Delayed flights in Rattle. Chapter 10.5

MEMORANDUM - CLASS NOTES

To: BDA 2018
From: Roger Bohn
Subject: Using Rattle to run a model and analyze results: case study of airplane delays
Date: April 18, 2018

In today's class we analyzed the airline data in real time. Using the data in Chapter 10.5 of DMBA textbook. This is a "dump" of the work I did with chapter 10.5 in Rattle, the day before our actual class. I have deliberately left in a few mistakes to demonstrate that they are inevitable, and the goal is just to notice them quickly and fix them.

It includes four kinds of information:

a. Screen shots of how Rattle looked
b. The output from running various Rattle methods
c. The LOG from Rattle of the R code it used to carry out the methods. The log starts and ends with.
   ==========
d. Some plots, captured from Rstudio, of other Rattle results.

Analysis of airline late data.

conducted April 17, 2018 by R Bohn

The basic flow of working with Rattle is:

A. Setup:

- Start Rstudio
- Start a word processor to keep a log of your work.
- Identify the folder you will use for data, results, etc.
- >library(rattle)
- >rattle()
- In Rattle, Menu/Project/New. This cleans out old information. Remember to SAVE it first if you may want it later.

B. Explore

1. Decide what you want to do next
2. In the DATA tab, set up the conditions you want in the data. (Hit Execute)
3. Go to Explore, Run various visual and table descriptions of the data
3b. To see figures, go to RStudio. They should appear in the Plots tab.
3c. Use Zoom, and adjust dimensions of the plot
3d. Use the blue "back" arrow in the Plots window, to see earlier plots
3e. Export the result to a file. Turn off the "Maintain aspect ratio" button in Rstudio
3f. Paste the results into your word processor.

4. Go to Transform, change variables e.g. numeric to categorical
Repeat this cycle multiple times.

Delayed flights in Rattle. Chapter 10.5

Turn DATA "partition" back on. Make a note of your seed.

Record at least your final configuration, in your word processing logbook.
Screen shot of the Data Tab
Copy and paste the results from Explore tab.
Copy and paste the LOG tab showing the R code for what you did. Into your word processing logbook

C. Model

Go to MODEL tab.
Set up the desired model configuration
Execute button.
Copy and paste the results from the MODEL tab.
Copy and paste the code from the LOG tab.
Go to EVALUATE.
Execute at least the Error Matrix (confusion matrix) on the Validation dataset.

Below we summarise the dataset.

Use a fixed pitch font for your word processing log. Helvetica 11 in this case.

==========
I first looked at the partitioned data. Then I realized that for descriptive work, I should look at the whole thing. (Or possibly everything but the Test set.)

Now working with 100% sample.
Explore/Summary/Describe

Below is a description of the dataset.

crs$dataset[, c(crs$input, crs$risk, crs$target)] #This is the R code, captured from the LOG tab of rattle

```
 13  Variables      2201  Observations
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
CRS_DEP_TIME
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2201        0       59    0.999     1372    496.2      645      730     1000     1455     1710     1900     2100

lowest :  600  630  640  645  700, highest: 2000 2030 2100 2120 2130
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
CARRIER
       n  missing distinct
    2201        0        8

Value        CO    DH    DL    MQ    OH    RU    UA    US
Frequency    94   551   388   295    30   408    31   404
Proportion 0.043 0.250 0.176 0.134 0.014 0.185 0.014 0.184
```


DEP_TIME
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95
2201 0 633 1 1369 507.4 644 728 1004 1450 1709 1937 2057

lowest : 10 109 548 550 552, highest: 2244 2300 2305 2312 2330

DEST
n missing distinct
2201 0 3

Value EWR JFK LGA
Frequency 665 386 1150
Proportion 0.302 0.175 0.522

DISTANCE
n missing distinct Info Mean Gmd
2201 0 7 0.904 211.9 12.23

Value 169 184 199 213 214 228 229
Frequency 115 30 256 443 965 207 185
Proportion 0.052 0.014 0.116 0.201 0.438 0.094 0.084

FL_DATE
n missing distinct
2201 0 31

lowest : 01/01/2004 01/02/2004 01/03/2004 01/04/2004 01/05/2004, highest: 1/27/2004 1/28/2004 1/29/2004
1/30/2004 1/31/2004

FL_NUM
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95
2201 0 103 1 3815 2614 816 1744 2156 2385 6155 7792 7810

lowest : 746 806 808 810 814, highest: 7812 7814 7816 7818 7924

ORIGIN
n missing distinct
2201 0 3

Value BWI DCA IAD
Frequency 145 1370 686
Proportion 0.066 0.622 0.312

Weather
n missing distinct Info Sum Mean Gmd

```
   2201       0       2    0.043       32  0.01454  0.02867

--------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
DAY_WEEK
      n  missing distinct     Info     Mean      Gmd
   2201       0       7    0.978    3.905    2.172

Value          1     2     3     4     5     6     7
Frequency    308   307   320   372   391   250   253
Proportion 0.140 0.139 0.145 0.169 0.178 0.114 0.115
--------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
DAY_OF_MONTH
      n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   2201       0      31    0.999    16.02    10.01        2        4        8       16       23       28       30

lowest :  1  2  3  4  5, highest: 27 28 29 30 31
--------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
TAIL_NUM
      n  missing distinct
   2201       0      549

lowest : N10323 N10575 N11107 N11113 N11119, highest: N986DL N987DL N994DL N995CA N997DL
--------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
Flight.Status
      n  missing distinct
   2201       0       2

Value      delayed  ontime
Frequency      428    1773
Proportion   0.194   0.806
--------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------
```

========== Still 100% sample. Summary.

Data frame:crs$dataset[, c(crs$input, crs$risk, crs$target)] 2201 observations and 13 variables   Maximum # NAs:0

```
              Levels Storage
CRS_DEP_TIME         integer
CARRIER            8 integer
DEP_TIME             integer
DEST               3 integer
DISTANCE             integer
FL_DATE           31 integer
FL_NUM               integer
```

```
ORIGIN            3 integer
Weather            integer
DAY_WEEK           integer
DAY_OF_MONTH       integer
TAIL_NUM         549 integer
Flight.Status      2 integer

+-------------+-----------------------------------------------+
|Variable     |Levels                                         |
+-------------+-----------------------------------------------+
|CARRIER      |CO,DH,DL,MQ,OH,RU,UA,US                        |
+-------------+-----------------------------------------------+
|DEST         |EWR,JFK,LGA                                    |
+-------------+-----------------------------------------------+
|FL_DATE      |01/01/2004,01/02/2004,01/03/2004,01/04/2004    |
|             |01/05/2004,01/06/2004,01/07/2004,01/08/2004    |
|             |01/09/2004,01/10/2004,01/11/2004,01/12/2004    |
|             |1/13/2004,1/14/2004,1/15/2004,1/16/2004        |
|             |1/17/2004,1/18/2004,1/19/2004,1/20/2004        |
|             |1/21/2004,1/22/2004,1/23/2004,1/24/2004        |
|             |1/25/2004,1/26/2004,1/27/2004,1/28/2004        |
|             |1/29/2004,1/30/2004,1/31/2004                  |
+-------------+-----------------------------------------------+
|ORIGIN       |BWI,DCA,IAD                                    |
+-------------+-----------------------------------------------+
|TAIL_NUM     |N10323,N10575,N11107,N11113,N11119,N11121,N11127|
|             |N11137,N11140,N11150,N11535,N11536,N11539,N11548|
|             |N11551,N11565,N11612,N11641,N11656,N12126,N12135|
| |           |N940CA,N940DL,N949CA,N949DL,N951DL,N954DL,N956CA|
|             |N958CA,N959CA,N962DL,N963DL,N964DL,N966CA,N967CA|
|             |N970DL,N973CA,N974DL,N975CA,N983CA,N986DL,N987DL|
|             |N994DL,N995CA,N997DL                           |
+-------------+-----------------------------------------------+
|Flight.Status|delayed,ontime                                 |
+-------------+-----------------------------------------------+
```

For the simple distribution tables below the 1st and 3rd Qu.
refer to the first and third quartiles, indicating that 25%
of the observations have values of that variable which are
less than or greater than (respectively) the value listed.

```
 CRS_DEP_TIME     CARRIER       DEP_TIME     DEST
Min.   : 600   DH     :551   Min.   :  10   EWR: 665
1st Qu.:1000   RU     :408   1st Qu.:1004   JFK: 386
Median :1455   US     :404   Median :1450   LGA:1150
Mean   :1372   DL     :388   Mean   :1369
3rd Qu.:1710   MQ     :295   3rd Qu.:1709
Max.   :2130   CO     : 94   Max.   :2330
               (Other): 61
   DISTANCE          FL_DATE         FL_NUM     ORIGIN
Min.   :169.0   1/22/2004 :  86   Min.   : 746   BWI: 145
1st Qu.:213.0   01/06/2004:  85   1st Qu.:2156   DCA:1370
Median :214.0   01/08/2004:  85   Median :2385   IAD: 686
```

```
 Mean   :211.9   1/13/2004 :  85   Mean   :3815
 3rd Qu.:214.0   1/20/2004 :  85   3rd Qu.:6155
 Max.   :229.0   1/21/2004 :  85   Max.   :7924
                 (Other)   :1690
   Weather          DAY_WEEK      DAY_OF_MONTH
 Min.   :0.00000   Min.   :1.000   Min.   : 1.00
 1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 8.00
 Median :0.00000   Median :4.000   Median :16.00
 Mean   :0.01454   Mean   :3.905   Mean   :16.02
 3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:23.00
 Max.   :1.00000   Max.   :7.000   Max.   :31.00

   TAIL_NUM    Flight.Status
 N225DL :  65   delayed: 428
 N242DL :  56   ontime :1773
 N223DZ :  50
 N221DL :  45
 N241DL :  36
 N722UW :  36
 (Other):1913

Rattle timestamp: 2018-04-18 00:02:52 Rbohn
============================================================
```

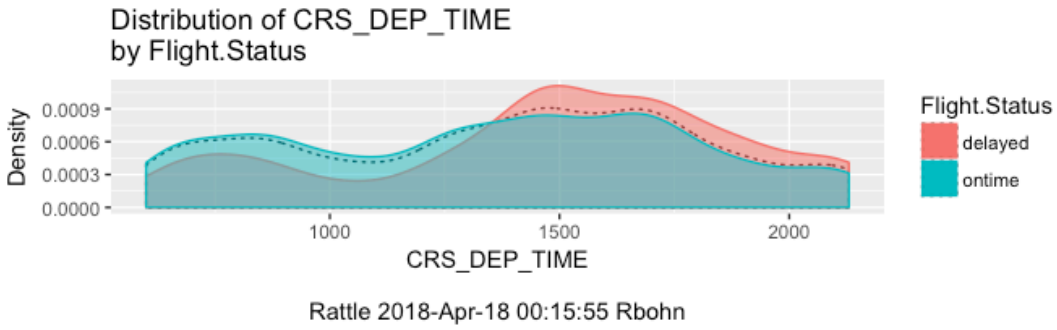
Now do variable distributions.
Save the results in various formats and file names.

Here is the code:

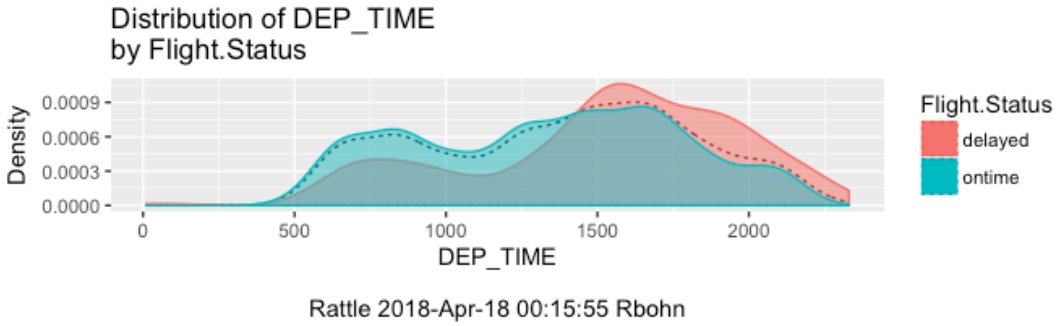
```
# Rattle timestamp: 2018-04-18 00:08:41 x86_64-apple-darwin15.6.0

# Display histogram plots for the selected variables.

# Use ggplot2 to generate histogram plot for Weather

# Generate the plot.

p01 <- crs %>%
  with(dataset[,]) %>%
  dplyr::mutate(Flight.Status=as.factor(Flight.Status)) %>%
  dplyr::select(Weather, Flight.Status) %>%
  ggplot2::ggplot(ggplot2::aes(x=Weather)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::geom_density(ggplot2::aes(fill=Flight.Status, colour=Flight.Status), alpha=0.55) +
  ggplot2::xlab("Weather\n\nRattle 2018-Apr-18 00:08:41 Rbohn") +
  ggplot2::ggtitle("Distribution of Weather\nby Flight.Status") +
  ggplot2::labs(fill="Flight.Status", y="Density")
```

```
# Use ggplot2 to generate histogram plot for DAY_WEEK

# Generate the plot.

p02 <- crs %>%
  with(dataset[,]) %>%
  dplyr::mutate(Flight.Status=as.factor(Flight.Status)) %>%
  dplyr::select(DAY_WEEK, Flight.Status) %>%
  ggplot2::ggplot(ggplot2::aes(x=DAY_WEEK)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::geom_density(ggplot2::aes(fill=Flight.Status, colour=Flight.Status), alpha=0.55) +
  ggplot2::xlab("DAY_WEEK\n\nRattle 2018-Apr-18 00:08:42 Rbohn") +
  ggplot2::ggtitle("Distribution of DAY_WEEK\nby Flight.Status") +
  ggplot2::labs(fill="Flight.Status", y="Density")

# Use ggplot2 to generate histogram plot for DAY_OF_MONTH

# Generate the plot.

p03 <- crs %>%
  with(dataset[,]) %>%
  dplyr::mutate(Flight.Status=as.factor(Flight.Status)) %>%
  dplyr::select(DAY_OF_MONTH, Flight.Status) %>%
  ggplot2::ggplot(ggplot2::aes(x=DAY_OF_MONTH)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::geom_density(ggplot2::aes(fill=Flight.Status, colour=Flight.Status), alpha=0.55) +
  ggplot2::xlab("DAY_OF_MONTH\n\nRattle 2018-Apr-18 00:08:42 Rbohn") +
  ggplot2::ggtitle("Distribution of DAY_OF_MONTH\nby Flight.Status") +
  ggplot2::labs(fill="Flight.Status", y="Density")

# Display the plots. In Rstudio

gridExtra::grid.arrange(p01, p02, p03)
```

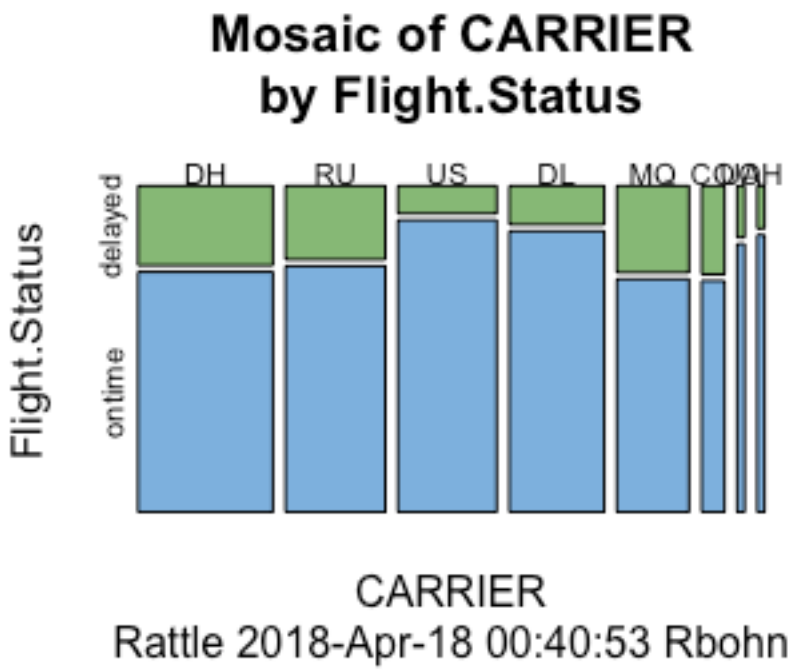
More useful plots:



Rattle 2018-Apr-18 00:15:55 Rbohn



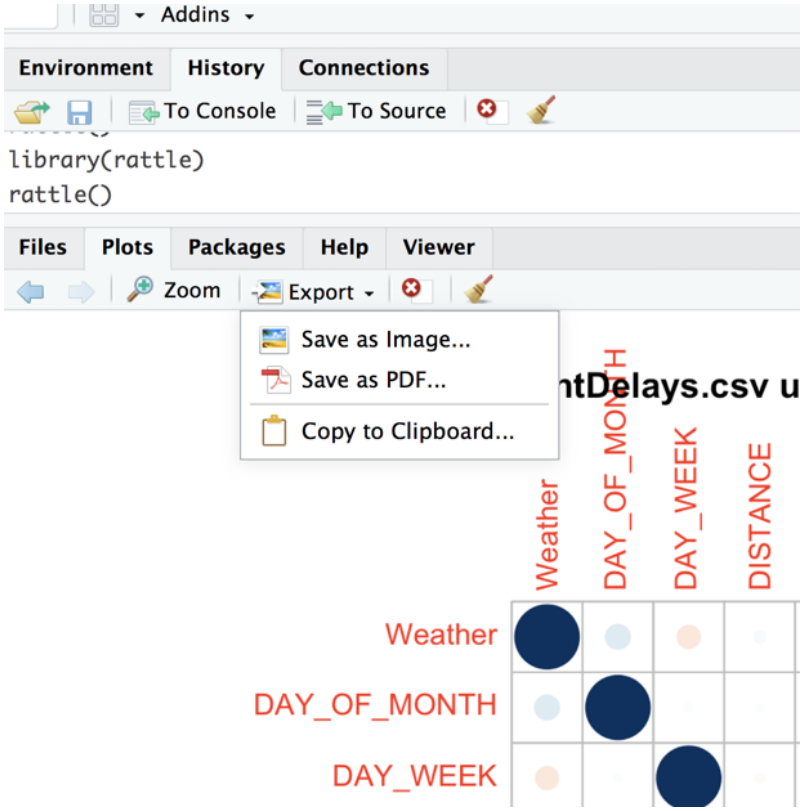
Rattle 2018-Apr-18 00:15:55 Rbohn

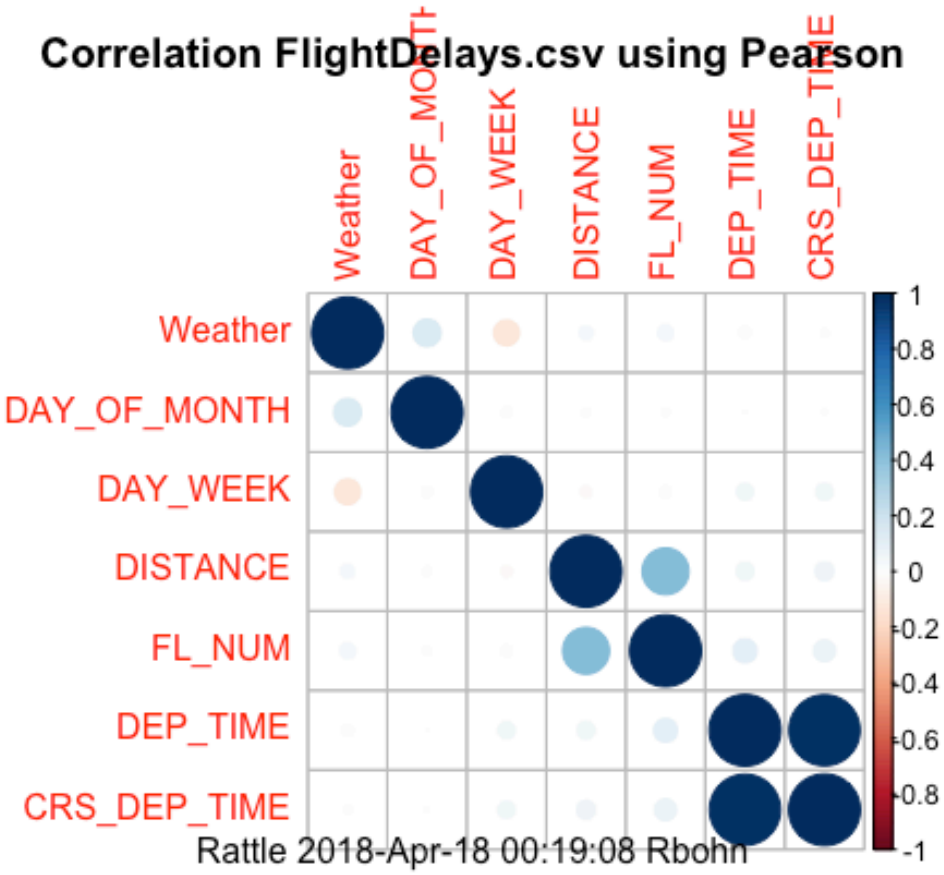


Correlation plot

How to create the correlation plot:

Delayed flights in Rattle. Chapter 10.5

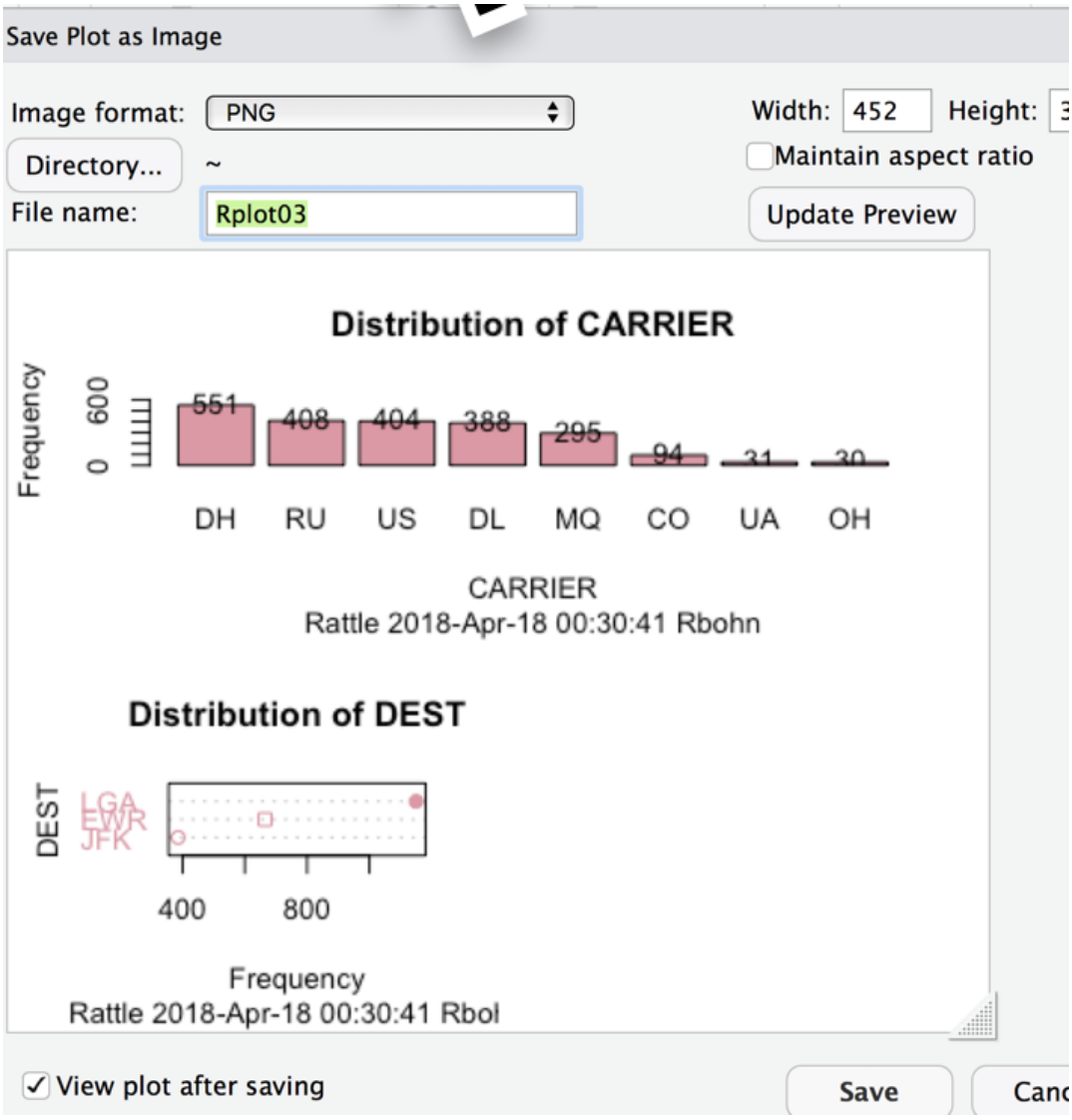




Working with plots

The plots appear in Rstudio, not in Rattle!
In lower right, under *Plots*
Once you get what you want, Use Zoom
and then Export

By turning OFF the "maintain aspect ratio," you can enlarge along X or Y axis to emphasize what you want.



OOPS! Translate day of week into a categorical variable.

Delayed flights in Rattle. Chapter 10.5

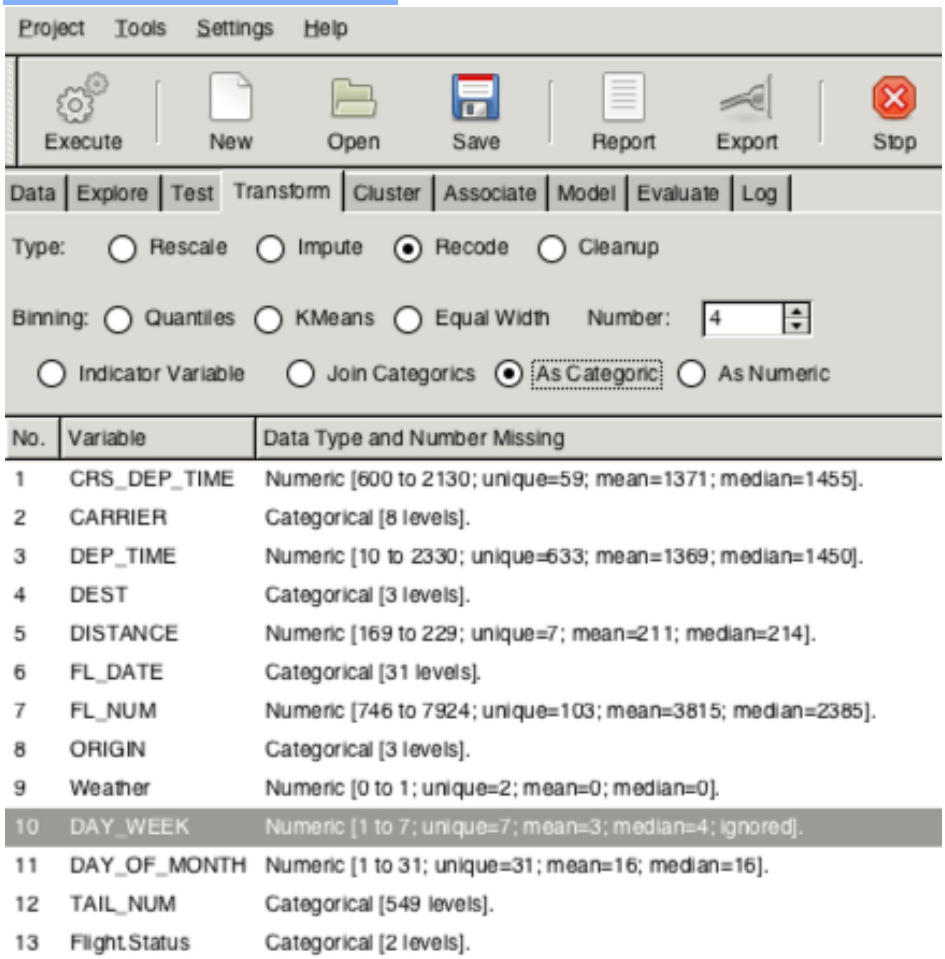
Rattle timestamp: 2018-04-18 00:43:44 x86_64-apple-darwin15.6.0

Remap variables.

Transform into a factor. Day of week; flight number (if it's included at all). Weather??

```
crs$dataset[["TFC_DAY_WEEK"]] <- as.factor(crs$dataset[["DAY_WEEK"]])

ol <- levels(crs$dataset[["TFC_DAY_WEEK"]])
lol <- length(ol)
nl <- c(sprintf("[%s,%s]", ol[1], ol[1]), sprintf("(%s,%s]", ol[-lol], ol[-1]))
levels(crs$dataset[["TFC_DAY_WEEK"]]) <- nl
```

here is how to do it in Rattle:

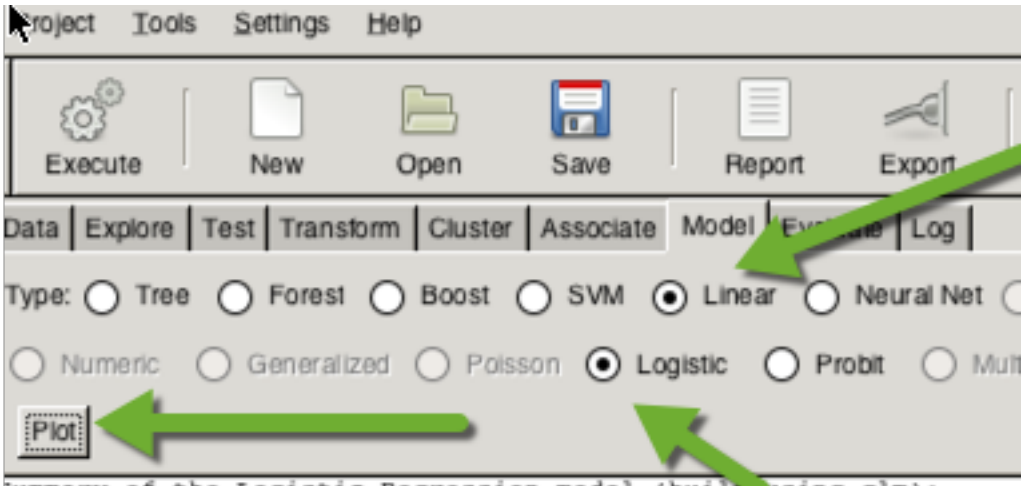


No.	Variable	Data Type and Number Missing
1	CRS_DEP_TIME	Numeric [600 to 2130; unique=59; mean=1371; median=1455].
2	CARRIER	Categorical [8 levels].
3	DEP_TIME	Numeric [10 to 2330; unique=633; mean=1369; median=1450].
4	DEST	Categorical [3 levels].
5	DISTANCE	Numeric [169 to 229; unique=7; mean=211; median=214].
6	FL_DATE	Categorical [31 levels].
7	FL_NUM	Numeric [746 to 7924; unique=103; mean=3815; median=2385].
8	ORIGIN	Categorical [3 levels].
9	Weather	Numeric [0 to 1; unique=2; mean=0; median=0].
10	DAY_WEEK	Numeric [1 to 7; unique=7; mean=3; median=4; ignored].
11	DAY_OF_MONTH	Numeric [1 to 31; unique=31; mean=16; median=16].
12	TAIL_NUM	Categorical [549 levels].
13	Flight Status	Categorical [2 levels].

Then go. back to DATA tab, and re-execute. (Write down the SEED you used)

Variables to include and leave out:

Actually running a model




```
Summary of the Logistic Regression model (built using glm):

***Note*** Singularities were found in the modeling
and are indicated by an NA in the following table.
This is often the case when variables are linear
combinations of other variables, or the variable
has a constant value.  These variables will be ignored
when using the model to score new data and will not be
included as parameters in the exported scoring routine.

Call:
glm(formula = Flight.Status ~ ., family = binomial(link = "logit
    data = crs$dataset[crs$train, c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904   0.1840   0.3403   0.4811   1.6291

Coefficients: (7 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       17.182593 700.990902   0.025 0.980444
CRS_DEP_TIME       0.033061   0.002731  12.108  < 2e-16 ***
CARRIERDH          2.461424   0.644253   3.821 0.000133 ***
CARRIERDL          1.446465   0.552663   2.617 0.008864 **
CARRIERMQ          0.036246   0.548188   0.066 0.947283
CARRIEROH          3.160753   1.203465   2.626 0.008630 **
CARRIERRU          0.855170   0.469516   1.821 0.068548 .
CARRIERUA          3.501608   1.251020   2.799 0.005126 **
CARRIERUS          1.026982   0.587437   1.748 0.080422 .
DEP_TIME          -0.033741   0.002724 -12.386  < 2e-16 ***
DESTJFK            0.178648   0.380373   0.470 0.638594
DESTLGA           -0.055384   0.389838  -0.142 0.887025
FL_DATE01/02/2004 -14.582760 700.990911  -0.021 0.983403
FL_DATE01/03/2004 -14.843857 700.990956  -0.021 0.983106
FL_DATE01/04/2004 -15.852190 700.990775  -0.023 0.981958
FL_DATE01/05/2004 -16.088873 700.990723  -0.023 0.981689
FL_DATE01/06/2004 -15.128291 700.990737  -0.022 0.982782
FL_DATE01/07/2004 -15.558256 700.990736  -0.022 0.982293
FL_DATE01/08/2004 -14.678790 700.990802  -0.021 0.983293
FL_DATE01/09/2004 -15.457318 700.990727  -0.022 0.982408
FL_DATE01/10/2004 -14.967652 700.990887  -0.021 0.982965
FL_DATE01/11/2004 -14.847006 700.990818  -0.021 0.983102
```

It's clear we misunderstood the date variable. Get rid of it, or make it non-categorical. Then repeat the model run.

Run the Confusion Matrix.
Error matrix for the Linear model on FlightDelays.csv [validate] (counts):

```
        Predicted
Actual    delayed ontime
  delayed       6     55
  ontime        0    269
         Predicted
Actual    Error
  delayed  90.2
  ontime    0.0

Error matrix for the Linear model on FlightDelays.csv [validate]
(proportions):

         Predicted
Actual    delayed ontime
  delayed     1.8   16.7
  ontime      0.0   81.5
         Predicted
Actual    Error
  delayed  90.2
  ontime    0.0

Overall error: 16.7%, Averaged class error: 45.1%

Rattle timestamp: 2018-04-18 12:21:50 Rbohn
```

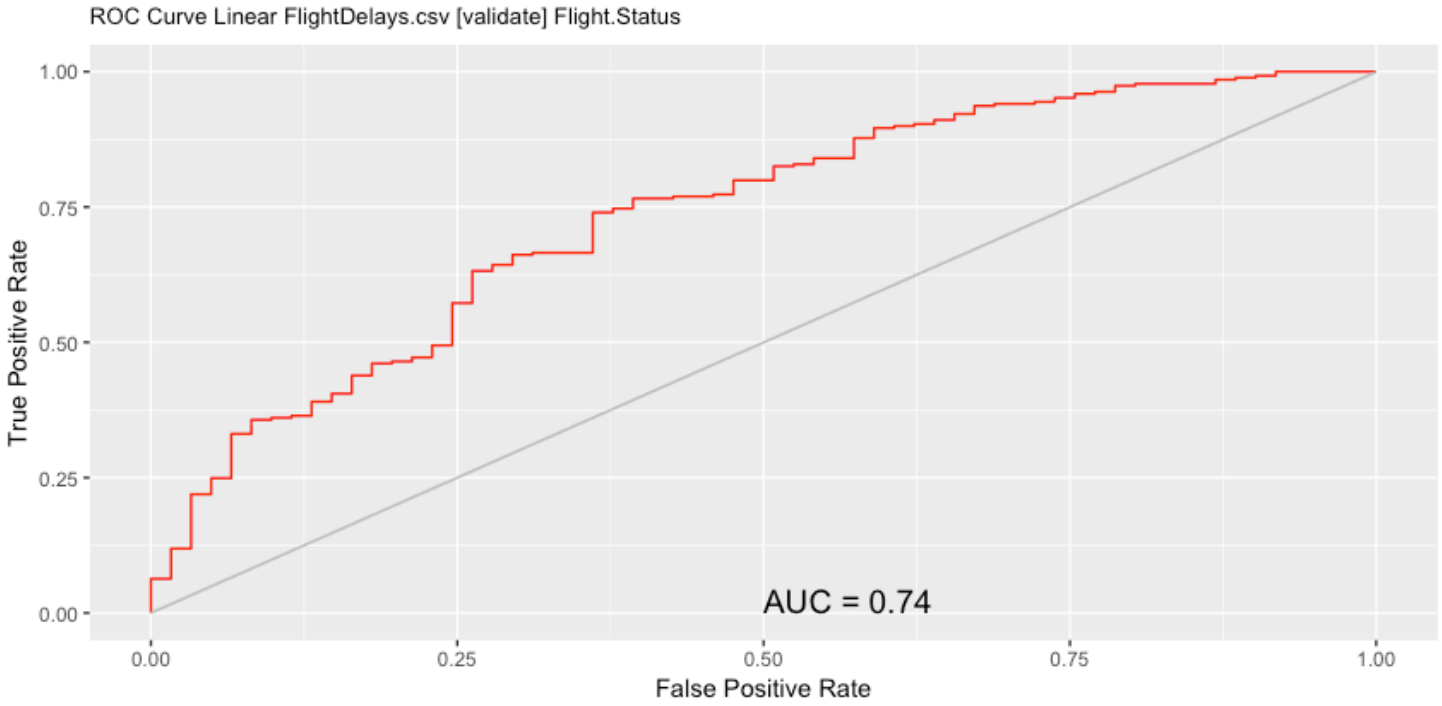
Here is the R code for the confusion matrix

```
#=====================================================================
# Rattle timestamp: 2018-04-18 12:21:50 x86_64-apple-darwin15.6.0

# Evaluate model performance on the validation dataset.

# Generate an Error Matrix for the Linear model.

# Obtain the response from the Linear model.

crs$pr <- as.vector(ifelse(predict(crs$glm,
  type    = "response",
  newdata = crs$dataset[crs$validate, c(crs$input, crs$target)]) > 0.5, "ontime", "delayed"))

# Generate the confusion matrix showing counts.

rattle::errorMatrix(crs$dataset[crs$validate, c(crs$input, crs$target)]$Flight.Status, crs$pr, count=TRUE)
```

```
# Generate the confusion matrix showing proportions.

(per <- rattle::errorMatrix(crs$dataset[crs$validate, c(crs$input, crs$target)]$Flight.Status, crs$pr))

# Calculate the overall error percentage.

cat(100-sum(diag(per), na.rm=TRUE))

# Calculate the averaged class error percentage.

cat(mean(per[,"Error"], na.rm=TRUE))

# Plot the model evaluation.

ttl <- genPlotTitleCmd("Linear Model",crs$dataname,vector=TRUE)
plot(crs$glm, main=ttl[1])
```

==========

Create the ROC curve, and other displays of results.



Remember to SAVE your work from Rattle. It saves as a file ending in .rattle
Give it a useful name. Keep this file with the rest of your documentation for what you have done.