<u>MS9004 Introduction to Statistical Modelling Assignment 2019/ 2020 Semester 2</u>

Name: Wong Qi Yuan, Jeffrey
Class: PA-01
Student ID: P7359567

<u>Part I: Explore Data</u>

- **Response:** arrive delay (in mins), quantitative, continuous
- **Predictors:**
    - **Qualitative & Ordinal:** month, day, weekend and quarter
    - **Qualitative & Nominal:** airline and low cost
    - **Quantitative & Continuous:** distance (in km), sched time (in mins), depart delay (in mins), taxi_out (in mins), airtime (in mins), taxi_in (in mins), system delay (in mins), security delay (in mins), airline delay (in mins) and aircraft delay (in mins)
- n = 4486, p = 16
- The descriptive statistics showing the spread of the data are as follows:

	month	day	weekend	quarter	airline	lowcost	distance	schedtime	departdelay	taxi_out	airtime	taxi_in
count	4486	4486	4486	4486	4486	4486	4486.000000	4486.000000	4486.000000	4486.000000	4486.000000	4486.000000
unique	12	7	2	4	8	2	NaN	NaN	NaN	NaN	NaN	NaN
top	Aug	Tue	No	Q4	WN	No	NaN	NaN	NaN	NaN	NaN	NaN
freq	418	683	3318	1205	1270	2275	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	879.646456	148.504681	9.754570	15.959206	119.676995	7.191039
std	NaN	NaN	NaN	NaN	NaN	NaN	618.742669	76.781549	34.803134	8.900635	73.092112	5.305332
min	NaN	NaN	NaN	NaN	NaN	NaN	31.000000	23.000000	-23.000000	3.000000	9.000000	1.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	406.000000	90.000000	-4.000000	11.000000	64.000000	4.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	731.000000	130.000000	-1.000000	14.000000	102.000000	6.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	1138.250000	183.000000	8.000000	18.000000	153.000000	8.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	4962.000000	604.000000	821.000000	134.000000	559.000000	99.000000

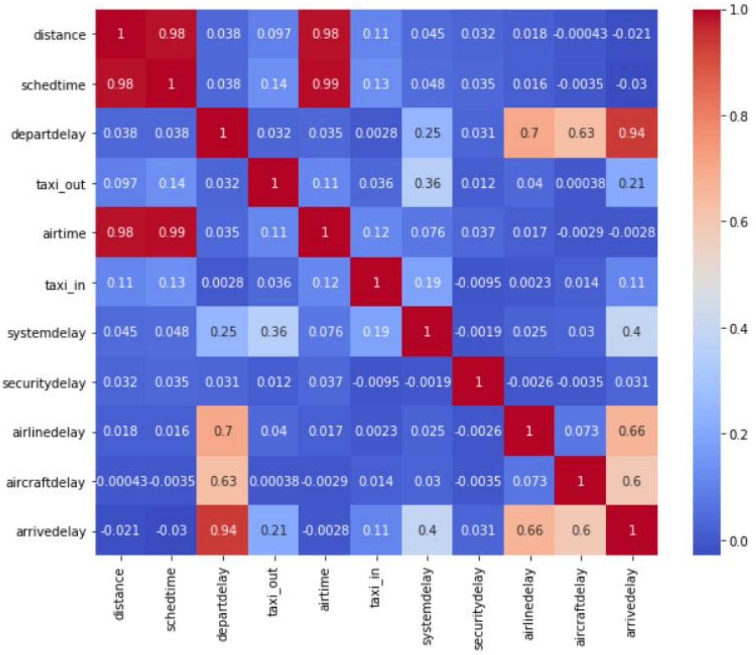
	systemdelay	securitydelay	airlinedelay	aircraftdelay	arrivedelay
count	4486.000000	4486.000000	4486.000000	4486.000000	4486.000000
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	2.269059	0.019394	3.495542	4.391217	4.077129
std	12.036198	1.198970	22.068079	18.257372	37.130369
min	0.000000	0.000000	0.000000	0.000000	-60.000000
25%	0.000000	0.000000	0.000000	0.000000	-13.000000
50%	0.000000	0.000000	0.000000	0.000000	-5.000000
75%	0.000000	0.000000	0.000000	0.000000	8.000000
max	381.000000	80.000000	801.000000	228.000000	801.000000

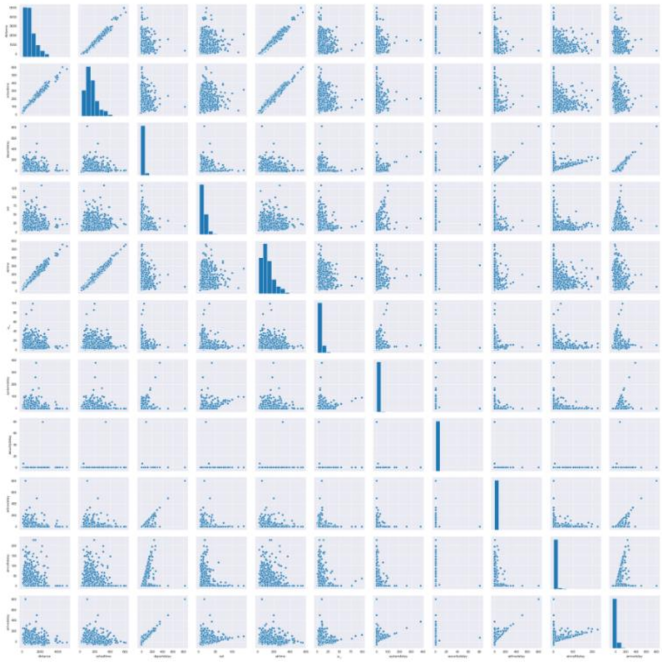
- The table below shows the Pearson's correlation between predictor variables.

	distance	schedtime	departdelay	taxi_out	airtime	taxi_in	systemdelay	securitydelay	airlinedelay	aircraftdelay	arrivedelay
distance	1.000000	0.984129	0.038370	0.096971	0.984891	0.109723	0.045297	0.031738	0.018460	-0.000430	-0.021400
schedtime	0.984129	1.000000	0.038271	0.137918	0.990605	0.133994	0.048299	0.034802	0.015988	-0.003508	-0.029780
departdelay	0.038370	0.038271	1.000000	0.032250	0.035429	0.002798	0.254376	0.031431	0.698442	0.631293	0.936055
taxi_out	0.096971	0.137918	0.032250	1.000000	0.113618	0.035989	0.355886	0.012151	0.039620	0.000384	0.213545
airtime	0.984891	0.990605	0.035429	0.113618	1.000000	0.116934	0.076061	0.036609	0.016645	-0.002892	-0.002784
taxi_in	0.109723	0.133994	0.002798	0.035989	0.116934	1.000000	0.188025	-0.009486	0.002288	0.014494	0.107236
systemdelay	0.045297	0.048299	0.254376	0.355886	0.076061	0.188025	1.000000	-0.001860	0.024925	0.030053	0.400459
securitydelay	0.031738	0.034802	0.031431	0.012151	0.036609	-0.009486	-0.001860	1.000000	-0.002563	-0.003464	0.031119
airlinedelay	0.018460	0.015988	0.698442	0.039620	0.016645	0.002288	0.024925	-0.002563	1.000000	0.072656	0.664193
aircraftdelay	-0.000430	-0.003508	0.631293	0.000384	-0.002892	0.014494	0.030053	-0.003464	0.072656	1.000000	0.595450
arrivedelay	-0.021400	-0.029780	0.936055	0.213545	-0.002784	0.107236	0.400459	0.031119	0.664193	0.595450	1.000000

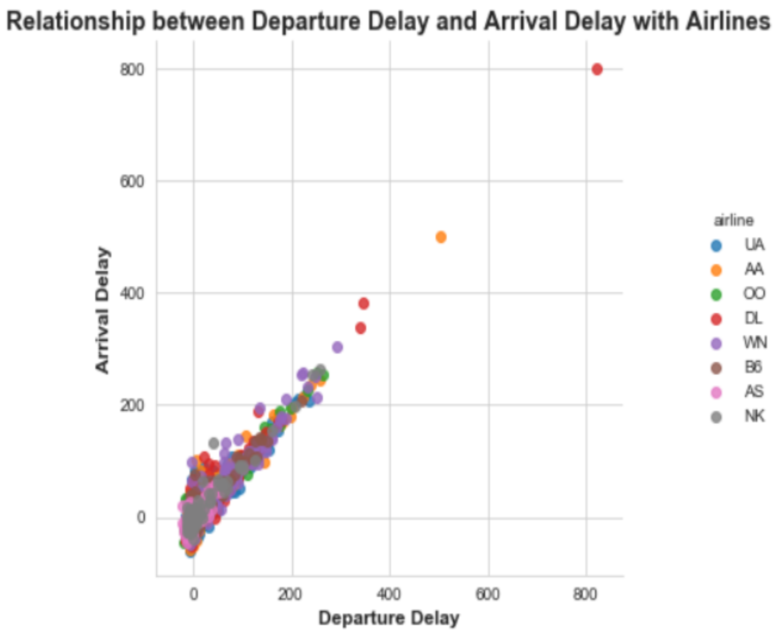
- According to the heatmap and pair plot, it is observed that arrival delay has strong positive linear relation with the departure delay, then followed by moderate positive linear relation with the both airline delay and aircraft delay and finally weak positive linear relation with taxi_out, taxi_in, system delay and security delay variables. Similarly, the arrival delay involved weak negative linear relation with the distance, scheduled time and airtime variables, respectively.

- In addition, there is a slight hint of multicollinearity between variables can be gained from the plot below and also see that the relation between distance and scheduled time, distance and airtime, airtime and scheduled time, departure delay and airline delay, departure delay and aircraft delay might have serious multicollinearity problems. However, further tests should be carried out to confirm the same.
- Also, the distribution of all variables is seen in the diagonal of the plot. The response variable as well as the rest of the quantitative variable are highly skewed to the right. This might lead to the violation of the normality rule after the basic regression model is formed. However, further tests should be carried out to confirm the distribution.





- Based on the above preliminary analysis, there is a very significant correlation between departure and arrival delays. However, correlation does not imply causation. Airline delay and aircraft delay have a moderately significant impact on arrival delays while other delays like security and system delay have least impact on flight arrival delay.
- One discovery made here was that departure delay is most impacted by the same airline delay and aircraft delays, which in turn impacts arrival delay.
- The scatterplot provides view of density as well as distribution of departure versus arrival delay, proving that most of the arrival delays occurs due to departure delays for all 8 airlines in 2015.



Part II: Build & Evaluate Model

Model 1: Additive MLR model consisting of significant predictor only

- Split the data into 80: 20, with seed 9567
- Actual ratio is 80.41%: 19.59%, that is n = 3607 data for training, and n = 879 for testing.
- A **stepwise regression** was applied and the output of the optimal regression on the training data are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            arrivedelay   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                     6215.
Date:                Fri, 24 Jan 2020   Prob (F-statistic):               0.00
Time:                        09:03:46   Log-Likelihood:                -12985.
No. Observations:                3607   AIC:                         2.599e+04
Df Residuals:                    3596   BIC:                         2.606e+04
Df Model:                          10
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        -17.5257      0.516    -33.962      0.000     -18.537     -16.514
lowcost[T.Yes]     3.4481      0.312     11.055      0.000       2.837       4.060
weekend[T.Yes]    -0.7944      0.338     -2.352      0.019      -1.457      -0.132
distance          -0.0039      0.000    -15.717      0.000      -0.004      -0.003
departdelay        0.6892      0.013     51.794      0.000       0.663       0.715
taxi_out           0.5743      0.019     31.040      0.000       0.538       0.611
taxi_in            0.5618      0.031     17.886      0.000       0.500       0.623
airlinedelay       0.3240      0.015     21.547      0.000       0.295       0.353
systemdelay        0.4956      0.016     30.373      0.000       0.464       0.528
aircraftdelay      0.3323      0.017     19.887      0.000       0.300       0.365
securitydelay      0.4292      0.112      3.838      0.000       0.210       0.648
==============================================================================
Omnibus:                      227.727   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              985.878
Skew:                          -0.108   Prob(JB):                    8.31e-215
Kurtosis:                       5.552   Cond. No.                     4.03e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.03e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
MSE =  78.66892207078736
```

Additive MLR Model:

$$\begin{aligned}\widehat{arrivedelay} &= -17.5257 + 3.4481\ lowcost - 0.7944\ weekend - 0.0039\ distance + 0.6892\ departdelay \\ &+ 0.5743\ taxi_{out} + 0.5618\ taxi_{in} + 0.3240\ airlinedelay + 0.4956\ systemdelay + 0.3323\ aircraftdelay \\ &+ 0.4292\ securitydelay\end{aligned}$$

Interpretation on the Constant Term:

- The average overall arrival delay for the airlines with high-cost carrier on weekdays is estimated to be -17.5257 minutes.

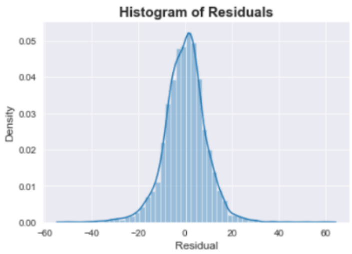
Interpretation on the Coefficients:

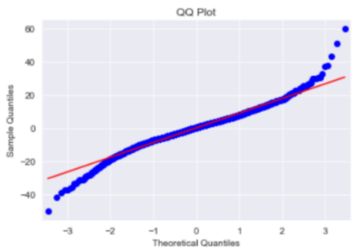
- **Baseline:** Airline is not a low-cost carrier and the day is not a weekend
- While holding other factors constant, the overall arrival delay for airlines with low-cost carrier is expected to increase with additional of 3.4481, in minutes, that those airlines with high-cost carrier, on average.
- While holding other factors constant, the overall arrival delay on weekend is estimated to be -0.7944, in minutes, earlier than those on weekdays, on average.
- While holding other factors constant, for every additional of 1 km on distance travelled between two airports, the overall arrival delay is estimated to be decrease by 0.0039, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on the overall departure delay, the overall arrival delay is estimated to be increase by 0.6892, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on taxi out, the overall arrival delay is estimated to be increase by 0.5743, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on taxi in, the overall arrival delay is estimated to be increase by 0.5618, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on airline delay, the overall arrival delay is estimated to be increase by 0.3240, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on system delay, the overall arrival delay is estimated to be increase by 0.4956, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on aircraft delay, the overall arrival delay is estimated to be increase by 0.3323, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on security delay, the overall arrival delay is estimated to be increase by 0.4292, in minutes, on average.

Evaluate Model on Training Data and Conduct Diagnostics

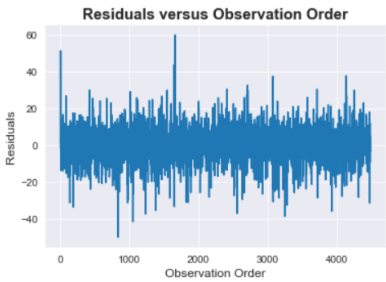
- **Coefficient of Determination, R^2** = 94.5%, indicating 94.5% of the total variability in the overall arrival delay can be accounted for by the additive MLR model. This model is in good fit.
- **Adjusted R^2** = 94.5%, which is the same as the R^2.
- **MSE =** 78.67%
- **F-test for overall model:** p-value ≈ 0.000 < 5%
    - At least one predictor contributes significantly to the model.
- **t-test for coefficients:**
    - The associated p-values of all quantitative coefficients (i.e. distance, departdelay, taxi_out, taxi_in, airlinedelay, systemdelay, aircraftdelay, securitydelay) ≈ 0.000 < 5%, tested to be statistically significant predictors to the overall arrival delay.
    - The associated p-value of weekend coefficient ≈ 0.019 < 5% tested to be statistically significant predictors and there is a statistical evidence of difference in overall arrival delay between the weekend and non-weekend.
    - The associated p-value of low-cost coefficient ≈ 0.000 < 5% tested to be statistically significant predictors and there is a statistical evidence of difference in overall arrival delay between the low-cost and high-cost carriers.
- **Checking for Normality:** The QQ plot seem to indicate that the residuals are not normally distributed as the two tails end deviate from the probability line, and the outliers are very evident. Similarly, the following tests also indicate that the residuals are not normal. Thus, the normality assumption is not valid.

Omnibus Test	**Anderson-Darling Test**	**Jacque-Bera Test**	**Shapiro-Wilks Test**
p-value ≈ 0.000 < 5%	test-statistic = 13.05 > critical value = 0.786	p-value ≈ 0.000 < 5%	p-value ≈ 0.000 < 5%

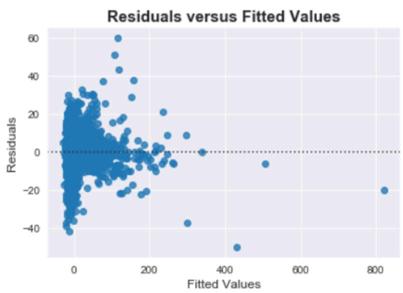




- **Checking for Independence:** The plot of residuals by observation order shows that the residuals are in random pattern and no systematic pattern observed. Furthermore, the Durbin-Watson Test statistic = 1.960 ≈ 2. This indicates that the residuals are not auto-correlated. Thus, independence assumption is valid.



- **Checking for Homoscedasticity:** The plot of residuals vs. fitted values shows that residuals are not randomly dispersed, suggesting that the residuals are non-constant variance. Furthermore, the Breusch-Pagan test (p-value ≈ 0.000), confirms that the residuals are heteroscedastic. Thus, constant variance assumption is not valid. From the plot, it could be observed that there are a couple of outliers could influence the reliability of regression analysis.



- **Checking for Multicollinearity:** The Variance Inflation Factors (VIF) show that there might be some moderate to serious multicollinearity issue (VIF > 5 or VIF > 10) within the regression function. Thus, multicollinearity assumption is not valid.

VIF Results of Model 1			
Predictor Variable	**VIF**	**Predictor Variable**	**VIF**
lowcost	1.115	taxi in	1.087
weekend	1.005	airlinedelay	5.641
distance	1.084	systemdelay	1.926
departdelay	10.344	securitydelay	1.017
taxi out	1.248	aircraftdelay	4.395

Evaluate & Deploy Model on Testing Data

- **Testing Data:** R^2 = 93.34%; MSPE = 76.57%, similar as Training Data.
- Based on the above training data diagnostics, this regression model could not be deployed for prediction purpose with due to some strict regression criteria are not met. Although the goodness-of-fit and the predictors are showing satisfactory, and if it is chosen to deployed for prediction, then there will be a very high chance of getting prediction errors and leads to an unreliable analysis.
- This regression model can be seen to be further improved by adding possible interaction terms, transformation on variables, perform PCA analysis or re-investigate the whole dataset on the influence outliers, whichever is applicable.

Model 2: Investigate possible interactions and/ or transformation of variables

- There are many choices for possible interaction effects. For this analysis, it seems logical that the effect of between depart delay, airline delay, aircraft delay as well as system delay may have some kind of joint impact with one another on the overall arrival delay, and hence this can be added to the extension of model 1.
- Split the data into 80: 20, with seed 9567
- Actual ratio is 80.41%: 19.59%, that is n = 3607 data for training, and n = 879 for testing.
- A **stepwise regression** was applied and the output of the optimal regression on the training data are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           arrivedelay   R-squared:                       0.948
Model:                           OLS   Adj. R-squared:                  0.948
Method:                Least Squares   F-statistic:                     5041.
Date:               Wed, 29 Jan 2020   Prob (F-statistic):               0.00
Time:                       08:09:20   Log-Likelihood:                -12894.
No. Observations:               3607   AIC:                         2.582e+04
Df Residuals:                   3593   BIC:                         2.590e+04
Df Model:                         13
Covariance Type:           nonrobust
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  -16.2668      0.514    -31.620      0.000     -17.275     -15.258
lowcost[T.Yes]               3.0765      0.306     10.070      0.000       2.477       3.675
weekend[T.Yes]              -0.8327      0.330     -2.527      0.012      -1.479      -0.187
distance                    -0.0040      0.000    -16.259      0.000      -0.004      -0.003
departdelay                  0.7268      0.013     54.306      0.000       0.701       0.753
airlinedelay                 0.3238      0.018     18.255      0.000       0.289       0.359
systemdelay                  0.6769      0.021     31.674      0.000       0.635       0.719
securitydelay                0.3973      0.109      3.641      0.000       0.183       0.611
aircraftdelay                0.3400      0.025     13.666      0.000       0.291       0.389
taxi_in                      0.5150      0.031     16.682      0.000       0.455       0.576
taxi_out                     0.5038      0.019     26.747      0.000       0.467       0.541
departdelay:airlinedelay  -8.138e-05   2.11e-05     -3.849      0.000      -0.000   -3.99e-05
departdelay:aircraftdelay   -0.0004      0.000     -2.910      0.004      -0.001      -0.000
departdelay:systemdelay     -0.0012   9.42e-05    -12.687      0.000      -0.001      -0.001


==============================================================================
Omnibus:                      180.263   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              543.319
Skew:                          -0.200   Prob(JB):                    1.05e-118
Kurtosis:                       4.859   Cond. No.                     4.71e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.71e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
MSE =  74.82618630551103
```

MLR Model with possible Interaction Effects:

$$\widehat{arrivedelay} = -16.2668 + 3.0765\ lowcost - 0.8327\ weekend - 0.0040\ distance + 0.7268\ departdelay + 0.3238\ airlinedelay + 0.6769\ systemdelay + 0.3973\ securitydelay + 0.3400\ aircraftdelay + 0.5150\ taxi_{in} + 0.5038\ taxi_{out} - 0.00008138\ departdelay * airlinedelay - 0.0004\ departdelay * aircraftdelay - 0.0012\ departdelay * systemdelay$$

Interpretation on the Constant Term:

- The average overall arrival delay for the airlines with high-cost carrier on weekdays is estimated to be -16.2668 minutes.

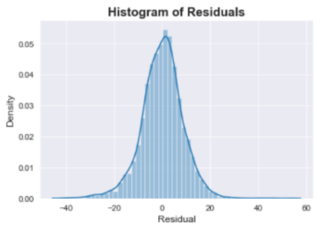
Interpretation on the Coefficients:

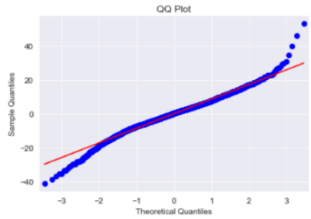
- **Baseline:** Airline is not a low-cost carrier and the day is not a weekend
- While holding other factors constant, the overall arrival delay for airlines with low-cost carrier is estimated to increase with additional of 3.0765, in minutes, that those airlines with high-cost carrier, on average.
- While holding other factors constant, the overall arrival delay on weekend is estimated to be -0.8327, in minutes, earlier than those on weekdays, on average.
- While holding other factors constant, for every additional of 1 km on distance travelled between two airports, the overall arrival delay is expected to be decrease by 0.0040, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on security delay, the overall arrival delay is estimated to be increase by 0.3973, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on taxi out, the overall arrival delay is estimated to be increase by 0.5038, in minutes, on average.
- While holding other factors constant, for every additional of 1 minute on taxi in, the overall arrival delay is estimated to be increase by 0.5150, in minutes, on average.
- The average overall arrival delay can be expected to change by (0.7268 – 0.00008138*airlinedelay), in minutes, when the overall depart delay increases with every additional of 1 minute, given airline delay (or vice-versa).
- The average overall arrival delay can be expected to change by (0.7268 – 0.004*aircraftdelay), in minutes, when the overall depart delay increases with every additional of 1 minute, given aircraft delay (or vice-versa).
- The average overall arrival delay can be expected to change by (0.7268 – 0.0012*systemdelay), in minutes, when the overall depart delay increases with every additional of 1 minute, given systemdelay (or vice-versa).

Evaluate Model on Training Data and Conduct Diagnostics

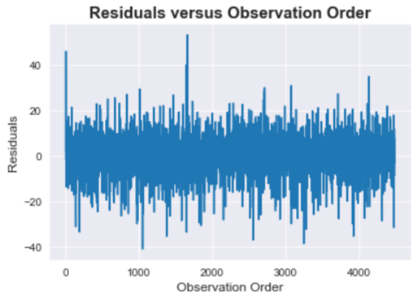
- **Coefficient of Determination, R^2** = 94.8%, indicating 94.8% of the total variability in the overall arrival delay can be accounted for by the MLR model. This model is in good fit.
- **Adjusted R^2** = 94.8%, which is the same as the R^2.
- **MSE =** 74.83%
- **F-test for overall model:** p-value ≈ 0.000 < 5%
  - At least one predictor contributes significantly to the model.
- **t-test for coefficients:**
  - The associated p-values of all quantitative coefficients (i.e. distance, depart delay, taxi_out, taxi_in, airline delay, system delay, aircraft delay, security delay) ≈ 0.000 < 5%, tested to be statistically significant predictors to the overall arrival delay.
  - The associated p-values of all interaction coefficients (i.e. depart delay and airline delay, depart delay and system delay, depart delay and aircraft delay) ≈ 0.000 < 5%, tested to be statistically significant predictors to the overall arrival delay.
  - The associated p-value of weekend coefficient ≈ 0.012 < 5% tested to be statistically significant predictors and there is a statistical evidence of difference in overall arrival delay between the weekend and non-weekend.
  - The associated p-value of low-cost coefficient ≈ 0.000 < 5% tested to be statistically significant predictors and there is a statistical evidence of difference in overall arrival delay between the low-cost and high-cost carriers.
- **Checking for Normality:** QQ plot seem to indicate that the residuals are not normally distributed as the two tails end deviate from the probability line, and the outliers are very evident. But however, the following tests also indicate that the residuals are not normal. Thus, normality assumption is not valid.

Omnibus Test	**Anderson-Darling Test**	**Jacque-Bera Test**	**Shapiro-Wilks Test**
p-value ≈ 0.000 < 5%	test-statistic = 10.56 > critical value = 0.786	p-value ≈ 0.000 < 5%	p-value ≈ 0.000 < 5%

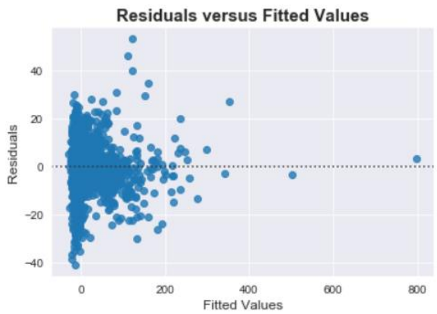




- **Checking for Independence:** The plot of residuals by observation order shows that the residuals are in random pattern, and no systematic pattern is observed. Furthermore, the Durbin-Watson Test statistic = 1.963 ≈ 2. Therefore, the residuals are not auto-correlated. Thus, independence assumption is valid.



- **Checking for Homoscedasticity:** The plot of residuals vs. fitted values shows that residuals are not randomly dispersed, suggesting that the residuals are non-constant variance. Furthermore, the Breusch-Pagan test (p-value ≈ 0.000), confirms that the residuals are heteroscedastic. Thus, constant variance assumption is not valid. From the plot, it can be observed that there are a couple of outliers could influence the regression analysis.



- **Checking for Multicollinearity:** The Variance Inflation Factors (VIF) show that there are some multicollinearity issue (VIF > 5 or VIF > 10) within the regression function. Thus, multicollinearity assumption is not valid.

Predictor Variable	VIF	Predictor Variable	VIF	Predictor Variable	VIF
lowcost	1.124	airlinedelay	8.253	departdelay*systemdelay	2.909
weekend	1.005	systemdelay	3.473	departdelay*airlinedelay	3.249
distance	1.086	securitydelay	1.018	departdelay*aircraftdelay	6.104
departdelay	11.000	aircraftdelay	10.245		
taxi out	1.361	taxi in	1.103		

Evaluate & Deploy Model on Testing Data

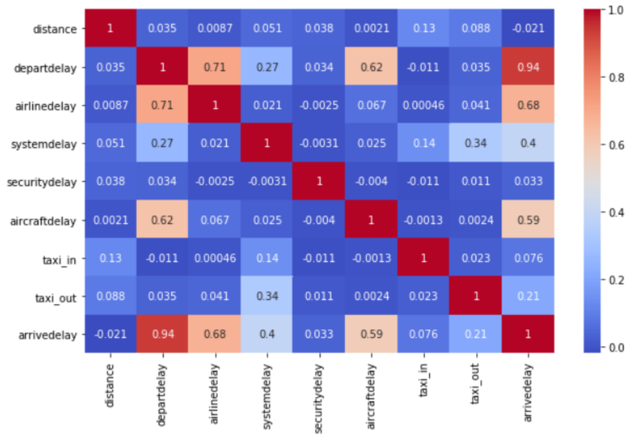
- **Testing Data:** R^2 = 93.43%; MSPE = 75.55%, similar as Training Data.
- Based on the above training data diagnostics, this regression model could not be deployed for prediction purpose with due to some strict regression criteria are not met. Although the goodness-of-fit and the predictors are showing satisfactory, and if it is chosen to deployed for prediction, then there will be a very high chance of getting prediction errors and leads to an unreliable analysis.
- This regression model can be seen to be further improved by transformation on variables, perform PCA analysis or re-investigate the whole dataset on the influence outliers, whichever is applicable.

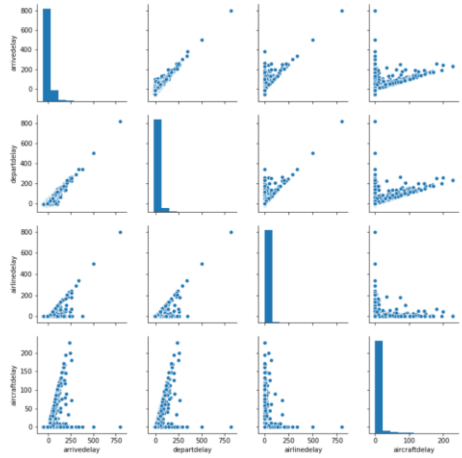
Model 3: Investigate Multicollinearity and perform principal component regression

Investigate Data-based Multicollinearity

- From Model 1, based on variance inflation factors (VIF), we see that the depart delay, airline delay and aircraft delay predictors are highly correlated with each other on arrival delay.
- This can also be seen from the following data visualization: heat map and pair plot.

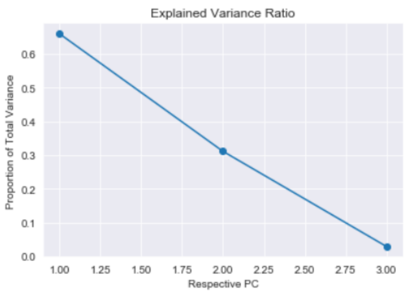
VIF Results of Model 1			
Predictor Variable	**VIF**	**Predictor Variable**	**VIF**
lowcost	1.115	taxi in	1.087
weekend	1.005	airlinedelay	5.641
distance	1.084	systemdelay	1.926
departdelay	10.344	securitydelay	1.017
taxi out	1.248	aircraftdelay	4.395





Perform Principal Component Regression

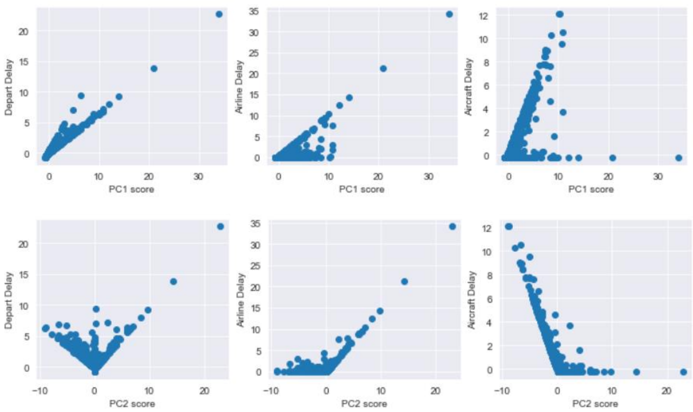
- Split the data into 80: 20, with seed 9567
- Actual ratio is 80.41%: 19.59%, that is n = 3607 data for training, and n = 879 for testing.
- Based on the proportion of total variance, PC1 (65.95%) capture the most variability/ information contained in the predictors, whereas PC3 (2.93%) capture little variability/ information contain in the predictors. PC2 (31.12%) capture about half of the PC1 of the variability/ information contained in the predictors.



Predictors	**Principal Components:**		
	PC1	**PC2**	**PC3**
Depart Delay	0.6946	0.0093	0.7194
Airline Delay	0.5367	0.6591	-0.5267
Aircraft Delay	0.4709	-0.7520	-0.4528

Standardized Mean: Depart Delay (9.717), Airline Delay (3.495), Aircraft Delay (4.408)
Standardized SD: Depart Delay (35.695), Airline Delay (23.326), Aircraft Delay (18.532)

- Plot of respective PC scores with respective predictors



- A **stepwise regression** was applied and the output of the optimal regression on the training data are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            arrivedelay   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                     6215.
Date:                Thu, 06 Feb 2020   Prob (F-statistic):               0.00
Time:                        15:54:20   Log-Likelihood:                -12985.
No. Observations:                3607   AIC:                         2.599e+04
Df Residuals:                    3596   BIC:                         2.606e+04
Df Model:                          10
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept         -8.2315      0.513    -16.057      0.000      -9.237      -7.226
weekend[T.Yes]    -0.7944      0.338     -2.352      0.019      -1.457      -0.132
lowcost[T.Yes]     3.4481      0.312     11.055      0.000       2.837       4.060
distance          -0.0039      0.000    -15.717      0.000      -0.004      -0.003
systemdelay        0.4956      0.016     30.373      0.000       0.464       0.528
securitydelay      0.4292      0.112      3.838      0.000       0.210       0.648
taxi_in            0.5618      0.031     17.886      0.000       0.500       0.623
taxi_out           0.5743      0.019     31.040      0.000       0.538       0.611
pc1               24.0927      0.107    225.106      0.000      23.883      24.303
pc2                0.5796      0.153      3.787      0.000       0.280       0.880
pc3               10.9278      0.640     17.075      0.000       9.673      12.183
==============================================================================
Omnibus:                      227.727   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              985.878
Skew:                          -0.108   Prob(JB):                    8.31e-215
Kurtosis:                       5.552   Cond. No.                     4.77e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
MSE =  78.66892207078739
```

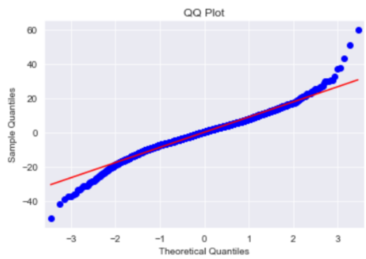
MLR with Principal Component Regression:

$$\widehat{arrivedelay} = -8.2315 + 3.4481\ lowcost - 0.338\ weekend - 0.0039\ distance + 0.5743\ taxi_{out} + 0.5618\ taxi_{in} + 0.4956\ systemdelay + 0.4292\ securitydelay + 24.0927\ pc1 + 0.5796\ pc2 + 10.9278\ pc3$$

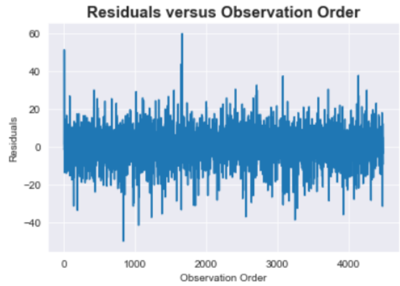
Evaluate Model on Training Data and Conduct Diagnostics

- **Coefficient of Determination, R^2** = 94.5%, indicating 94.5% of the total variability in the overall arrival delay can be accounted for by the MLR model. This model is in good fit.
- **Adjusted R^2** = 94.5%, which is the same as the R^2.
- **MSE =** 78.67%
- **F-test for overall model:** p-value ≈ 0.000 < 5%
    - At least one predictor contributes significantly to the model.
- **t-test for coefficients:**
    - The associated p-values of all quantitative coefficients (i.e. distance, taxi_out, taxi_in, systemdelay, securitydelay) ≈ 0.000 < 5%, tested to be statistically significant predictors to the overall arrival delay.
    - The associated p-values of all PC coefficients (i.e. pc1, pc2, and pc3) ≈ 0.000 < 5%, tested to be statistically significant predictors to the overall arrival delay.
    - The associated p-value of weekend coefficient ≈ 0.019 < 5% tested to be statistically significant predictors and there is a statistical evidence of difference in overall arrival delay between the weekend and non-weekend.
    - The associated p-value of low-cost coefficient ≈ 0.000 < 5% tested to be statistically significant predictors and there is a statistical evidence of difference in overall arrival delay between the low-cost and high-cost carriers.
- **Checking for Normality:** QQ plot seem to indicate that the residuals are not normally distributed as the two tails end deviate from the probability line, and the outliers are very evident. But however, the following tests also indicate that the residuals are not normal. Thus, normality assumption is not valid.

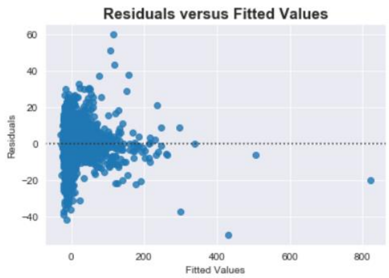
Omnibus Test	**Anderson-Darling Test**	**Jacque-Bera Test**	**Shapiro-Wilks Test**
p-value ≈ 0.000 < 5%	test-statistic = 13.05 > critical value = 0.786	p-value ≈ 0.000 < 5%	p-value ≈ 0.000 < 5%



- **Checking for Independence:** The plot of residuals by observation order shows that the residuals are in random pattern, and no systematic pattern is observed. Furthermore, the Durbin-Watson Test statistic = 1.960 ≈ 2. Therefore, the residuals are not auto-correlated. Thus, independence assumption is valid.



- **Checking for Homoscedasticity:** The plot of residuals vs. fitted values shows that residuals are not randomly dispersed, suggesting that the residuals are non-constant variance. Furthermore, the Breusch-Pagan test (p-value ≈ 0.000), confirms that the residuals are heteroscedastic. Thus, constant variance assumption is not valid. From the plot, it can be observed that there are a couple of outliers could influence the regression analysis.



- **Checking for Multicollinearity:** The variance inflation factors (VIF) show that there is no multicollinearity issue (VIF < 5 or VIF < 10) within the regression function. Thus, multicollinearity assumption is valid.

VIF Results of Model 3					
Predictor Variable	**VIF**	**Predictor Variable**	**VIF**	**Predictor Variable**	**VIF**
lowcost	1.11	taxi_in	1.09	pc1	1.04
weekend	1.00	taxi_out	1.25	pc2	1.00
distance	1.08	systemdelay	1.93	pc3	1.65
securitydelay	1.02				

<u>Evaluate & Deploy Model on Testing Data</u>

- **Testing Data:** R^2 = 93.2%; MSPE = 78.7%, similar as Training Data.
- Based on the above training data diagnostics, this regression model could not be deployed for prediction purpose with due to some strict regression criteria are not met. Although the goodness-of-fit and the predictors are showing satisfactory, and if it is chosen to deployed for prediction, then there will be a very high chance of getting prediction errors and leads to an unreliable analysis.
- This regression model can be seen to be further improved by using modelling approach as well as re-investigate the whole dataset on the influence outliers, whichever is applicable.