Preparación de datos

DAIH:Arturo Laflor

2017-07-27

Contents

Chapter 1

Data preparation

```
library(ggplot2)
library(bitops)
library(RCurl)
#funciones online
script <- getURL("https://raw.githubusercontent.com/arturo-laflor/util-R-codes/master/QOfCategoricalF.R
eval(parse(text = script),envir=.GlobalEnv)

script <- getURL("https://raw.githubusercontent.com/arturo-laflor/util-R-codes/master/QOfContinuousF.R"
eval(parse(text = script),envir=.GlobalEnv)

script <- getURL("https://raw.githubusercontent.com/arturo-laflor/util-R-codes/master/multiplot.R", ssl
eval(parse(text = script),envir=.GlobalEnv)

knitr::opts_chunk$set(echo = TRUE,root.dir="C:/Master/Libro-Pearson-CETYS/",fig.pos = 'H')
```

Hasta aquí ya sabes que los datos son de suma importancia para tomar decisiones que pueden hacer crecer tu capacidad de ingresos en un negocio. Cómo ya se te mencionó, dichos datos normalmente se encontrarán almacenados en el disco duro de alguna computadora destinada para este fin (cuando las empresas son grandes, a las computadoras que almacenan los datos se conocen como "servidores de base de datos"). Sin embargo, antes de que los datos estén almacenados y listos para ser utilizados en el análisis que permitirá extraer información útil para el negocio, los datos normalmente deben recibir cierto tratamiento. Este proceso se denomina "preparación de datos" porque verifica que los datos con los que se realizarán los análisis para la toma de decisiones estén "libres" de imperfecciones hasta donde sea posible. De esta manera, se puede decir que los datos "se preparan" previamente para el análisis, de tal forma que los algoritmos (programas de computadora especializado en tareas que generan información relevante), trabajen con datos cuya calidad les permita un desempeño óptimo. Las principales imperfecciones de los datos son: Registros con datos faltantes, registros con datos en formato incorrecto, registros con valores atípicos y variables con datos irrelevantes. Estas imperfecciones en los datos mayormente se deben a omisiones de datos al llenar los campos de un formulario manual o electrónico, errores humanos en la captura y programas sin validaciones rigurosas de entrada que eviten su ingreso.

El proceso de preparación de los datos inicia con una exploración visual de los datos que permite identificar algunos problemas que son evidentes y que se pueden solucionar, antes de que los procesos automatizados entren en acción. Por ejemplo, en una exploración visual de un conjunto de datos, es posible identificar si existen registros donde ninguno de los campos tenga datos, así mismo, podría advertirse de alguna variable donde todos los datos se encuentren almacenados en un formato no conveniente para hacer análisis. Las acciones a tomar en estos casos particulares, pueden ser: Eliminar los registros sin datos y hacer un proceso de transformación de formato a uno que convenga para el análisis. Una vez hecho esto, se sigue con un

reporte de calidad de los datos (Kelleher et al., 2015). Este reporte permite tener una vista general de los datos que permite tomar decisiones inmediatas para iniciar el proceso de depuración de la información, hasta tener un conjunto de datos de calidad para el análisis.

Vamos a ver un ejemplo que permita clarificar lo que hasta aquí se ha escrito. Se cuenta con un conjunto de datos de una tienda departamental con la que se pretende saber los factores que influyen para que el primer cliente que llega a la tienda compre algún producto. Las variables con las que se cuenta para este análisis son: id, edad, genero, ocupación, estado.civil, hora.entrada, tiempo.at, vis.por.mes, venta.

Table 1.1: Resumen técnico de la estructura de los datos

Nombre	Tipo	Rol	Valores	Descripción
ID	continua	predictiva	CLI{3}[0-9]	Identificador del cliente
EDAD	continua	predictiva	{*3}[0-9]	Edad en años
GENERO	categórica	predictiva	Femenino	Masculino
OCUPACIÓN	categórica	predictiva	Empleado	estudiante
H.ENT	continua	predictiva	{*2}[0-9].{*2}[0-9]	Hora del día a la que entró el primer cliente a la tienda
TI.AT	continua	predictiva	{2}[0-9]	Tiempo en minutos en el que se atendió al cliente
VIS.POR.MES	categórica	predictiva	Nunca	menos de una vez por semana
VENTA	categórica	Objetivo	Si	No

La Tabla 1.1 muestra el resumen técnico de las variables. Se puede observar que se tiene ocho variables, siete de ellas son variables predictivas y una es la variable dependiente o variable objetivo. La variable objetivo es la variable de más interés en este conjunto de datos puesto que es la variable que se desea predecir a partir de las otras variables. También se puede observar que la variable objetivo es de tipo categórica, llamada en muchas ocasiones dicotómica, puesto que sólo puede tomar una de dos posibles valores (SI | NO). De las variables predictivas, cuatro son continuas y tres son categóricas, en cada una se especifica los posibles valores que la variable puede tomar. Por ejemplo, que la variable Hora.entrada puede tomar un valor con formato {*2}[0-9].{*2}[0-9], se lee {*2}=uno o dos números del cero al nueve [0-9], seguido de un punto (.), seguido de uno o dos números más {*2} del cero al nueve [0-9].

Los datos originales como fueron capturados se muestran en la tabla 1.2. Al hacer una exploración rápida a la tabla, es posible observar algunas irregularidades que pueden ser corregidas de inmediato y otras que requieren de un proceso computacional más elaborado, debido a que manualmente tomaría mucho tiempo hacerlo (pensemos en que en este ejemplo son pocos datos, sin embargo, podría tenerse una base de datos con cientos o miles de registros). La primera irregularidad que se observa es que el registro del cliente CLI011 no cuenta con datos, lo cual es indicativo de que este registro puede eliminarse del conjunto de datos sin ningún problema, de hecho, es beneficioso eliminarlo. Otro factor importante a resaltar es la captura de los datos en H.ENT y TI.AT, en el caso de H.ENT, ningún dato está en el formato adecuado ({2}[0-9].{2}[0-9]), mientras que en TI.AT algunos datos no están en el formato adecuado. Además, algunos registros cuentan con valores faltantes.

```
dcrudos<-read.csv(file="C:/Master/Libro-Pearson-CETYS/Datos-ejemplo-preproceso.csv",header = T,sep = ",
```

```
knitr::kable(
  dcrudos, caption = 'Datos de los primeros clientes entrando a la tienda.',
  booktabs = TRUE
)
```

```
#se leimina el registro número 11 que no tiene datos
dprepro_1<-dcrudos[-11,]
#se reconstruyen los indices del conjunto de datos
```

Table 1.2: Datos de los primeros clientes entrando a la tienda.

ID	EDAD	GENERO	OCUPACION	H.ENT	TI.AT	VIS.POR.MES	VENTA
CLI001	25	Femenino	Empleado	10:30	15	Una o dos veces por semana	SI
CLI002	26	Masculino	Empleado	11:00	3	Nunca	SI
CLI003	22	Masculino	Empleado	11:00	NA	Menos de una vez por semana	NO
CLI004	41	Masculino	Empleado	10:30am	15	Una o dos veces por semana	NO
CLI005	38	Masculino	Empleado	NA	10	Nunca	SI
CLI006	NA	Masculino	Empleado	12:00 AM	15	Una o dos veces por semana	NO
CLI007	22	Masculino	Empleado	11:00	5	Nunca	SI
CLI008	40	Femenino	Empleado	11:30	23 min	Tres o más veces por semana	NO
CLI009	34	Masculino	Empleado	10:00	12	Una o dos veces por semana	NO
CLI010	27	Masculino	Empleado	10:30	13	Una o dos veces por semana	NO
CLI011	NA	NA	NA	NA	NA	NA	NO
CLI012	52	Femenino	Empleado	11:00 PM	15	Menos de una vez por semana	SI
CLI013	48	Femenino	Empleado	11:50 PM	60 min	Tres o más veces por semana	NO
CLI014	51	Masculino	Empleado	12:30	5	Nunca	SI
CLI015	26	Femenino	Empleado	10:00pm	18	Una o dos veces por semana	NO
CLI016	48	Masculino	Empleado	12.3	2	Nunca	SI
CLI017	35	Femenino	Empleado	11:30	10	Nunca	SI
CLI018	41	Femenino	Empleado	9:30	15	Una o dos veces por semana	SI
CLI019	53	Femenino	Empleado	11:00	NA	Menos de una vez por semana	NO
CLI020	36	Femenino	Empleado	NA	10	Nunca	SI
CLI021	51	Femenino	Empleado	11:00	15	Una o dos veces por semana	SI

```
rownames(dprepro_1)<-1:dim(dprepro_1)[1]
```

```
####se valida y transforman los datos de tiempo y edad###

#se cargan las funciones que se utilizará

source(file="C:/Master/Libro-Pearson-CETYS/material-complementario/code/valida_edad.R",encoding = "UTF8
source(file="C:/Master/Libro-Pearson-CETYS/material-complementario/code/valida_tiempo.R",encoding = "UTF
source(file="C:/Master/Libro-Pearson-CETYS/material-complementario/code/valida_minutos.R",encoding = "UT
source(file="C:/Master/Libro-Pearson-CETYS/material-complementario/code/corrige_tiempo.R",encoding = "UT
source(file="C:/Master/Libro-Pearson-CETYS/material-complementario/code/regeneraImputedDS.R",encoding =


dprepro_1$H.ENT<-sapply(dprepro_1$H.ENT,valida_tiempo)
dprepro_1$EDAD<-sapply(dprepro_1$EDAD,valida_edad)
dprepro_1$TI.AT<-sapply(dprepro_1$TI.AT,valida_minutos)
dprepro_1$EDAD<-as.numeric(dprepro_1$EDAD)
dprepro_1$TI.AT<-as.numeric(dprepro_1$TI.AT)
```

Después de estos cambios el conjunto de datos luce de la siguiente forma

```
knitr::kable(
  dprepro_1, caption = 'Datos después de la validación y transformación',
  booktabs = TRUE
)
```

Al hacer el reporte de calidad del conjunto de datos se podrán ver estas irregularidades y otras de las que no

Table 1.3: Datos después de la validación y transformación

ID	EDAD	GENERO	OCUPACION	H.ENT	TI.AT	VIS.POR.MES	VENTA
CLI001	25	Femenino	Empleado	10.50	15	Una o dos veces por semana	SI
CLI002	26	Masculino	Empleado	11.00	3	Nunca	SI
CLI003	22	Masculino	Empleado	11.00	NA	Menos de una vez por semana	NO
CLI004	41	Masculino	Empleado	10.50	15	Una o dos veces por semana	NO
CLI005	38	Masculino	Empleado	NA	10	Nunca	SI
CLI006	NA	Masculino	Empleado	12.00	15	Una o dos veces por semana	NO
CLI007	22	Masculino	Empleado	11.00	5	Nunca	SI
CLI008	40	Femenino	Empleado	11.50	23	Tres o más veces por semana	NO
CLI009	34	Masculino	Empleado	10.00	12	Una o dos veces por semana	NO
CLI010	27	Masculino	Empleado	10.50	13	Una o dos veces por semana	NO
CLI012	52	Femenino	Empleado	11.00	15	Menos de una vez por semana	SI
CLI013	48	Femenino	Empleado	11.83	60	Tres o más veces por semana	NO
CLI014	51	Masculino	Empleado	12.50	5	Nunca	SI
CLI015	26	Femenino	Empleado	10.00	18	Una o dos veces por semana	NO
CLI016	48	Masculino	Empleado	12.05	2	Nunca	SI
CLI017	35	Femenino	Empleado	11.50	10	Nunca	SI
CLI018	41	Femenino	Empleado	9.50	15	Una o dos veces por semana	SI
CLI019	53	Femenino	Empleado	11.00	NA	Menos de una vez por semana	NO
CLI020	36	Femenino	Empleado	NA	10	Nunca	SI
CLI021	51	Femenino	Empleado	11.00	15	Una o dos veces por semana	SI

Table 1.4: Reporte de calidad de datos: Variables continuas

	Count	Miss	Card	Min	Qrt1	Median	Qrt3	Max	Mean	Sdev
EDAD	20	1	14	22.0	26.5	38	48.0	53.0	37.68	10.80
H.ENT	20	2	9	9.5	10.5	11	11.5	12.5	11.02	0.79
TI.AT	20	2	10	2.0	10.0	14	15.0	60.0	14.50	12.59

nos percatamos a simple vista. Es importante notar que en este ejemplo, se tienen pocos datos, sin embargo en un escenario real se tendrían tantos datos que las irregularidades advertidas de forma simple, podrían fácilmente pasarse por alto.

```
VARCONT<-QOfContinuousF(cbind.data.frame(EDAD=dprepro_1$EDAD,H.ENT=dprepro_1$H.ENT,TI.AT=dprepro_1$TI.A
VARCAT<-QOfCategoricalF(cbind.data.frame(ID=dprepro_1$ID,dprepro_1[,3:4],dprepro_1[,7:8]))
```

```
knitr::kable(
  VARCONT, caption = 'Reporte de calidad de datos: Variables continuas',
  booktabs = FALSE)
```

```
#write.csv(file = "C:/Master/Libro-Pearson-CETYS/material-complementario/tables/repcontvar.csv",VARCONT
```

En la Tabla 1.4, se pueden observar algunos detalles a considerar para hacer tratamiento de los datos antes de que estén listos para ser guardados en la base de conocimientos. La columna de datos faltantes muestra que existen cinco registros que no tiene valores en las variables continuas (1 en edad, 2 en hora de entrada y 2 en tiempo de atención). Se puede optar por eliminar los registros con datos faltantes o bien, utilizar una técnica de imputación[1]. También puede observarse que en el tiempo de atención existe una desviación

[1]La imputación es un proceso que permite suplir los datos faltantes con datos aproximados a los valores reales que se calculan por métodos estadísticos y/o matemáticos a partir de los existentes.

Table 1.5: Reporte de calidad de datos: Variables continuas

	Count	Miss	Card	Mode	ModeFrec	ModePerc	Mode2	Mode2Frec
ID	20	0	20	CLI001	1	5%	CLI002	1
GENERO	20	0	2	Femenino	10	50%	Masculino	10
OCUPACION	20	0	1	Empleado	20	100%	NA	NA
VIS.POR.MES	20	0	4	Una o dos veces por semana	8	40%	Nunca	7
VENTA	20	0	2	SI	11	55%	NO	9

estándar bastante amplia (15.18). Tomando en cuenta que la media es de 12.59, el valor mínimo 2 y el valor máximo 60, puede anticiparse la posibilidad de valores atípicos, es decir valores que son excepcionales y que muchas veces no conviene tomarlos en cuenta para los análisis predictivos porque ocurren con muy poca frecuencia en la realidad y pueden ocasionar que los modelos no se ajusten a los datos de forma debida. En resumen, se debe dar tratamiento a los datos faltantes y se debe buscar si existen valores atípicos para dejarlos fuera de la generación de los modelos[2].

```
knitr::kable(
  VARCAT, caption = 'Reporte de calidad de datos: Variables continuas',
  booktabs = FALSE)
```

```
#write.csv(file = "C:/Master/Libro-Pearson-CETYS/material-complementario/tables/repcatvar.csv",VARCAT)
```

En el reporte de calidad de las variables categóricas mostrado en la Tabla 1.5, no existen datos faltantes, sin embargo, la columna "cardinalidad", aporta información relevante que debe tomarse en consideración. En primer lugar, la variable ID tiene cardinalidad 20, significa que existen 20 ID distintos en 20 registros. La tienda fue visitada por 20 clientes distintos, de esta forma se evita pensar que algunos resultados de VENTA/NO VENTA se debe a un determinado cliente (BUEN COMPRADOR/MAL COMPRADOR), así que, esta columna podría eliminarse del conjunto de datos puesto que no aporta información relevante al análisis. Se tiene el caso contrario, la cardinalidad de la variable OCUPACIÓN, tiene cardinalidad uno, es decir todos los registros tienen el mismo valor en esta variable ("Empleado"). Esta columna también puede eliminarse debido a que la VENTA/NO VENTA en este caso no tiene que ver con la ocupación de las personas. Las demás variables se observan con buena calidad y serán utilizadas para el análisis.

Con la examinación de los datos en las tablas 1.4 y 1.5, se genera el plan para atender las irregularidades en la tabla 1.6

Table 1.6: Plan para atención de irregularidades en los datos

Variable	Irregularidad	Plan
EDAD, H.ENT, TI.AT	No todos los datos están en formato correcto	En el caso de la edad existen registros con la palabra "años" agregada, en H.ENT los datos están separados por (:) en lugar de (.), estos datos serán convertidos al nuevo formato (por ejemplo, 11:30 pasará a 11.5 en formato decimal). La variable TI.AT será validada y transformada según el formato especificado.
EDAD, H.ENT, TI.AT	Valores faltantes	Se utilizará un método de imputación para suplir los valores faltantes
TI.AT	Desviación estándar amplia	Se utilizará un método para identificar valores atípicos, en el caso de existir, los datos serán excluidos del análisis.

[2]Cabe aclarar que estos datos no se eliminan completamente de las bases de datos, es útil reportarlos como existentes porque proveen información valiosa para los analistas, sin embargo, si se dejan fuera de muchos procesos para evitar sesgos o como se dijo anteriormente, para evitar modelos predictivos menos eficientes.

Table 1.7: Dato atípico identificado

	ID	EDAD	GENERO	OCUPACION	H.ENT	TI.AT	VIS.POR.MES	VENTA
11	CLI013	48	Femenino	Empleado	11.83	60	Tres o más veces por semana	NO

Variable	Irregularidad	Plan
ID	Cardinalidad muy alta	La columna será excluida de los datos para el análisis, no aporta información relevante.
OCUPACION	Cardinalidad muy baja	La columna será excluida de los datos para el análisis, no aporta información relevante.

El siguiente paso en la depuración de la base de datos es la identificación de valores atípicos que serán descartados del conjunto de datos que servirá para la obtención de conjeturas o la generación modelos que permitan hacer pronósticos para beneficio de la empresa.

```
library(outliers)
library(magrittr)

#se descartan los registros con valores faltantes en la variable TIEMPO.AT para el
#analisis de los valores atípicos
nasdp<-which(is.na(dprepro_1$TI.AT))
dptemp<-dprepro_1[-nasdp,]
rownames(dptemp)<-1:nrow(dptemp)

ic<-0.997
#hist(dprepro_1$TIEMPO.AT)
outTAT<-scores(dptemp$TI.AT,type = "z",prob = ic)%>%which(.)
outTAT<-unique(outTAT)

datipico<-dptemp[outTAT,]

#elimina los registros donde hay valores atípicos
dptemp<-dptemp[-outTAT,]
#reestructura los índices del dataset
rownames(dptemp)<-1:nrow(dptemp)

#se restructuran los registros con valores faltantes en d_preproc1 descartando el registo con valor atí
dprepro_1<-rbind.data.frame(dprepro_1[nasdp,],dptemp)
rownames(dprepro_1)<-1:nrow(dprepro_1)

dordered<-dprepro_1[order(dprepro_1$ID),]
```

Existe un registro con un dato atípico en la variable TIEMPO.AT (ver Tabla 1.7), este registro se encuentra por arriba de las tres desviaciones estándar, así que será excluido de los datos que serán considerados para análisis posteriores.

```
knitr::kable(datipico,caption = 'Dato atípico identificado',booktabs = FALSE)
```

Como penúltimo paso, se inputarán los valores faltantes mediante regresión logítica multivariable proporcionada por el paquete *mice* (van Buuren and Groothuis-Oudshoorn, 2011).

```
library("mice")
```

```
##
```

```
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:RCurl':
##
##     complete
```

```
  #busca las columnas que tienen valores faltantes
  colNA<-(c(which(is.na(dordered)))%/%dim(dordered)[1])+1

  #elimina los índices repetidos
  colNA<-unique(colNA)
  #construye el dataset con las columnas que tienen datos faltantes para hacer la imputacion
  qs_dataforimp<-dordered[,colNA]
```

```
  #imputa los datos con el dataset #dos en este caso elegido arbitrariamente
  completeSH<-complete(imputedSH,2)

  dpreproc1_imputed<-regeneraImputedDS(dordered,completeSH,colNA,colnames(dprepro_1))
```

Las Tabla 1.8, muestra dos tablas con las columnas que tiene los datos faltantes y las columnas que tienen los datos imputados.

```
variablesconNa<-dprepro_1[,c(2,5,6)]
conNA<-variablesconNa[order(rownames(variablesconNa)),]
sinNA<-completeSH[order(rownames(completeSH)),]

knitr::kable(
  list(
    conNA,sinNA
  ),
  caption = 'Variables con datos faltantes y sus respectivas imputaciones',booktabs=TRUE
)
```

Como último paso, se eliminan las columnas con datos irrelevantes por su cardinalidad como se explicó anteriormente.

```
dpreproc1_imputed<-dpreproc1_imputed[,-c(1,4)]
#write.csv(file = "C:/Master/Libro-Pearson-CETYS/material-complementario/tables/datos-listos.csv",dprep
```

```
knitr::kable(
  dpreproc1_imputed,
  caption = 'Conjunto de datos preparado para análisis',booktabs=TRUE
)
```

En el conjunto de datos de la Tabla 1.9, se han excluido las variables ID y OCUPACION, se han sustituido los valores faltantes por nuevos valores haciendo uso del método de regresión logística multi-variable mediante el uso del software "mice" (van Buuren and Groothuis-Oudshoorn, 2017). Se eliminó un registro que contenía un dato atípico que se identificó fuera de tres desviaciones estándar de la media, mediante el software "outliers" (Komsta, 2011). Los datos de las variables EDAD, H.ENT y TI.AT han sido validados y transformados a valores válidos para el proceso de análisis. El listado completo de paquetes que se han ocupado para realizar este ejercicio incluye: (Xie, 2016b, Bache and Wickham (2014), Temple Lang and the CRAN team (2016), Xie (2016a) and Allaire et al. (2016))

Table 1.8: Variables con datos faltantes y sus respectivas imputaciones

	EDAD	H.ENT	TI.AT		EDAD	H.ENT	TI.AT
1	22	11.00	NA	1	22	11.00	2
10	34	10.00	12	10	34	10.00	12
11	27	10.50	13	11	27	10.50	13
12	52	11.00	15	12	52	11.00	15
13	51	12.50	5	13	51	12.50	5
14	26	10.00	18	14	26	10.00	18
15	48	12.05	2	15	48	12.05	2
16	35	11.50	10	16	35	11.50	10
17	41	9.50	15	17	41	9.50	15
18	36	NA	10	18	36	11.00	10
19	51	11.00	15	19	51	11.00	15
2	53	11.00	NA	2	53	11.00	12
3	25	10.50	15	3	25	10.50	15
4	26	11.00	3	4	26	11.00	3
5	41	10.50	15	5	41	10.50	15
6	38	NA	10	6	38	11.00	10
7	NA	12.00	15	7	40	12.00	15
8	22	11.00	5	8	22	11.00	5
9	40	11.50	23	9	40	11.50	23

Table 1.9: Conjunto de datos preparado para análisis

	EDAD	GENERO	H.ENT	TI.AT	VIS.POR.MES	VENTA
3	25	Femenino	10.50	15	Una o dos veces por semana	SI
4	26	Masculino	11.00	3	Nunca	SI
1	22	Masculino	11.00	2	Menos de una vez por semana	NO
5	41	Masculino	10.50	15	Una o dos veces por semana	NO
6	38	Masculino	11.00	10	Nunca	SI
7	40	Masculino	12.00	15	Una o dos veces por semana	NO
8	22	Masculino	11.00	5	Nunca	SI
9	40	Femenino	11.50	23	Tres o más veces por semana	NO
10	34	Masculino	10.00	12	Una o dos veces por semana	NO
11	27	Masculino	10.50	13	Una o dos veces por semana	NO
12	52	Femenino	11.00	15	Menos de una vez por semana	SI
13	51	Masculino	12.50	5	Nunca	SI
14	26	Femenino	10.00	18	Una o dos veces por semana	NO
15	48	Masculino	12.05	2	Nunca	SI
16	35	Femenino	11.50	10	Nunca	SI
17	41	Femenino	9.50	15	Una o dos veces por semana	SI
2	53	Femenino	11.00	12	Menos de una vez por semana	NO
18	36	Femenino	11.00	10	Nunca	SI
19	51	Femenino	11.00	15	Una o dos veces por semana	SI

Bibliography

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., and Hyndman, R. (2016). *rmarkdown: Dynamic Documents for R*. R package version 1.3.

Bache, S. M. and Wickham, H. (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: algorithms, worked examples, and case studies*. Number 1. The MIT Press, London.

Komsta, L. (2011). *outliers: Tests for outliers*. R package version 0.14.

Temple Lang, D. and the CRAN team (2016). *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. R package version 1.95-4.8.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

van Buuren, S. and Groothuis-Oudshoorn, K. (2017). *mice: Multivariate Imputation by Chained Equations*. R package version 2.30.

Xie, Y. (2016a). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.3.

Xie, Y. (2016b). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.15.1.