FLIGHT DELAY PREDICTION

統碩一 賴東昇 2018.01.04

目錄

使用資料

資料集1 : 航班資料

資料概述 : 2008年美國國內航班之統計資料，計7,009,728筆

資料出處 : U.S. Department of Transportation's Bureau of Transportation Statistics

資料集2 : 氣象資料

資料概述 : 航空站之歷史氣象資料，去除遺漏值後，每天每地有20~74次紀錄

資料出處 : https://www.wunderground.com/

變數概觀 (航班)

Name	Description	Name	Description	Name	Description
1	Year	11	TailNum	21	TaxiOut
2	Month	12	ActualElapsedTime	22	Cancelled
3	DayofMonth	13	CRSElapsedTime	23	CancellationCode
4	DayOfWeek	14	AirTime	24	Diverted
5	DepTime	15	ArrDelay	25	CarrierDelay
6	CRSDepTime	16	DepDelay	26	WeatherDelay
7	ArrTime	17	Origin	27	NASDelay
8	CRSArrTime	18	Dest	28	SecurityDelay
9	UniqueCarrier	19	Distance	29	LateAircraftDelay
10	FlightNum	20	TaxiIn		

資料概觀 (氣象)

Hourly Weather History & Observations

時間 (EDT)	溫度	露點	濕度	氣壓	能見度	Wind Dir	風速	最大陣風	Precip	天氣情況	狀況
12:52 AM	15.6 °C	14.4 °C	93%	1012.8 百帕	14.5 公里	南	5.6 公里每小時 / 1.5 m/s	-	0.8 毫米		陰天
	METAR KATL 050452Z 17003KT 9SM FEW005 SCT055 BKN070 OVC200 16/14 A2992 RMK AO2 RAE28 SLP128 P0003 T01560144 402170078										
1:52 AM	15.6 °C	14.4 °C	93%	1012.2 百帕	8.0 公里	靜止	靜止	-	0.0 毫米		陰天
	METAR KATL 050552Z 00000KT 5SM BR SCT001 BKN050 OVC100 16/14 A2990 RMK AO2 RAB17E37 SLP122 P0000 60041 T01560144 10172 20150 58006										
2:04 AM	15.0 °C	14.0 °C	94%	1012.1 百帕	6.4 公里	東南偏東	7.4 公里每小時 / 2.1 m/s	-	N/A		陰天
	SPECI KATL 050604Z 11004KT 4SM BR BKN001 BKN050 OVC100 15/14 A2989 RMK AO2										
2:28 AM	15.0 °C	14.0 °C	94%	1012.4 百帕	4.8 公里	靜止	靜止	-	N/A		陰天
	SPECI KATL 050628Z 00000KT 3SM BR SCT001 OVC050 15/14 A2990 RMK AO2										

Source : https://www.wunderground.com/history/airport/KATL/2008/4/5/DailyHistory.html?req_city=KATL

定義問題

DepTime
實際起飛時間

DepTime
實際抵達時間

CRSDepTime
預計起飛時間

CRSArrTime
預計抵達時間

定義問題

DepTime
實際起飛時間

DepTime
實際抵達時間

CRSDepTime
預計起飛時間

CRSArrTime
預計抵達時間

預測目標：在 預計起飛 的時間點，預測班機是否抵累

CRSDepTime

研究方法

針對所有航班紀錄
建立預測模型(120萬筆)

利用爬蟲抓取天氣資料
並與航班資料合併

選取 ATL, WRD 等前5大機場
利用天氣資料建立預測模型(25萬筆)

結合兩者，建立最終模型

研究方法

針對所有航班紀錄
建立預測模型

利用爬蟲抓取天氣資料
並與航班資料合併

選取 ATL, WRD 等前5大機場
利用天氣資料建立預測模型(25萬筆)

結合兩者，建立最終模型

資料縮減(1)

去除ArrDelay為遺漏值的觀測值(154,699筆)，接著以分層隨機抽樣的方式

每月各抽10萬筆，計120萬筆觀測值，作為第一階段建模依據

資料縮減(2)

接著選取飛機到站數最多的五個機場 : ATL,DEN DFW, LAX, ORD，計256004筆資料

利用當地天氣資訊建模

地圖 : https://hyades910739.github.io/flight_research/Airport_Map.html

資料合併

以CRSDepTime為基準，以最近一次的觀測資料作為當地之氣象資料

航班資料

Date	Origin	Dest	CRSDepTime	CRSArrTime
08-04-05	EWR	ATL	845	1118
08-04-05	IAD	ATL	1220	1420

氣象資料

Date	Time
08-04-05	0752
08-04-05	0852
…	…
08-04-05	12:12
08-04-05	12:34

資料合併

以CRSDepTime為基準，以最近一次的觀測資料作為當地之氣象資料

航班資料

Date	Origin	Dest	CRSDepTime	CRSArrTime
08-04-05	EWR	ATL	845	1118
08-04-05	IAD	ATL	1220	1420

氣象資料

Date	Time
08-04-05	0752
08-04-05	0852
…	…
08-04-05	12:12
08-04-05	12:34

變數性質(航班)

Name	Description	Info	Name	Description	
1	Year	日期變數	9	UniqueCarrier	航空公司，20類
2	Month		10	FlightNum	航線編號，7480類
3	DayofMonth		11	TailNum	飛機編號，5338類
4	DayOfWeek		17	Origin	起飛地，301類
5	DepTime	時間變數	18	Dest	目的地，301類
6	CRSDepTime		19	Distance	飛行距離
8	CRSArrTime		15	ArrDelay	抵累時間
13	CRSElapsedTime		NEW	DelayOver60	抵累超過60分鐘

時間變數
類別變數
連續變數

變數性質(天氣)

Name	Description	Info	Name	Description	
1	Date	日期變數	9	Gust.Speed	最大陣風
2	Time	時間變數	10	Precip	降水(mm)
3	Temp	溫度(攝氏)	11	Wind.Dir	風向
4	DewPoint	露點溫度(攝氏)	12	Events	天氣狀況
5	Humidity	濕度(%)	13	Conditions	狀況
6	Pressure	氣壓(hPa)	14	Dest	目的地
7	Visibility	能見度(Km)	15	ArrDelay	抵累時間
8	Wind.Speed	風速(Km/h)	NEW	DelayOver60	抵累超過60分鐘

時間變數
類別變數
連續變數

遺漏值處理 :

Case I : 大部分col皆為遺漏值

直接刪去該期資料

Case II : 少部分col為遺漏值

個別處理，若變動幅度不大則以前一期與後一期的平均值取代

變數研究 : ArrDelay

Summary

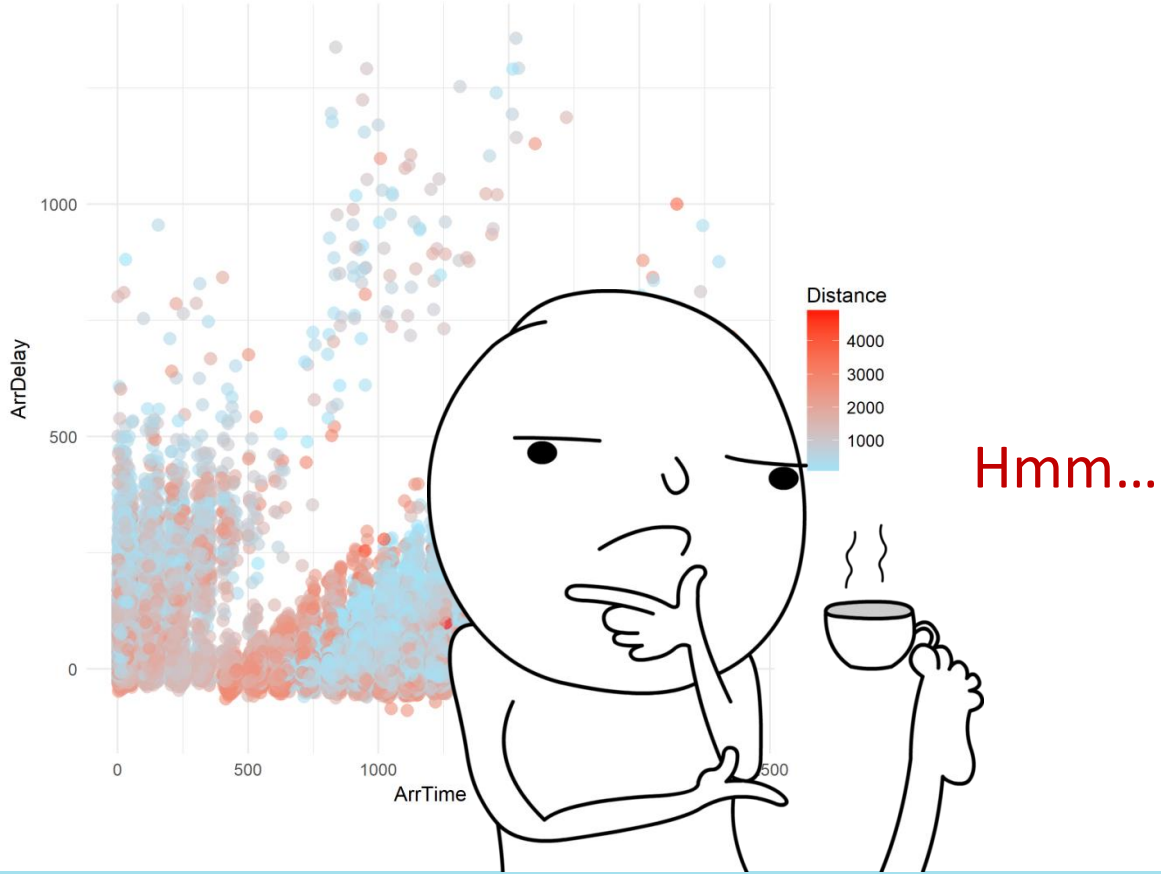
Min	Q1	Median	Mean	Q3	Max
-109	-10	-2	8.07	12	1357

percentile

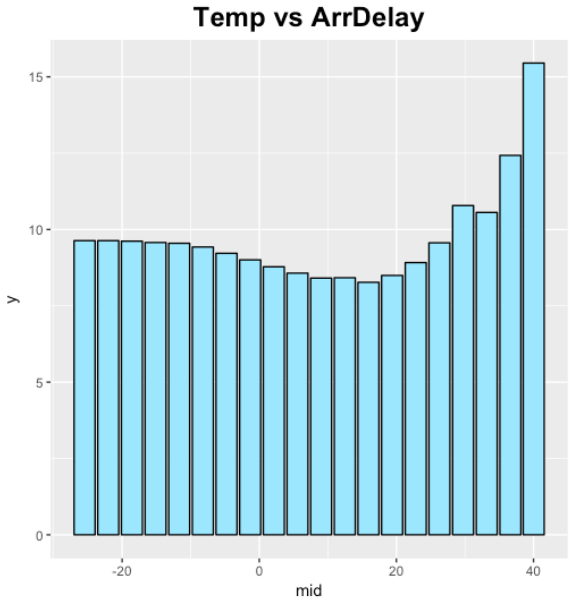
x	10 min	30 min	60 min	120 min
P(X>x)	27%	14%	6.6%	2.3%

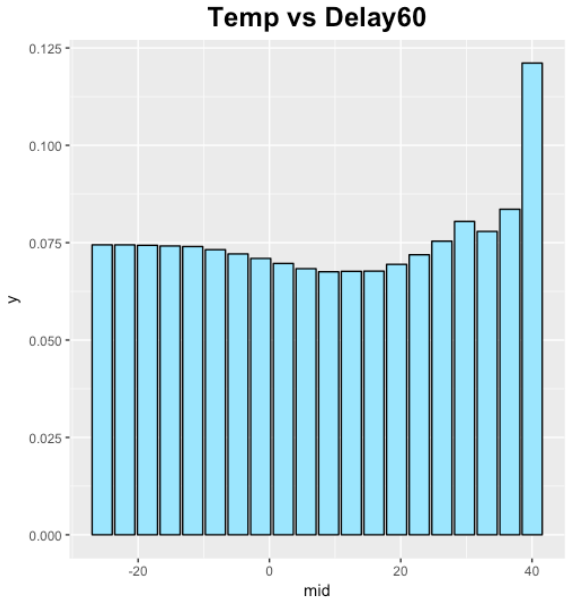
變數研究 : ArrTime

變數研究 : ArrTime

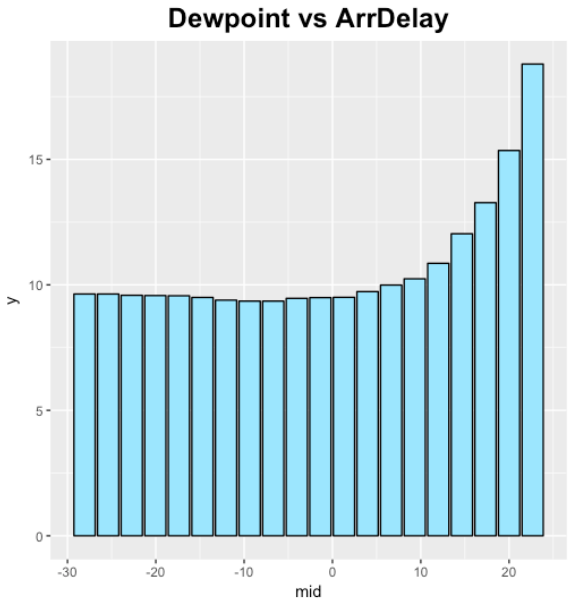


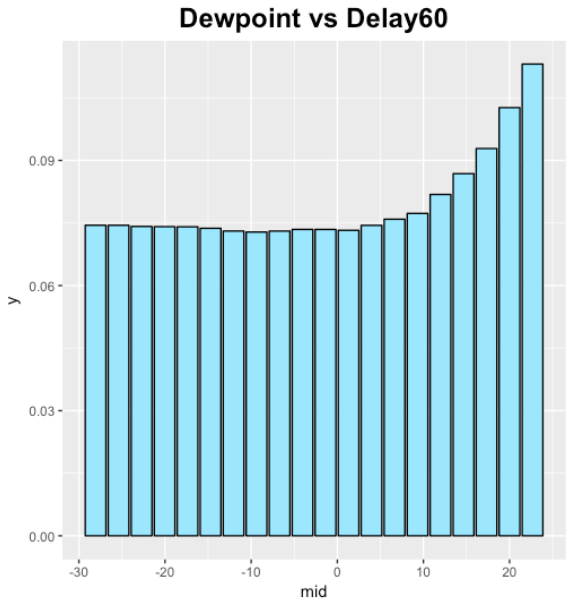
變數研究 : Temp


Temp vs ArrDelay
15
10
5
0
y
-20
0
20
40
mid

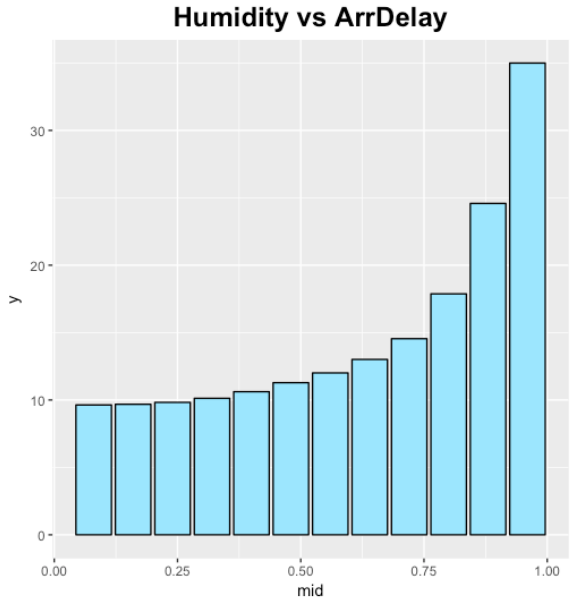

Temp vs Delay60
0.125
0.100
0.075
0.050
0.025
0.000
y
-20
0
20
40
mid

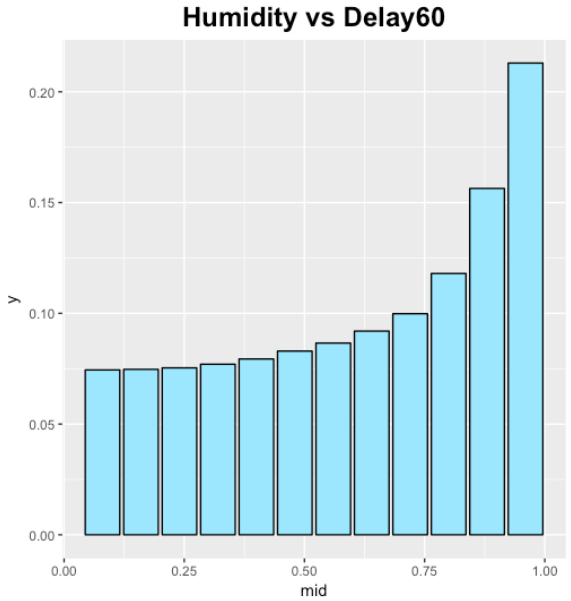
變數研究 : Dew.Temp



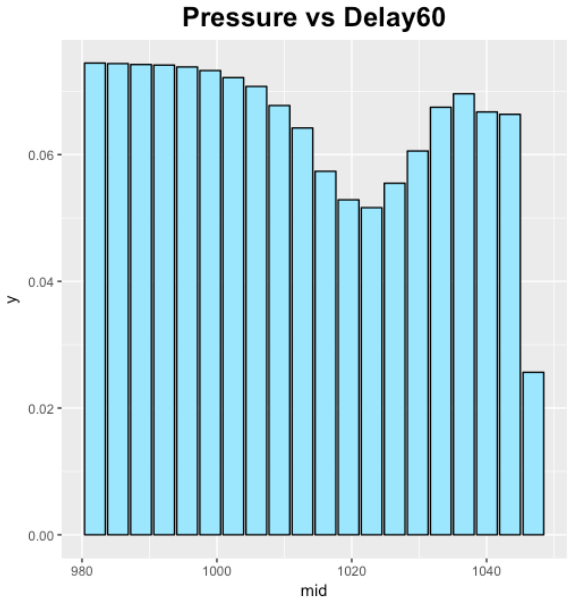


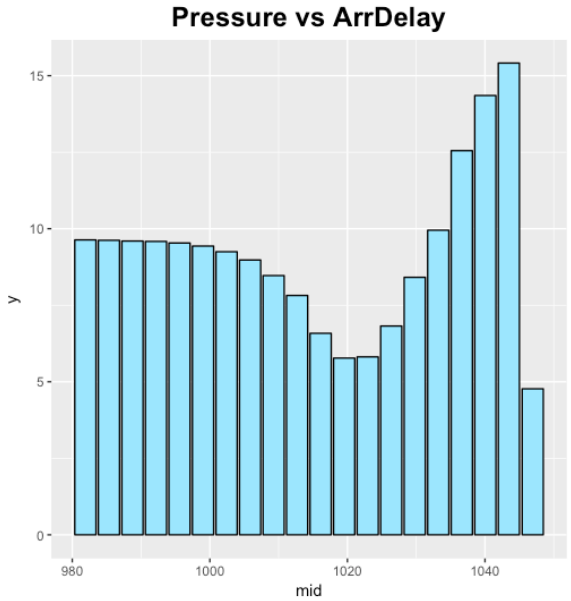
變數研究 : Humidity



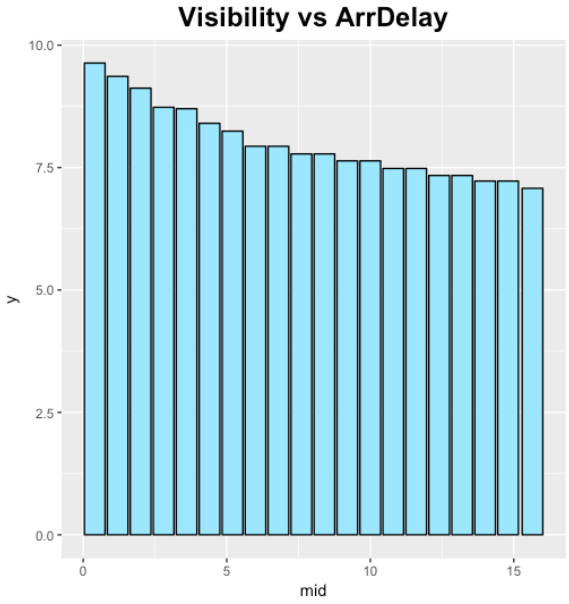


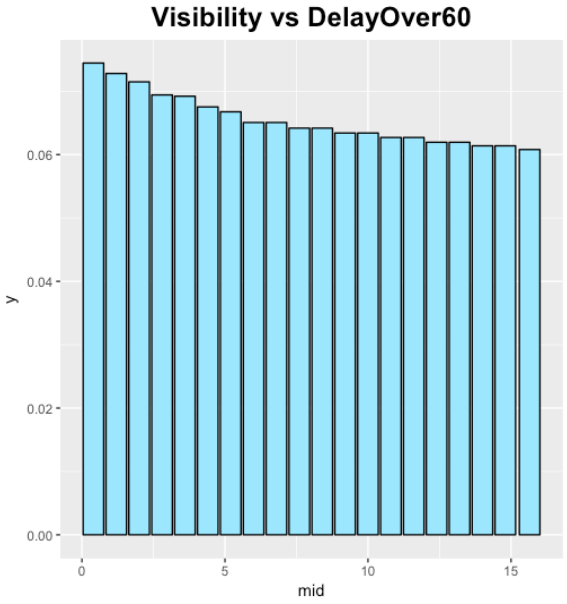
變數研究 : Pressure



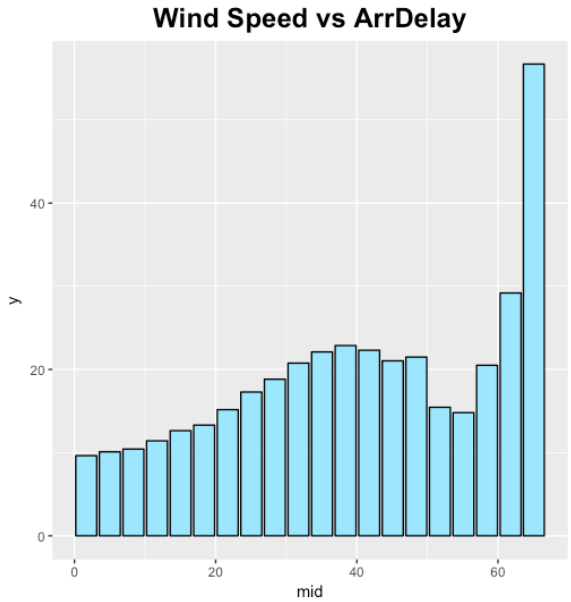


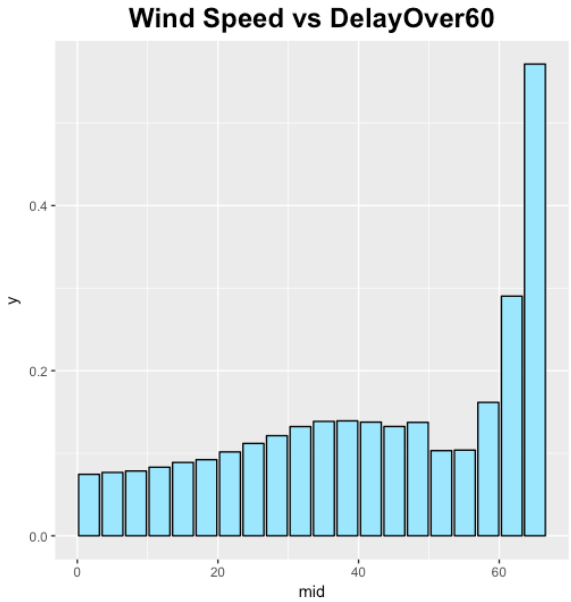
變數研究 : Visibility



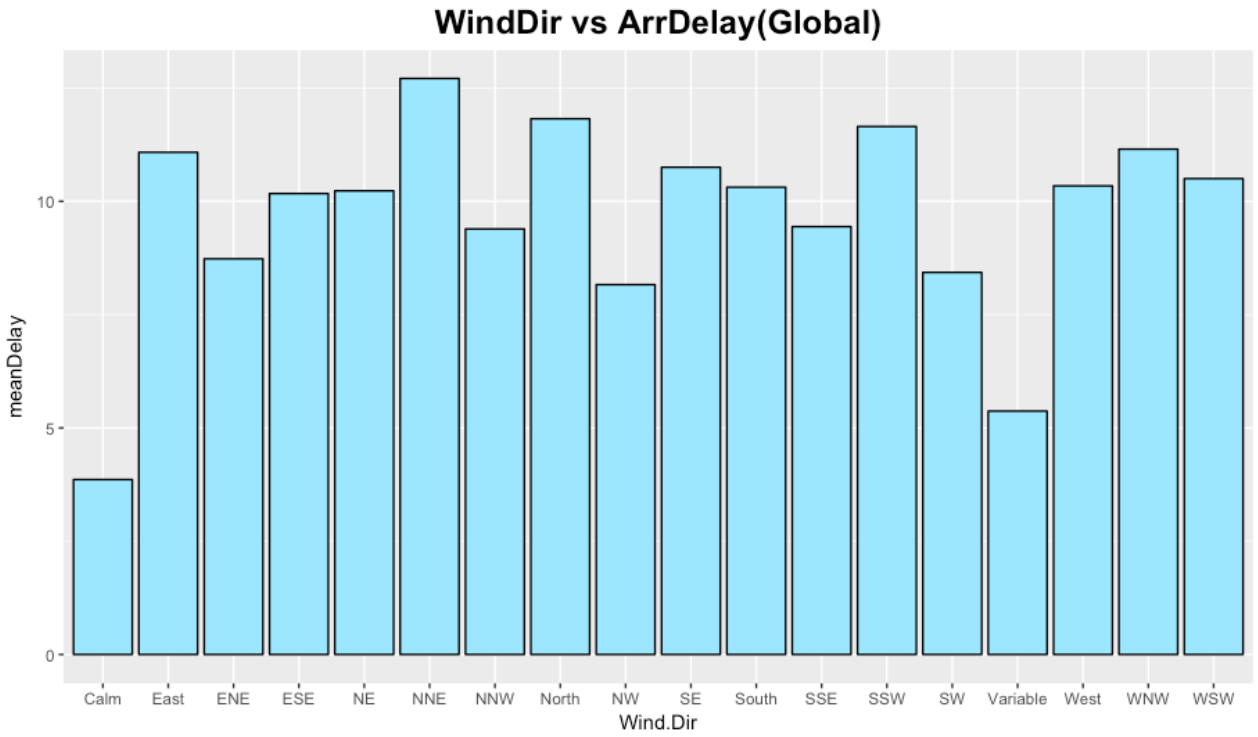


變數研究 : Wind Speed

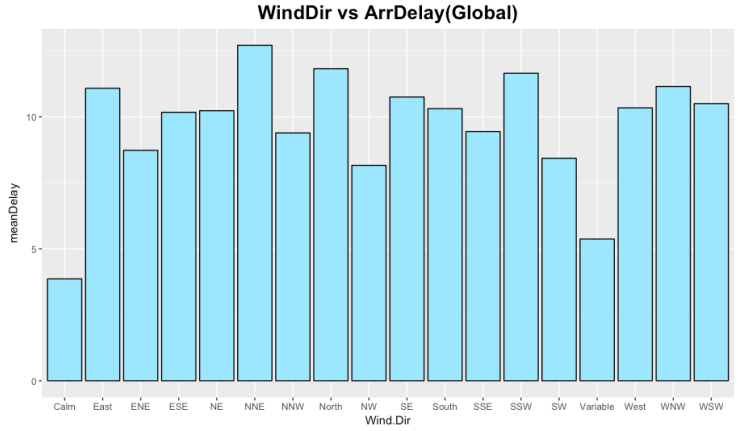


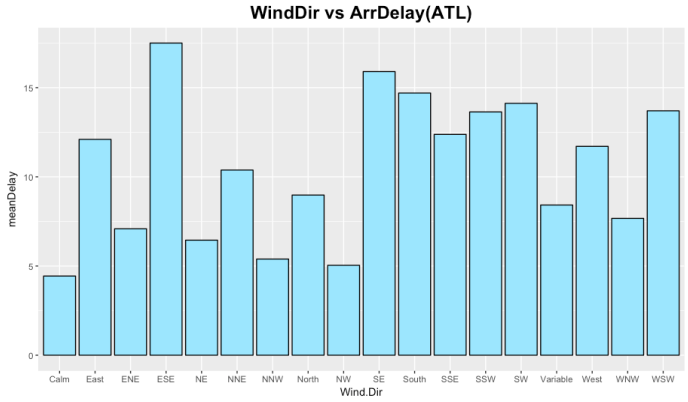


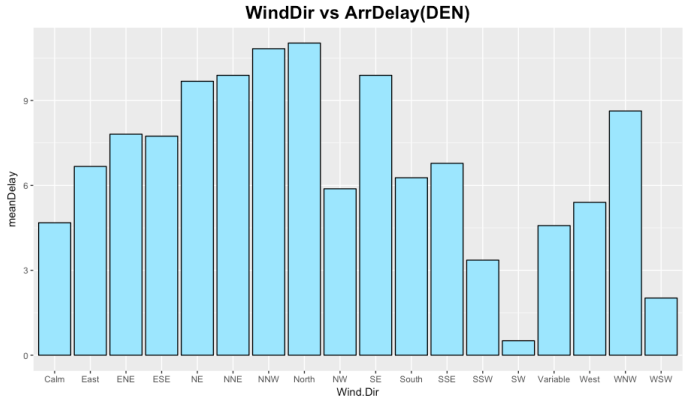
變數研究 : Wind Dir



變數研究 : Wind Dir







變數研究 : Conditions

Conditions:	Count
Funnel Cloud	1
Heavy Small Hail	1
Heavy Thunderstorms with Small Hail	1
Small Hail	1
Squalls	1
Light Hail	2
Light Thunderstorms and Snow	2
Thunderstorms with Hail	2
…	…

變數共有40種類別

其中21個類別，觀測值個數少於50個

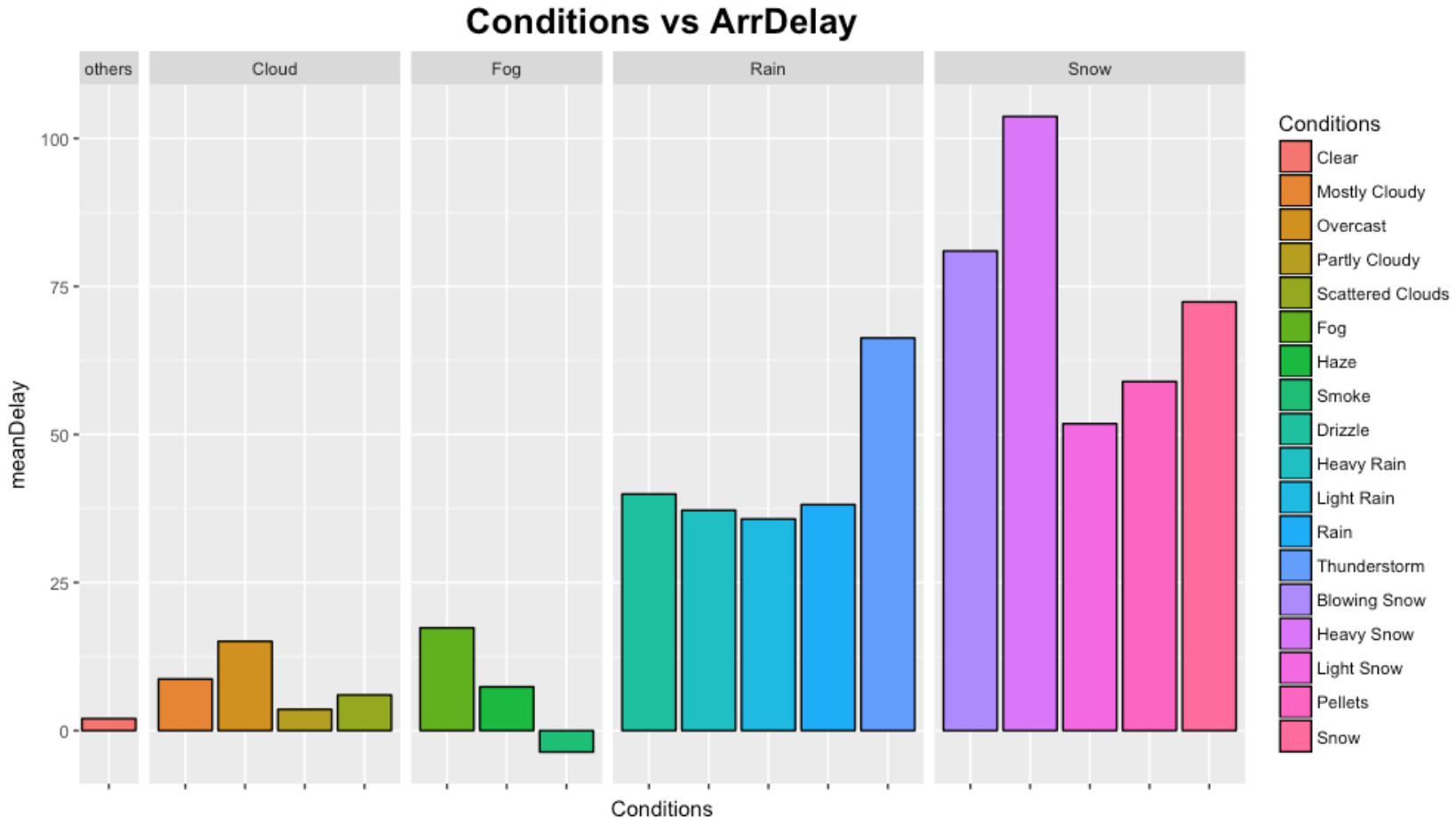
變數研究 : Conditions

Conditions:	Count
Funnel Cloud	1
Heavy Small **Hail**	1
Heavy **Thunderstorms** with Small **Hail**	1
Small **Hail**	1
Squalls	1
Light **Hail**	2
Light **Thunderstorms** and Snow	2
Thunderstorms with **Hail**	2
…	…

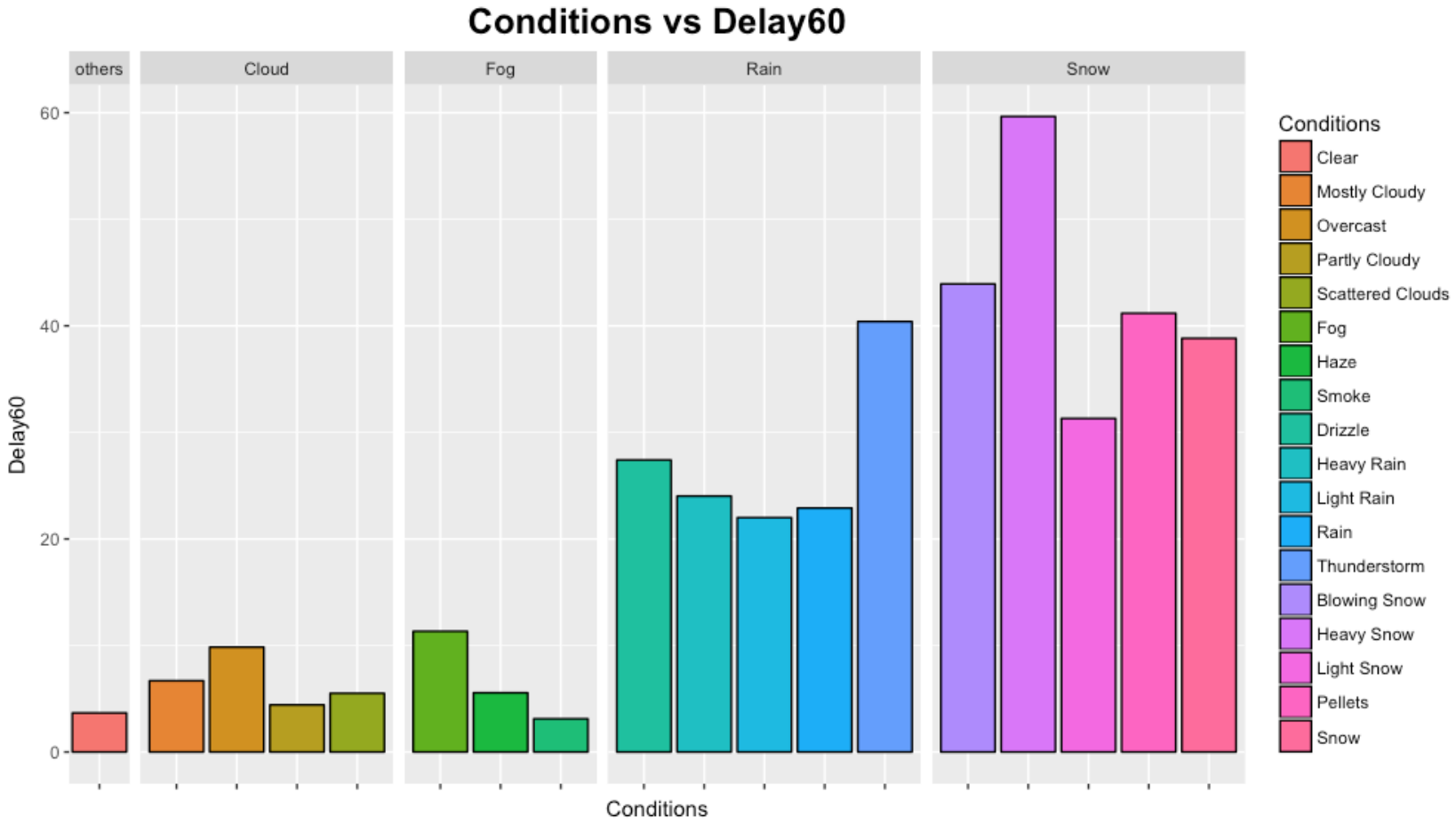
對於觀測值數不足50之類別進行整併

整併後剩下18個類別

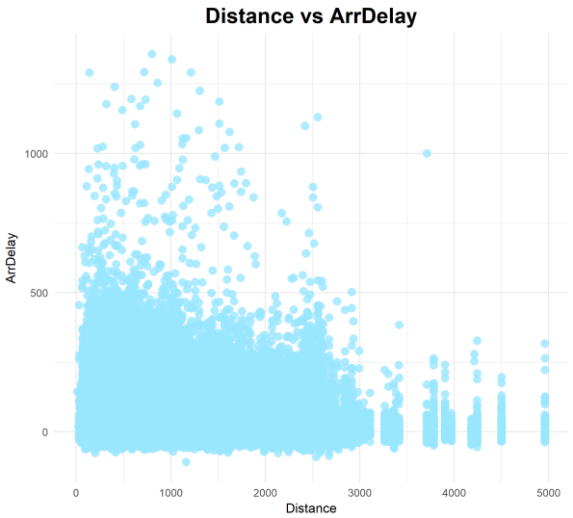
變數研究 : Conditions

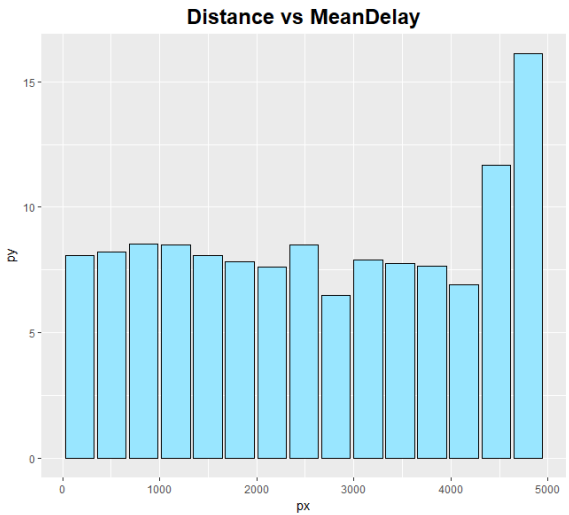


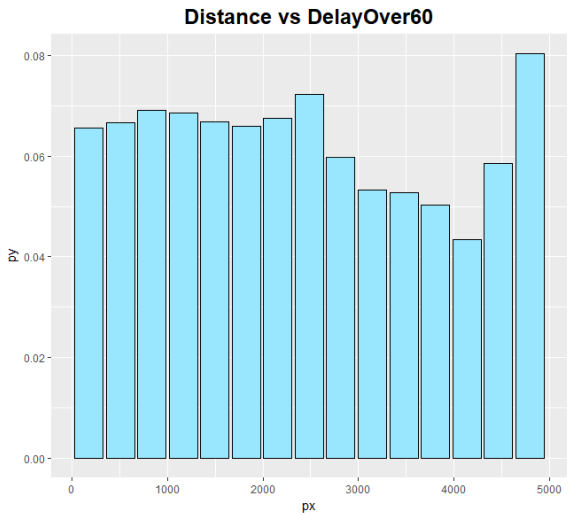
變數研究 : Conditions



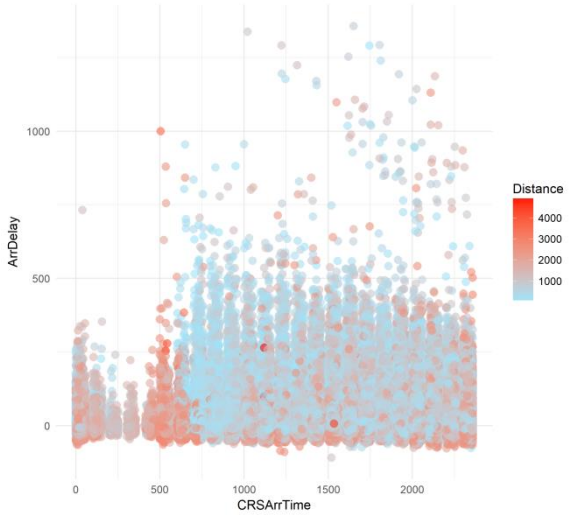
變數研究 : Distance

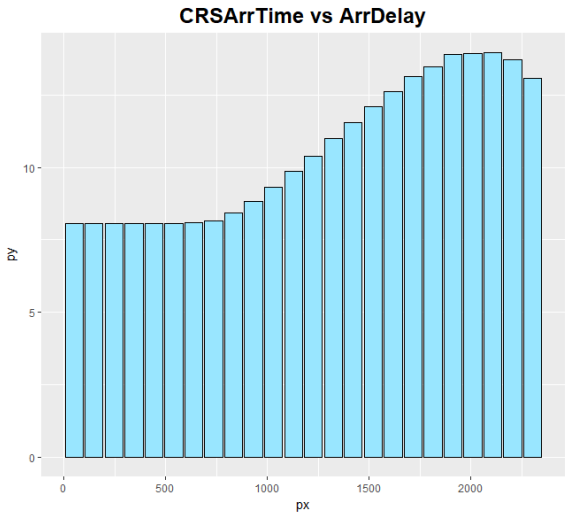


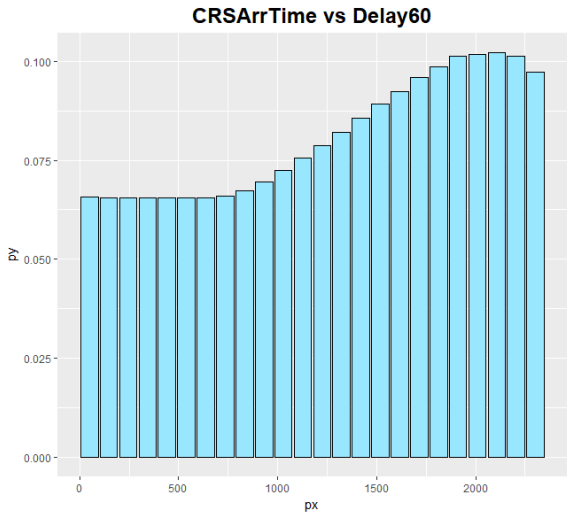




變數研究 : CRSArrTime

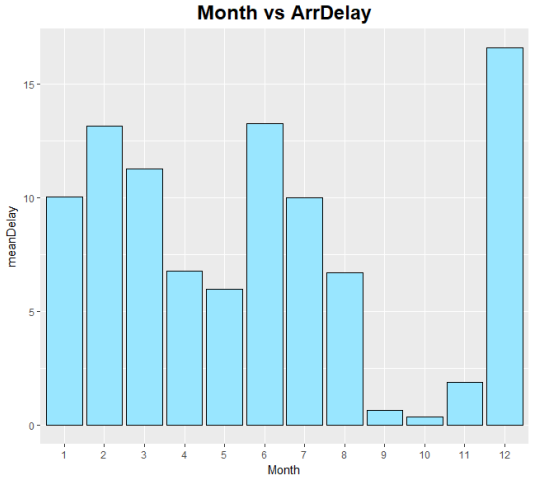


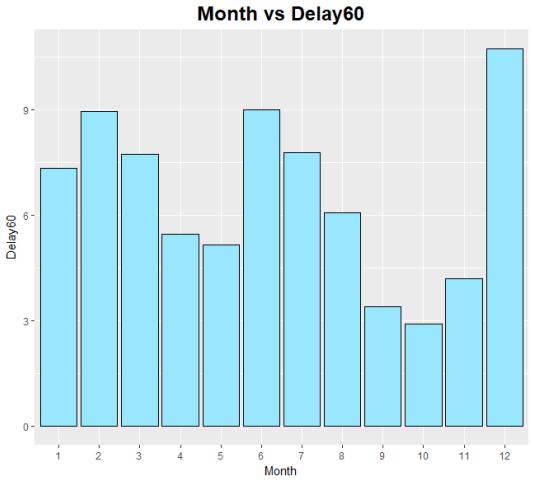




變數研究 : Month

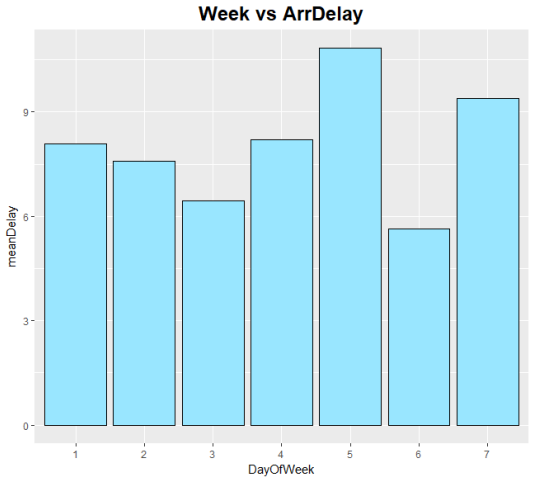
Month	Mean Delay	Delay60
1	10.07	7.34
2	13.17	8.95
3	11.3	7.74
4	6.79	5.46
5	6.01	5.15
6	13.27	9
7	10	7.77
8	6.72	6.08
9	0.65	3.41
10	0.39	2.91
11	1.88	4.21
12	16.63	10.74

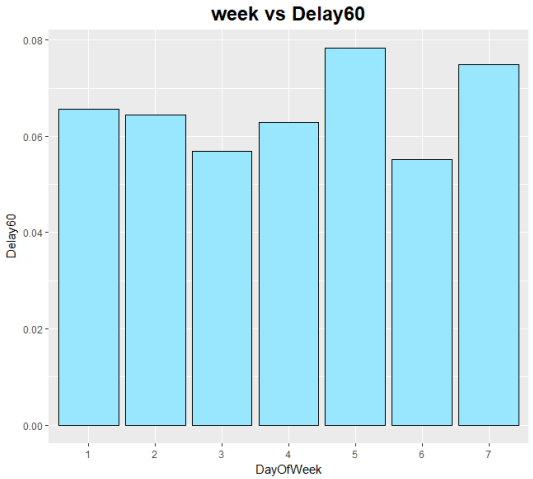




變數研究 : DayOfWeek

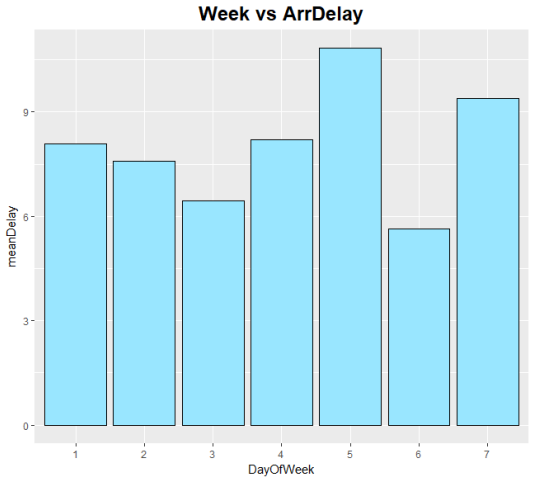
Week	meanDelay	Delay60
1	8.09	6.56
2	7.58	6.44
3	6.44	5.69
4	8.2	6.29
5	10.83	7.83
6	5.65	5.51
7	9.4	7.5

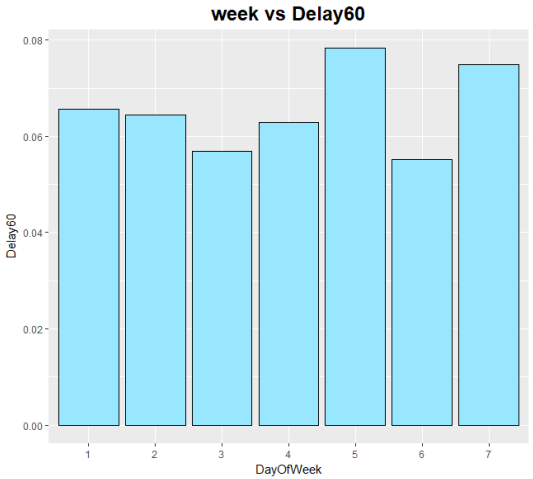




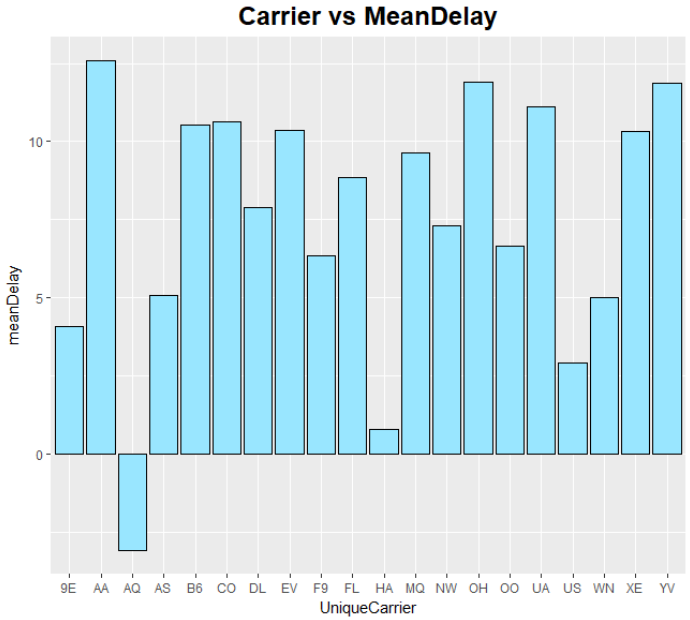
變數研究 : DayOfWeek

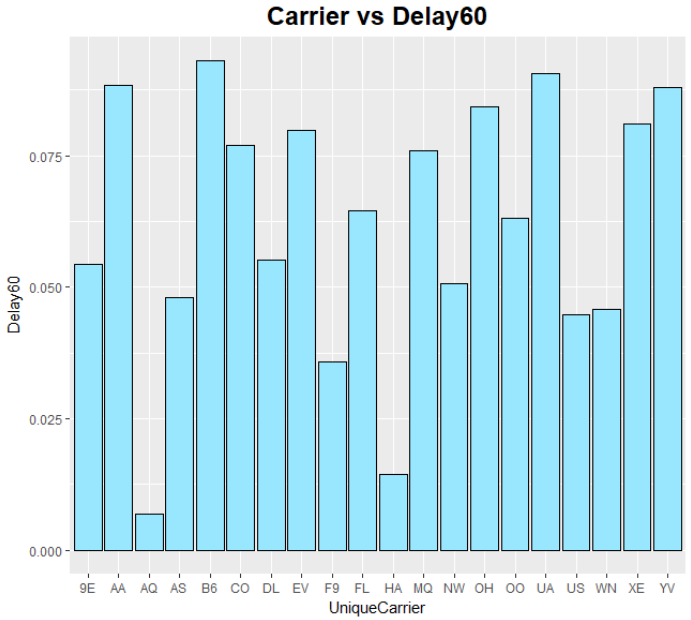
Week	meanDelay	Delay60
1	8.09	6.56
2	7.58	6.44
3	6.44	5.69
4	8.2	6.29
5	10.83	7.83
6	5.65	5.51
7	9.4	7.5





變數研究 : UniqueCarrier





總結

1. 大多數連續變數有線性趨勢，但變動幅度小，只在極端值時有較高的平均延誤

2. Conditions 能定位出平均延誤時間較長的班次

3. 風向要在各別地區看較有意義

4. 大抵而言，**分得很爛**

模型配置摘要

1. 主要使用 logistic regression, Ridge Regression
2. 將資料切為 第一部分 : 100萬筆訓練集與20萬筆測試集

   第二部分 : 25.6萬筆訓練集與5萬筆測試集

1. 預測變數為「**延誤超過一小時之類別**」
2. 模型目標為預測，因此較不關注於係數解釋與共線性問題
3. 模型選擇標準為 AUC，相近則比較準確率(Accuracy)

模型歷史(第一部分)

Model no.	Description	AUC	
		Logistic	Ridge
0-1	僅使用原資料集數據	0.707	
1-1	加入 mean delay 3	0.748	0.750
1-2.1	ArrDelay 改為每小時一類別	0.754	0.756
1-2.2	加入ArrDelay 平方項	0.742 (低)	
1-3	刪去Dest 項	0.742	
1-4.1	加入DepDelayCateg	0.889	0.888
1-4.2b	以平衡訓練集配適1-4.1	0.889(略高)	
1-5	加入mean delay 3變數的平方項	0.892	0.891

Model : Logistic w0-1

模型摘要

僅使用氣象變數，如下：

Temp. + DewPoint + Humidity + Pressure + Visibility +
Wind.Dir + Wind.Speed + Gust.Speed + Conditions

未使用變數，如下：

Events : 遺漏值過多，且與Condtions相似

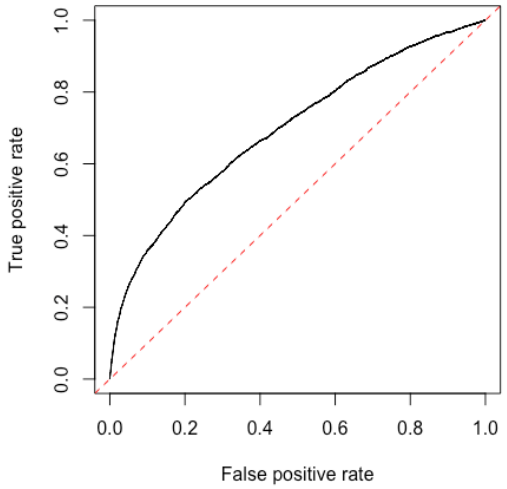
Precip : 遺漏值過多，不符現實(大雷雨中，雨量0?)

Model : Logistic 0-1

模型成效

AUC	0.693
Accuracy	0.736
Recall	0.535
Precision	0.148

ROC curve



報表: https://hyades910739.github.io/flight_research/weather_research.html#model-w0-1-fit-only-weather-variables

模型改進(1)

刪除變數： DewPoint

1. 露點溫度為濕度與溫度的函數，放入模型會造成共線性問題
2. 模型改進過程中，在不同階段移除DewPoint，有時增加AUC，有時減少
3. 基於以上兩點，最後決定從最終模型中移除

相關矩陣

	Temp	DewPoint	Humidity
Temp	1	0.787	-0.308
DewPoint	0.787	1	0.319
Humidity	-0.308	0.319	1

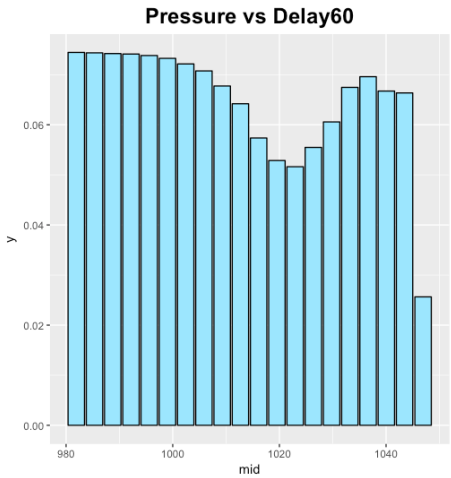
模型改進(2) : Bad Try

增加變數： Pressure^2

1. Pressure對Delay60之關係不適合用線性解釋，因而加入二次項
2. 雖然變數是顯著的，但造成AUC略為的降低
3. 由於AUC降低，最後決定從最終模型中移除

改變變數： DustSpeedCateg

1. 最高風速變數本質上更像是類別變數
2. 加入模型造成AUC降低
3.由於AUC降低，最後決定從最終模型中移除



模型改進(3)

交互作用項:

1. Wind.Dir 與 Dest 有明顯之交互關係
2. Wind.Speed 應該也有影響
3. 如上述，因而加入交互作用項，至三階皆為顯著
4. 係數顯著外，AUC亦上升，因而保留交互作用項

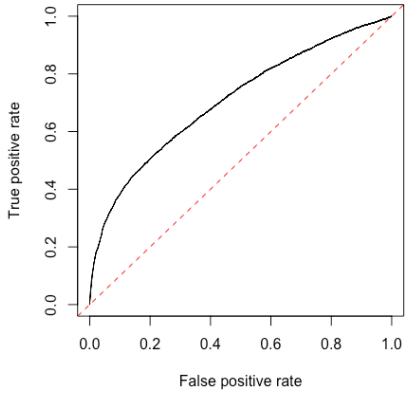
目前模型：model W0-9

模型變數：

Temp. + Humidity + Pressure + Visibility + Wind.Dir +
Wind.Speed + Gust.Speed + Conditions + Dest +
Dest * Wind.Dir + Dest * Wind.Speed + Wind.Speed * Wind.Dir + Dest * Wind.Dir * Wind.Speed

模型成效：

AUC	0.708
Accuracy	0.745
Recall	0.541
Precision	0.156

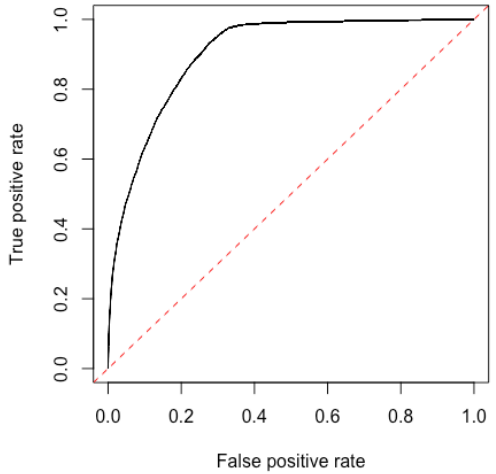


結合後的模型：model W1-1

結合航班模型與氣象模型所含變數所配適之模型

模型成效：

AUC	0.903
Accuracy	0.748
Recall	0.917
Precision	0.219



模型歷史：

model W0-1 : fit with only unhandled weather col.

model W0-2 : remove Dewpoint. (lower)

model W0-3 : add Dest with Dewpoint (higher)

model W0-4 : add Dest without Dewpoint (higher)

model W0-5 : change DustSpeed to categ

model W0-6 : 0-4 add Pressure^2

model W0-7 : 0-4 add Dest*Wind.Dir

model W0-8 : 0-7 add Wind.Speed*Wind.Dir

model W0-9 : 0-8 add Wind.Speed*Wind.Dir*Dest

model W1-1 : 1-5 + W0-9

model W1-2 : W1-1 +DustSpeedCateg

model W1-4 : W1-1 +without md3

Logistic:

No.	AUC	accuracy	recall	precision	AIC	Deviance
W0-0	0.5				109205	109203
W0-1	0.6929374	0.736080	0.535109	0.147804	100316	100230
W0-2	0.6928029	0.737160	0.534840	0.148347	100321	100237
W0-3	0.6997253	0.729780	0.548023	0.146885	99385	99291
W0-4	0.6998183	0.730040	0.548292	0.147074	99385	99293
W0-5	0.6995962	0.731760	0.546592	0.149179	99238	99122
W0-6	0.699414					
W0-7	0.7038518	0.732320	0.555911	0.151261	98916	98664
W0-8	0.7048907	0.737120	0.553248	0.153426	98813	98529
W0-9	0.7077657	0.744840	0.541267	0.155571	98664	98244
1-5	0.8956611	0.731180	0.939031	0.210706	72562	72418
W1-1	0.903333	0.748480	0.917465	0.219323	70405	69851
W1-1b	0.9019124	0.65814	0.983493	0.178239	不比較	不比較
W1-2	0.9033084	0.748680	0.917998	0.219534		
W1-3	0.9032782	0.748440	0.916400	0.219152	70398	69842
W1-4	0.8963565	0.742620	0.918797	0.215485	72554	72040

Ridge:

W1-1	0.9022838	0.748020	0.912939	0.218394		

待改進的議題：

1. 需要氣象相關的Domain Knowledge
2. 是否該對Condition 與 Wind.Dir 做處理
3. 仍有許多未能解釋之變異
4. 嘗試以不同切點，做多類別的預測