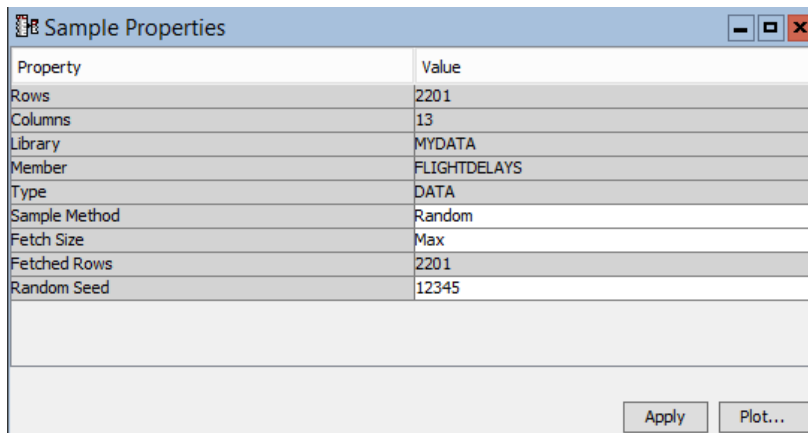


Exploration of flight delay data on commercial flights from the Washington, DC area to the New York area during January, 2004.

4. Variable Exploration

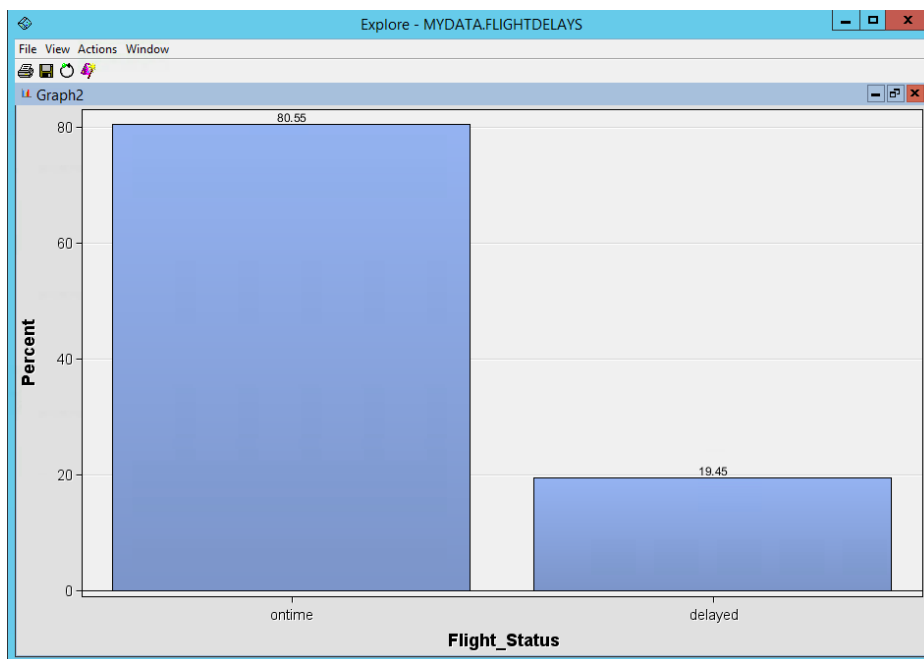
- a. Number of flights in the dataset = **2,201**



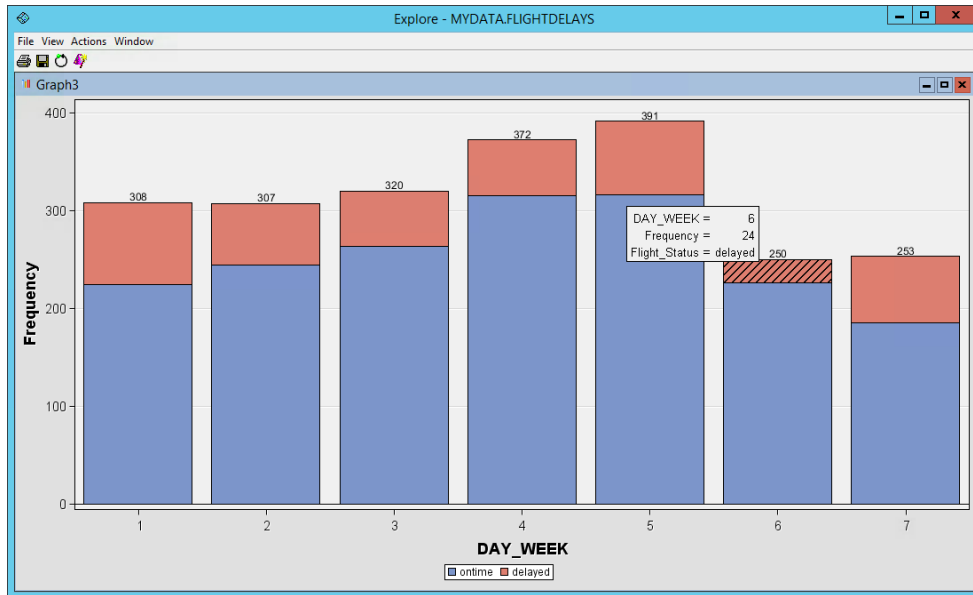
The 'Sample Properties' dialog box displays the following information:

Property	Value
Rows	2201
Columns	13
Library	MYDATA
Member	FLIGHTDELAYS
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetches Rows	2201
Random Seed	12345

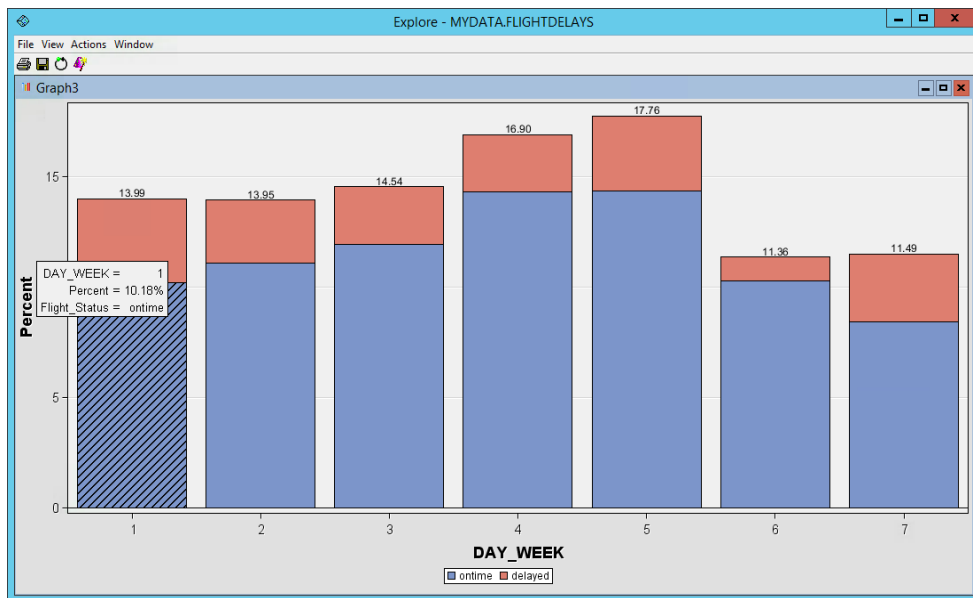
- b. Percentage of flights on time = **80.55%**



- c. Day of the week data
 - i. Day of the week with the least number of delayed flights = **Saturday (6)**
 - ii. Total number of flights for Saturday = **250**
- d. How many flights delayed on Saturday = **24**



- e. Percentage of flights on Mondays which were on time = **10.18%**

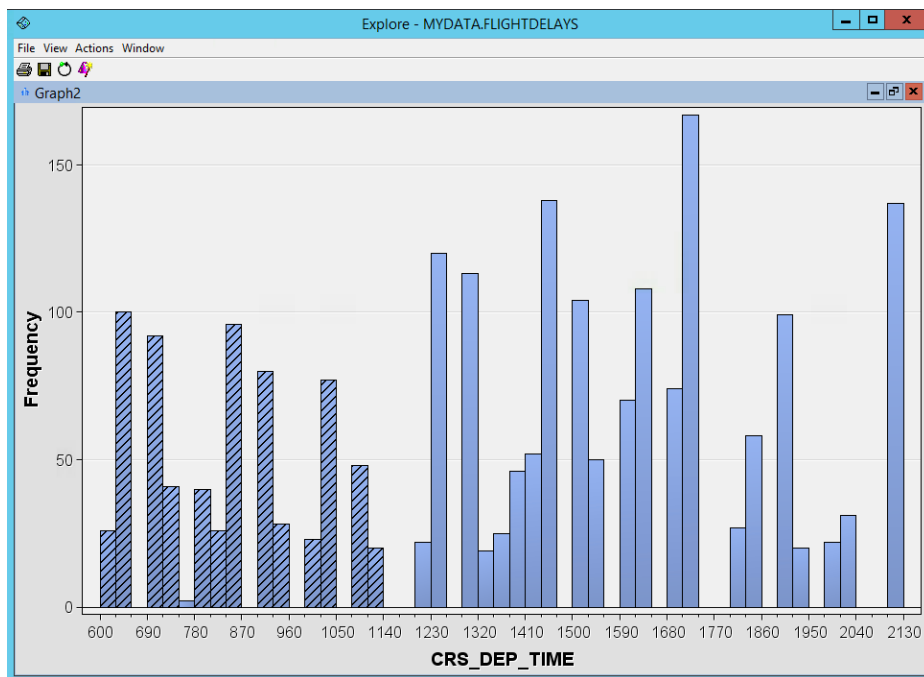


f. Departure time data

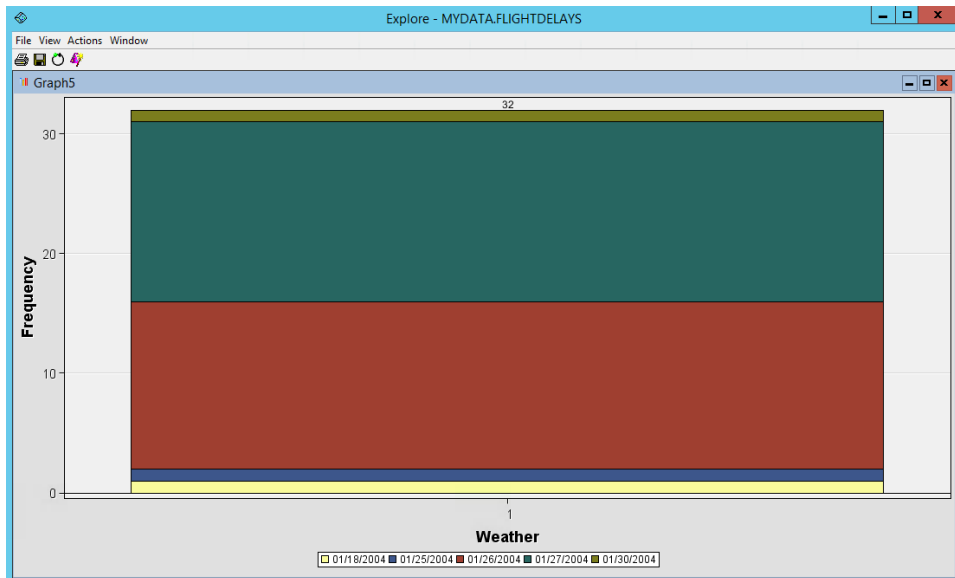
- i. Earliest scheduled departure time = **6:00 am (600)**
- ii. Latest scheduled departure time = **9:30 pm (2130)**

Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum
1	CARRIER	CARRIER	CLASS	0	.	.
2	DEST	DEST	CLASS	0	.	.
3	Flight_Status	Flight_Status	CLASS	0	.	.
4	ORIGIN	ORIGIN	CLASS	0	.	.
5	TAIL_NUM	TAIL_NUM	CLASS	0	.	.
6	CRS_DEP_...	CRS_DEP_...	VAR	0	600	2130
7	DAY_OF_M...	DAY_OF_M...	VAR	0	1	31
8	DAY_WEEK	DAY_WEEK	VAR	0	1	7
9	DEP_TIME	DEP_TIME	VAR	0	10	2330
10	DISTANCE	DISTANCE	VAR	0	169	229
11	FL_DATE	FL_DATE	VAR	0	16071	16101
12	FL_NUM	FL_NUM	VAR	0	746	7924
13	Weather	Weather	VAR	0	0	1

- iii. More flights scheduled to depart in the = **afternoon (dep time > 1200)**

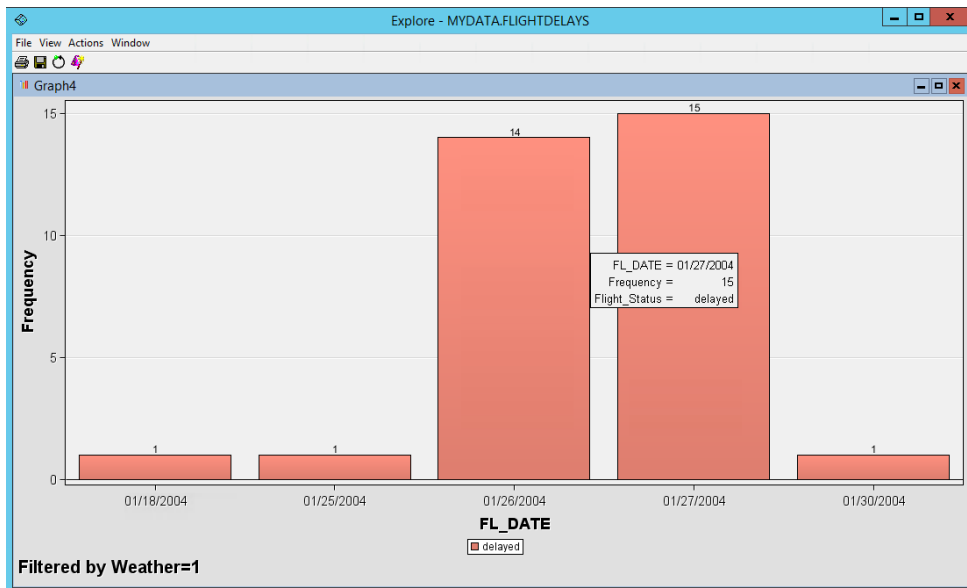


iv. Dates which had inclement weather = **1/18, 1/25, 1/26, 1/27, 1/30**



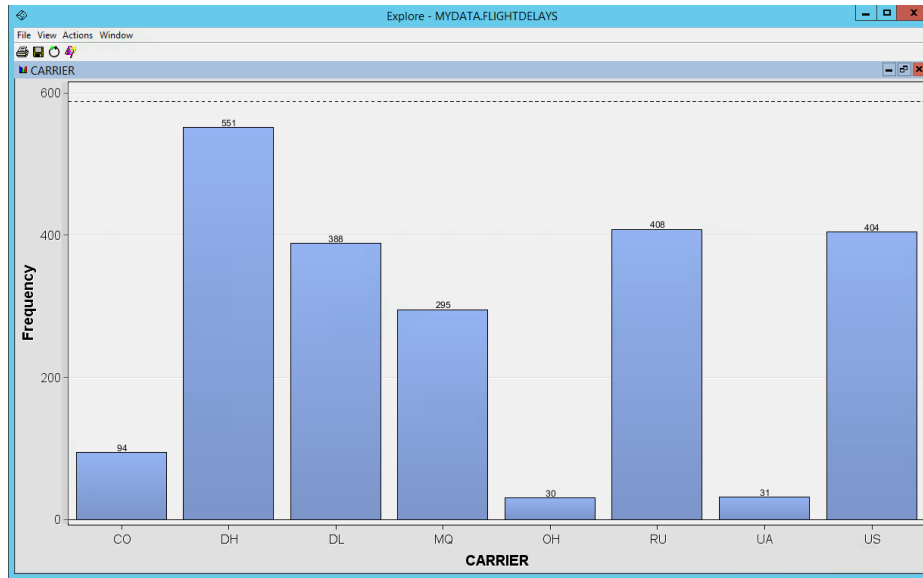
g. Inclement weather data

- i. Date with the most number of flights affected by inclement weather = **1/27**
- ii. How many flights affected on January 27 = **15**



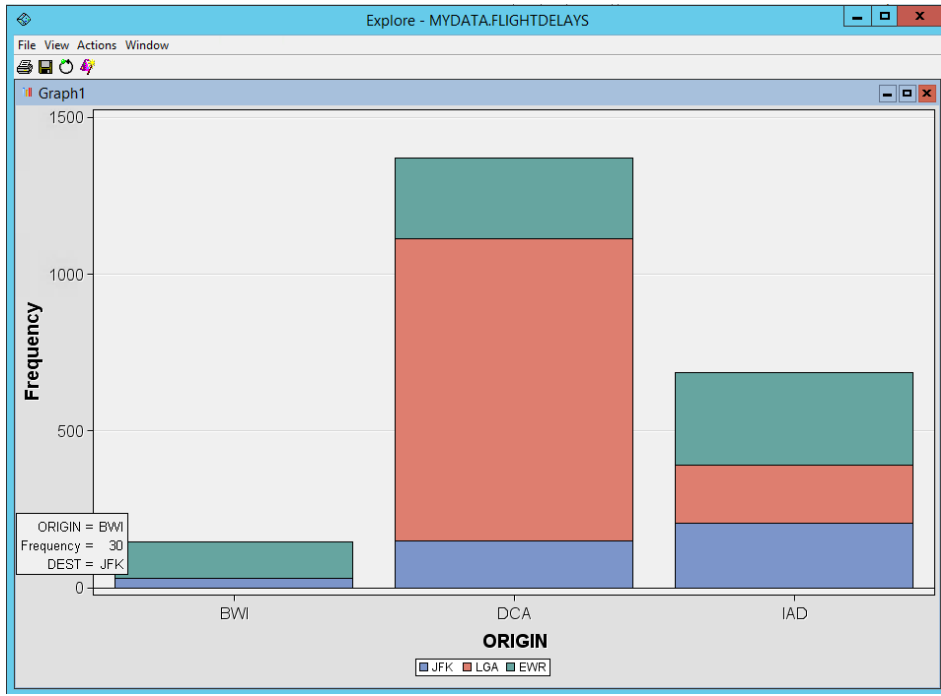
h. Carrier data – Number of Flights > 400

Carrier Code	Carrier Name	Number of Flights
DH	Atlantic Coast Airlines	551
RU	Continental Express	408
US	US Airways	404
Total		1,363



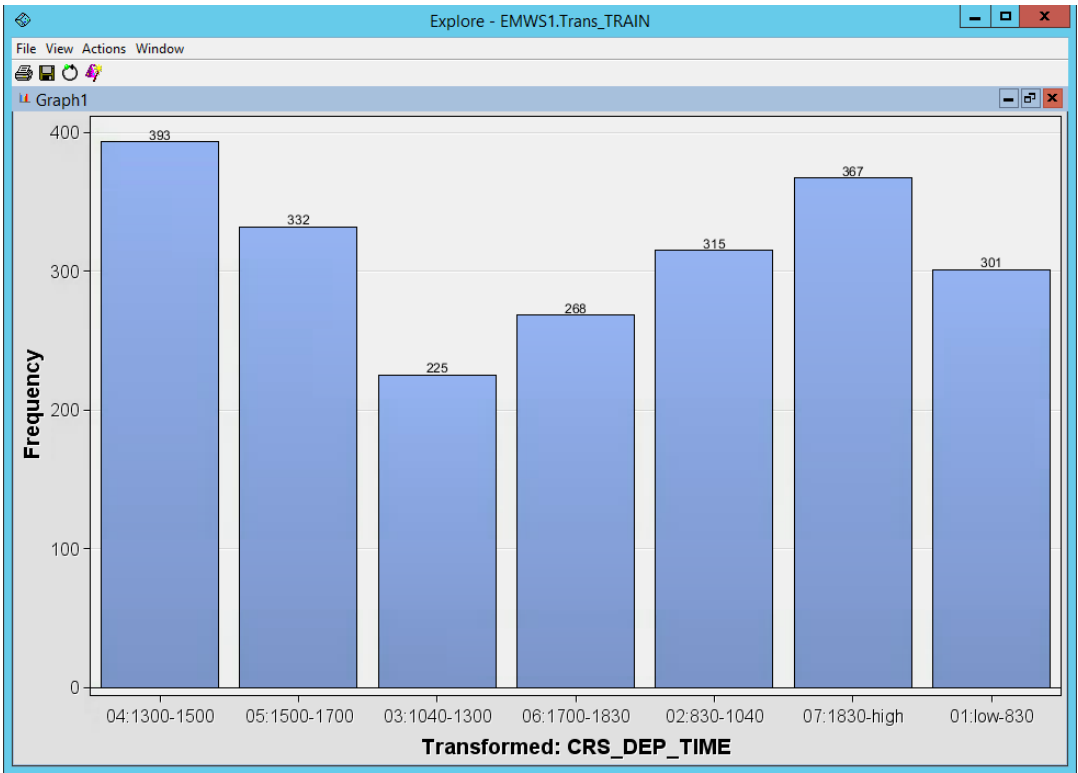
i. Flight route data

- i. Flight route with the least number of flights = **Baltimore-Washington to Kennedy (BWI -> JFK)**
- ii. Number of flights for this route = **30**



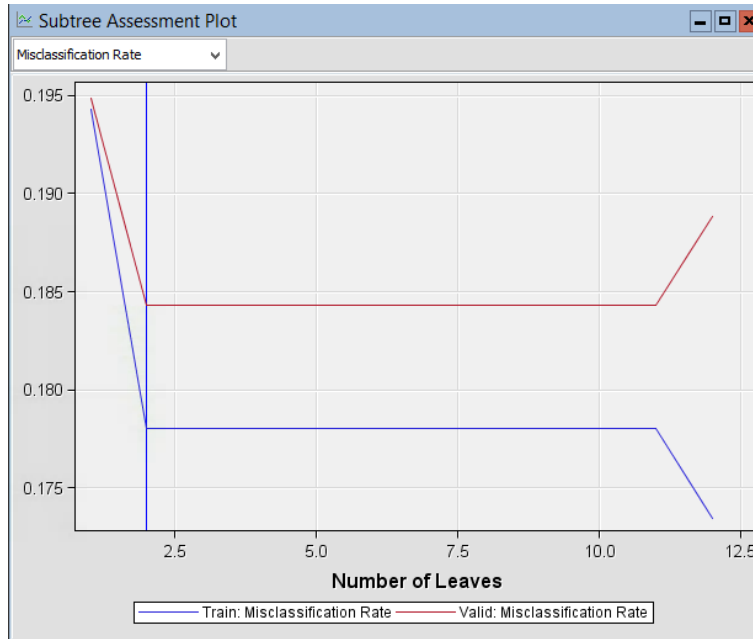
7. Scheduled Departure Time (CRS_DEP_TIME) Transform

Bin	Min	Max
1	Low	8:30 am
2	8:30 am	10:40 am
3	10:40 am	1:00 pm
4	1:00 pm	3:00 pm
5	3:00 pm	5:00 pm
6	5:00 pm	6:30 pm
7	6:30 pm	high

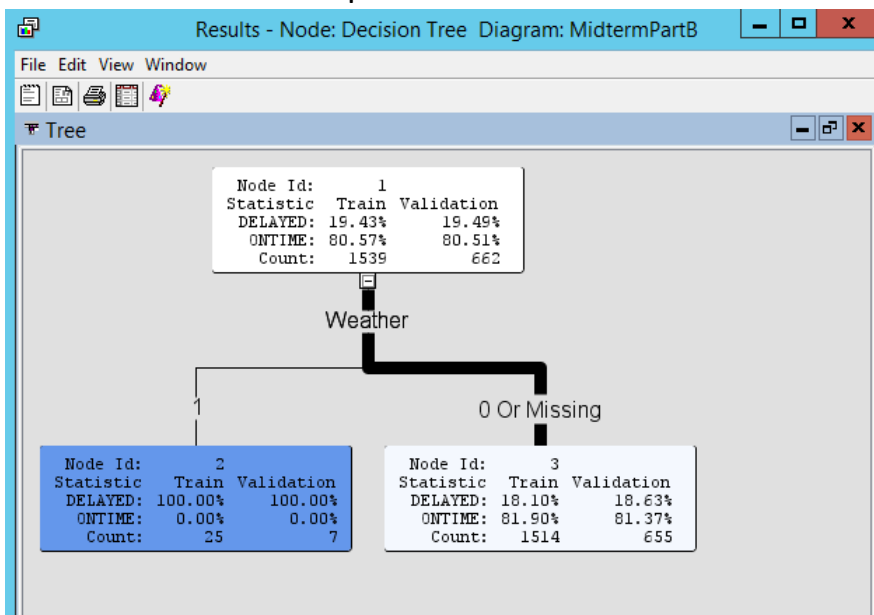


10. Misclassification Decision Tree

a. Number of leaves in optimal tree = 2



b. Variable used for first split = **Weather**



c. Rates

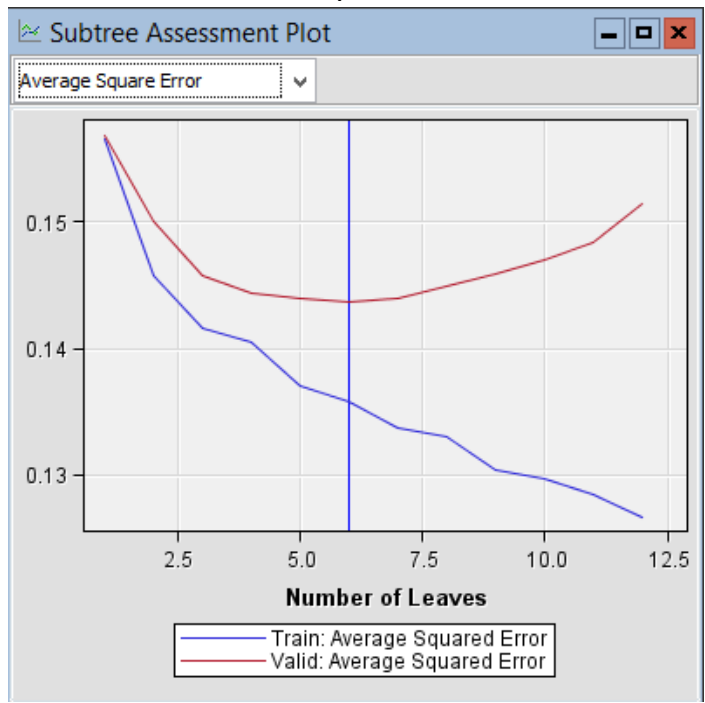
- i. Misclassification Rate: (Train = 0.178038) **Validation = 0.18429**
- ii. Average Square Error: (Train = 0.145817) **Validation = 0.149992**

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Flight_Status	Flight_Status	_NOBS_	Sum of Fre...	1539	662	.
Flight_Status	Flight_Status	_MISC_	Misclassific...	0.178038	0.18429	.
Flight_Status	Flight_Status	_MAX_	Maximum A...	0.819022	0.819022	.
Flight_Status	Flight_Status	_SSE_	Sum of Squ...	448.8243	198.5892	.
Flight_Status	Flight_Status	_ASE_	Average Sq...	0.145817	0.149992	.
Flight_Status	Flight_Status	_RASE_	Root Avera...	0.38186	0.387288	.
Flight_Status	Flight_Status	_DIV_	Divisor for A...	3078	1324	.
Flight_Status	Flight_Status	_DFT_	Total Degre...	1539	.	.

d. **We could not** use this model to predict whether or not the flight will be delayed because the model only uses the Weather variable for prediction. We would need information on inclement weather before that particular case could be predicted.

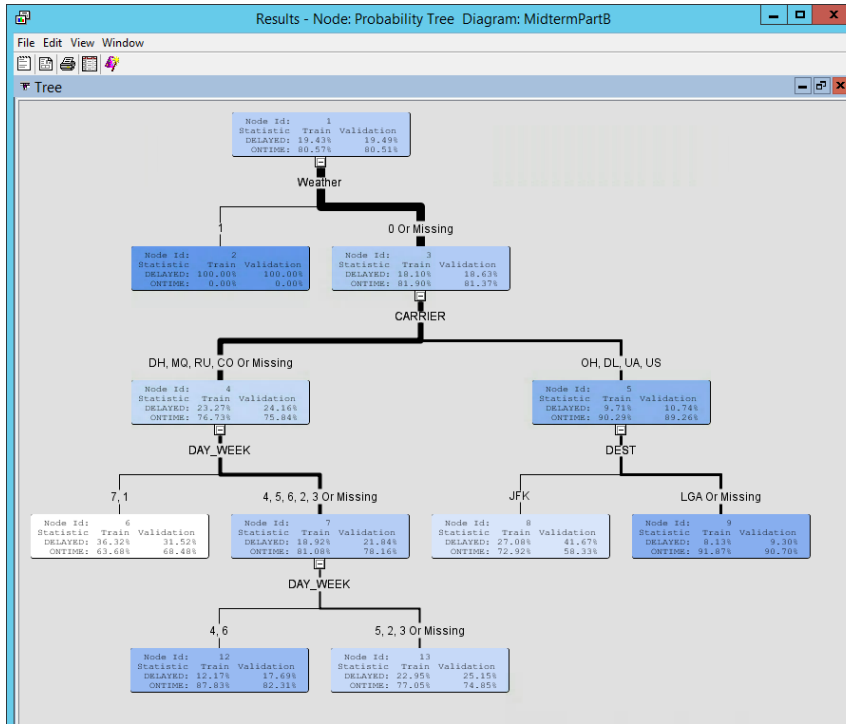
12. Probability Decision Tree

- a. Number of leaves in optimal tree = 6



b. First split

i. Variable = **Weather**



ii. Logworth = **23.9988**

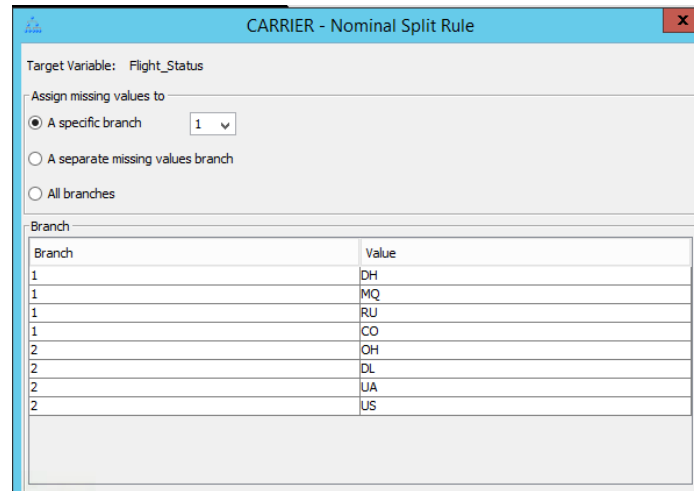
Variable	Variable Description	-Log(p)	Branches
Weather	Weather	23.9988	2
CARRIER	CARRIER	9.9174	2
PCTL_CRS_DEP_TIME	Transformed: CRS_DEP...	6.4825	2
ORIGIN	ORIGIN	4.3911	2
DAY_WEEK	DAY_WEEK	4.3778	2
DEST	DEST	3.9759	2

c. Closest Competitor

i. Variable = **Carrier**

ii. Logworth = **9.9174**

- iii. Split rule = **Branch 1: {DH, MQ, RU, CO} / Branch 2: {OH, DL, UA, US}**



d. Rates

- i. Misclassification Rate: (Train = 0.178038) **Validation = 0.18429**
- ii. Average Square Error: (Train = 0.135822) **Validation = 0.143715**

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Flight_Status	Flight_Status	_NOBS_	Sum of Freq...	1539	662	.
Flight_Status	Flight_Status	_MISC_	Misclassific...	0.178038	0.18429	.
Flight_Status	Flight_Status	_MAX_	Maximum A...	0.918715	0.918715	.
Flight_Status	Flight_Status	_SSE_	Sum of Squ...	418.0604	190.2786	.
Flight_Status	Flight_Status	_ASE_	Average Squ...	0.135822	0.143715	.
Flight_Status	Flight_Status	_RASE_	Root Averag...	0.36854	0.379098	.
Flight_Status	Flight_Status	_DIV_	Divisor for A...	3078	1324	.
Flight_Status	Flight_Status	_DFT_	Total Degre...	1539	.	.

e. Explanatory Variables

- i. Four most important = **Weather, DAY_WEEK, CARRIER, DEST**
- ii. Variable with most number of split rules = **DAY_WEEK**
- iii. Number of DAY_WEEK split rules = **2**

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Weather	Weather	1	1.0000	1.0000	1.0000
DAY_WEEK	DAY_WEEK	2	0.6622	0.2963	0.4475
CARRIER	CARRIER	1	0.6309	0.7850	1.2442
DEST	DEST	1	0.3096	0.4534	1.4647

- f. If there is inclement weather, the probability of a delayed flight = **100%**
If there is not inclement weather, the probability of a delayed flight = **25.15%**
- g. Better model = **Misclassification Decision Tree** (modestly smaller validation ASE).