Data Import and Submission Export Tutorial

This tutorial will use R to import the data and output a sample submission.

To begin, make sure your R session has its working directory set to the same directory where your data is located. To view your current working directory, run the command `getwd()` in the R console. Use the R options `File -> Change dir...` in the RGui to set your working directory or use the `setwd('insert/your/path/to/data.csv')` command to set your working directory.

To double check that `train.csv` and `test.csv` are in your current working directory, the following command should return `TRUE` twice, as seen in the output below.

```
In [1]: c('train.csv', 'test.csv') %in% list.files()
```

```
Out[1]:     TRUE  TRUE
```

Data Import

Now that we have an R session with our two data files in the working directory, read the comma separated data using the `read.csv()` function and view a few predictor summaries.

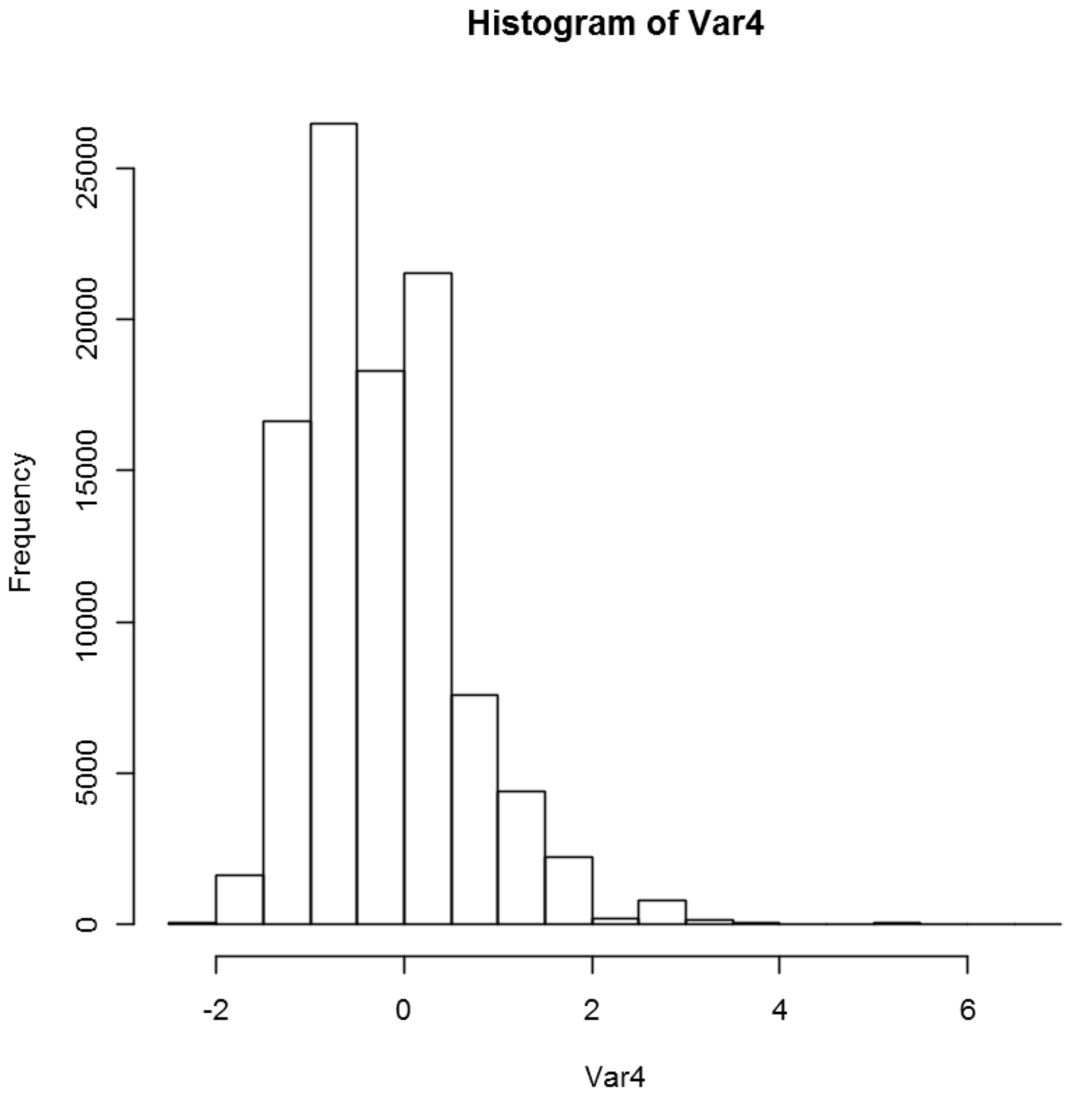
```
In [2]: train = read.csv('train.csv')
```

```
In [3]: head(train)
```

Out[3]:

	RowID	CalendarYear	ModelYear	Make	Model	Cat1	Cat2	Cat3	Cat4	Ca
1	418079	2005	2004	AU	AU.14	B	C	A	A	A
2	232625	2006	2003	R	R.30	B	C	B	A	A
3	379029	2006	2006	AU	AU.14	B	A	A	A	A
4	181458	2007	2000	BU	BU.38	F	C	A	C	A
5	192434	2005	1999	BU	BU.38	F	A	A	C	A
6	443321	2007	2005	AU	AU.11	B	C	B	A	A

```
In [4]: hist(train$Var4, main = "Histogram of Var4", xlab = "Var4")
```



```
In [5]: summary(train$Var8)

Out[5]:     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
        -1.48500 -0.52320 -0.24210 -0.00946  0.16210 33.90000
```

Also read the test set into your R session via the `read.csv()` function.

```
In [6]: test = read.csv('test.csv')
```

In [7]:

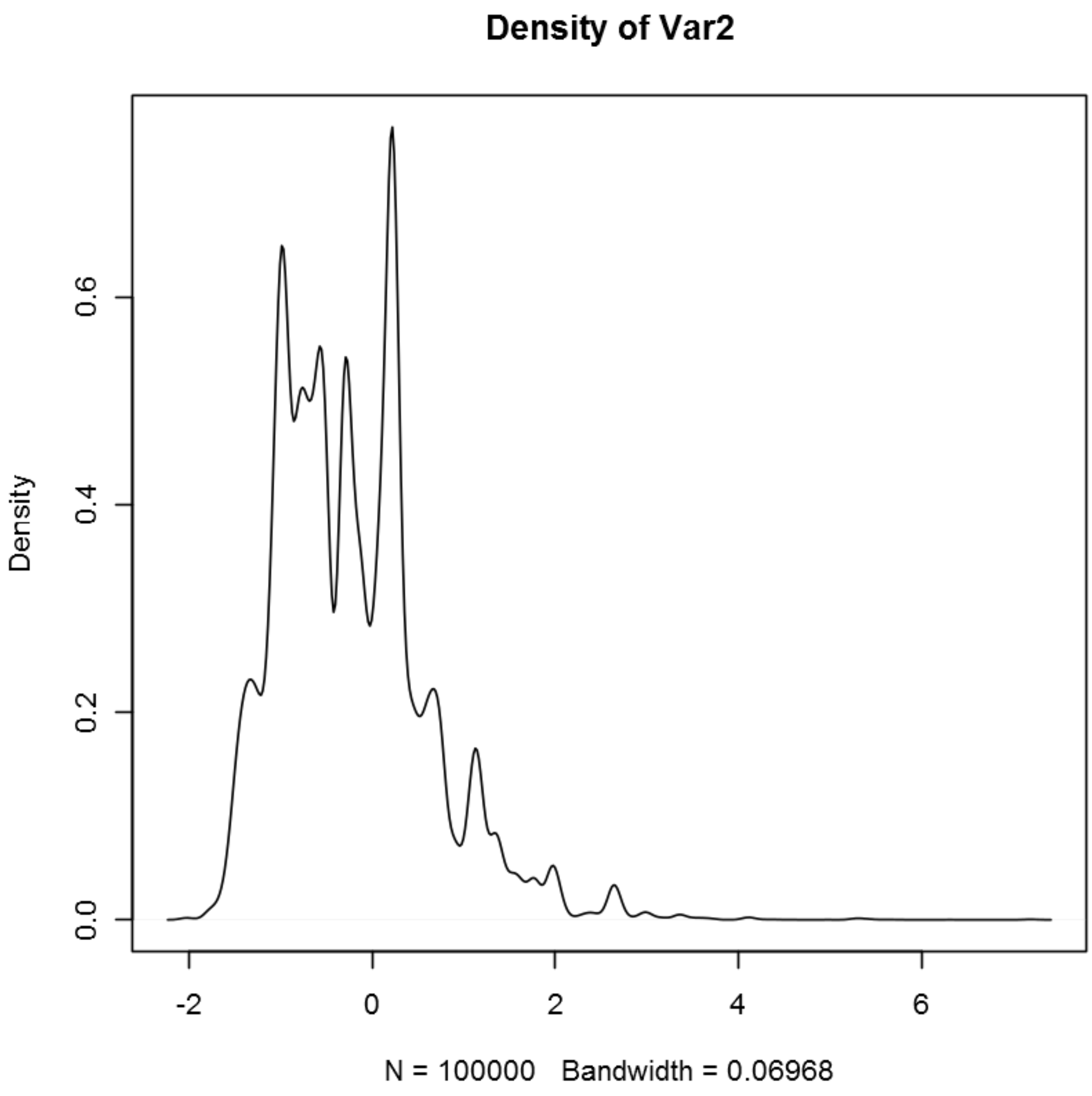
```
head(test)
```

Out[7]:

	RowID	CalendarYear	ModelYear	Make	Model	Cat1	Cat2	Cat3	Cat4	Cat5
1	491169	2009	2006	AN	AN.4	B	B	B	A	A
2	661907	2009	2001	BU	BU.5	B	C	B	A	A
3	489459	2009	2005	Y	Y.21	B	C	A	A	A
4	496824	2009	2006	BF	BF.18	B	C	A	A	A
5	464567	2009	2008	K	K.40	E	C	A	A	A
6	307296	2009	2007	K	K.40	E	C	A	A	A

In [8]:

```
plot(density(train$Var2), main = "Density of Var2")
```



In [9]:

```
dim(train)
```

Out[9]:

```
100000   32
```

In [10]:

```
dim(test)
```

Out[10]:

```
40000   31
```

The training and testing sets have a different number of columns. This is, of course, because the test set does not contain the response variable. The following command will tell us which column is contained in the training set and not in the testing set.

In [11]:

```
setdiff(names(train), names(test))
```

Out[11]: "Response"

Data Export

When making a submission, the predictions need to be exported in a certain fashion. The example below will generate random uniform numbers and use them as our predictions.

In [12]:

```
numberOfObservationsInTestSet = nrow(test)
vectorOfPredictions = runif(numberOfObservationsInTestSet, 0, 1)
summary(vectorOfPredictions)
```

Out[12]:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0000013 0.2508000 0.4968000 0.4987000 0.7473000 0.9999000 
```

In [13]:

```
outputDataSet = data.frame("RowID" = test$RowID,
                           "ProbabilityOfResponse" = vectorOfPrediction
s)
```

Inspect data set before export

In [14]:

```
head(outputDataSet)
```

Out[14]:

	RowID	ProbabilityOfResponse
1	491169	0.5298258
2	661907	0.7990218
3	489459	0.42184
4	496824	0.7455896
5	464567	0.1310719
6	307296	0.8185932

The following command will output a comma separated file to the current working directory. Find your current working directory again by executing the `getwd()` command.

In [15]:

```
write.csv(outputDataSet, "submissionExample.csv", row.names = FALSE)
```