A Preprocessing Pipeline for Exposome Data

Jeff Sorbo

19th December 2016

1 Introduction

This project will include multiple preprocessing techniques for exposome data.

2 Business Problem

Describe discussions with client (business experts) and record decisions made and shared understanding of the business problem.

3 Data Sources

Identify the data sources and discuss access with the data owners. Document data sources, integrity, providence, and dates.

4 Data Preparation

Load the data into R and perform various operations on the data to shape it for modelling.

5 Data Exploration

We should always understand our data by exploring it in various ways. Include data summaries and various plots that give insights.

6 Model Building

Include all models built and parameters tried. Include R code and model evaluations.

7 Deployment

Choose the model to deploy and export it, perhaps as PMML.

8 Echoing Code

```
x <- runif(1000) * 1000
head(x)

## [1] 591.14071 102.23089 547.29147  23.40307 627.79111 332.79199

mean(x)

## [1] 504.001
```

9 Non-Echoing Code

```
## [1] 253.3932 855.8366 215.1099 665.8442 929.3086 674.9027
## [1] 508.6685
```

10 Inline Code

Today's date is Monday, 19 December 2016.

The weather dataset from rattle (Williams, 2014) has 366 (i.e., 366) observations including observations of the following 4 variables: MinTemp, MaxTemp, Rainfall, Evaporation (i.e., MinTemp, MaxTemp, Rainfall, Evaporation).

11 Table with kable

```
#library(rattle)
library(dplyr)

## Warning:  package 'dplyr' was built under R version 3.2.5
##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union

set.seed(42)

dsname <- "weatherAUS"
ds <- tbl_df(get(dsname))
nobs <- nrow(ds)
obs <- sample(nobs, 20)
vars <- 2:7
```

```
ds <- ds[obs, vars]
kable(ds, row.names=FALSE, digits=0, booktabs=TRUE)
```

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	14	22	0	6	9
Launceston	-3	11	0	NA	NA
Williamtown	11	16	32	NA	NA
PerthAirport	9	20	1	1	4
GoldCoast	10	21	0	NA	NA
Portland	5	16	0	3	12
Woomera	18	34	0	NA	NA
NorahHead	19	27	0	NA	NA
Townsville	16	30	0	7	11
MountGambier	6	20	0	1	6
MelbourneAirport	5	21	0	4	9
Nuriootpa	17	31	0	10	13
Launceston	9	15	0	NA	NA
WaggaWagga	10	31	0	11	14
MelbourneAirport	8	20	0	5	6
AliceSprings	23	37	0	13	10
Darwin	18	34	0	6	9
Newcastle	7	19	0	NA	NA
Melbourne	13	20	0	6	6
Dartmoor	10	17	2	7	8

12 Table with xtable

```
#library(rattle)
library(xtable)

## Warning:  package 'xtable' was built under R version 3.2.5

dst <- weatherAUS[sample(nobs, 20), vars]
xtable(dst)
```

```
print(xtable(ds), include.rownames=FALSE)
```

```
print(xtable(ds, digits=1), include.rownames=FALSE)
```

```
dst <- ds
dst[-1] <- sample(10000:99999, nrow(dst)) * dst[-1]
print(xtable(dst, digits=0), include.rownames=FALSE)
```

	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
108592	Hobart	12.60	20.80	0.00	2.40	5.70
16662	NorahHead	14.70	19.80	0.00		
118783	Katherine	15.30	34.90	0.00	7.20	
113711	AliceSprings	4.10	23.20	0.00	4.00	4.80
9902	CoffsHarbour	17.70	27.20	0.00		
61765	Nhil	5.80	19.90	0.00		
46869	Ballarat	7.30	26.10	0.00		
108791	Hobart	10.60	20.40	0.00	4.00	9.30
53686	MelbourneAirport	18.30	21.80	0.00	15.20	0.50
100413	Perth	13.20	19.60	1.00	4.20	11.00
88592	Woomera	15.10	22.60	15.40	1.80	
97415	PearceRAAF	8.10	26.00	0.00		12.90
46615	Ballarat	11.80	28.00	0.00		
82293	Adelaide	8.80	14.30	0.40		
475	Albury	16.70	31.90	0.00		
100037	Perth	1.00	19.00	5.40	1.60	11.10
881	Albury	3.60	15.90	0.00		
24941	Richmond	18.10	39.70	0.00		
108884	Hobart	15.80	26.30	0.60	5.60	2.40
73475	Cairns	15.80	25.00	0.00	5.40	10.60

```r
print(xtable(dst, digits=0),
include.rownames=FALSE,
format.args=list(big.mark=","))
```

```r
print(xtable(ds,
digits=0,
caption="Selected observations from \\textbf{weatherAUS}."),
include.rownames=FALSE)
```

```r
print(xtable(ds,
digits=0,
caption="Selected observations from \\textbf{weatherAUS}.",
label="MyTable"),
include.rownames=FALSE)
```

```r
print(xtable(ds,
digits=0,
caption=paste("Here we include in the caption a sample of \\LaTeX{}",
"symbols that can be included in the string, and note that the",
"caption string can be the result of R commands, using paste()",
"in this instance. Some sample symbols include:",
"$\\alpha$ $\\longrightarrow$ $\\wp$.",
"We also get a timestamp from R:",
Sys.time()),
```

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	13.60	22.40	0.00	5.60	8.70
Launceston	-2.80	10.80	0.20		
Williamtown	11.10	15.70	31.60		
PerthAirport	9.00	20.10	0.80	1.20	4.00
GoldCoast	9.70	21.10	0.00		
Portland	5.00	15.80	0.00	3.20	12.50
Woomera	18.50	34.30	0.00		
NorahHead	18.80	26.70	0.00		
Townsville	15.80	29.70	0.00	7.00	10.70
MountGambier	6.00	20.10	0.00	1.00	6.10
MelbourneAirport	5.20	20.90	0.00	4.20	9.30
Nuriootpa	17.40	31.40	0.20	9.80	13.40
Launceston	8.80	14.60	0.00		
WaggaWagga	9.90	30.60	0.00	10.80	13.70
MelbourneAirport	8.00	20.00	0.00	4.80	5.50
AliceSprings	23.00	37.30	0.20	13.20	9.60
Darwin	18.50	33.90	0.00	6.20	8.60
Newcastle	6.90	19.20	0.00		
Melbourne	13.10	20.20	0.00	5.80	6.00
Dartmoor	9.70	16.70	2.40	6.60	7.70

```
label="SymbolCaption"),
include.rownames=FALSE)
```

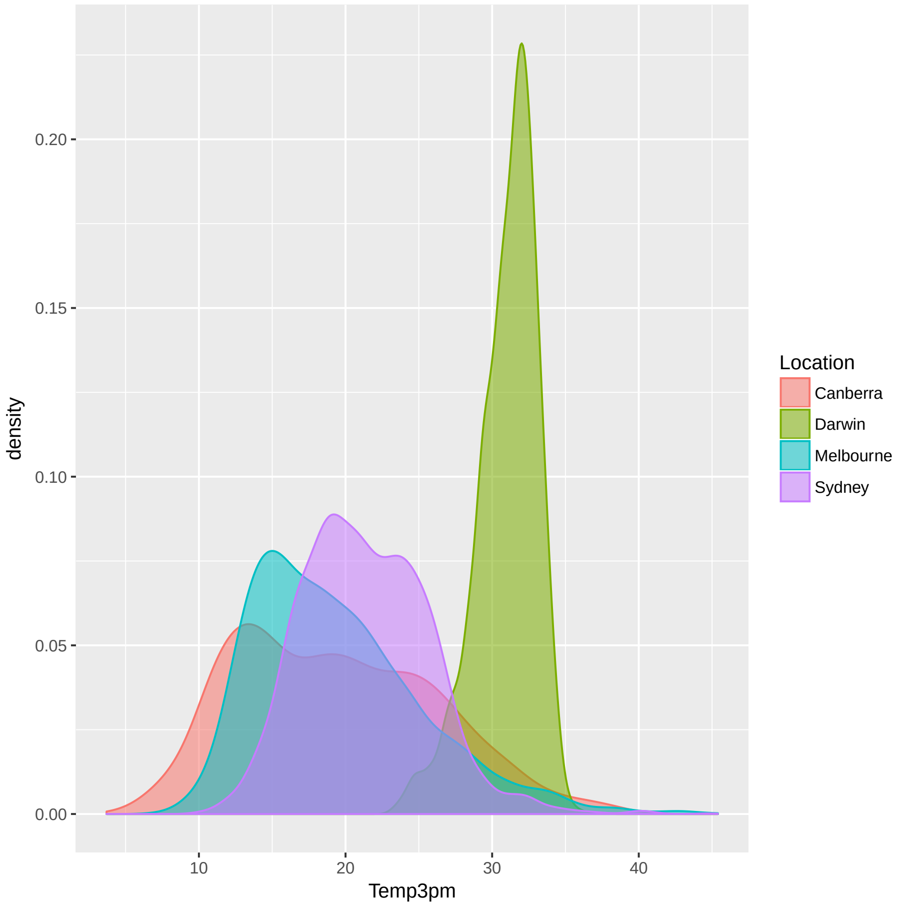
13 Figure

```
library(rattle) # For the weatherAUS dataset.
library(ggplot2) # To generate a density plot.

## Warning:  package 'ggplot2' was built under R version 3.2.5

cities <- c("Canberra", "Darwin", "Melbourne", "Sydney")
ds <- subset(weatherAUS, Location %in% cities & ! is.na(Temp3pm))
p <- ggplot(ds, aes(Temp3pm, colour=Location, fill=Location))
p <- p + geom_density(alpha=0.55)
p
```

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	13.6	22.4	0.0	5.6	8.7
Launceston	-2.8	10.8	0.2		
Williamtown	11.1	15.7	31.6		
PerthAirport	9.0	20.1	0.8	1.2	4.0
GoldCoast	9.7	21.1	0.0		
Portland	5.0	15.8	0.0	3.2	12.5
Woomera	18.5	34.3	0.0		
NorahHead	18.8	26.7	0.0		
Townsville	15.8	29.7	0.0	7.0	10.7
MountGambier	6.0	20.1	0.0	1.0	6.1
MelbourneAirport	5.2	20.9	0.0	4.2	9.3
Nuriootpa	17.4	31.4	0.2	9.8	13.4
Launceston	8.8	14.6	0.0		
WaggaWagga	9.9	30.6	0.0	10.8	13.7
MelbourneAirport	8.0	20.0	0.0	4.8	5.5
AliceSprings	23.0	37.3	0.2	13.2	9.6
Darwin	18.5	33.9	0.0	6.2	8.6
Newcastle	6.9	19.2	0.0		
Melbourne	13.1	20.2	0.0	5.8	6.0
Dartmoor	9.7	16.7	2.4	6.6	7.7



Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	600576	989184	0	247296	384192
Launceston	-137813	531565	9844		
Williamtown	148385	209878	422429		
PerthAirport	878535	1962062	78092	117138	390460
GoldCoast	473893	1030841	0		
Portland	480885	1519597	0	307766	1202213
Woomera	1663002	3083296	0		
NorahHead	1270748	1804733	0		
Townsville	1538588	2892156	0	681653	1041955
MountGambier	394134	1320349	0	65689	400703
MelbourneAirport	208026	836105	0	168021	372047
Nuriootpa	716932	1293774	8241	403789	552120
Launceston	403550	669527	0		
WaggaWagga	798059	2466727	0	870610	1104384
MelbourneAirport	108024	270060	0	64814	74267
AliceSprings	1779740	2886274	15476	1021416	742848
Darwin	1312464	2405002	0	439853	610118
Newcastle	175329	487872	0		
Melbourne	438758	676559	0	194259	200958
Dartmoor	545984	939993	135089	371494	433410

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	600,576	989,184	0	247,296	384,192
Launceston	-137,813	531,565	9,844		
Williamtown	148,385	209,878	422,429		
PerthAirport	878,535	1,962,062	78,092	117,138	390,460
GoldCoast	473,893	1,030,841	0		
Portland	480,885	1,519,597	0	307,766	1,202,213
Woomera	1,663,002	3,083,296	0		
NorahHead	1,270,748	1,804,733	0		
Townsville	1,538,588	2,892,156	0	681,653	1,041,955
MountGambier	394,134	1,320,349	0	65,689	400,703
MelbourneAirport	208,026	836,105	0	168,021	372,047
Nuriootpa	716,932	1,293,774	8,241	403,789	552,120
Launceston	403,550	669,527	0		
WaggaWagga	798,059	2,466,727	0	870,610	1,104,384
MelbourneAirport	108,024	270,060	0	64,814	74,267
AliceSprings	1,779,740	2,886,274	15,476	1,021,416	742,848
Darwin	1,312,464	2,405,002	0	439,853	610,118
Newcastle	175,329	487,872	0		
Melbourne	438,758	676,559	0	194,259	200,958
Dartmoor	545,984	939,993	135,089	371,494	433,410

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	14	22	0	6	9
Launceston	-3	11	0		
Williamtown	11	16	32		
PerthAirport	9	20	1	1	4
GoldCoast	10	21	0		
Portland	5	16	0	3	12
Woomera	18	34	0		
NorahHead	19	27	0		
Townsville	16	30	0	7	11
MountGambier	6	20	0	1	6
MelbourneAirport	5	21	0	4	9
Nuriootpa	17	31	0	10	13
Launceston	9	15	0		
WaggaWagga	10	31	0	11	14
MelbourneAirport	8	20	0	5	6
AliceSprings	23	37	0	13	10
Darwin	18	34	0	6	9
Newcastle	7	19	0		
Melbourne	13	20	0	6	6
Dartmoor	10	17	2	7	8

Table 1: Selected observations from **weatherAUS**.

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	14	22	0	6	9
Launceston	-3	11	0		
Williamtown	11	16	32		
PerthAirport	9	20	1	1	4
GoldCoast	10	21	0		
Portland	5	16	0	3	12
Woomera	18	34	0		
NorahHead	19	27	0		
Townsville	16	30	0	7	11
MountGambier	6	20	0	1	6
MelbourneAirport	5	21	0	4	9
Nuriootpa	17	31	0	10	13
Launceston	9	15	0		
WaggaWagga	10	31	0	11	14
MelbourneAirport	8	20	0	5	6
AliceSprings	23	37	0	13	10
Darwin	18	34	0	6	9
Newcastle	7	19	0		
Melbourne	13	20	0	6	6
Dartmoor	10	17	2	7	8

Table 2: Selected observations from **weatherAUS**.

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
Hobart	14	22	0	6	9
Launceston	-3	11	0		
Williamtown	11	16	32		
PerthAirport	9	20	1	1	4
GoldCoast	10	21	0		
Portland	5	16	0	3	12
Woomera	18	34	0		
NorahHead	19	27	0		
Townsville	16	30	0	7	11
MountGambier	6	20	0	1	6
MelbourneAirport	5	21	0	4	9
Nuriootpa	17	31	0	10	13
Launceston	9	15	0		
WaggaWagga	10	31	0	11	14
MelbourneAirport	8	20	0	5	6
AliceSprings	23	37	0	13	10
Darwin	18	34	0	6	9
Newcastle	7	19	0		
Melbourne	13	20	0	6	6
Dartmoor	10	17	2	7	8

Table 3: Here we include in the caption a sample of LaTeX symbols that can be included in the string, and note that the caption string can be the result of R commands, using paste() in this instance. Some sample symbols include: $\alpha \longrightarrow \wp$. We also get a timestamp from R: 2016-12-19 23:40:20