Introduction:

Airport operation as on-timer performance, fares for travelling to or from the airport, certain connection facilities as train, bus to and from the airport are related to how travelers decide to travel through the airport. At any given airport, the airport revenue is based on flights being flown in and out of the airport. However, it also depends on how many travelers have travelled through the airport to provide added revenue by utilizing different services at the airport.

Descriptive Statistics:

The data being used in this study is collected from US Department of Transportation available at http://www.transtats.bts.gov for following:

- US domestic airports on-time performance for domestic travel as reported by major airlines on monthly basis
- US domestic traffic as flights were scheduled for domestic travel plus number of seats available and number of passenger being travelled. Data is available on monthly basis.
- US domestic average fare based on airport from where travel has originated. This is based on round trip fare if round trip was purchased and one-way fare if one-way trip was purchased. Data is only available on quarterly basis as finance reports are available on quarterly basis. I have applied the fares to each month in the years based on the quarter of the years. For example, the average fare reported in 1st Quarter of 2014 is applied to month 1, 2, and 3 in 2014.
- Other inter-connection services available at US domestic airports as intercity connection through rail, bus, airline, ferry and airport official website in order to provide certain travel information prior to travel planning. Data is available as up-to-date information, and information is not available on historical basis. I have applied this data to all the months for given airport based on airport code.

This study is lacking to gather data for security checkpoint wait time at the airport. It was challenging and manual process to gather historical data from Transportation Security Administration site https://apps.tsa.dhs.gov/mytsa/status_home.aspx.

Data selection:

I have collected data for year 2014 and 2015. As Average fare quarterly report for 3Q of 2015 is still not available, I have removed the data for 3Q of 2015.

I have selected data for airports that have network with at least 10 different airport for inbound and outbound flights. Additionally I have only included airports with at least 5000 departures and arrival scheduled per month. This will reduce the possibility of any outliers due to very small airport operations.

Goal:

The goal of this study is to analyze data using data reduction models and analyze the variable that are correlated to either passengers being travelled to or from the airport.

Explanatory variables: Sums are aggregated on month except for categorical (Yes/No) and Numerical data types

Variable	Abbreviation	Data type	Used in Analysis
Count of different airlines flying out of the airport	outbound_carrier_cnt	Numerical	Removed from initial analysis as it is mostly same as inbound carrier count
Count of different airlines flying out of the airport	inbound_carrier_cnt	Numerical	Yes
Count of different airport that are connected through outbound flights from the airport	inbound_network_cnt	Numerical	Yes
Count of different airport that are connected through inbound flights to the airport	outbound_network_cnt	Numerical	Yes
Is other connection service by rail, bus, ferry, air is available to/from the airport to/from city	INTERCITY_SERVICE	Yes/No	Removed after initial analysis
Is other connection service by rail, bus, ferry, air is available to/from the airport to/from another airport in the area	transit_service	Yes/No	Removed after initial analysis
How many different services available either as intercity service or transit service	modes_serving	Numerical	Removed after initial analysis for PC
Does the airport has official website	website_avail	Yes/No	Removed after initial analysis
Average fare from origination airport	fare	Continuous	Yes
Sum of number of Departure delays >= 15 minutes	DEP_DEL15	Continuous	Yes
Sum of cancelled flights	CANCELLED		Yes
Sum of number of Arrival delay >= 15 minutes	ARR_DEL15		Yes
Sum of delays due to carrier's operation	carrier_delay	Continuous	Yes
Sum of delays due to incoming aircraft being late causing the on-going flight being late >= 15 minutes	LATE_AIRCRAFT_DELAY	Continuous	Yes
Sum of delays or cancellation attributed to National Aviation System	nas_delay	Continuous	Yes
Sum of delays and cancellation due to security issues as re-boarding, evacuation.	SECURITY_DELAY	Continuous	Yes
Sum of delays due to weather delays on either origin or destination	WEATHER_DELAY	Continuous	Yes
Sum of departures scheduled as planned	departures_scheduled	Continuous	Yes
Sum of departures actually performed	departures_performed	Continuous	Yes

Sum of arrivals actually performed	arrivals_performed	Continuous	Yes
Sum of arrivals scheduled as planned	arrivals_scheduled	Continuous	Yes
Sum of seats available on flights departing from the airport	outbound_capacity	Continuous	Yes
Sum of seats available on flights arriving at the airport	inbound_capacity	Continuous	Yes

Table 1

Response variables:

Variable	Abbreviation	Data Type	Used in Analysis
Number of passengers boarded on flights flying out from the airport	passengers_enplaned	Continuous	Yes
Number of passengers arrived at the airport from incoming flights	passengers_deplaned	Continuous	Yes

Table 2

After some initial analysis as finding the Means and SD as shown in Figure 1, I have decided to remove **outbound_carrier_cnt** from the analysis as it is almost similar to **inbound_carrier_cnt**. Usually airline that has arrived at the airport, will depart too.

Variable	N	Mean	Std Dev	Minimum	Maximum
outbound_carrier_cnt	5354	8.5293239	5.9484035	2.0000000	26.0000000
inbound_carrier_cnt	5354	8.5517370	5.9737918	2.0000000	26.0000000
inbound_network_cnt	5354	23.5823683	29.9279045	1.0000000	170.0000000
outbound_network_cnt	5354	23.6186029	30.7003536	1.0000000	175.0000000
INTERCITY_SERVICE	5354	0.9607770	0.1941433	0	1.0000000
transit_service	5354	0.4747852	0.4994105	0	1.0000000
modes_serving	5354	1.5528577	0.7059224	0	3.0000000
website_avail	5354	0.8117295	0.3909645	0	1.0000000
fare	5354	432.6848991	113.1860903	109.5900000	1592.90
DEP_DEL15	5354	392.2891296	961.4642346	0	14336.00
CANCELLED	5354	39.2342174	134.1659962	0	3596.00
ARR_DEL15	5354	402.0067239	895.5722000	0	13102.00
carrier_delay	5354	119.4090400	344.8930227	0	5154.00
LATE_AIRCRAFT_DELAY	5354	164.7480388	377.7338623	0	5352.00
nas_delay	5354	124.3875607	295.1923128	0	5016.00
SECURITY_DELAY	5354	0.5603287	2.2826195	0	50.0000000
WEATHER_DELAY	5354	15.7919313	76.6730258	0	3144.00
departures_scheduled	5354	2484.33	4864.09	5.0000000	35610.00
departures_performed	5354	2474.22	4775.63	5.0000000	35115.00
arrivals_performed	5354	2474.11	4772.53	4.0000000	35036.00
arrivals_scheduled	5354	2474.11	4772.53	4.0000000	35036.00
outbound_capacity	5354	269092.94	564585.85	362.0000000	4563349.00
inbound_capacity	5354	269095.65	564416.60	312.0000000	4564270.00
passengers_enplaned	5354	221469.33	474187.76	154.0000000	4030512.00
passengers_deplaned	5354	221489.77	474756.20	125.0000000	4052964.00

Figure 1

Variable	N	Mean	Std Dev	Minimum	Maximum
linbound_carrier_cnt	5354	1.8768299	0.7626006	0.6931472	3.2580965
linbound_network_cnt	5354	2.5347790	1.1229849	0	5.1357984
loutbound_network_cnt	5354	2.5002470	1.1600083	0	5.1647860
INTERCITY_SERVICE	5354	0.9607770	0.1941433	0	1.0000000
transit_service	5354	0.4747852	0.4994105	0	1.0000000
modes_serving	5354	1.5528577	0.7059224	0	3.0000000
website_avail	5354	0.8117295	0.3909645	0	1.0000000
lfare	5354	6.0399478	0.2472559	4.6967461	7.3733115
lDEP_DEL15	5322	4.3538590	1.7926113	0	9.5705291
lCANCELLED	4776	2.4845788	1.4956990	0	8.1875774
lARR_DEL15	5328	4.5685000	1.7174124	0	9.4805202
lcarrier_delay	5044	2.9758784	1.8476456	0	8.5475284
lLATE_AIRCRAFT_DELAY	5222	3.7324707	1.6793615	0	8.5852256
lnas_delay	5187	3.3858493	1.7177480	0	8.5203881
lSECURITY_DELAY	890	0.7507215	0.8546796	0	3.9120230
lWEATHER_DELAY	3643	1.6773939	1.4759609	0	8.0532512
ldepartures_scheduled	5354	6.6111994	1.5346055	1.6094379	10.4803818
ldepartures_performed	5354	6.6601704	1.4883335	1.6094379	10.4663837
larrivals_performed	5354	6.6607201	1.4884291	1.3862944	10.4641314
larrivals_scheduled	5354	6.6607201	1.4884291	1.3862944	10.4641314
loutbound_capacity	5354	10.9614992	1.7591243	5.8916442	15.3335673
linbound_capacity	5354	10.9621478	1.7591242	5.7430032	15.3337691
lpassengers_enplaned	5354	10.6968086	1.8283676	5.0369526	15.2094040
lpassengers_deplaned	5354	10.6912736	1.8325203	4.8283137	15.2149590

Figure 2

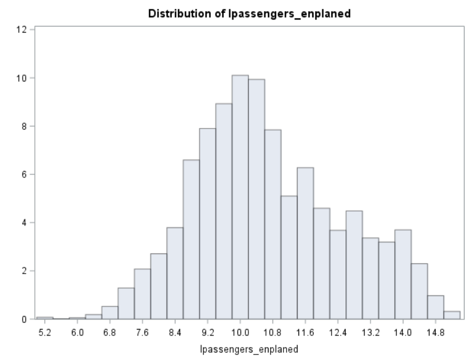
As standard deviation is large on most of the continuous variables, I have decided to take log transformation on continuous variables and the in/outbound network counts and inbound carrier count. New logged transformed data is displayed in Figure 2 above.

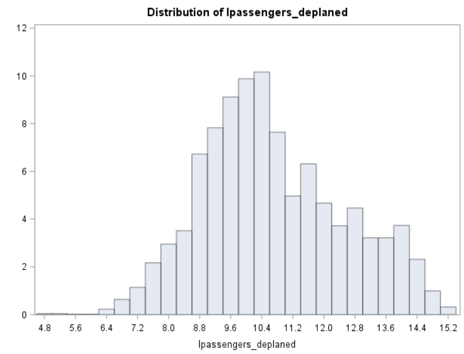
Initial observation for normal distribution is done by generating histograms. Generating scatter plot was not very helpful with large number of variables and not being able to visualize it clearly.

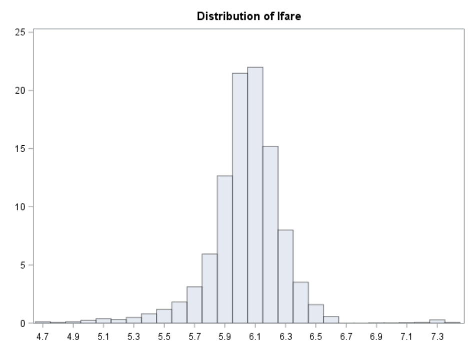
Data exception from normality check:

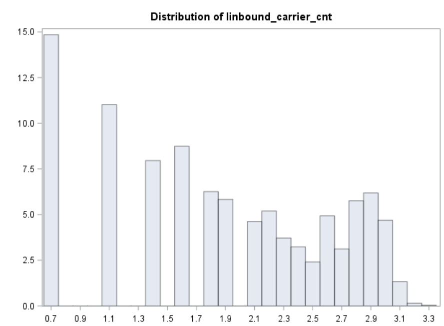
First histograms for categorical variables as website_avail, transit_service, INTERCITY_SERVICE would not be applicable to normality as they have just two values. For modes_serving that I have not transformed to log data as it is not a continuous variable so its histogram doesn't apply.

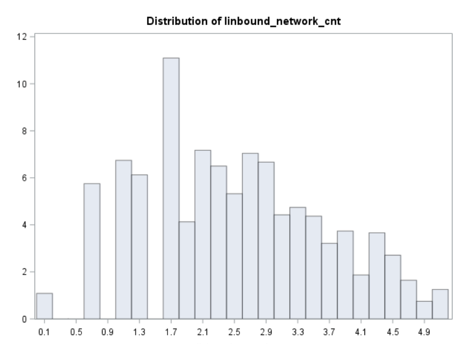
__Data included in normality check:__

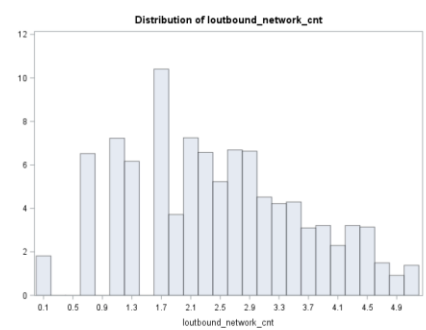


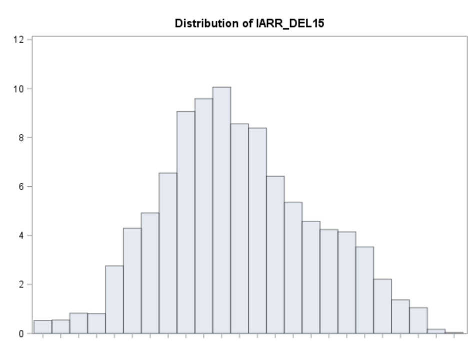


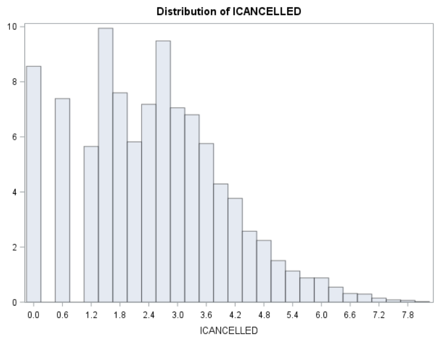


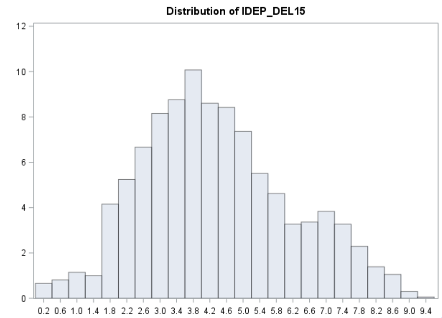


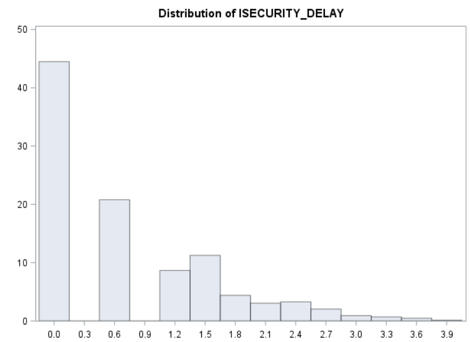


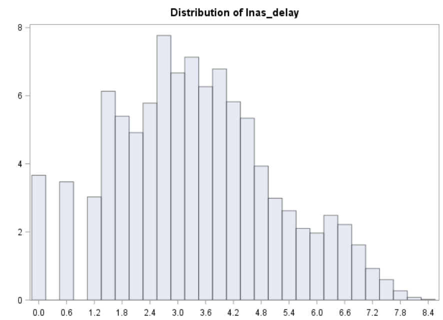


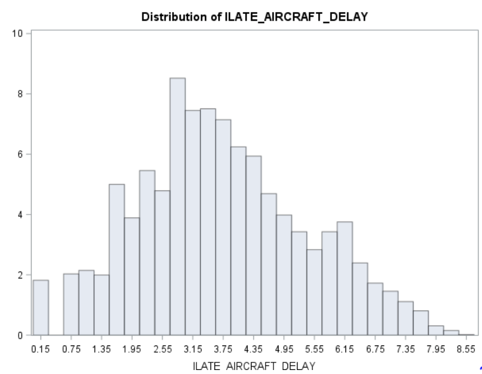


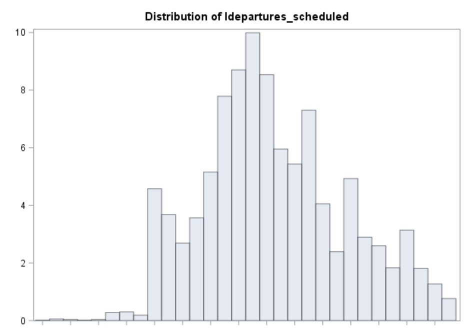

Distribution of ICANCELLED
ICANCELLED

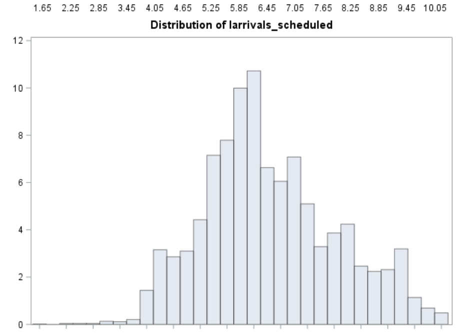

Distribution of IDEP_DEL15

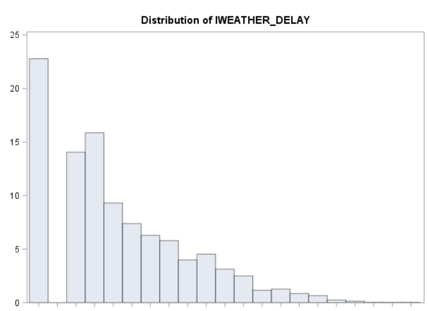

Distribution of ISECURITY_DELAY

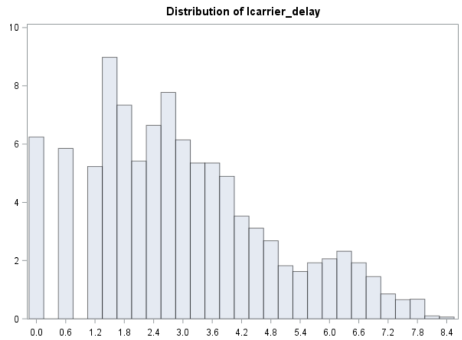

Distribution of lnas_delay

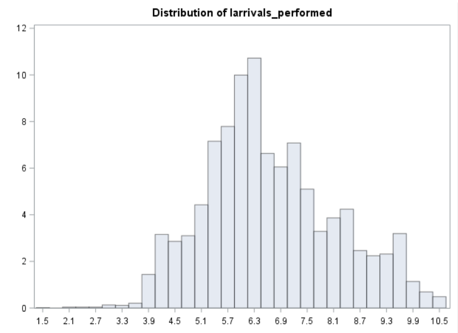

Distribution of ILATE_AIRCRAFT_DELAY
ILATE_AIRCRAFT_DELAY

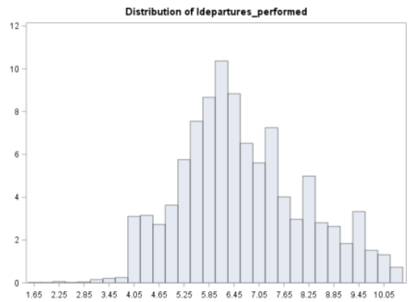

Distribution of Idepartures_scheduled


Distribution of Iarrivals_scheduled


Distribution of IWEATHER_DELAY







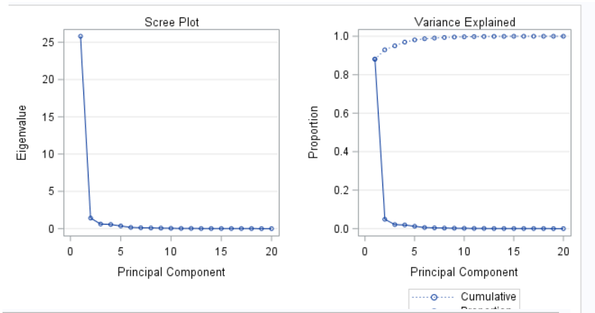
As evident from histograms, most of the continuous variables are normally distributed as log transformed, some are skewed, and less has exceptions as not being normally distributed.

Analysis:

I have decided to first try PCA to see if I can eliminate more variables before running canonical correlation analysis CCA. As PCA can take one response variable, I have perform PCA for both response variables lpassengers_enplaned and lpassengers_deplaned separately. As discussed in the class about PCA with categorical variables, I have removed categorical variable website_avail, transit_service, INTERCITY_SERVICE from PCA analysis. As data is already been adjusted using log transformed, I have used covariance option with PCA analysis using SAS procedure princomp.

First Performed analysis for lpassenger_enplanded, and it shows that two PC should be enough to get over 90% variance covered. PC1: It seems to be correlated on most of the variables:

	Prin1
lcarrier_delay	0.30633
lpassengers_enplaned	0.27897
loutbound_capacity	0.27358
linbound_capacity	0.27346
lDEP_DEL15	0.26877
lnas_delay	0.26078
lWEATHER_DELAY	0.25709
lARR_DEL15	0.25426
lLATE_AIRCRAFT_DELAY	0.24846
ldepartures_scheduled	0.24027
ldepartures_performed	0.23398
larrivals_scheduled	0.23381
larrivals_performed	0.23381
lCANCELLED	0.22362



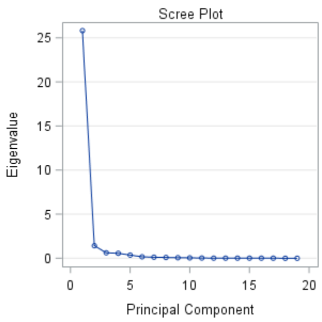
Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	25.9447388	24.5273792	0.8706	0.8706
2	1.4173596	0.7618233	0.0476	0.9182

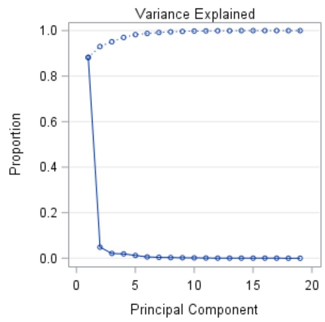
AS modes_serving is not very correlated, I will leave it out from analysis going forward. It is shown that most of the variables as correlated in PC1 (Prin1).

For PC2 (Prin2), flight cancellation and weather delays seems to be much correlation and it is evident historically.

	Prin1	Prin2
lCANCELLED	0.22362	0.62097
lWEATHER_DELAY	0.25709	0.5113

From PCA for passenger_deplanded, again two PC are enough to get more than 90% of variance covered.





Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	25.8212083	24.4039779	0.8819	0.8819
2	1.4172305	0.8097836	0.0484	0.9303

	Prin1
lcarrier_delay	0.30709
lpassengers_deplaned	0.28062
loutbound_capacity	0.27423
linbound_capacity	0.27411
lDEP_DEL15	0.26942
lnas_delay	0.26137
lWEATHER_DELAY	0.25758
lARR_DEL15	0.25486
lLATE_AIRCRAFT_DELAY	0.24908
ldepartures_scheduled	0.24091
ldepartures_performed	0.23458
larrivals_scheduled	0.23441
larrivals_performed	0.23441
lCANCELLED	0.22416

Again it is evident that most of the variables are correlated in PC1 (Prin1) for response variable of lpassenger_deplanded.

For PC2 (Prin2), seems like three variables are correlated mostly as shown below:

	Prin1	Prin2
lpassengers_deplaned	0.28062	0.2361
loutbound_capacity	0.27423	0.22085
linbound_capacity	0.27411	0.22064

From the separate PCA for both response variable, it is evident that carrier count and both inbound and outbound network count is not very correlated. Fare is not very correlated either. So moving forward I will drop linbound_carrier_cnt, linbound_network_count, loutbound_network_cnt and lfare from further analysis.

As we have multiple response variables, and still large number of explanatory variables, I have decide to perform Cannonical Component analysis. MANOVA cannot be applied here as the explanatory variables are correlated. I have large number of sample as 5000+. CCA is suggested with medium size sample as 50 to 100. To limit the sample size, I have selected data for some of the busy airports as following:

DFW (Dallas Fort Worth), ATL (Atlanta), ORD (Chicago), LAX (Los Angeles), JFK (New York)

The sample size now is about 105 that is acceptable for CCA. I have processed the CCA using SAS procedure cancorr.

Hypothesis: **Test of H0: The canonical correlations in the current row and all that follow are zero**

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.997363	0.997027	0.000521	0.994733	188.8770	188.7325	0.9992	0.9992
2	0.355318	0.212208	0.086514	0.126251	0.1445		0.0008	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.00460166	85.39	28	174	<.0001
0.87374885	0.98	13	88	0.4791

From the output from SAS as shown above, it is evident that one variate is good enough to explain the variability in the model. First canonical variate in the result is explaining about 99.4% of variability in the model. First variate is also supported by having very Eigenvalue. Also from the hypothesis test, it is again evident that first canonical variate is significant with p-value < 0.0001. On the other hand second variate is not significant with p-value of 0.4791.

Standardized Canonical Coefficients for the WITH Variables		
	W1	W2
IDEP_DEL15	-0.0273	4.4175
ICANCELLED	-0.0067	-0.6035
IARR_DEL15	-0.1198	-2.4193
Icarrier_delay	0.0246	-2.0854
ILATE_AIRCRAFT_DELAY	0.0836	1.3495
Inas_delay	0.0261	0.3417
ISECURITY_DELAY	-0.0008	0.2995
IWEATHER_DELAY	-0.0074	-0.2544
Idepartures_scheduled	4.6495	-28.3788
Idepartures_performed	-0.8765	-243.716
Iarrivals_scheduled	1.2302	238.0315
Iarrivals_performed	-4.9451	32.0457
Ioutbound_capacity	-7.6146	288.7981
Iinbound_capacity	8.5754	-287.808

Standardized Canonical Coefficients for the VAR Variables		
	V1	V2
Ipassengers_deplaned	0.7478	-24.9148
Ipassengers_enplaned	0.2523	24.9247

I will only consider the variate V1 and W1 as response variate and explanatory variate following from the hypothesis test.

As discussed in the class lectures, only loading > 0.4 should be considered. So I have highlighted in yellow the explanatory variables that are mostly defining the response variable. From response variables, V1, lpassengers_deplaned is selected as > 0.4 that is passengers arriving at the airport by incoming flights. I have also circled the canonical variate W1 for IARR_DEL15 as it should be included in the model as it defines passengers arriving at the airport.

I still think that IDEP_DEL15 and ICANCELLED as flights delayed to depart > 15 minutes and flights being cancelled should be included in the model. However, as I are looking from the airport perspective and flight might be more of the planning controlled by airlines and not by airport.

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables		
	W1	W2
lpassengers_deplaned	0.9973	-0.0036
lpassengers_enplaned	0.9969	0.0107

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables		
	V1	V2
lDEP_DEL15	0.7809	0.0679
lCANCELLED	0.1341	0.0187
lARR_DEL15	0.7231	0.0646
lcarrier_delay	0.7578	0.0304
lLATE_AIRCRAFT_DELAY	0.6943	0.0996
lnas_delay	0.4167	0.0360
lSECURITY_DELAY	-0.2164	0.1418
lWEATHER_DELAY	0.3229	0.0100
ldepartures_scheduled	0.9298	-0.0026
ldepartures_performed	0.9473	-0.0069
larrivals_scheduled	0.9476	-0.0071
larrivals_performed	0.9294	-0.0027
loutbound_capacity	0.9962	0.0043
linbound_capacity	0.9963	0.0040

Flights scheduled to arrive and depart is the coordination between airport and airlines. Thus it make more sense to add it to the model. loutbound_capacity and linbound_capacity are representing the log value of total seat capacity for flights coming in and going out of the airport. As seats are based on flight aircraft being big or small with more seats, it is partially related to airport as how many big and small aircafts can be handled at the airport.

From the correlation between response variables and variates, departure delay and arrival delays seems more correlated to response along with delays related to carrier operations. It does seems logical as more passengers are being handled, it might be possible to get delayed for various reasons; however it should be already in the flight plan.

Conclusion:

I have analyzed the dataset for on-time performance in regards to airport and airline operations, average fares summary and other intercity and transit services for the airport. Provided given data, it is evident that passenger traffic for in/out of the airport is highly based on planning of flight schedules vs. actual flight operations performed as arrival/departure. Plus it is also based on total seat capacity that will refer back to what kind of aircraft being used by airlines, as bigger aircraft has more seats available as compare to smaller aircraft. It is a question if airport is capable of handling small or big aircrafts. I would also include that flight arrival/departure delays are also correlated, however the impact of current on-time performance may affect future travelers in order to choose airports as origin and destination for next travel.

References:

Data: https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time

Database to hold data and reformat for analysis: MySql Database plus references operations on tables.

Class Lectures – MSDS 6372