Data Quality Report

Customer Churn Data

Dr. Simon Caton

13/9/2018

Intro to the method of reporting

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

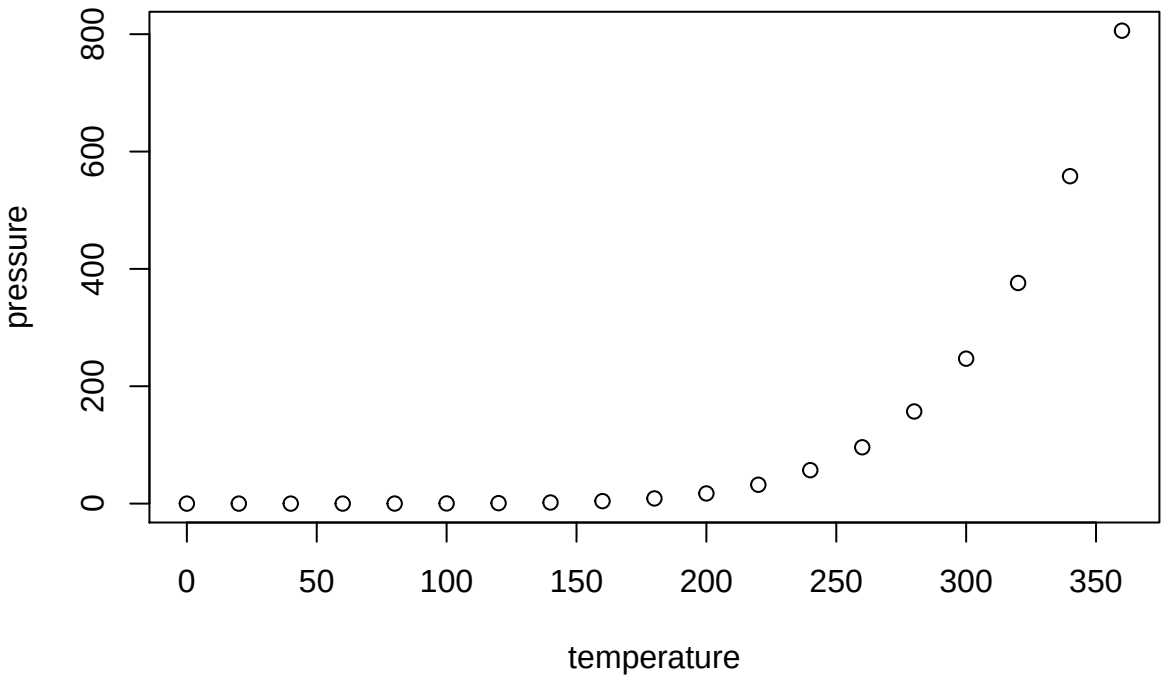
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

DQR Continuous Data

From the code provided in the slides, we can rebuild our table. Either copy and paste all the code into a code block like the ones above, or just reference the .R file, like this (make sure you comment out anything you don't want to run). Note that I have set *eval=FALSE* that's because I don't want it to run – it will make a mess of the document.

```
source("1-cleaned.R")
```

So let's rebuild the table:

Table 1: Continuous Data Quality Report

Feature	Instances	Missing	Cardinality	Min	FirstQuantile	Median	ThirdQuantile	Max	Mean
Contract Len	4014	0	2	18.00	18.00	24.00	24.00	24.00	21.23
Avg Bill	4014	0	3633	35.00	45.82	55.78	74.18	138.62	58.80
Avg OOB TC	4014	0	4009	0.00	1.48	3.93	9.43	38.61	5.99
Avg OOB TA	4014	0	4009	0.00	0.03	0.07	0.13	0.34	0.09
Avg Voice	4014	0	3985	45.62	143.70	322.56	483.25	1162.60	346.82
Avg SMS	4014	0	3978	64.71	189.64	484.37	722.78	1499.82	478.17
Avg Data	4014	0	4014	199.21	991.21	3849.56	6055.09	11705.82	3801.59
Age	4014	0	52	14.00	27.00	36.00	47.00	74.00	37.75

Oops, our table is too wide! Either we would need to break it down a little like this:

```
kable(df_numeric[, c(1:4)], caption = "Continuous Data Quality Report", digits = 2)
```

Table 2: Continuous Data Quality Report

Feature	Instances	Missing	Cardinality
Contract Len	4014	0	2
Avg Bill	4014	0	3633
Avg OOB TC	4014	0	4009
Avg OOB TA	4014	0	4009
Avg Voice	4014	0	3985
Avg SMS	4014	0	3978
Avg Data	4014	0	4014
Age	4014	0	52

```
kable(df_numeric[, c(1,5:8)], caption = "Continuous Data Quality Report", digits = 2)
```

Table 3: Continuous Data Quality Report

Feature	Min	FirstQuantile	Median	ThirdQuantile
Contract Len	18.00	18.00	24.00	24.00
Avg Bill	35.00	45.82	55.78	74.18
Avg OOB TC	0.00	1.48	3.93	9.43
Avg OOB TA	0.00	0.03	0.07	0.13
Avg Voice	45.62	143.70	322.56	483.25
Avg SMS	64.71	189.64	484.37	722.78
Avg Data	199.21	991.21	3849.56	6055.09
Age	14.00	27.00	36.00	47.00

Or display it in landscape:

```
library(kableExtra)
landscape(knitr::kable(df_numeric, caption = "Continuous Data Quality Report", digits = 2))
```

Table 4: Continuous Data Quality Report

Feature	Instances	Missing	Cardinality	Min	FirstQuantile	Median	ThirdQuantile	Max	Mean	Stdev
Contract Len	4014	0	2	18.00	18.00	24.00	24.00	24.00	21.23	2.99
Avg Bill	4014	0	3633	35.00	45.82	55.78	74.18	138.62	58.80	18.69
Avg OOB TC	4014	0	4009	0.00	1.48	3.93	9.43	38.61	5.99	5.69
Avg OOB TA	4014	0	4009	0.00	0.03	0.07	0.13	0.34	0.09	0.06
Avg Voice	4014	0	3985	45.62	143.70	322.56	483.25	1162.60	346.82	224.18
Avg SMS	4014	0	3978	64.71	189.64	484.37	722.78	1499.82	478.17	299.66
Avg Data	4014	0	4014	199.21	991.21	3849.56	6055.09	11705.82	3801.59	2755.96
Age	4014	0	52	14.00	27.00	36.00	47.00	74.00	37.75	13.57