On Time Flight Performance: SFO Airport

Dijana Obralic

2016-11-19

Introduction

The dataset contains data 15962 flights from San Francisco airport for August 2016. In total there were 3931 departure delayed flights.

This report analyzes the most common factors that cause departure delay, whether there is a signfificant difference between the mentioend airlines and a regression model for prediction probability of future delayed flights.

Airlines that are analyzed in the report are: American Airlines (AA), Alaska (AS), JetBlue (B6), Delta (DL), United Airlines (UA), SkyWest (OO), Virgin America (VX), Southwest (WN), Frontier (F9) and Hawaiian (HA).

Descriptive Analysis

The most frequent reason for a delayed flight is late aircraft. On average, flights delayed due to late aircraft arrival are 52.2 minutes. The least frequent reason for delay is security. The data bellow suggests also that many times there are more than one reason causing flight departure delay because the total number of delayed flights is 3931 and summing all reasons individally exceeds 3931.

The least predictable time of delay is carrier issues. It has the largest range of 1192 mins (about 19 hrs) of delay time and largest standard deviation (58.6 min) which suggests that the data is farther away from the mean.

```
##
## ----------------------------------------------------------------------
##                    Carrier   Weather   NAS    Security   LateAircraft
## ----------------- --------- --------- ------ ---------- --------------
##      nbr.val         1770       55      1101      4           2267
##
##      nbr.null          0         0       0        0           0
##
##       nbr.na           0         0       0        0           0
##
##        min             1         1       1        11          1
##
##        max           1193       252     232       37          683
##
##       range          1192       251     231       26          682
##
##        sum           65922     1947    25600     85         118340
##
```

```
##       median          20        18        11      18.5          38
##
##         mean       37.24      35.4     23.25     21.25        52.2
##
##      SE.mean       1.393     6.576    0.9523     5.721       1.068
##
##  CI.mean.0.95       2.732     13.18     1.869     18.21       2.095
##
##          var        3434      2379     998.5     130.9        2586
##
##      std.dev        58.6     48.77      31.6     11.44       50.86
##
##     coef.var       1.573     1.378     1.359    0.5384      0.9742
## ------------------------------------------------------------------------
##
## Table: Reasons for departure delay
```

Is there a difference between the airlines: ANOVA

From table bellow:Froniter Airline has the highest percentage of delayed flights (35.7 %) and the longest delay time (69.8 min).

Table 1: Information about airines

Carrier	DelayedFlights	TotalFlights	PercentDelayed	AverageDelay
AA	336	1454	23.11	69.32
AS	92	481	19.13	55.46
B6	150	500	30	61.46
DL	235	1357	17.32	81.72
F9	89	249	35.74	69.81
HA	2	62	3.226	42
OO	906	3071	29.5	61.11
UA	1363	5386	25.31	59.86
VX	402	1835	21.91	45.57
WN	356	1398	25.46	51.68

Let's check if there is a significant difference betweeen the airlines:
H0 There is no significant airlines between the airlines
Ha There is a significant difference between the airlines
Decision Reject Null H0 if p-value < 0.05

Table 2: Analysis of Variance Model

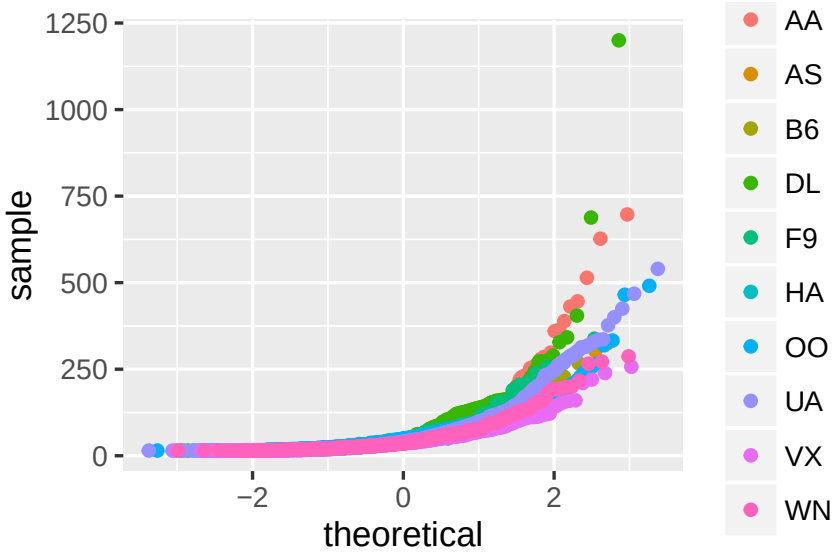
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Carrier	9	260770	28974	8.245	3.01e-12
Residuals	3921	13778320	3514	NA	NA

P-value is less than 0.05 therefore we reject the Null Hypothesis and conclude that there are significant differneces between the airlines reagarding departure delay time. The next step is to analyze where are the difference.

Based on the Tukey Kramer test [1] the major differences are (p-value < 0.05) are between the following airlines:

Virgin America (VX) and American Airlines (AA)
Southwest (WN) and American Airlines (AA)
Delta (DL) and Alaska Airlines (AS)
Delta (DL) and JetBlue (B6)
Skywest (OO) and Delta (DL)
United Airlines (UA) and Delta (DL)
Virgin America (VX) and Delta (DL)
Southwest (WN) and Delta (DL)
Virgin America (VX) and Froniter

[1] Detailed analysis of Tukey Kramer test in the Appendix



[2]

[2] The graph shows dispersion of departure flight delay time among different airlines

What is the probability that my flight is going to be delayed: Logistic regressions

The goal is to predict whether the flight is going to be delayed or not based on the delayed flights in the dataset, day of the week, distance and carrier.

The reason these variables are chosen is because we don't encounter many missing in which case we would need to deal with averages or meadian instead of raw data.

First model

The first model includes following variables:

DepDel15 - indicates whether the flight is delayed (if delay time > 15 min)

Carrier - converted into numerical binary values

Day of week - represented as numbers 1-7 (Monday to Sunday)

H0 Delayed flights are not significantly affected by Day of Week, Distance and Airline **Ha** Delayed flights are significantly affected by Day of Week, Distance and Airline

	Estimate	Std. Error	z value	Pr(>|z|)
Day	0.008198	0.0139	0.5896	0.5554
Distance	0.001541	0.008469	0.182	0.8556
AA	-0.1358	0.09503	-1.429	0.1529
AS	-0.3682	0.1312	-2.807	0.005007
B6	0.2184	0.1239	1.762	0.07802
DL	-0.3758	0.1021	-3.681	0.0002324
UA	0.4235	0.1108	3.822	0.0001321
OO	0.7734	0.1761	4.392	1.122e-05
VX	-0.2014	0.08607	-2.34	0.01929
(Intercept)	-1.109	0.08426	-13.16	1.433e-39

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	8285 on 7499 degrees of freedom
Residual deviance:	8191 on 7490 degrees of freedom

The data above suggests that we should drop Day, Distance and AA from the model because p-value < 0.05. This takes all of the variables from the model. Before making a conclusion, let's look at McFadden "pseudo" T-squared fit.

According to McFadden "pseudo" R-squared fit, values between 0.2 and 0.4 represent an excellent fit. McFadden R-squared value is between 0 and 1 and represents likelihood of the dependent variable happening.

```
##           llh       llhNull            G2
## -4.095274e+03 -4.142298e+03  9.404794e+01
##      McFadden          r2ML          r2CU
##  1.135215e-02  1.246143e-02  1.863640e-02
```

Second model

MacFadden R-squared value < 0 which suggests that this is not a good fit.

Let's see what happens if we only drop Day of Week:

```
##           llh       llhNull            G2
## -4.095274e+03 -4.142298e+03  9.404794e+01
##      McFadden          r2ML          r2CU
##  1.135215e-02  1.246143e-02  1.863640e-02
```

Still, McFadden "pseudo" R-squared fit < 0.

Therefore we can reject the Alternative Hypothesis and accept the Null Hypothesis and conclude that Day of Week, Distance and Airline do not significantly affect delayed time.

Next step would be to obtain additional data about weather and see if it significantly affects the departure flight delay.

Appendix

Tukey Kramer Test

```
## Warning in pander.default(SFO_flights.tukey):
## No pander.method for "TukeyHSD", reverting
## to default.No pander.method for "multicomp",
## reverting to default.
```

- **Carrier**:

	diff	lwr	upr	p adj
AS-AA	-13.86	-35.95	8.215	0.6078
B6-AA	-7.861	-26.29	10.57	0.9416
DL-AA	12.4	-3.555	28.36	0.2906
F9-AA	0.4876	-21.88	22.86	1
HA-AA	-27.32	-160.4	105.8	0.9997
OO-AA	-8.211	-20.2	3.775	0.4788
UA-AA	-9.458	-20.89	1.972	0.2087
VX-AA	-23.75	-37.62	-9.879	2.833e-06
WN-AA	-17.64	-31.91	-3.369	0.003683
B6-AS	6.003	-18.85	30.85	0.999
DL-AS	26.27	3.189	49.34	0.01182
F9-AS	14.35	-13.55	42.25	0.8344
HA-AS	-13.46	-147.6	120.7	1
OO-AS	5.654	-14.88	26.19	0.9973

	diff	lwr	upr	p adj
UA-AS	4.407	-15.81	24.62	0.9996
VX-AS	-9.884	-31.57	11.8	0.9137
WN-AS	-3.777	-25.72	18.17	0.9999
DL-B6	20.26	0.6526	39.87	0.03615
F9-B6	8.349	-16.76	33.46	0.9889
HA-B6	-19.46	-153	114.1	1
OO-B6	-0.3496	-16.89	16.19	1
UA-B6	-1.596	-17.74	14.55	1
VX-B6	-15.89	-33.84	2.066	0.1362
WN-B6	-9.78	-28.05	8.486	0.7987
F9-DL	-11.91	-35.27	11.44	0.8412
HA-DL	-39.72	-173	93.53	0.995
OO-DL	-20.61	-34.35	-6.876	9.162e-05
UA-DL	-21.86	-35.11	-8.606	8.325e-06
VX-DL	-36.15	-51.56	-20.74	2.073e-08

	diff	lwr	upr	p adj
WN-DL	-30.04	-45.82	-14.27	1.011e-07
HA-F9	-27.81	-162	106.4	0.9997
OO-F9	-8.699	-29.54	12.15	0.949
UA-F9	-9.945	-30.48	10.58	0.8788
VX-F9	-24.24	-46.22	-2.254	0.01755
WN-F9	-18.13	-40.37	4.109	0.2271
OO-HA	19.11	-113.7	151.9	1
UA-HA	17.86	-114.9	150.6	1
VX-HA	3.572	-129.4	136.6	1
WN-HA	9.68	-123.4	142.7	1
UA-OO	-1.247	-9.29	6.797	1
VX-OO	-15.54	-26.78	-4.293	0.0005289
WN-OO	-9.431	-21.17	2.307	0.2457
VX-UA	-14.29	-24.94	-3.641	0.0009208

	diff	lwr	upr	p adj
WN-UA	-8.184	-19.35	2.985	0.3764
WN-VX	6.108	-7.549	19.76	0.9227

Desination Info for Regression Analysis

Dest	DestCityName	Distance	DistanceGroup
MRY	Monterey, CA	77	1
SMF	Sacramento, CA	86	1
FAT	Fresno, CA	158	1
SBP	San Luis Obispo, CA	190	1
RNO	Reno, NV	192	1
RDD	Redding, CA	199	1
SMX	Santa Maria, CA	216	1
BFL	Bakersfield, CA	238	1
ACV	Arcata/Eureka, CA	250	2
SBA	Santa Barbara, CA	262	2
BUR	Burbank, CA	326	2
MFR	Medford, OR	329	2
LAX	Los Angeles, CA	337	2
LGB	Long Beach, CA	354	2
ONT	Ontario, CA	363	2
SNA	Santa Ana, CA	372	2
OTH	North Bend/Coos Bay, OR	412	2
LAS	Las Vegas, NV	414	2
PSP	Palm Springs, CA	421	2
SAN	San Diego, CA	447	2
EUG	Eugene, OR	451	2
RDM	Bend/Redmond, OR	462	2
BOI	Boise, ID	522	3
PDX	Portland, OR	550	3
SUN	Sun Valley/Hailey/Ketchum, ID	587	3
SLC	Salt Lake City, UT	599	3

Dest	DestCityName	Distance	DistanceGroup
PSC	Pasco/Kennewick/Richland, WA	620	3
PHX	Phoenix, AZ	651	3
SEA	Seattle, WA	679	3
JAC	Jackson, WY	737	3
TUS	Tucson, AZ	751	4
MSO	Missoula, MT	769	4
BZN	Bozeman, MT	807	4
ASE	Aspen, CO	848	4
ABQ	Albuquerque, NM	896	4
DEN	Denver, CO	967	4
OKC	Oklahoma City, OK	1384	6
DFW	Dallas/Fort Worth, TX	1464	6
DAL	Dallas, TX	1476	6
SAT	San Antonio, TX	1482	6
MCI	Kansas City, MO	1499	6
AUS	Austin, TX	1504	7
XNA	Fayetteville, AR	1550	7
MSP	Minneapolis, MN	1589	7
IAH	Houston, TX	1635	7
STL	St. Louis, MO	1735	7
MKE	Milwaukee, WI	1845	8
ORD	Chicago, IL	1846	8
MDW	Chicago, IL	1855	8
IND	Indianapolis, IN	1943	8
BNA	Nashville, TN	1969	8
ANC	Anchorage, AK	2018	9
CVG	Cincinnati, OH	2036	9
DTW	Detroit, MI	2079	9
ATL	Atlanta, GA	2139	9
CLE	Cleveland, OH	2161	9
PIT	Pittsburgh, PA	2254	10
CLT	Charlotte, NC	2296	10
OGG	Kahului, HI	2338	10
KOA	Kona, HI	2367	10
HNL	Honolulu, HI	2398	10
RDU	Raleigh/Durham, NC	2400	10
IAD	Washington, DC	2419	10
DCA	Washington, DC	2442	10

Dest	DestCityName	Distance	DistanceGroup
MCO	Orlando, FL	2446	10
LIH	Lihue, HI	2447	10
BWI	Baltimore, MD	2457	10
PHL	Philadelphia, PA	2521	11
EWR	Newark, NJ	2565	11
FLL	Fort Lauderdale, FL	2584	11
MIA	Miami, FL	2585	11
JFK	New York, NY	2586	11
BOS	Boston, MA	2704	11

References