124***
	(0.042)	(0.041)	(0.041)	(0.041)	(0.040)
precip	103.194***	89.925***	86.999***	87.629***	90.994***
	(2.760)	(2.715)	(2.713)	(2.694)	(2.685)
originJFK			−5.705***	−4.411***	−3.377***
			(0.187)	(0.253)	(0.355)
originLGA			−3.355***	−1.807***	−3.064***
			(0.187)	(0.231)	(0.348)
distance					0.062***
					(0.019)
Constant	17.748***	13.082***	15.684***	17.627***	−114.615***
	(0.471)	(1.104)	(1.106)	(1.169)	(33.674)
Time FE?	No	Yes	Yes	Yes	Yes
Airline FE?	No	No	No	Yes	Yes
Destination FE?	No	No	No	No	Yes
Observations	325,356	325,356	325,356	325,356	325,356
R^2	0.031	0.070	0.072	0.085	0.088
Adjusted R^2	0.031	0.070	0.072	0.085	0.088
Note:					*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:				
	dep_delay				
	(1)	(2)	(3)	(4)	(5)
temp	0.173***	0.104***	0.098***	0.098***	0.143***
	(0.004)	(0.007)	(0.007)	(0.007)	(0.004)
wind_speed	0.453***	0.257***	0.319***	0.320***	0.341***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
visib	−1.740***	−1.999***	−2.024***	−2.034***	−2.031***
	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)
precip	81.137***	70.120***	67.791***	68.092***	70.676***
	(2.489)	(2.441)	(2.439)	(2.428)	(2.422)
originJFK			−4.712***	−2.064***	−1.675***
			(0.168)	(0.228)	(0.320)
originLGA			−4.591***	−1.367***	−2.327***
			(0.168)	(0.208)	(0.314)
distance					0.056***
					(0.017)
Constant	13.522***	9.404***	11.687***	14.571***	−101.354***
	(0.425)	(0.993)	(0.994)	(1.053)	(30.377)
Time FE?	No	Yes	Yes	Yes	Yes
Airline FE?	No	No	No	Yes	Yes
Destination FE?	No	No	No	No	Yes
Observations	325,356	325,356	325,356	325,356	325,356
R^2	0.021	0.066	0.069	0.078	0.078
Adjusted R^2	0.021	0.066	0.069	0.077	0.078
Note:					*p<0.1; **p<0.05; ***p<0.01

Prediction - Probit Model

Probit model is a type of regression where the dependent variable can take only two values

- Mark the cancellation flight and delay flight as ‘1’
- Otherwise, mark as ‘0’

Red line - true random given data breakdown

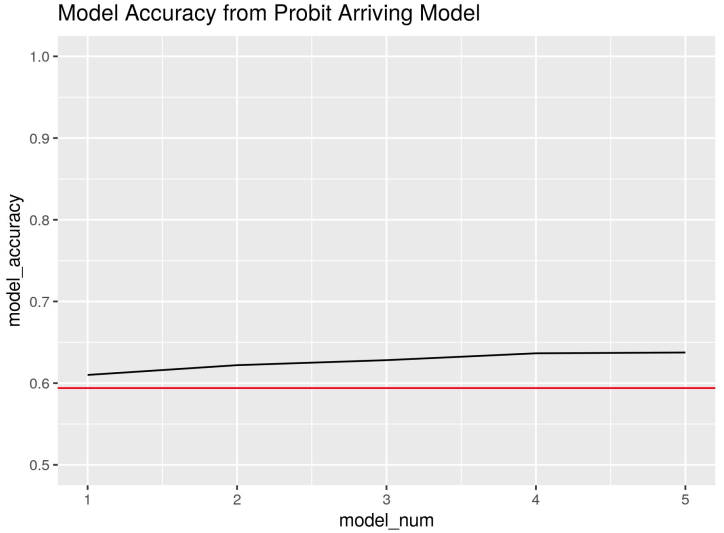
Model 1: weather data

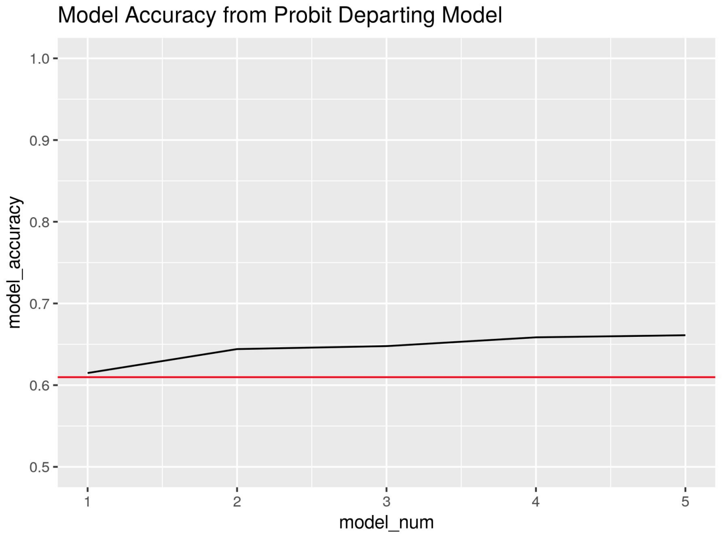
Model 2: model-1 + time/day of week/quarter

Model 3: model-2 + origin-airport

Model 4: model-3 + airline-carrier

Model 5: model-4 + destination-airport





Prediction (Departure Delay) - Logistic Regression

- Logistic Regression
  - classification algorithm
  - Maximum Likelihood Estimation - coefficients
- Target variables: Departure Delay
  - Delay - 1
  - Ontime - 0
- Independent variables:
  - Distance, departure airport, carrier, time period
  - Weather :
    - wind_speed, precipitation, pressure, visibility
    - Exclude: dewpoint, wind_dir, wind_gust

Prediction - Logistic Regression 1

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)      1.977e+01  7.505e-01  26.342  < 2e-16 ***
originJFK       -1.666e-01  1.667e-02  -9.999  < 2e-16 ***
originLGA       -1.295e-01  1.488e-02  -8.702  < 2e-16 ***
carrierAA       -5.201e-01  2.758e-02 -18.856  < 2e-16 ***
carrierAS       -8.505e-01  1.203e-01  -7.067 1.58e-12 ***
carrierB6       -1.521e-01  2.350e-02  -6.473 9.63e-11 ***
carrierDL       -5.661e-01  2.561e-02 -22.100  < 2e-16 ***
carrierEV        2.684e-01  2.591e-02  10.358  < 2e-16 ***
carrierF9        3.843e-02  1.066e-01   0.361 0.718402    
carrierFL        2.126e-01  5.187e-02   4.098 4.16e-05 ***
carrierHA       -9.386e-01  2.346e-01  -4.000 6.33e-05 ***
carrierMQ       -2.536e-01  2.725e-02  -9.307  < 2e-16 ***
carrierOO       -6.548e-01  4.253e-01  -1.540 0.123595    
carrierUA       -1.317e-01  2.764e-02  -4.763 1.91e-06 ***
carrierUS       -6.957e-01  3.140e-02 -22.159  < 2e-16 ***
carrierVX       -1.842e-01  4.742e-02  -3.884 0.000103 ***
carrierWN        2.718e-01  3.337e-02   8.144 3.83e-16 ***
carrierYV       -2.528e-02  1.052e-01  -0.240 0.810096    
distance         4.922e-05  8.531e-06   5.770 7.93e-09 ***
temp             2.187e-03  2.878e-04   7.600 2.95e-14 ***
humid            1.309e-02  3.409e-04  38.406  < 2e-16 ***
wind_speed       2.267e-02  9.881e-04  22.944  < 2e-16 ***
precip           3.122e+00  3.954e-01   7.897 2.85e-15 ***
pressure        -2.078e-02  7.263e-04 -28.614  < 2e-16 ***
visib           -2.265e-02  3.796e-03  -5.967 2.42e-09 ***
labelEvening     3.959e-01  1.150e-02  34.430  < 2e-16 ***
labelLate Night -4.690e-03  8.066e-02  -0.058 0.953632    
labelmorning    -9.694e-01  1.367e-02 -70.933  < 2e-16 ***
---
```

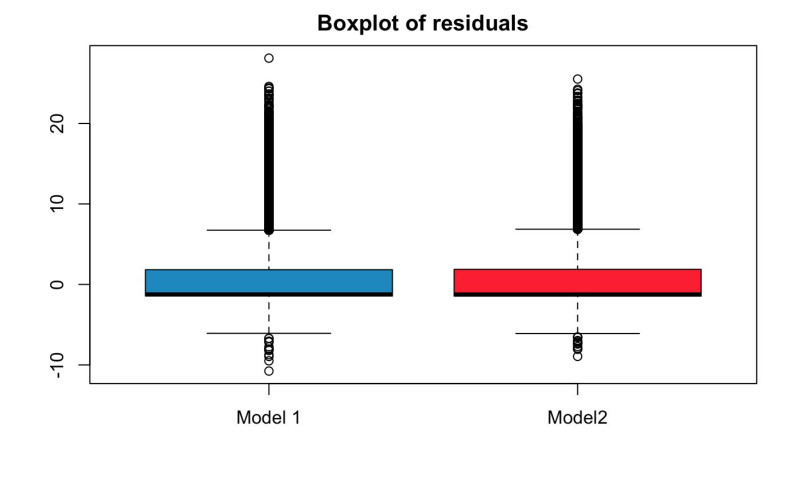
- P-value:

carrierF9	0.718402
carrierOO	0.123595
carrierYV	0.810096
labelLate Night	0.953632

- R square = 0.19
- Accuracy = 71.1%
- Improving model:
    - Add ‘Month’

Prediction - Logistic Regression 2

	Model 1	Model 2
AIC	252942	186927
Accuracy	71.1%	72.8%
R^2	0.19	0.21



- AIC:
  - Provides a method for assessing the quality of your model through comparison of related models.
  - Model 2 is the parsimonious model
- Accuracy
  - $$\frac{TP+TN}{TP+TN+FP+FN}$$
  - Model 2 has better performance
- Boxplot
  - Model2 has smaller outlier range.

Prediction - Naive Bayes

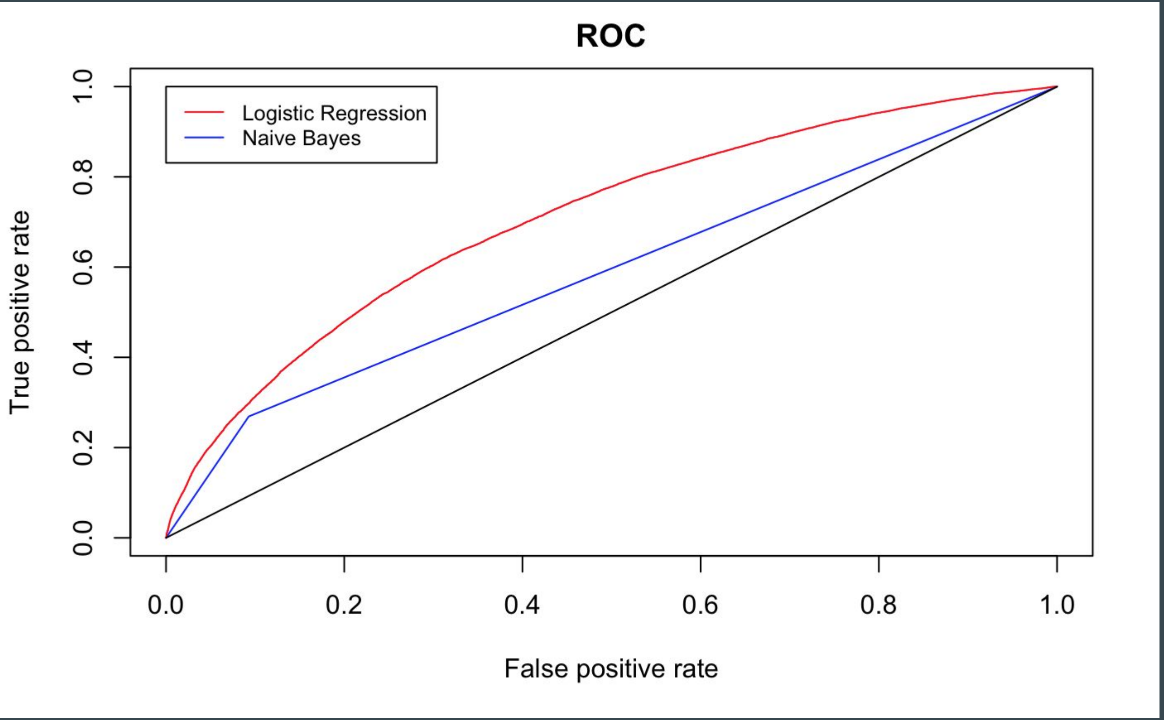
- Naive Bayes
  - classification algorithm
  - $$P(delay \& conditions)$$
    $$= P(x_1 \mid delay) \ldots P(x_n \mid delay) P(delay)$$
- Data processing
  - Binning all continuous variables to four quartiles

Prediction - Naive Bayes

- Result:
  - Confusion Matrix

	0	1
0	39260	13818
1	4019	5086

  - Accuracy : 71.3%
- ROC plot
  - Logistic regression has a better performance.



Summary

Time Period: Departing on morning has the highest probability of getting on time.

Origin Airport: LGA has the highest delay and EWR has the lowest.

Weather: Wind direction and wind speed do not have any correlation with delay.

Carrier: Delays are statistically different based on the carrier.

Prediction Model: Probit Regression 66%; Logistic Regression 72.8%; Naive Bayes 71.3%