Harish Morekonda
M12728920

BANA5043 - Statistical Computing - FAA Landing Overrun Analysis

executive summary

The goal of this analysis is to study the impact of factors that could be causing landing overruns. 950 rows of data with flight information of two different aircrafts is examined, and filtered based on the conditions in the variable dictionary. The distribution and spread of all estimator variables are graphically explored and found to be closely normal. Correlation tests are run to show that the no_passenger, duration, height, and pitch are independent of landing distance. Evaluating the aircraft makes against the distance indicates that airbus flights have a significantly lower landing distance and pitch. Tests also show that the variables speed_ground and speed_air variables contribute the most. The regression model for the landing distance predictor variable is built with the two velocity variables and has an r^2 of 0.7479. Introducing the height and pitch as dependent variables slightly increases the r^2, but renders the speed_ground variable unresponsive.

direct answers to questions

1. **How many observations (flights) do you use to fit your final model? If not all 950 flights, why**

   The model uses 840 records. Please see page 14 for a detailed breakdown of row deletion at different steps.

2. **What factors and how they impact the landing distance of a flight?**

   The landing distances are affected linearly by speed_air and speed_ground. Also to keep in mind is the height, which should not be below 6 meters. Please see page 10-13 for more thorough analysis.

3. **Is there any difference between the two makes Boeing and Airbus?**

   Yes, it is observed that the landing distances are distributed differently between the two aircraft makers, with a lower mean for airbus. The pitch is also slightly lower in airbus. Please see page 10 for more details.

open points

The first impressions on looking at the dataset is that it is too generic and contains very less information. Having answers to the following questions would help to understand and make better sense of the data.

weather

Weather can always be the biggest issue when it comes to landing overruns. Since the dataset does not contain any weather information, the model built using this dataset would assume ideal weather.

airport

Different airports have different runway configurations making it easier/difficult to land. The model will assume that the data was all collected from a single airport and hence it will apply to only that airport.

visibility

Visibility plays an important role and flights landing in the night is riskier than the ones landing with sufficient light. Fog could cause a big problem as well.

equipment malfunctions

The model we are building does not take into account any malfunctions of sensors and transducers. They are strictly out of scope of the model.

pilot error-human factor

Some pilots are better than others. But since the experience/expertise is not taken into account, the model assumes that all the data was collected from a single pilot.

runway contamination

Research suggests that runway contamination is the largest contributor of runway overruns in modern planes. Again, this factor is out of scope of the model.

data preparation

importing and initial quality check

To start off, the two Excel files are imported using PROC IMPORT statement into two datasets and then combined together into one larger dataset.

```
proc import datafile="/folders/myfolders/statcomp/faa2.xlsx"
            out=statcomp.excelimport2 dbms=xlsx replace;
run;
```

On importing, it's noticed 50 empty rows with null values from file2 require to be deleted. This is done while combining the data with an if statement to check the number of passenger and distance variables.

```
data statcomp.faacombinedBackUp;
      set statcomp.excelimport1 statcomp.excelimport2;
      if no_pasg='.' & distance = '.' then
            delete;
run;
```

Also deleted are 100 duplicate lines across the data from the two files.

```
/* Remove duplicate lines */
proc sort data=statComp.faaCombinedbackup nodupkey;
      by aircraft height pitch speed_ground distance;
run;
```

The 850 datalines are hardcoded with a serial number to track the datalines and copied to a back up variable.

handling missing values

To look for missing values across the dataset, proc means statement with nmiss argument is used.

```
proc means data=statComp.faaCombined nmiss;
run;
```

The results show that there exists 50 lines with missing duration and 642 lines with missing air speed of the aircraft.

The MEANS Procedure

Variable	Label	N Miss
duration	duration	50
no_pasg	no_pasg	0
speed_ground	speed_ground	0
speed_air	speed_air	642
height	height	0
pitch	pitch	0
distance	distance	0
serialNo		0

speed_air

From the variable dictionary, if it can be assumed that the speed_air and speed_ground variables are related and indicate the same behavior of the aircraft, then we can still work with datalines with missing speed_air variable. This has to be verified with the provider of the dataset.

This assumption should be confirmed statistically as well.

duration

The entries present in file2 do not contain data for duration variable. This could be because the collected file did not have a means to include the duration data. But, it could also mean that the flight did not take off. A bad landing could have also hypothetically damaged the sensor which lead to blank values. The variable dictionary also specifies that the flight should always be greater than 40 minutes. Ignoring the entire file of data does not make sense as well. For now, these lines are to be considered for the final model, but this could be changed based on what the data owner says.

validity check

aircraft

There are 400 lines of Boeing and 450 lines of Airbus lines in the dataset, after removing the duplicate entries

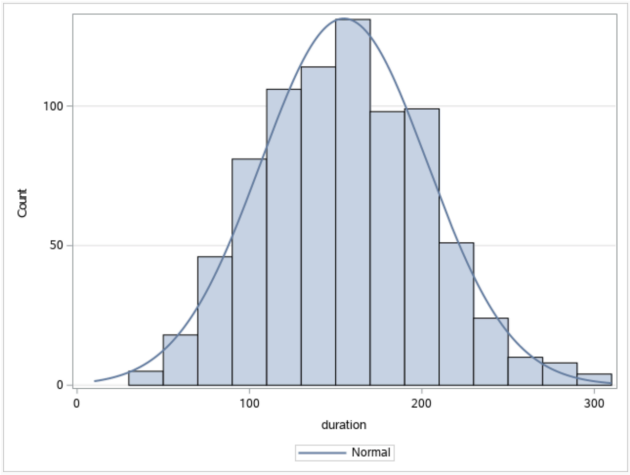
duration

On sorting the data based on duration, it's observed that there are five datalines with durations less than 40. Following the dictionary requirements, these 5 entries are to be deleted.

The duration, when plotted as a histogram looks normally distributed with mean at about 151 minutes. There doesn't appear to be any outliers as the maximum is about 300 minutes, a normal flight time.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14.7642	831	289.320	18
16.8935	346	293.230	722
17.3755	634	298.522	590
31.3910	845	302.967	345
31.7017	234	305.622	56

Moments			
N	800	Sum Weights	800
Mean	154.006538	Sum Observations	123205.231
Std Deviation	49.2592338	Variance	2426.47211
Skewness	0.12147943	Kurtosis	-0.0551851
Uncorrected SS	20913162.3	Corrected SS	1938751.22
Coeff Variation	31.9851574	Std Error Mean	1.74157691

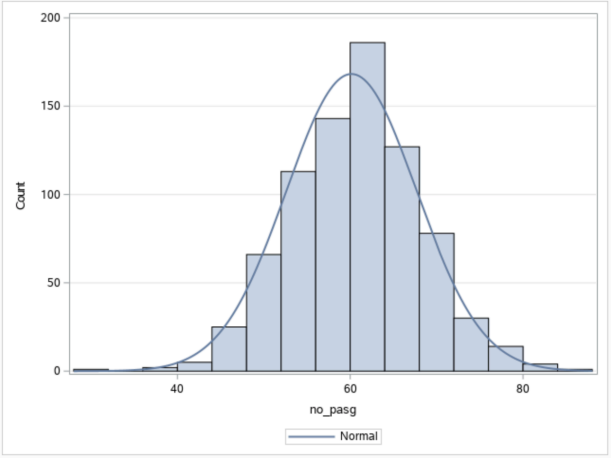



```
data statComp.faaCombinedStep2;
      set statComp.faacombinedstep1;
      if duration <> '.' then
            if duration < 40 then
                  delete;
run;
proc univariate data=statComp.faaCombinedStep1;
      var duration;  run;
proc sgplot data=StatComp.FaaCombinedStep1;
      histogram duration / scale=count;
      density duration;
      yaxis grid; run;
```

no_pasg

Nothing looks out of ordinary. The data is again spread out normally at an average of 60 passengers.

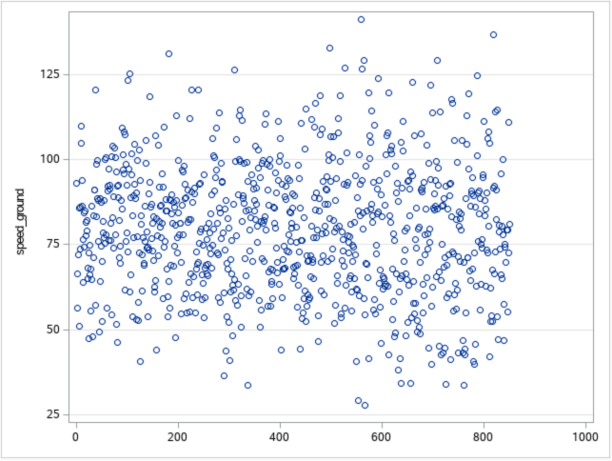
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
29	604	80	378
36	24	80	743
38	338	82	271
40	194	82	547
41	355	87	410



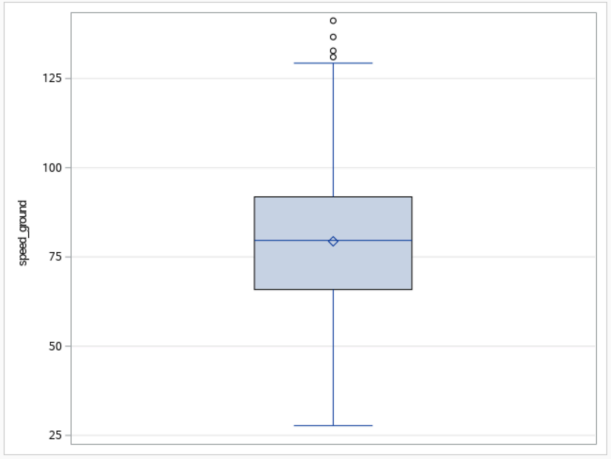
speed_ground

On throwing the data into a scatter plot, it's observed that the speed is spread between 27 mph, all the way up to 141 mph.

Moments			
N	845	Sum Weights	845
Mean	79.400624	Sum Observations	67093.5273
Std Deviation	19.0638891	Variance	363.431868
Skewness	0.12145975	Kurtosis	-0.0956391
Uncorrected SS	5634004.43	Corrected SS	306736.497
Coeff Variation	24.0097472	Std Error Mean	0.65581772



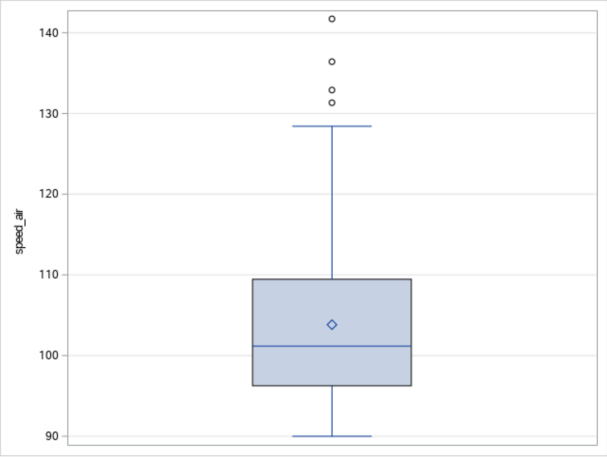
The box plot however, does show some outliers. As stated in the dictionary, any speed below 30 mph and above 140 mph could cause abnormal landings. There are three datalines with this outliers.

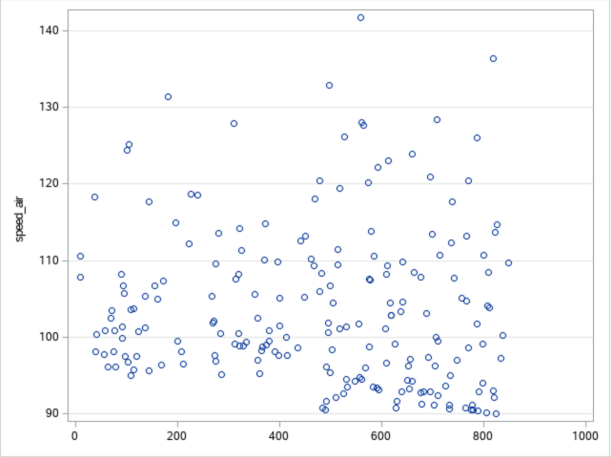


speed_air

Applying the same steps, we observe that the lowest speed on air is 90 mph and crosses 140 mph limit once.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
90.0029	822	128.418	705
90.1110	803	131.338	181
90.3674	785	132.911	495
90.4767	777	136.423	815
90.5033	487	141.725	557

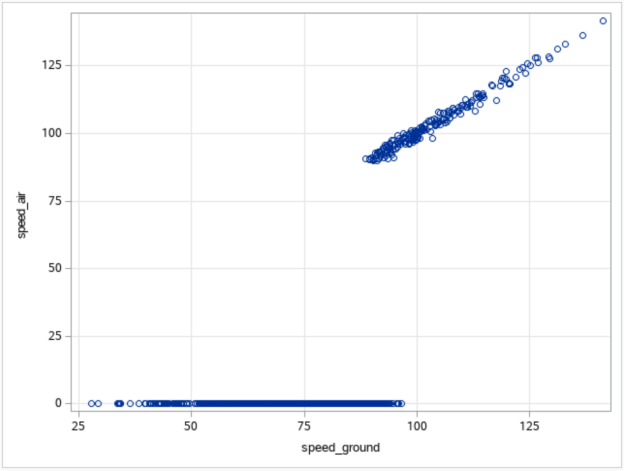




comparing speed_ground and speed_air

On graphically exploring both the speed variables, it's observed that speed_air is empty for speed_ground less than around 90 mph. This could indicate a strong correlation between the two variables and there could be a specific reason for blank speed_air cells.

The reason for this observation has to be explored thoroughly since it could throw valuable insights.

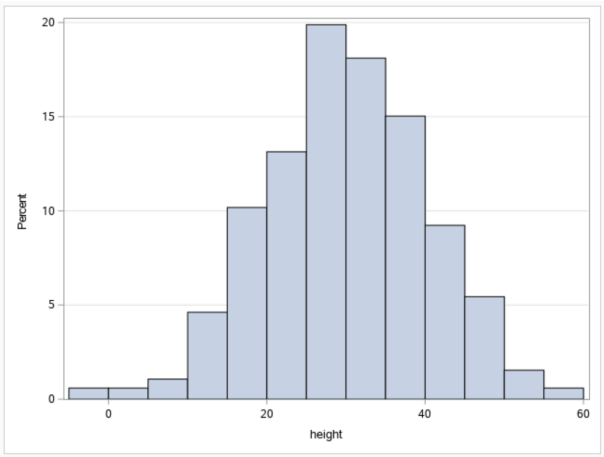


Both the speed variables contain very few outliers above the 120 mph limit. Although the dictionary says that 140 mph is the limit, bringing it down to 120 mph could help build a safer model that reduces the risk of overruns.

height

Throwing the height data on a scatter plot shows negative entries, which clearly does not sound plausible. The 5 lines with negative values should be ignored from the final model. Also, according to the dictionary, a height below 6 meters can cause landing problems and we notice 5 datalines that matches this criterion. Otherwise, the data looks normally distributed around a mean height of around 31 meters.

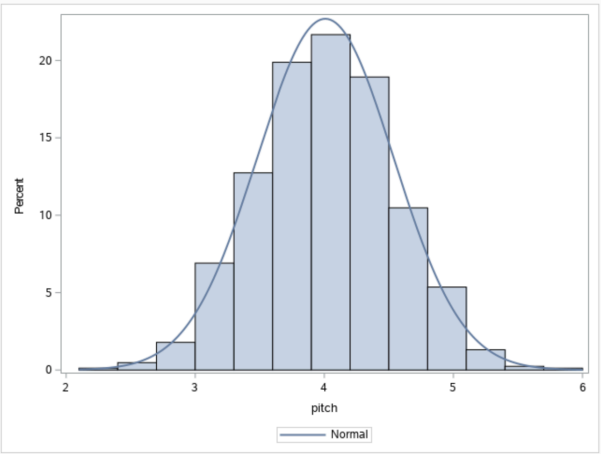
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-3.5462524	449	55.0935	842
-3.3323880	1	58.0818	843
-2.9153359	2	58.0835	844
-1.5281292	450	58.2278	448
-0.0677586	3	59.9460	845



pitch

Nothing abnormal is observed. The mean of the pitch data is about 4 degrees with a standard deviation of 0.52 degrees.

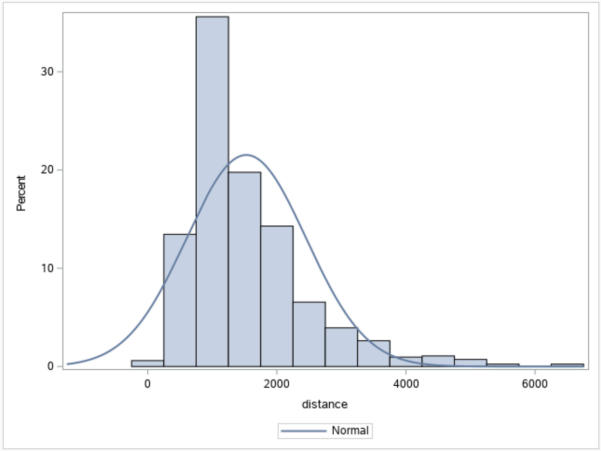
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2.28448	284	5.31068	571
2.66891	399	5.32475	805
2.67133	201	5.52678	30
2.67599	379	5.55640	819
2.68955	128	5.92678	470



distance

The distance between the threshold of the runway and the point where the aircraft can be fully stopped varies between 34 feet to 6400 feet, with the majority of the data around 1000 feet. Clearly there are a lot of outliers.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
41.7223	29	5147.41	779
133.0869	75	5343.20	490
180.5652	25	5381.96	700
241.1610	177	6309.95	810
242.5959	3	6533.05	552



The dictionary says that airport runways are typically 6000 feet long. Taking this definition at its word would mean that any row with a distance below 6000 feet landed before the strip even starts. At the same time, the lower this number, the closer it got to going out the runway.

Setting the limits at 200 feet on the lower end and 6000 m on the upper end, we see 5 datalines that fall out of this range. While modeling the algorithm, these data lines with serial numbers 559, 818, 32, 78, 28 would be great indicators for overrun prediction.

dependency tests

Testing each variable against the target variable to check for relations between the two.

aircraft

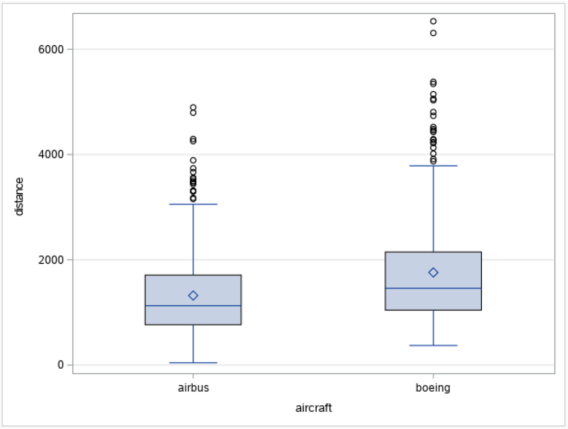
To test if the distances vary across Airbus and Boeing, a ttest is run. Comparing the 445 airbus and 395 Boeing records, it's 99.99% certain that the mean distances are vary between the two. The Boeing flights have an average 440 feet higher landing distance than airbus.

aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		445	1321.6	791.9	37.5402	41.7223	4896.3
boeing		395	1759.8	1009.1	50.7742	371.3	6533.0
Diff (1-2)	Pooled		-438.3	900.6	62.2565		
Diff (1-2)	Satterthwaite		-438.3		63.1450		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	838	-7.04	<.0001
Satterthwaite	Unequal	744.95	-6.94	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	394	444	1.62	<.0001

On exploring graphically, it's clear that airbus aircrafts have a lower distribution compared to Boeing.



It is also observed that the airbus flights stop closer to the end of the runway (distance too low) and the Boeing flights land way before the runway strip starts (distance>6000).

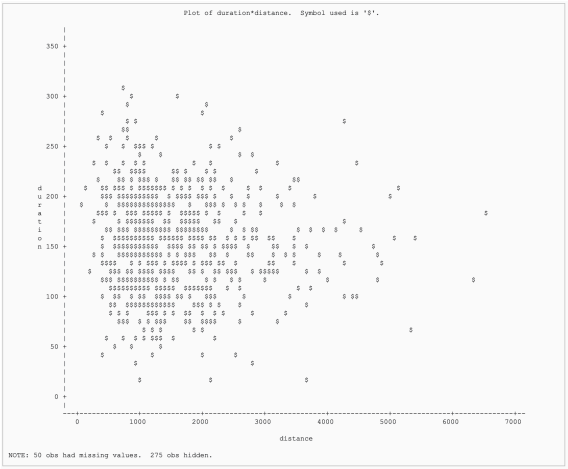
Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	serialNo
1	airbus	128.37336566	64	55.461625107	.	14.65127605	3.9792117538	180.56522534	28
2	airbus	190.7394255	77	47.882117055	.	14.835964361	2.7322842836	41.722312733	32
3	airbus	212.05403613	63	51.587044527	.	20.451285811	3.063686215	133.08690985	78
4	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506	559
5	boeing	119.92455279	64	136.65915832	136.42342138	44.286109179	4.1694037368	6309.9459762	818

duration

To test the dependancies of duration and landing distance, a correlation test is run. As evident in the x-y plot below, the duration and the landing distances are not related.

The correlation coefficient is too small at -0.05287 to be related. The p-value is greater than 0.05, which tells us to accept the null hypothesis that there is no correlation. The imaginary points are too spread out around the linear line.

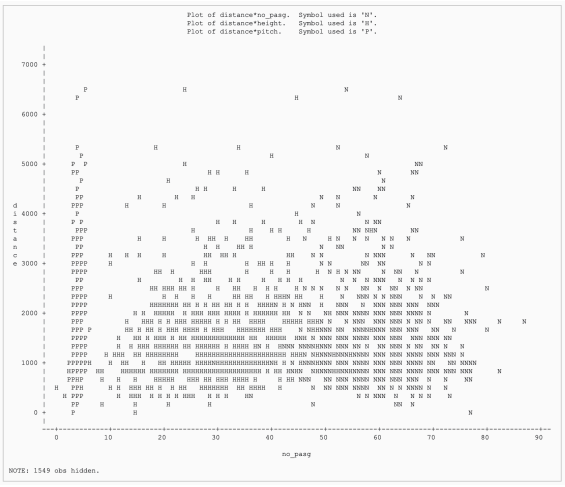
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	duration
distance distance	-0.05287 0.1376 790



no_pasg, height, pitch

Running correlation tests for number of passengers, height and pitch, it's pretty clear that they do not influence the landing distance.

```
proc corr data=statComp.faaCombinedStep3;
      var no_pasg height pitch;
      with distance;
run;
proc plot data=statComp.faaCombinedStep3;
      plot distance*no_pasg='N'
           distance*height='H'
           distance*pitch='P'/overlay;
run;
```



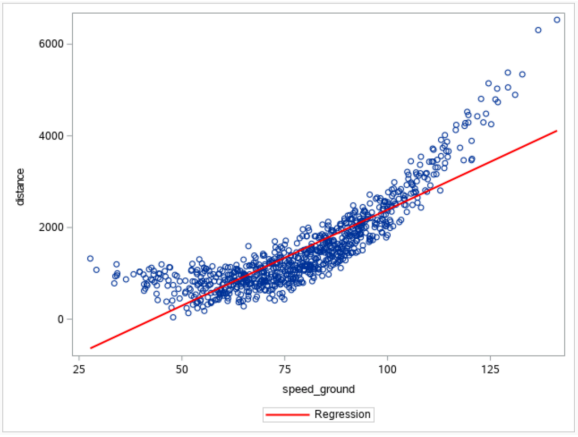
Pearson Correlation Coefficients, N = 840 Prob > |r| under H0: Rho=0			
	no_pasg	height	pitch
distance distance	-0.02457 0.4770	0.11543 0.0008	0.09726 0.0048

The correlation coefficient numbers do not show any statistical significance. The p-value is less than 0.05 for all three parameters. However, one must keep in mind that the height has a 6 meter limitation for a safe landing.

speed_ground

Evidently, speed_ground and landing distance are related as it can be seen from the results of correlation check and scatter plot. The correlation factor is pretty high and the dependent values are densely packed.

Pearson Correlation Coefficients, N = 840 Prob > |r| under H0: Rho=0	
	speed_ground
distance distance	0.86167 <.0001



Another insight that one can see in the distribution graph is that at extreme values of speed ground, the distance dependence become more pronounced.

speed_air

As it can be seen in the results below, speed_air very much influences the landing distance of the flight.

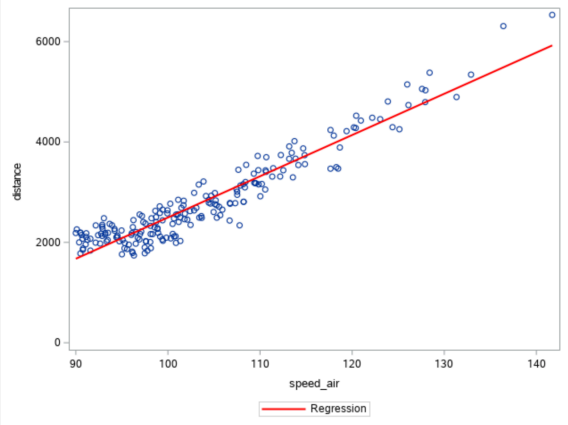
But, it has already been observed that, of the 840 datalines that we are have, only 205 lines contain speed_air. With this high a correlation, it makes sense to keep the speed_air variable for analysis when developing the model.

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	840	1528	926.27751	1283220	41.72231	6533	distance
speed_air	205	103.83225	10.30684	21286	90.00286	141.72494	speed_air

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	speed_air
distance	0.94787
distance	<.0001
	205



As seen in the scatter plot above, the speed_air and distance variables have a very high linear relationship.

data engineering

As seen earlier above, speed_air and speed_ground are linearly correlated.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	speed_air
speed_ground speed_ground	0.98930 <.0001 205

We have also established that speed_air is a dependent variable to predict overruns. Thus to make use of speed_air more effectively, it can be calculated from speed_ground for rows with null values. This is done through regression analysis.

```
proc reg data=statComp.faaCombinedStep3;
    model speed_air=speed_ground;
run;
```

205 observations, out of the 840 rows were used to generate the plot and as expected, we get a very linear relationship with an r^2 0.9787.

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	2.58580	1.05339	2.45	0.0149
speed_ground	speed_ground	1	0.97580	0.01010	96.60	<.0001

This model is used to replace null values of speed_air.

The univariate processes before and after engineering is available below. The mean has decreased significantly and this was expected as speed_air had null values only for speed_ground less than 90mph.

Moments			
N	205	**Sum Weights**	205
Mean	103.832246	**Sum Observations**	21285.6105
Std Deviation	10.3068422	**Variance**	106.230995
Skewness	1.04901428	**Kurtosis**	0.87026547
Uncorrected SS	2231803.87	**Corrected SS**	21671.123
Coeff Variation	9.92643665	**Std Error Mean**	0.71986108

Moments			
N	840	**Sum Weights**	840
Mean	80.1308288	**Sum Observations**	67309.8962
Std Deviation	18.6218748	**Variance**	346.774222
Skewness	0.12636849	**Kurtosis**	-0.0576897
Uncorrected SS	5684541.34	**Corrected SS**	290943.573
Coeff Variation	23.2393389	**Std Error Mean**	0.64251572

This can be observed again when we look at the extreme values. Only the lower portion of the data has been replaced.

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
90.0029	817	128.418	700
90.1110	798	131.338	178
90.3674	780	132.911	490
90.4767	772	136.423	810
90.5033	482	141.725	552

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
29.6503	560	128.418	700
31.1061	547	131.338	178
35.3474	332	132.911	490
35.5902	753	136.423	810
35.8779	718	141.725	552

Following what we have observed above and the information available in the data dictionary, there are four variables that could indicate landing overruns.

Variable	Conditions	Number of rows
distance	If > 6000 or < 200	5
speed_ground	If > 140 or < 30	3
speed_air	If > 140 or < 30	2
height	If < 6 and > 0	5

Using these conditions, a new categorical variable overrunYesNo is created. It's observed that there are 12 lines that returned a TRUE.

```
data statComp.FaaCombinedStep4;
    set statComp.faaCombinedStep3;
        overrunYesNo='FALSE';
    if distance < 200 or distance > 6000 then
        overrunYesNo='TRUE';
    if speed_ground < 30 or speed_ground > 140 then
        overrunYesNo='TRUE';
    if speed_air < 30 or speed_air2 > 140 then
        overrunYesNo='TRUE';
    if height < 6 then
        overrunYesNo='TRUE';
run;
```

Before going ahead and building the model, below is a review of the datelines deleted during different steps of the analysis.

1. Started with 950 rows.
2. Deleted 100 duplicate records - 850
3. Deleted 5 rows where duration is less than 40 minutes - 845
4. Deleted 5 rows where height was negative - 840

Final model will have 840 rows.

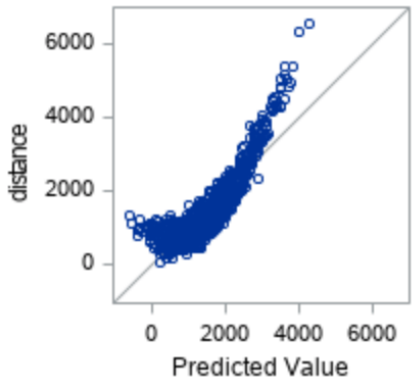
modelling

The distance directly depends upon the ground and air velocities, with a limitation on height at 6 meter. The regression model will hence only contain the velocities as estimators.

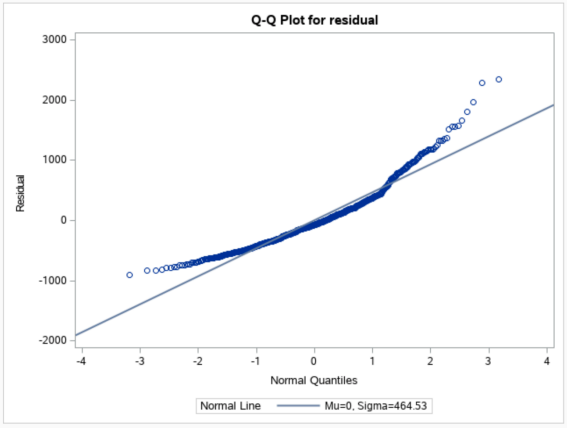
```
proc reg data=statComp.faaCombinedStep5;
  model distance = speed_ground speed_air;
run;
```

The generated model has an r^2 value of 0.7479, with loss in precision when the speed is too low or high, as we already expected it to happen while checking dependencies.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2049.28340	88.71689	-23.10	<.0001
speed_ground	speed_ground	1	-52.72828	21.14420	-2.49	0.0128
speed_air	speed_air	1	96.93086	21.65153	4.48	<.0001

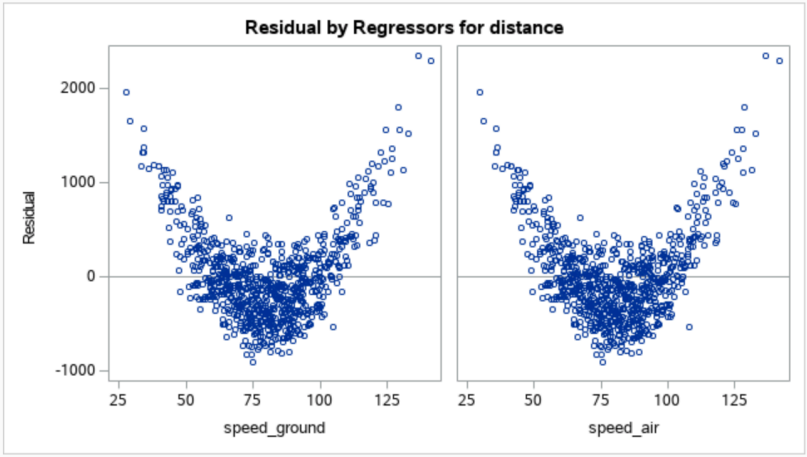


Running a Q-Q and normality tests on the residual data shows that the residuals are not normally distributed. Although the Q-Q plot may almost look straight, the test for normality contains p-value all less than 0.05, which indicates that the residual errors are not normally distributed.



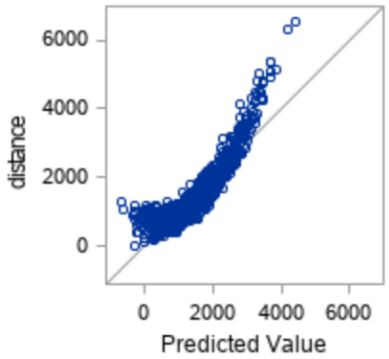
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.929902	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.084489	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.022038	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	12.94733	Pr > A-Sq	<0.0050

The residual plot also indicates a high variance in the data.
On a positive note, the mean of the residuals is zero.



Although it has been established that height and pitch do not play a role in the model, it is tempting prospect to include them to see if the model improves. Surprisingly, the r^2 value does improve to 0.7826

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-3296.36351	146.22588	-22.54	<.0001
speed_ground	speed_ground	1	-38.06407	19.68998	-1.93	0.0536
speed_air	speed_air	1	82.35832	20.15991	4.09	<.0001
height	height	1	13.16833	1.49401	8.81	<.0001
pitch	pitch	1	212.22676	28.36391	7.48	<.0001



The residuals, have a mean zero and fail the normality test. The variances are larger for the velocity variables and comparatively smaller for the height and pitch. This model looks more promising than the previous one.

```
proc univariate data=statComp.faaResiduals normaltest;
var residual;
qqplot residual/Normal(mu=est sigma=est color=red l=1);
run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.897113	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.105625	Pr > D	<0.0100
Cramer-von Mises	W-Sq	3.367659	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	20.84926	Pr > A-Sq	<0.0050

The problem with the model, however, is that the speed_ground variable takes a p-value greater than 0.05, meaning it no longer exerts a strong correlation. This is against what we have observed so far, casting doubts on the strength of the model.

The increased R^2 value is only minimal and the residual normality test pass can be attributed to smaller parameter estimates of height and pitch. Taking into consideration the height and pitch would probably be overfitting the model.

Thus the previous model with only the velocity variables as estimators would be the best fit for the given dataset.