Code ▾

Donations and Projects

Data Pre-processing

Mujeer M.

Required packages

Hide

```
library("readr")
library("magrittr")
library("dplyr")
library("forecast")
```

Executive Summary

Please find below the list of actions taken as part of this assignment:

- Projects and Donations datasets are merged on Project ID column.
- Subset of first 10000 records from this dataset was taken to perform preprocessing.
- Data structure is checked and then the data type were changed as per the context of the data.
- Checked if there are any missing values and replaced them accordingly.
- Required columns were factorised and the necessary data type conversions have been done.
- Data was in a tidy format hence no actions were performed.
- New columns were mutated in the data frame.
- Missing values were removed using na.omit function.
- There were outliers in project funding balance which were dealt with using capping, after which the results were shown in a boxplot.
- Histogram of unnormalised data total project donated amount was shown before applying the transformation. We used boxcox transformation on project amount donated and showed the normalisation using histogram.

Data

A clear description of data sets, their sources, and variable descriptions should be provided. In this section, you must also provide the R codes with outputs (head of data sets) that you used to import/read/scrape the data set. You need to fulfil the minimum requirement #1 and merge at least two data sets to create the one you are going to work on. In addition to the R codes and outputs, you need to explain the steps that you have taken.

The data has been taken from metadata located at https://www.kaggle.com/donorschoose/io (https://www.kaggle.com/donorschoose/io), we have considered the datasets Donations and Projects from the metadata for performing pre-processing activities. We are using the following columns for this assignment:

- Project ID (chr)
- Donation ID(chr)
- Donation Included Optional Donation(factor)
- Donation Amount (num)
- Donation Received Date(Date)
- Project Type(factor)
- Project Title (chr)
- Project Grade Level Category (factor)
- Project Resource Category(factor)
- Project Cost (num)
- Project Posted Date (Date)
- Project Expiration Date (Date)
- Project Current Status(Ord Factor)
- Project Fully Funded Date (Date)

Hide

```
donations <- read_csv("Donations.csv") %>% as.data.frame()
projects <- read_csv("Projects.csv") %>% as.data.frame()

par(mfcol=c(1,2))
head(donations)
head(projects)
```

Selecting necessary columns for pre-processing from Projects and Donations dataset

Hide

```
projects <- within(projects,rm("School ID","Teacher ID","Project Essay","Project Short Description",
                               "Project Need Statement","Project Subject Category Tree","Project Subject Subcategory Tre
e",
                               "Teacher Project Posted Sequence"))
donations <- within (donations,rm("Donor ID","Donor Cart Sequence"))
```

Merged donations and projects using left merge and named the new data set as project_donations

Hide

```
project_donations <- merge(donations,projects,"Project ID",all.x=T)
```

Took a subset of first 10000 records from this dataset and named it as project_donations_final

Hide

```
project_donations_final <- project_donations[1:10000,]
head(project_donations_final)
```

	Project ID <chr>	Donation ID <chr> ▸
1	000009891526c0ade7180f8423792063	38d2744bf9138b0b57ed581c76c0e2da
2	000009891526c0ade7180f8423792063	dcf1071da3aa3561f91ac689d1f73dee
3	000009891526c0ade7180f8423792063	688729120858666221208529ee3fc18e
4	000009891526c0ade7180f8423792063	8cea27f0cc03f41f66aab96b284ae6a1
5	000009891526c0ade7180f8423792063	5a032791e31167a70206bfb86fb60035
6	000009891526c0ade7180f8423792063	18a234b9d1e538c431761d521ea7799d

6 rows | 1-3 of 14 columns

Understand

Summarise the types of variables and data structures, check the attributes in the data. In addition to the R codes and outputs, explain briefly the steps that you have taken. In this section, show that you have fulfilled minimum requirements 2-4.

From the output of structure below, we have idenified that columns `Donation Included Optional Donation` , `Project Type` , `Project Grade Level Category` , `Project Resource Category` , `Project Current Status` are supposed to be as factors and ’Donation Received Date` is supposed to be a date

Hide

```
str(project_donations_final)
```

```
'data.frame':   10000 obs. of  14 variables:
 $ Project ID                         : chr  "000009891526c0ade7180f8423792063" "000009891526c0ade7180f8423792063" "00000
9891526c0ade7180f8423792063" "000009891526c0ade7180f8423792063" ...
 $ Donation ID                        : chr  "38d2744bf9138b0b57ed581c76c0e2da" "dcf1071da3aa3561f91ac689d1f73dee" "68872
912085866622120 8529ee3fc18e" "8cea27f0cc03f41f66aab96b284ae6a1" ...
 $ Donation Included Optional Donation: chr  "Yes" "Yes" "No" "Yes" ...
 $ Donation Amount                    : num  25 25 178 15 25 ...
 $ Donation Received Date             : POSIXct, format: "2016-05-15 10:23:04" "2016-06-06 20:05:23" "2016-08-23 13:15:5
7" "2016-06-04 17:58:55" ...
 $ Project Type                       : chr  "Teacher-Led" "Teacher-Led" "Teacher-Led" "Teacher-Led" ...
 $ Project Title                      : chr  "OHMS Musician Chair Cart" "OHMS Musician Chair Cart" "OHMS Musician Chair C
art" "OHMS Musician Chair Cart" ...
 $ Project Grade Level Category       : chr  "Grades 6-8" "Grades 6-8" "Grades 6-8" "Grades 6-8" ...
 $ Project Resource Category          : chr  "Other" "Other" "Other" "Other" ...
 $ Project Cost                       : num  530 530 530 530 530 ...
 $ Project Posted Date                : Date, format: "2016-05-13" "2016-05-13" "2016-05-13" "2016-05-13" ...
 $ Project Expiration Date            : Date, format: "2016-09-12" "2016-09-12" "2016-09-12" "2016-09-12" ...
 $ Project Current Status             : chr  "Fully Funded" "Fully Funded" "Fully Funded" "Fully Funded" ...
 $ Project Fully Funded Date          : Date, format: "2016-08-23" "2016-08-23" "2016-08-23" "2016-08-23" ...
```

To identify the levels of all the factor columns we have used the distinct function as shown below

Hide

```
project_donations_final %>% distinct(`Donation Included Optional Donation`)
```

Donation Included Optional Donation <chr>
Yes
No

2 rows

Hide

```
project_donations_final %>% distinct(`Project Type`)
```

Project Type <chr>
Teacher-Led
NA
Professional Development
Student-Led

4 rows

Hide

```
project_donations_final %>% distinct(`Project Grade Level Category`)
```

Project Grade Level Category <chr>
Grades 6-8
Grades PreK-2
Grades 3-5
NA
Grades 9-12

5 rows

Hide

```
project_donations_final %>% distinct(`Project Resource Category`)
```

Project Resource Category <chr>
Other
Technology
Supplies
Books
Classroom Basics
NA
Art Supplies
Musical Instruments
Computers & Tablets
Instructional Technology

1-10 of 18 rows Previous **1** 2 Next

Hide

```
project_donations_final %>% distinct(`Project Current Status`)
```

Project Current Status <chr>
Fully Funded
Live
Expired
NA

4 rows

As we can see that there are missing values in the above columns, we have replaced the missing values as ‘Others’

Hide

```
project_donations_final$`Project Type`[is.na(project_donations_final$`Project Type`)] <- "Other"
project_donations_final$`Project Resource Category`[is.na(project_donations_final$`Project Resource Category`)] <- "Othe
r"
project_donations_final$`Project Current Status`[is.na(project_donations_final$`Project Current Status`)] <- "Other"
project_donations_final$`Project Grade Level Category`[is.na(project_donations_final$`Project Grade Level Category`)] <-
"Other"
```

As noted below, we have replaced the NA values as ‘Others’

Hide

```
project_donations_final %>% distinct(`Donation Included Optional Donation`)
```

Donation Included Optional Donation <chr>
Yes
No

2 rows

Hide

```
project_donations_final %>% distinct(`Project Type`)
```

Project Type <chr>
Teacher-Led
Other
Professional Development
Student-Led

4 rows

Hide

```
project_donations_final %>% distinct(`Project Grade Level Category`)
```

Project Grade Level Category <chr>
Grades 6-8
Grades PreK-2
Grades 3-5
Other
Grades 9-12

5 rows

Hide

```
project_donations_final %>% distinct(`Project Resource Category`)
```

Project Resource Category <chr>
Other
Technology
Supplies
Books
Classroom Basics
Art Supplies
Musical Instruments
Computers & Tablets
Instructional Technology
Trips

1-10 of 17 rows Previous **1** 2 Next

Hide

```
project_donations_final %>% distinct(`Project Current Status`)
```

Project Current Status <chr>
Fully Funded
Live
Expired
Other

4 rows

All the columns have been factorised

Hide

```
project_donations_final$`Donation Included Optional Donation` <- project_donations_final$`Donation Included Optional Dona
tion` %>% factor(levels = c("Yes","No"))
project_donations_final$`Project Type` <- project_donations_final$`Project Type` %>%
  factor(levels = c("Teacher-Led","Other","Professional Development","Student-Led"))
project_donations_final$`Project Grade Level Category` <- project_donations_final$`Project Grade Level Category` %>%
  factor(levels = c("Grades 6-8","Grades PreK-2","Grades 3-5","Grades 9-12","Other"))
project_donations_final$`Project Resource Category` <- project_donations_final$`Project Resource Category` %>% factor(lev
els = c("Other","Technology","Supplies","Books","Classroom Basics","Art Supplies","Musical Instruments","Computers & Tabl
ets","Instructional Technology","Trips","Lab Equipment","Flexible Seating","Educational Kits & Games","Food, Clothing & H
ygiene","Reading Nooks, Desks & Storage","Sports & Exercise Equipment","Visitors"))
project_donations_final$`Project Current Status`<- project_donations_final$`Project Current Status` %>% factor(levels = c
("Fully Funded","Live","Expired","Other"),ordered = T)
```

Converting donation received date attribute as date

Hide

```
project_donations_final$`Donation Received Date` <- as.Date(project_donations_final$`Donation Received Date`)
```

As per below output, we can see that the necessary data type conversions have been done

Hide

```
str(project_donations_final)
```

```
'data.frame':	10000 obs. of  14 variables:
 $ Project ID                      : chr  "000009891526c0ade7180f8423792063" "000009891526c0ade7180f8423792063" "00000
9891526c0ade7180f8423792063" "000009891526c0ade7180f8423792063" ...
 $ Donation ID                     : chr  "38d2744bf9138b0b57ed581c76c0e2da" "dcf1071da3aa3561f91ac689d1f73dee" "68872
9120858666221208529ee3fc18e" "8cea27f0cc03f41f66aab96b284ae6a1" ...
 $ Donation Included Optional Donation: Factor w/ 2 levels "Yes","No": 1 1 2 1 1 1 1 1 1 1 ...
 $ Donation Amount                 : num  25 25 178 15 25 ...
 $ Donation Received Date          : Date, format: "2016-05-15" "2016-06-06" "2016-08-23" "2016-06-04" ...
 $ Project Type                    : Factor w/ 4 levels "Teacher-Led",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Project Title                   : chr  "OHMS Musician Chair Cart" "OHMS Musician Chair Cart" "OHMS Musician Chair C
art" "OHMS Musician Chair Cart" ...
 $ Project Grade Level Category    : Factor w/ 5 levels "Grades 6-8","Grades PreK-2",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Project Resource Category       : Factor w/ 17 levels "Other","Technology",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ Project Cost                    : num  530 530 530 530 530 ...
 $ Project Posted Date             : Date, format: "2016-05-13" "2016-05-13" "2016-05-13" "2016-05-13" ...
 $ Project Expiration Date         : Date, format: "2016-09-12" "2016-09-12" "2016-09-12" "2016-09-12" ...
 $ Project Current Status          : Ord.factor w/ 4 levels "Fully Funded"<..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Project Fully Funded Date       : Date, format: "2016-08-23" "2016-08-23" "2016-08-23" "2016-08-23" ...
```

Tidy & Manipulate Data I

Our data is in tidy format. Hence, no action taken.

Tidy & Manipulate Data II

Created a new data frame with total donated amounts grouped by project titles and named it as donation_by_title

Hide

```
donation_by_title <- aggregate(project_donations_final$`Donation Amount`,list(project_donations_final$`Project Title`),su
m)
donation_by_title %>% as.data.frame()
```

Group.1 <chr>	x <dbl>
"16 Bars" Rapping Club	126.55
"App"ealing Learning in Kindergarten with an iPad	325.21
"Be The Change You Want To See In The World"(Gandhi).	358.27
"Calendar Math" For Kids	170.81
"Dot" & "Dash" Into STEM: Programmable Robots Make Learning Fun	268.98
"Every Child is an Artist" Pablo Picasso	75.00
"INCREASE THE PEACE" @ P. HIGH	50.00
"Involve Me and I Learn"	191.41
"iSee" Into the Future!	174.00
"Kinder-Techies" iPad Center	2132.48

1-10 of 1,807 rows Previous **1** 2 3 4 5 6 ... 100 Next

Renamed the column name of donation by title dataframe

Hide

```
colnames(donation_by_title) <- c("Project Title","Total Project donated amount")
head(donation_by_title)
```

	Project Title <chr>	Total Project donated amount <dbl>
1	"16 Bars" Rapping Club	126.55
2	"App"ealing Learning in Kindergarten with an iPad	325.21
3	"Be The Change You Want To See In The World"(Gandhi).	358.27
4	"Calendar Math" For Kids	170.81

	Project Title <chr>	Total Project donated amount <dbl>
5	"Dot" & "Dash" Into STEM: Programmable Robots Make Learning Fun	268.98
6	"Every Child is an Artist" Pablo Picasso	75.00

6 rows

Joined this data frame onto project_donation_final data frame by using left merge

Hide

```
project_donations_final <- merge(project_donations_final,donation_by_title,"Project Title",all.x=T)
```

Mutated the project donations final data frame by creating a new column called project_funding_balance

Hide

```
project_donations_final <- mutate(project_donations_final,Project_funding_balance= project_donations_final$`Project Cost`
-project_donations_final$`Total Project donated amount`)
head(project_donations_final)
```

	Project Title <chr>	Project ID <chr>	▶
1	"16 Bars" Rapping Club	006474794cf16ccc9bf06321e63eda19	
2	"16 Bars" Rapping Club	006474794cf16ccc9bf06321e63eda19	
3	"16 Bars" Rapping Club	006474794cf16ccc9bf06321e63eda19	
4	"App"ealing Learning in Kindergarten with an iPad	001ea38fca86b3d6b14300aaaf89fc51	
5	"App"ealing Learning in Kindergarten with an iPad	001ea38fca86b3d6b14300aaaf89fc51	
6	"App"ealing Learning in Kindergarten with an iPad	001ea38fca86b3d6b14300aaaf89fc51	

6 rows | 1-3 of 16 columns

Scan I

In the context of this dataset, removing the missing values was appropriate. Hence we used the na.omit function to remove the missing values

Hide

```
project_donations_final <- na.omit(project_donations_final)
```

All the missing values have been removed as per below

Hide

```
sum(is.na(project_donations_final))
```

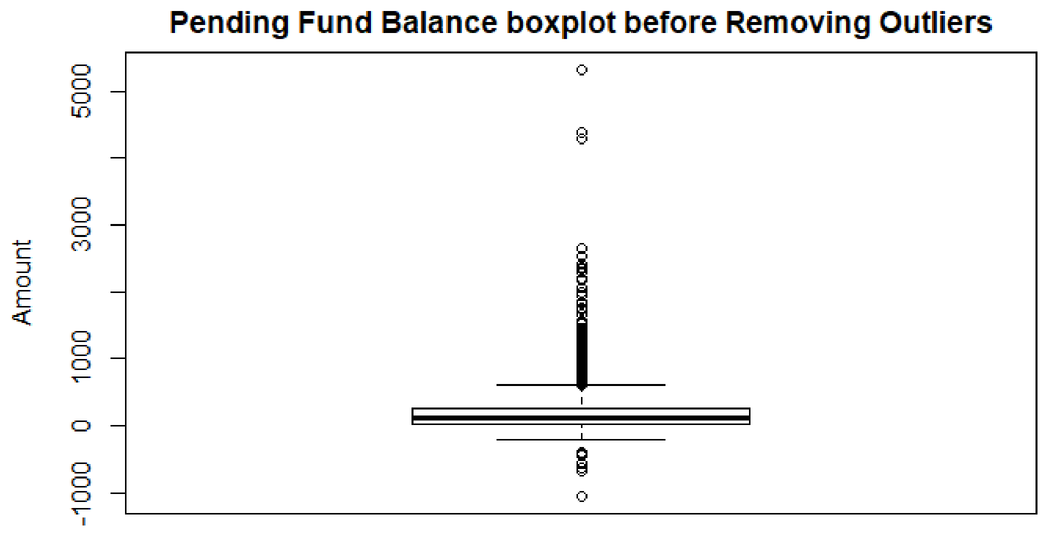
```
[1] 0
```

Scan II

As per the boxplot there are outliers in project funding balance which will be dealt with in the next step

Hide

```
boxplot(project_donations_final$Project_funding_balance,ylab="Amount",main="Pending Fund Balance boxplot before Removing
 Outliers")
```



Cap function has been created to identify and update the outliers to the nearest quartile

Hide

```
cap <- function(x){
  quantiles <- quantile( x, c(.05, 0.25, 0.75, .95 ))
  x[ x < quantiles[2] - 1.5*IQR(x) ] <- quantiles[1]
  x[ x > quantiles[3] + 1.5*IQR(x) ] <- quantiles[4]
  x}
```

Outliers for project funding attribute have been dealt with using capping

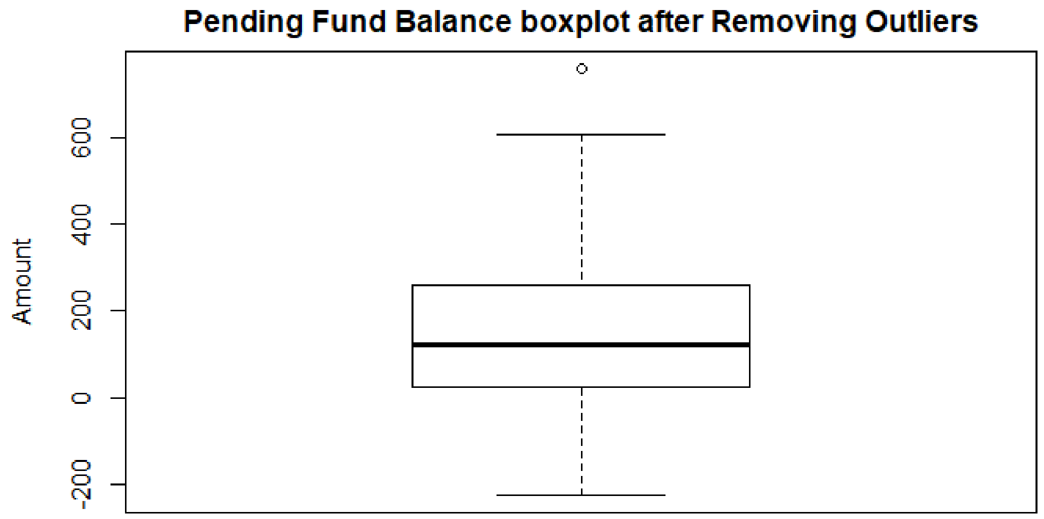
Hide

```
project_donations_final$Project_funding_balance <- project_donations_final$Project_funding_balance %>% cap()
```

Boxplot after dealing with the outliers

Hide

```
boxplot(project_donations_final$Project_funding_balance,ylab="Amount",main="Pending Fund Balance boxplot after Removing Outliers")
```

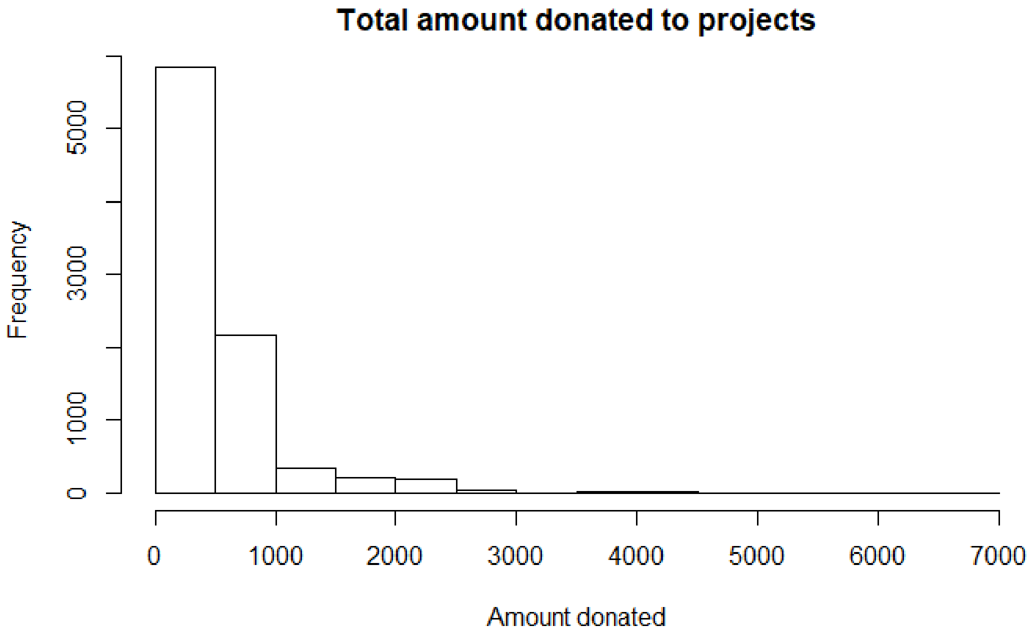


Transform

Histogram of total project donated amount before transformation (data is not normalised)

Hide

```
hist(project_donations_final$`Total Project donated amount`,main = "Total amount donated to projects",xlab = "Amount donated")
```



Applying boxcox transformation on project amount donated

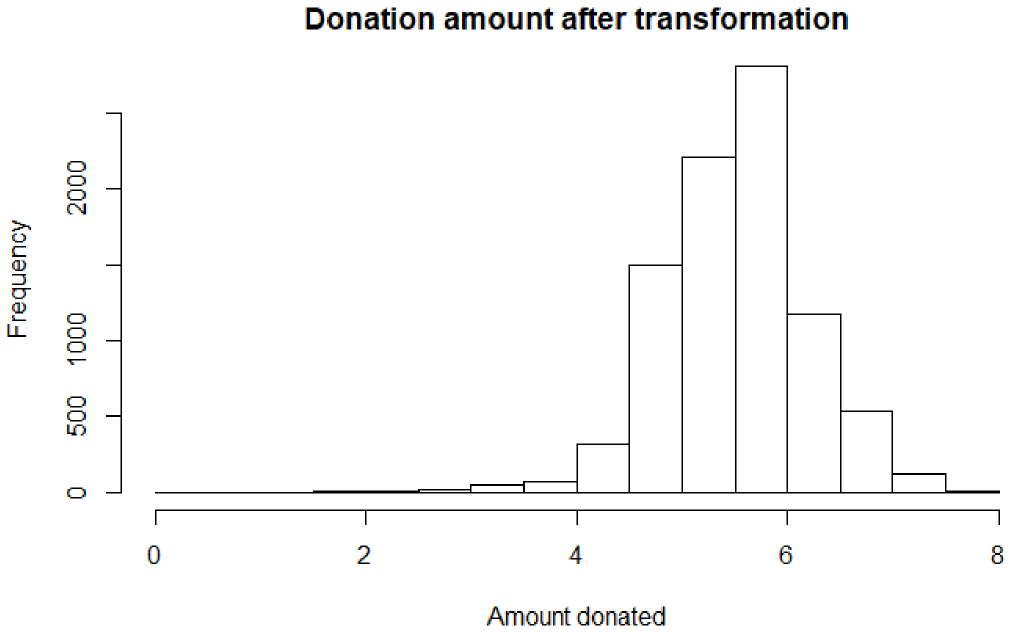
Hide

```
boxcox_tot_don_amt <- BoxCox(project_donations_final$`Total Project donated amount`,lambda = "auto")
```

Histogram after transformation (data is normalised)

Hide

```
hist(boxcox_tot_don_amt,main = "Donation amount after transformation",xlab = "Amount donated")
```


Donation amount after transformation
Frequency
Amount donated