Pierce Renio

STP 429

Schneider

18 September 2022

Lab 1

Abstract/Executive Summary

Airports try to limit the number and length of delays as much as possible. Using the flight data from the Bureau of Transportation Statistics, we are able to look into what could be the main source of arrival delays for Alaska airports. There were many observations in the data but the filtered data consisted of 12.2% of the original 6291 observations. With many variables in the data, there were many possible influences on the independent variable, arrival delays. Most of the individual variables were skewed right meaning that delays were mostly limited to little or none. However as the data skewed right, many of the delays increased to outlier potential. Looking at the correlation between the independent variables and the dependent variable discovered that only 3 variables were linked: departure delay, carrier delay, and late aircraft delay. Departure delay had the greatest correlation with arrival delay so this independent variable became the variable that the model used with the dependent variable. The model is:

ARR_DELAY = 2.86406 + 0.89595(DEP_DELAY)

The model for the data had great R^2 value and the p-values for each of the coefficient parameters were less than 0.0001 therefore the model was effective at predicting arrival delays.

Data

The data consists of all the flights that arrived and departed from Alaska in June 2022. There are 6291 observations in the data and 24 variables were used in the analysis. Day of the month was chosen to see if certain days/weeks had more influence on the arrival delays. Destination was used as we needed to see which airport in Alaska the most flights. We will analyze this airport exlusively. Departure delay was picked to see if these delays also affected arrival delays. Taxi in was picked to see if the airports had a longer time on the runway than other sources of the delay. Arrival delay will be our independent variable; we are analyzing what sources have more of an influence on arrival delays. Arrival time, Arrival Delay 15, Arrival Time Block, and Arrival delay group are variables that also describe Arrival delay. Canceled is to see if there were many canceled flights on the data as a whole. Air time and distance are two similar variables that could potentially affect arrival delays. Carrier delay, Weather delay, NAS delay, security delay, and late aircraft delay are all variables that describe some sort of delay that potentially affect the independent variable.

Methodology

For this data analysis, we will use exploratory data analysis first to analyze each of the variable distributions and look for outliers. We can use the PROC UNIVARIATE and HISTOGRAM statements to accomplish this task. This will give us statistics and histograms for each of the independent variables. Next, we need to determine which independent variables are strongly correlated with the dependent variable. We can use the PROC CORR statement to look at each variable and how they correlate with ARR_DELAY. We can look more into detail with the PROC SGPLOT statement as it will provide scatterplots and statistics on the data. Finally, we will perform simple linear regression with the strongest correlated independent variable.Use the

PROC REG statement, we will be able to determine what our model is for the data and how effective it is in determining ARR_DELAY.

Results

The SAS software was used to complete this project. In order to find a solution to the research question "Is there one variable more influential in causing arrival delays for the Ted Stevens Anchorage International Airport (ANC) than others?", we must first look at the data that we are working with. Only 77 of the 6291 flights were canceled (1.22%) (Figure 10) Next we will filter the data. Figure 6 displays a table for the number of observations for each Destination. Anchorage (ANC) is the most observed in this table so we will filter out any city that is not Anchorage. Next, we will filter out any observations that do not have arrival delays greater than 0 so we do not have observations where delays do not exist. Figure 8 shows the code for Figure 9, which is the filtered data. The filtered data consists of 768 observations, which is approximately 12.2% of the original data (6291 observations). Next, looking at each independent variable we would like to test is important as gaining more information for each will shed light upon which variables we would actually like to test. Figure 12 displays the code for finding information of each independent variable. For the variable DAY_OF_MONTH (Figure 13), it seems almost normally distributed but slightly skewed left as more flights took place in the second half of the month for June. For the variable DEP_DELAY (Figure 14), most of the data had little to no delay for flights, most of the delays appear to be within 120 minutes but it does appear to be a few outliers (Figure 14) such as 342 minutes, 374 minutes, and 481 minutes. For the variable TAXI_IN (Figure 15), most of the flights took within 10 minutes to taxi into the airport, however there are a few outliers here that took over 30 minutes. For the variable

ARR_DELAY (Figure 16), most of the delays are within an hour, however as the data is heavily skewed right there exists some data that extends to over 200 minutes for arrival delays. For the variable AIR_TIME (Figure 17), the data is distributed oddly but no signs of outliers. For the variable DISTANCE (Figure 18), the data distribution seems to match the AIR_TIME variable which makes sense as the greater the distance for a flight the greater the time taken to travel in the air. For the variable CARRIER_DELAY (Figure 19), the data is strongly skewed right so there are a few potential outliers that extend past 300 minutes. For the variable WEATHER_DELAY (Figure 20), most of the delays are only a few minutes, but there does seem to be a few outliers that extend past 50 minutes or more. For the variable NAS_DELAY (Figure 21), the data is again strongly skewed right, but there appear to be a few potential outliers past 50 minutes. For the variable SECURITY_DELAY (Figure 22), most of the data is at little to no delay, but there appears to be a few outliers at 45 minutes. Lastly for the variable LATE_AIRCRAFT_DELAY (Figure 23), the data is heavily skewed right with several potential outliers. Next, we will need to look how each individual independent variable correlates with the dependent variable (ARR_DELAY) so we can determine which variable we will use for creating a model. We can do this using a PROC CORR statement in SAS (Figure 24). Looking at the table in Figure 25, we are able to see how each variable correlates to ARR_DELAY. It would be useless to use any arriving variables to compare to ARR_DELAY since it is just another way to display arriving delays so we will not be using ARR_DELAY_GROUP. One would think that other delays would be more influential on ARR_DELAY such as NAS_DELAY and SECURITY_DELAY but seeing that they have a negative correlation, their National Air System and security are efficient and rarely cause delays. One also would think that distance would have an influence on delay time but when analyzing the correlation between distance and arrival time/depart time (Figure 26 and Figure 27), one can identify a stronger correlation between DISTANCE and DEP_DELAY (r = 0.21140) than DISTANCE and ARR_DELAY (r = 0.11408).The variables that correlate the most with ARR_DELAY are DEP_DELAY (r = 0.95901), CARRIER_DELAY (r = 0.63562), and LATE_AIRCRAFT_DELAY (r = 0.60976). The scatter plots for each of these variables line up with ARR_TIME decently but the one that stands out is DEP_DELAY (Figure 29, 30, 31). We will look at our strongest correlated variable (DEP_DELAY) by using the PROC REG statement in SAS. Using the PROC REG statement shown in Figure 32, we can create a model for our data using the DEP_DELAY independent variable. The least squares line equation for our model is:

ARR_DELAY = 2.86406 + 0.89595(DEP_DELAY)

We can interpret this as “for every 1 minute of DEP_DELAY, the ARR_DELAY increases by 0.89595 minutes. Additionally, if there is no DEP_DELAY, the ARR_DELAY is expected to be 2.86406 minutes. Both of the parameter estimates have a p-value of less than 0.001 and the R^2 value for the model is 0.9197 meaning that 91.197% of the variation is due to the independent variable, thus the model is effective. The QQPlot in Figure 34 also lines up well.

Final Conclusions and Next Steps

It was unknown prior to this analysis which variable had a greater influence on the dependent variable, arrival delay. While it seemed like many variables could have had a greater influence, only the departure delay had the most significant effect on arrival delay. While it is surprising that those are extremely well correlated, this is understandable as flights usually are late to arrive if they are late to depart. If this analysis was completed again, more variables would be used such as the number of diverted airport landings as more stops could possibly lead to greater arrival delays. Additionally, many of the variables were not numeric as they were character variables or other types. Being able to use these variables in the analysis would be useful as adding more variables could be informative of correlations and relationships that were not known.

<u>Appendix</u>

```
/* Generated Code (IMPORT) */
/* Source File: T_ONTIME_REPORTING.csv */
/* Source Path: /home/u62122616/sasuser.v94 */
/* Code generated on: 9/17/22, 6:57 PM */

%web_drop_table(WORK.IMPORT1);


FILENAME REFFILE '/home/u62122616/sasuser.v94/T_ONTIME_REPORTING.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=WORK.ANCDDT;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.ANCDDT; RUN;


%web_open_table(WORK.ANCDDT);
```

Figure 1

```
PROC CONTENTS DATA=ANCDDT;
    RUN;
```

Figure 2

The CONTENTS Procedure

Data Set Name	WORK.ANCDDT	Observations	6291
Member Type	DATA	Variables	24
Engine	V9	Indexes	0
Created	09/17/2022 19:00:13	Observation Length	184
Last Modified	09/17/2022 19:00:13	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	9
First Data Page	1
Max Obs per Page	711
Obs in First Data Page	684
Number of Data Set Repairs	0
Filename	/saswork/SAS_workB71500001443_odaws01-usw2-2.oda.sas.com/SAS_workE84600001443_odaws01-usw2-2.oda.sas.com/ancddt.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	1140850758
Access Permission	rw-r--r--
Owner Name	u62122616
File Size	1MB
File Size (bytes)	1310720

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
18	AIR_TIME	Num	8	BEST12.	BEST32.
14	ARR_DEL15	Num	8	BEST12.	BEST32.
12	ARR_DELAY	Num	8	BEST12.	BEST32.
15	ARR_DELAY_GROUP	Num	8	BEST12.	BEST32.
13	ARR_DELAY_NEW	Num	8	BEST12.	BEST32.
11	ARR_TIME	Num	8	BEST12.	BEST32.
16	ARR_TIME_BLK	Char	9	$9.	$9.
17	CANCELLED	Num	8	BEST12.	BEST32.
20	CARRIER_DELAY	Num	8	BEST12.	BEST32.
10	CRS_ARR_TIME	Num	8	BEST12.	BEST32.
1	DAY_OF_MONTH	Num	8	BEST12.	BEST32.
7	DEP_DELAY	Num	8	BEST12.	BEST32.
4	DEST	Char	3	$3.	$3.
5	DEST_CITY_NAME	Char	17	$17.	$17.
6	DEST_STATE_ABR	Char	2	$2.	$2.
19	DISTANCE	Num	8	BEST12.	BEST32.
24	LATE_AIRCRAFT_DELAY	Num	8	BEST12.	BEST32.
22	NAS_DELAY	Num	8	BEST12.	BEST32.
2	OP_UNIQUE_CARRIER	Char	2	$2.	$2.
3	ORIGIN	Char	3	$3.	$3.
23	SECURITY_DELAY	Num	8	BEST12.	BEST32.
9	TAXI_IN	Num	8	BEST12.	BEST32.
21	WEATHER_DELAY	Num	8	BEST12.	BEST32.
8	WHEELS_ON	Num	8	BEST12.	BEST32.

Figure 3

```
PROC PRINT DATA=ANCDDT (OBS=10);
RUN;
```

Figure 4

Obs	DAY_OF_MONTH	OP_UNIQUE_CARRIER	ORIGIN	DEST	DEST_CITY_NAME	DEST_STATE_ABR	DEP_DELAY	WHEELS_ON	TAXI_IN	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
1	1	QX	FAI	ANC	Anchorage, AK	AK	-6	1254	5	1304	1259	-5
2	1	QX	FAI	ANC	Anchorage, AK	AK	-9	2042	6	2059	2048	-11
3	1	QX	ANC	SCC	Deadhorse, AK	AK	-7	656	2	715	658	-17
4	1	QX	FAI	ANC	Anchorage, AK	AK	-14	1825	3	1845	1828	-17
5	1	QX	SCC	ANC	Anchorage, AK	AK	-12	929	3	950	932	-18
6	1	QX	ANC	FAI	Fairbanks, AK	AK	-9	1859	3	1918	1902	-16
7	1	QX	ANC	FAI	Fairbanks, AK	AK	-5	1656	3	1704	1659	-5
8	1	QX	FAI	SEA	Seattle, WA	WA	-4	931	23	935	954	19
9	1	QX	FAI	ANC	Anchorage, AK	AK	-7	1415	4	1409	1419	10
10	1	QX	ANC	FAI	Fairbanks, AK	AK	-5	1546	2	1554	1548	-6

Obs	ARR_DELAY_NEW	ARR_DEL15	ARR_DELAY_GROUP	ARR_TIME_BLK	CANCELLED	AIR_TIME	DISTANCE	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY	SECURITY_DELAY	LATE_AIRCRAFT_DELAY
1	0	0	-1	1300-1359	0	46	261	.	.	.	.	.
2	0	0	-1	2000-2059	0	43	261	.	.	.	.	.
3	0	0	-2	0700-0759	0	86	626	.	.	.	.	.
4	0	0	-2	1800-1859	0	45	261	.	.	.	.	.
5	0	0	-2	0900-0959	0	89	626	.	.	.	.	.
6	0	0	-2	1900-1959	0	39	261	.	.	.	.	.
7	0	0	-1	1700-1759	0	47	261	.	.	.	.	.
8	19	1	1	0900-0959	0	206	1533	0	0	19	0	0
9	10	0	0	1400-1459	0	54	261	.	.	.	.	.
10	0	0	-1	1500-1559	0	40	261	.	.	.	.	.

Figure 5

```
proc freq data=ANCDDT;
    table DEST;
run;
```

Figure 6

The FREQ Procedure

DEST	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ADK	9	0.14	9	0.14
ADQ	101	1.61	110	1.75
AKN	48	0.76	158	2.51
ANC	2174	34.56	2332	37.07
ATL	30	0.48	2362	37.55
BET	60	0.95	2422	38.50
BRW	30	0.48	2452	38.98
CDV	60	0.95	2512	39.93
DEN	82	1.30	2594	41.23
DFW	57	0.91	2651	42.14
DLG	44	0.70	2695	42.84
EWR	28	0.45	2723	43.28
FAI	495	7.87	3218	51.15
GST	30	0.48	3248	51.63
IAH	28	0.45	3276	52.07
JNU	475	7.55	3751	59.62
KTN	249	3.96	4000	63.58
LAS	9	0.14	4009	63.73
LAX	30	0.48	4039	64.20
MSP	118	1.88	4157	66.08
OME	59	0.94	4216	67.02
ORD	169	2.69	4385	69.70
OTZ	60	0.95	4445	70.66
PDX	27	0.43	4472	71.09
PHX	17	0.27	4489	71.36
PSG	60	0.95	4549	72.31
SCC	35	0.56	4584	72.87
SEA	1330	21.14	5914	94.01
SFO	49	0.78	5963	94.79
SIT	175	2.78	6138	97.57
SLC	33	0.52	6171	98.09
WRG	60	0.95	6231	99.05
YAK	60	0.95	6291	100.00

Figure 7

```
data ANC2;
    set ANCDDT;

    if DEST = "ANC" and ARR_DELAY > 0 then output ANC2;

proc print data=ANC2 (obs=20);
    run;

proc contents data=anc2;
    run;
```

Figure 8

Obs	DAY_OF_MONTH	OP_UNIQUE_CARRIER	ORIGIN	DEST	DEST_CITY_NAME	DEST_STATE_ABR	DEP_DELAY	WHEELS_ON	TAXI_IN	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
1	1	QX	FAI	ANC	Anchorage, AK	AK	-7	1415	4	1409	1419	10
2	1	AA	DFW	ANC	Anchorage, AK	AK	71	2137	3	2043	2140	57
3	1	AS	ADQ	ANC	Anchorage, AK	AK	11	801	6	757	807	10
4	1	AS	BET	ANC	Anchorage, AK	AK	10	1409	4	1405	1413	8
5	1	AS	BRW	ANC	Anchorage, AK	AK	-8	1826	6	1830	1832	2
6	1	AS	JNU	ANC	Anchorage, AK	AK	146	1123	6	908	1129	141
7	1	AS	SEA	ANC	Anchorage, AK	AK	15	838	5	837	843	6
8	1	AS	SEA	ANC	Anchorage, AK	AK	27	2221	4	2209	2225	16
9	1	AS	SEA	ANC	Anchorage, AK	AK	78	2358	3	2309	1	52
10	1	AS	SEA	ANC	Anchorage, AK	AK	2	254	6	258	300	2
11	1	AS	ADK	ANC	Anchorage, AK	AK	22	1913	4	1844	1917	33
12	1	AS	FAI	ANC	Anchorage, AK	AK	-7	1145	7	1149	1152	3
13	2	AA	DFW	ANC	Anchorage, AK	AK	46	2054	3	2043	2057	14
14	2	AS	JNU	ANC	Anchorage, AK	AK	131	14	4	2213	18	125
15	2	AS	JNU	ANC	Anchorage, AK	AK	69	1005	4	908	1009	61
16	2	AS	SEA	ANC	Anchorage, AK	AK	11	113	5	116	118	2
17	2	AS	SEA	ANC	Anchorage, AK	AK	62	1947	5	1904	1952	48
18	2	AS	SEA	ANC	Anchorage, AK	AK	8	1418	6	1416	1424	8
19	2	AS	SEA	ANC	Anchorage, AK	AK	23	26	5	11	31	20
20	2	AS	FAI	ANC	Anchorage, AK	AK	47	1923	5	1844	1928	44

Figure 9

```
proc freq data=ANCDDT;
    table cancelled;
run;
```

Figure 10

The FREQ Procedure

CANCELLED	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6214	98.78	6214	98.78
1	77	1.22	6291	100.00

Figure 11

```
proc univariate data = ANC2;
    HISTOGRAM;
    run;
```

Figure 12

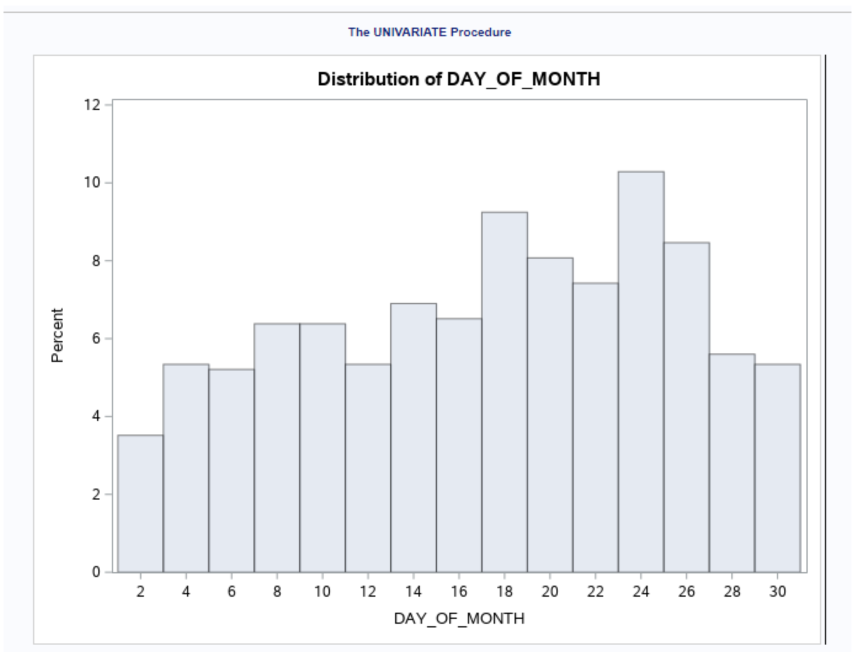

Figure 13

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-28	467	239	481
-19	313	265	480
-17	759	342	25
-16	506	374	323
-16	60	481	488

The UNIVARIATE Procedure

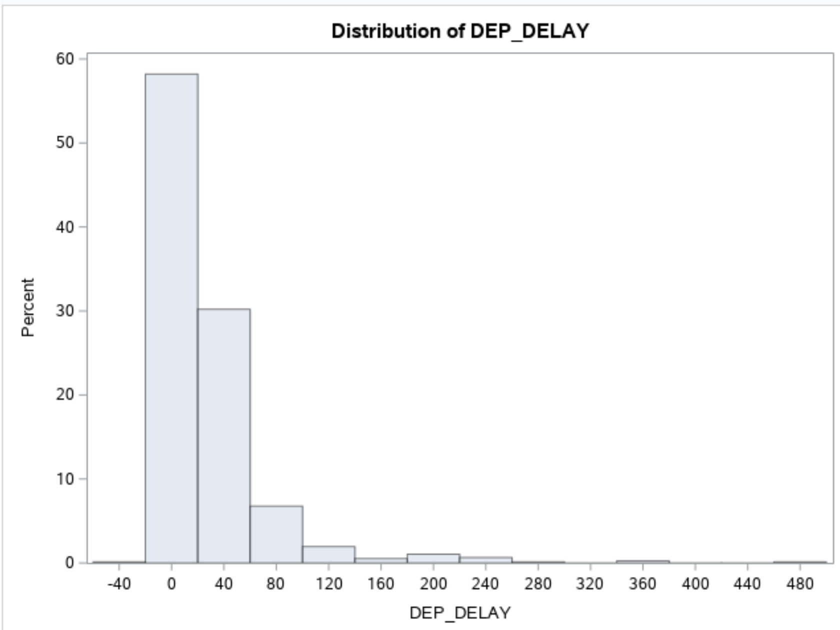

Figure 14

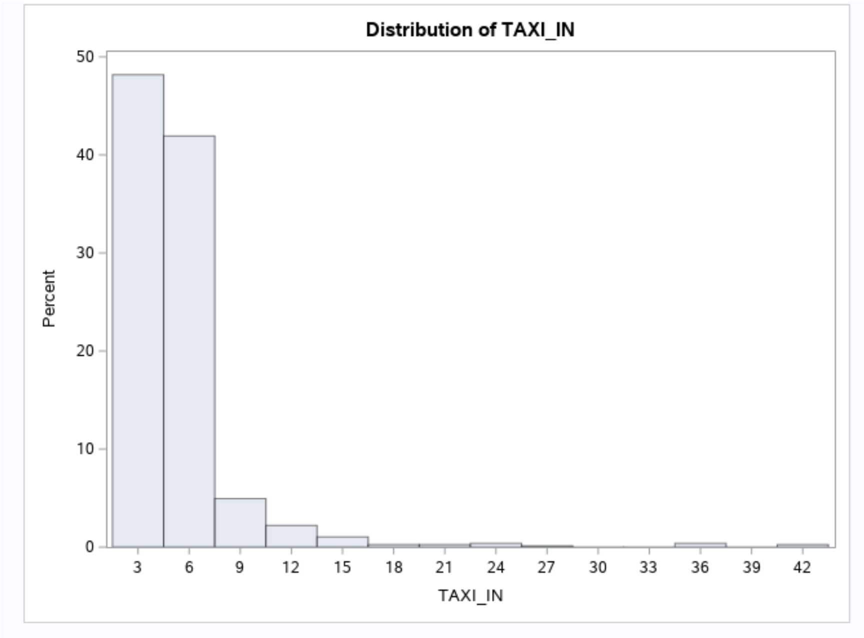

Figure 15

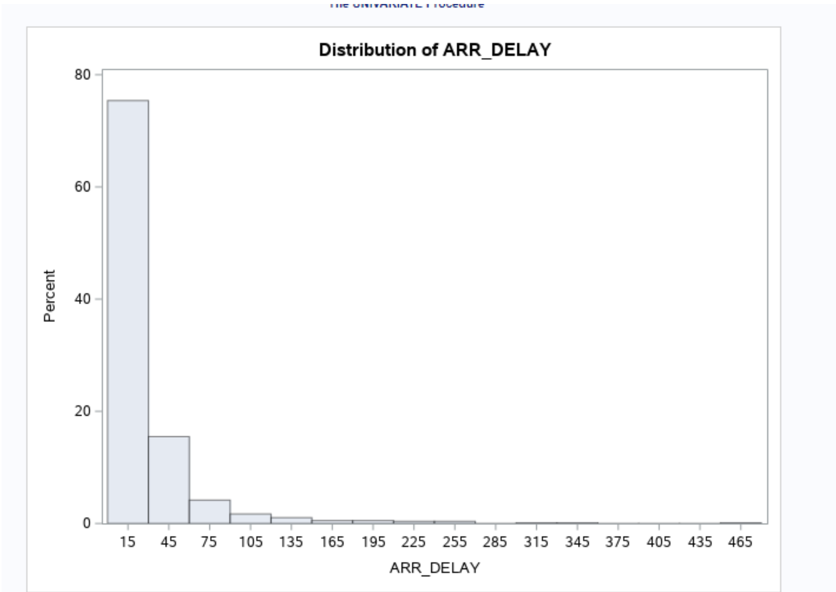

Figure 16

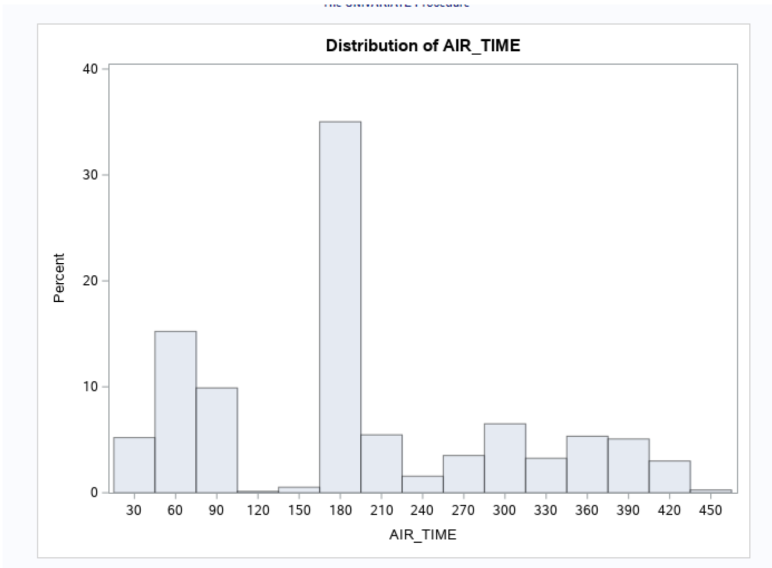

Figure 17

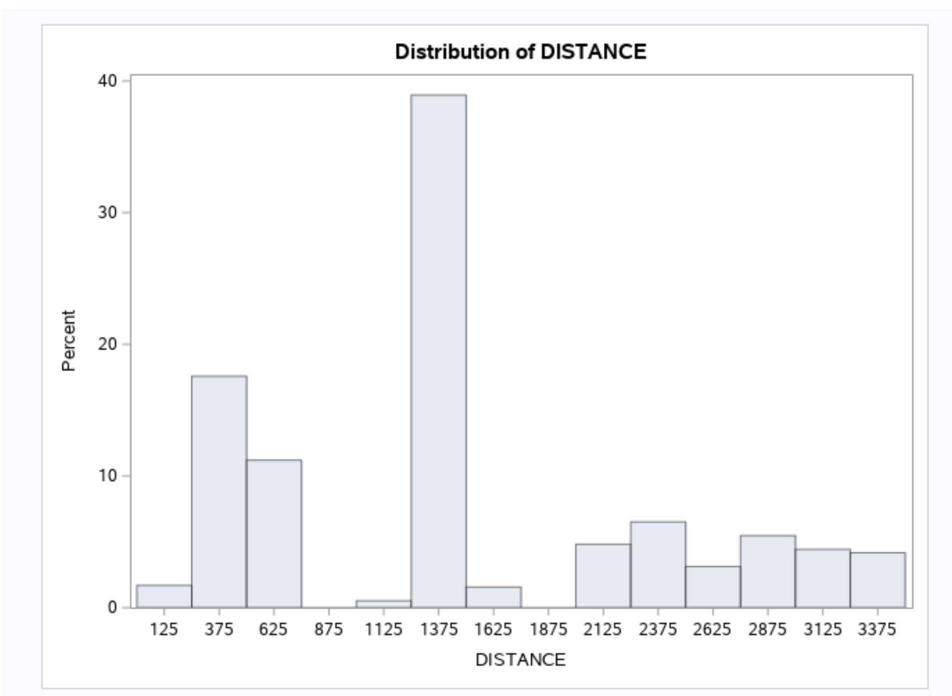

Figure 18

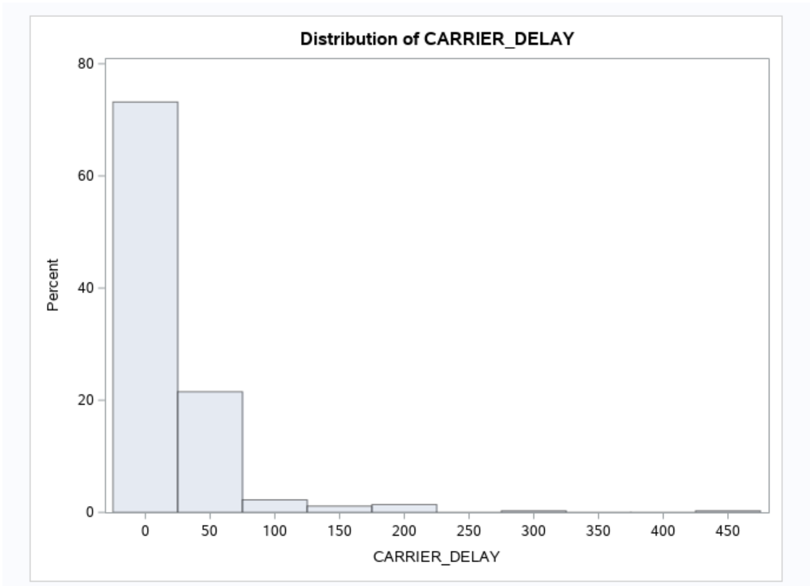

Figure 19

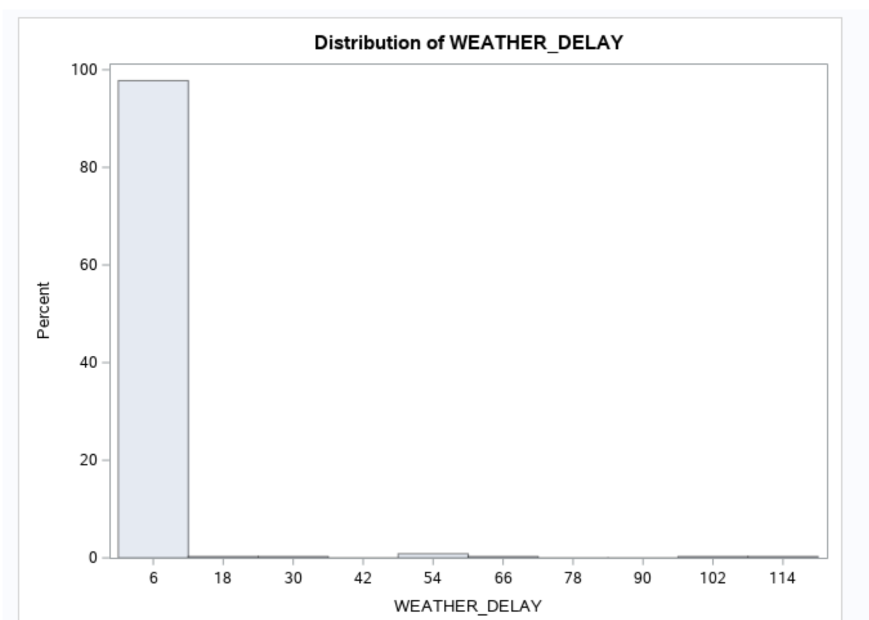

Figure 20

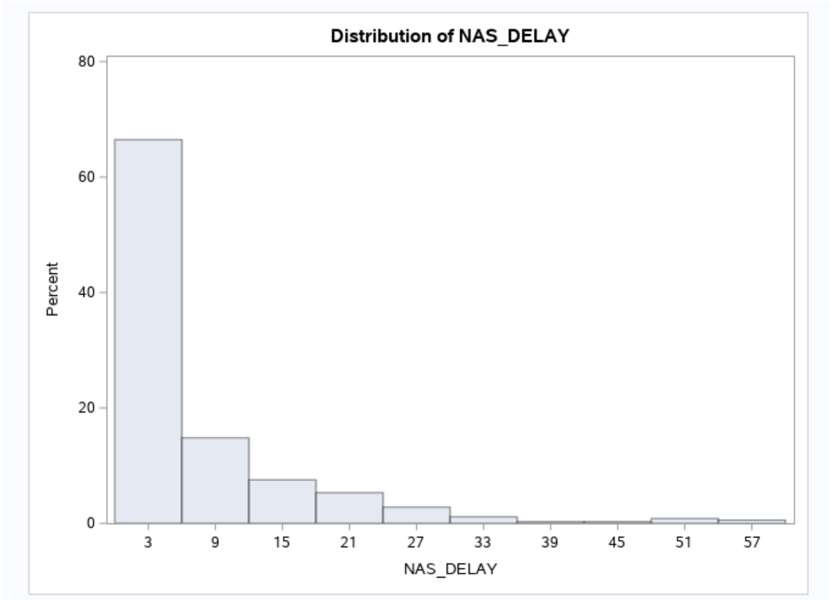

Figure 21

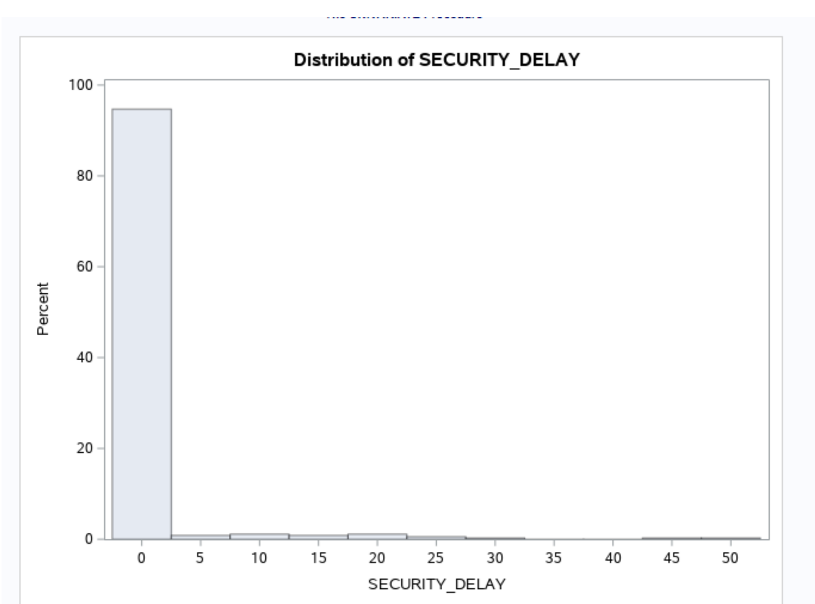

Figure 22

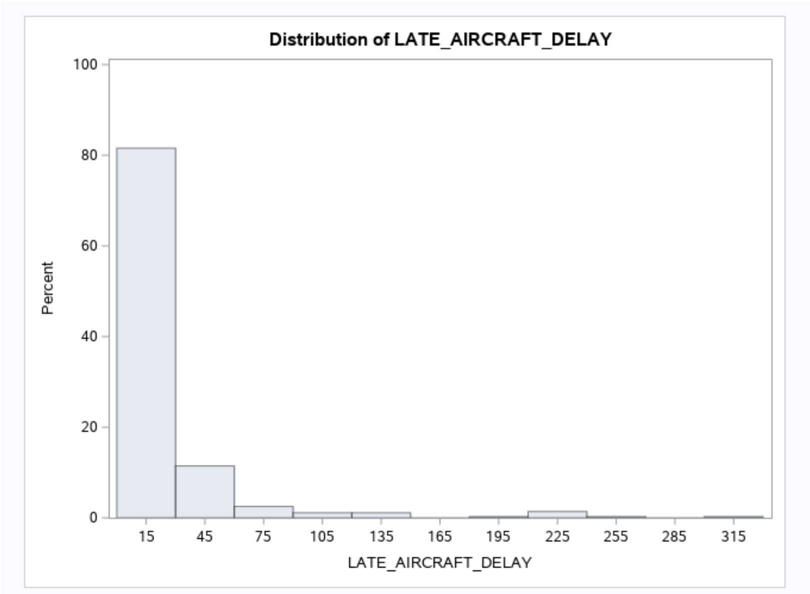

Figure 23

```
proc corr data=ANC2;
    var ARR_DELAY;
    with _numeric_;
    run;
```

Figure 24

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	ARR_DELAY
DAY_OF_MONTH	0.00224 0.9505 768
DEP_DELAY	0.95901 <.0001 768
WHEELS_ON	0.02235 0.5362 768
TAXI_IN	-0.04263 0.2380 768
CRS_ARR_TIME	0.05630 0.1190 768
ARR_TIME	-0.00896 0.8042 768
ARR_DELAY	1.00000 768
ARR_DELAY_NEW	1.00000 <.0001 768
ARR_DEL15	0.49974 <.0001 768
ARR_DELAY_GROUP	0.94800 <.0001 768
CANCELLED	. . 768
AIR_TIME	0.11003 0.0023 768
DISTANCE	0.11408 0.0015 768
CARRIER_DELAY	0.63562 <.0001 358
WEATHER_DELAY	0.11632 0.0278 358
NAS_DELAY	-0.09083 0.0862 358
SECURITY_DELAY	-0.05080 0.3379 358
LATE_AIRCRAFT_DELAY	0.60976 <.0001 358

Figure 25

```
proc corr data=ANC2;
    var distance;
    with arr_delay dep_delay;
    run;
```

Figure 26

The CORR Procedure

2 With Variables:	ARR_DELAY DEP_DELAY
1 Variables:	DISTANCE

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ARR_DELAY	768	25.83333	41.40197	19840	1.00000	473.00000
DEP_DELAY	768	25.63672	44.31571	19689	-28.00000	481.00000
DISTANCE	768	1481	907.04325	1137152	160.00000	3417

Pearson Correlation Coefficients, N = 768 Prob > |r| under H0: Rho=0	
	DISTANCE
ARR_DELAY	0.11408 0.0015
DEP_DELAY	0.21140 <.0001

Figure 27

```
proc sgplot data=ANC2;
   scatter y=arr_delay x=dep_delay;
run;

proc sgplot data=ANC2;
   scatter y=arr_delay x=carrier_delay;
run;

proc sgplot data=ANC2;
   scatter y=arr_delay x=late_aircraft_delay;
run;
```

Figure 28

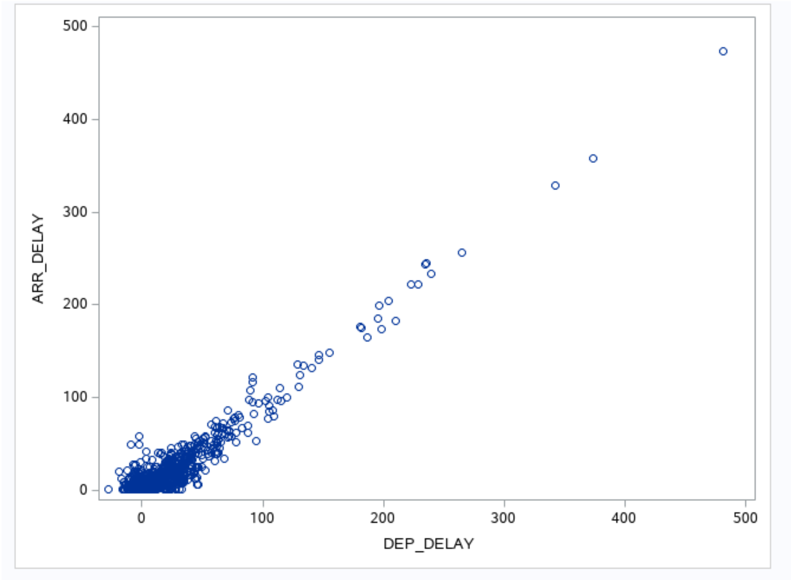

Figure 29

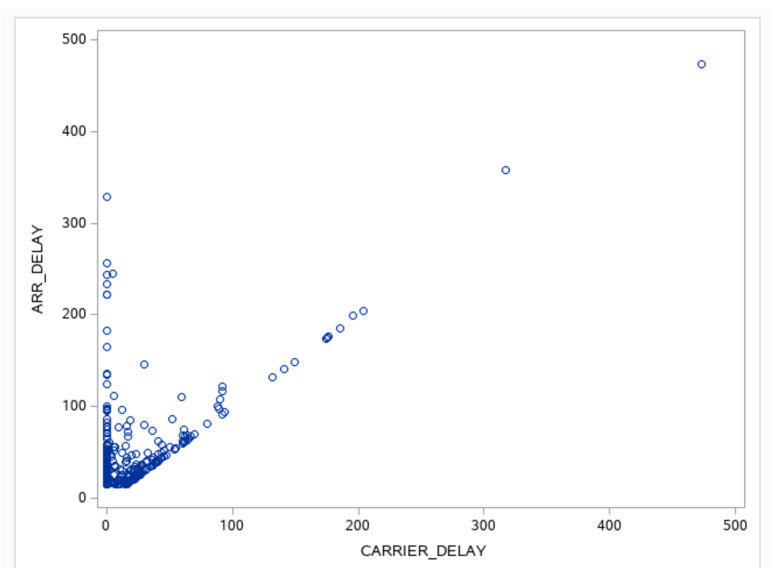

Figure 30

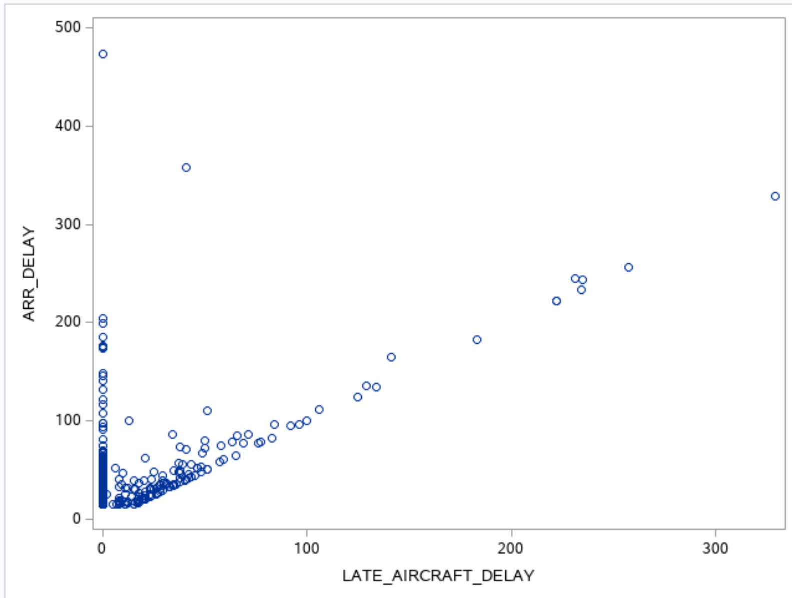

Figure 31

```
proc reg data=ANC2;
    model arr_delay=dep_delay;
    run;
```

Figure 32

The REG Procedure
Model: MODEL1
Dependent Variable: ARR_DELAY

Number of Observations Read	768
Number of Observations Used	768

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1209150	1209150	8772.40	<.0001
Error	766	105582	137.83574		
Corrected Total	767	1314733			

Root MSE	11.74035	R-Square	0.9197
Dependent Mean	25.83333	Adj R-Sq	0.9196
Coeff Var	45.44650		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.86406	0.48951	5.85	<.0001
DEP_DELAY	1	0.89595	0.00957	93.66	<.0001

Figure 33

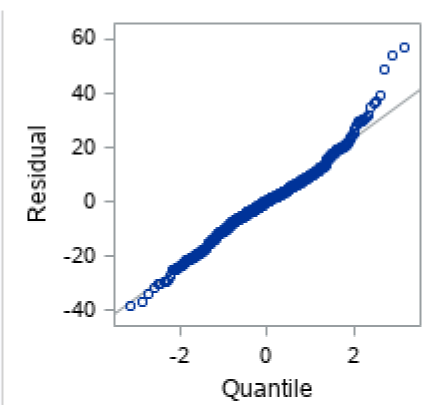

Figure 34