**Exploratory Data Analysis**

**# reading file**

data<-read.csv('DataSet for EDA.csv',stringsAsFactors = FALSE,colClasses =c("appearances"="numeric"))

**#columns of data**

> names(data)

[1] "name" "id" "align" "eye" "hair"

[6] "gender" "gsm" "alive" "appearances" "first\_appear"

[11] "publisher"

The data has one continuous variable: appearances

And other categorical variables.

**# number of records in data**

> nrow(data)

[1] 23272

**# calculating null values in each column**

> apply(data,2,function(x)sum(is.na(x)))

name id align eye hair gender

0 5783 3413 13395 6538 979

gsm alive appearances first\_appear publisher

23118 6 1451 884 0

**# visualizing the data**

library(ggplot2)

library(DataExplorer)

plot\_histogram(data$appearances)

boxplot(data$appearances)

library(ggplot2)

ggplot(data,aes(x=factor(id))) + geom\_bar()

ggplot(data,aes(x=factor(align))) + geom\_bar()

ggplot(data,aes(x=factor(hair))) + geom\_bar()

ggplot(data,aes(x=factor(gender))) + geom\_bar()

ggplot(data,aes(x=factor(gsm))) + geom\_bar()

ggplot(data,aes(x=factor(alive))) + geom\_bar()

ggplot(data,aes(x=factor(publisher))) + geom\_bar()

ggplot(data,aes(x=factor(eye))) + geom\_bar()

**# Bar chart side by side**

ggplot(data, aes(x = id, fill = align)) +

geom\_bar(position = position\_dodge()) +

theme\_classic()

**# we can drop eye column and gsm column as 60% and 90% of the data respectively is null in it.**

dataf$eye<-NULL

dataf$gsm<-NULL

**# outlier treatment for continuous variable - "appearance"**

boxplot(dataf$appearances)

**# an outlier is considered so if it is below the first quartile – 1.5·IQR or above third quartile + 1.5·IQR.**

qnt <- quantile(dataf$appearances, probs=c(.25, .75), na.rm = T)

caps <- quantile(dataf$appearances, probs=c(.05, .95), na.rm = T)

H <- 1.5 \* IQR(dataf$appearances, na.rm = T)

dataf$appearances[dataf$appearances < (qnt[1] - H)] <- caps[1]

dataf$appearances[dataf$appearances > (qnt[2] + H)] <- caps[2]

**#summary of appearance column**

> summary(dataf$appearances)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

1.00 1.00 4.00 12.15 10.00 63.00 1451

**# deleting rows having null values in first\_appear colum as it only has 884 missing values ( which are less than 5% of whole data)**

**and also deleting rows having missing values in alive and gender column ( they are also having 5% of missing values of whole data)**

data1<-dataf[!(is.na(dataf$first\_appear)),]

data2<-data1[!(is.na(data1$gender)),]

data3<-data2[!(is.na(data2$alive)),]

data4<-data3

nrow(data3)

apply(data3,2,function(x)sum(is.na(x)))

**# mean, median and mode of appearance column**

mean1<-mean(data3$appearances,na.rm=TRUE)

median1<-median(data3$appearances,na.rm=TRUE)

val1 <- unique(data3$appearances[!is.na(data3$appearances)])

mode1 <- val[which.max(tabulate(match(data3$appearances, val)))]

**# imputing missing values with median in appearance column**

library(dplyr)

**# imputed with median in main data**

data3<-data3 %>%

mutate(appearances

=replace(appearances,is.na(appearances),median(appearances,na.rm=TRUE)))

plot\_histogram(data3$appearances)

**# imputed with mean**

data5<-data4 %>%

mutate(appearances

=replace(appearances,is.na(appearances),mean(appearances,na.rm=TRUE)))

plot\_histogram(data5$appearances)

**# imputed with mode**

data6<-data4 %>%

mutate(appearances

=replace(appearances,is.na(appearances),mode1))

**# Imputing categorical variables with mode**

**# mode for id variable**

val2 <- unique(data3$id[!is.na(data3$id)])

mode2 <- val2[which.max(tabulate(match(data3$id, val2)))]

**# imputing id column with mode**

data3$id[is.na(data3$id)] <- mode2

**# mode for align variable**

val3 <- unique(data3$align[!is.na(data3$align)])

mode3 <- val3[which.max(tabulate(match(data3$align, val3)))]

mode3

**# imputing align column with mode**

data3$align[is.na(data3$align)] <- mode3

**# mode for hair variable**

val4 <- unique(data3$hair[!is.na(data3$hair)])

mode4 <- val4[which.max(tabulate(match(data3$hair, val4)))]

mode4

**# imputing id column with mode**

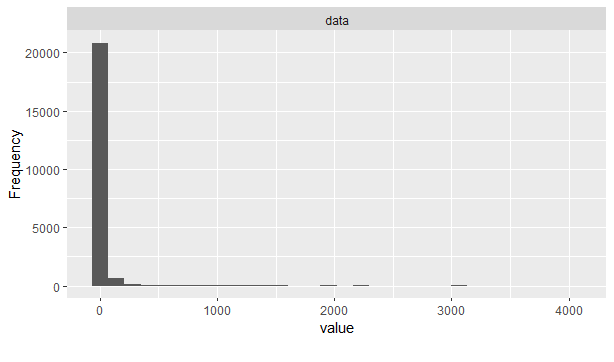
data3$hair[is.na(data3$hair)] <- mode4

|  |
| --- |
| **# now data has zero missing values in data**  > apply(data3,2,function(x)sum(is.na(x)))  name id align hair gender alive  0 0 0 0 0 0  appearances first\_appear publisher  0 0 0 |
|  |
| |  | | --- | |  | |

**# Data Visualization**

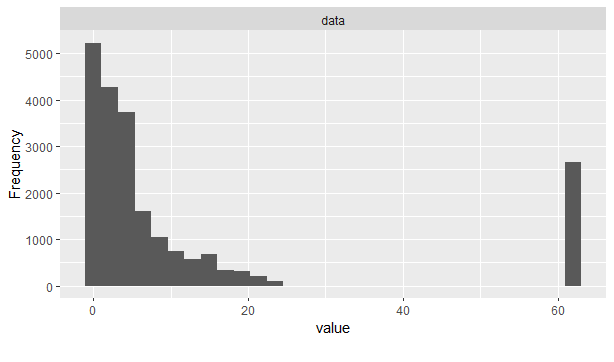
**# Histogram for appearance column**

Before outlier and missing value treatment

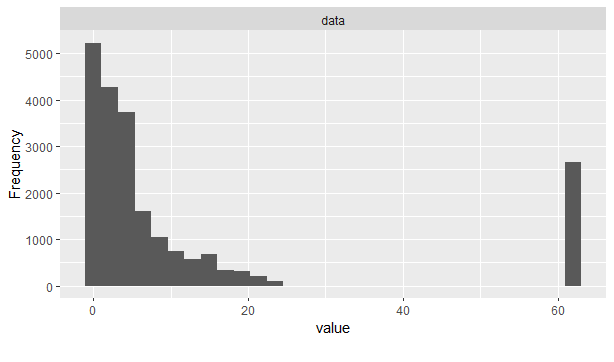


After outlier and missing value treatment

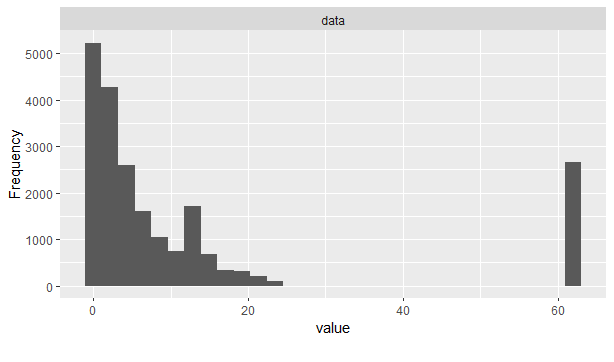
Outliers are capped at 60.



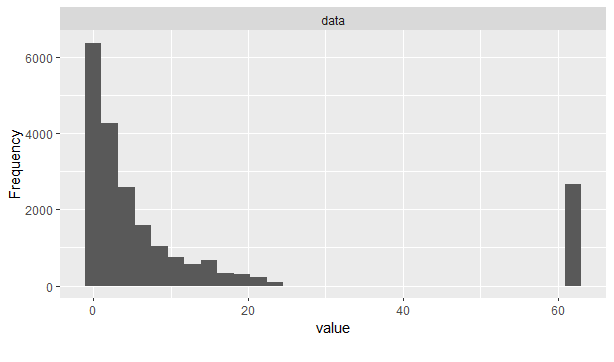
Missing values when imputed by mean for appearance column



Missing values when imputed by median for appearance column

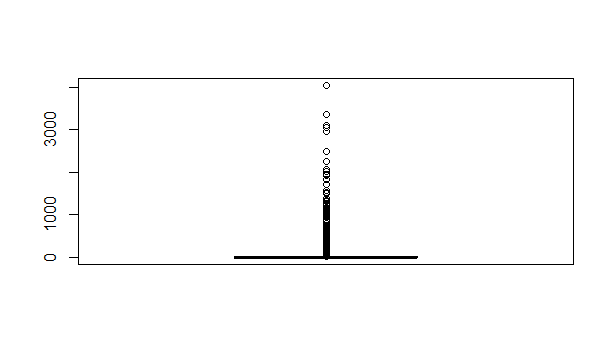


Missing values when imputed by mode for appearance column



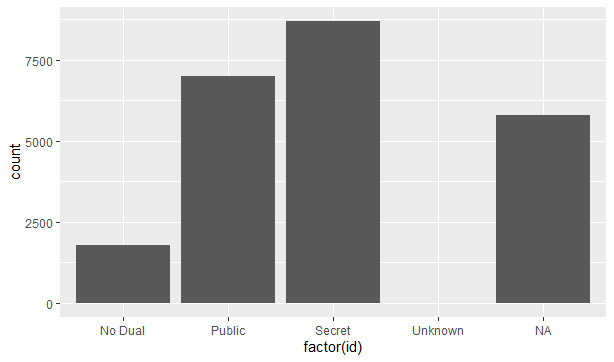
**# Box plot for appearance column**

After this the outliers are replaced ( flooring and capping)

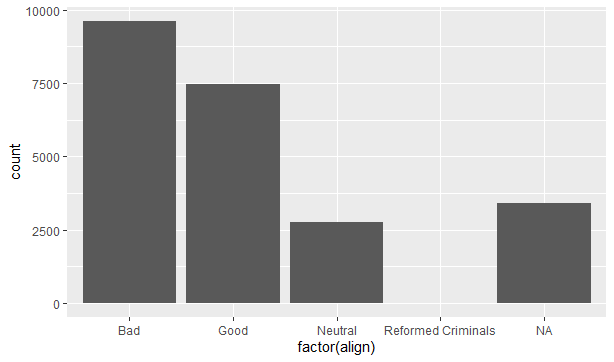


**# Bar plots for categorical variables**

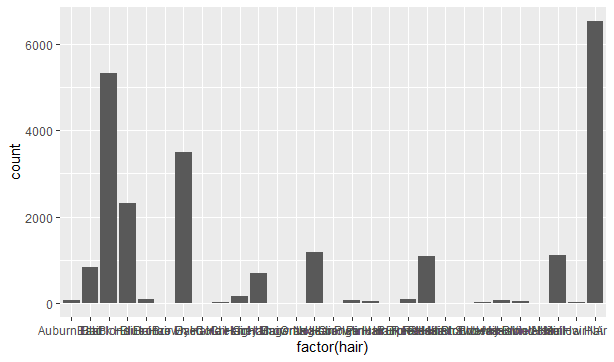
* For variable id



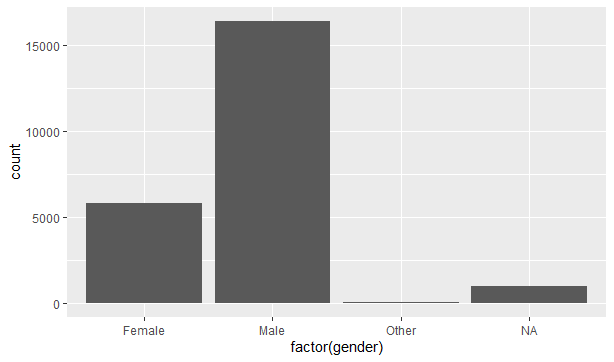
For variable align



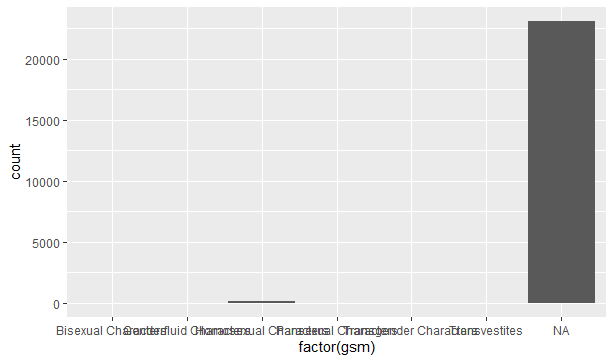
For variable hair



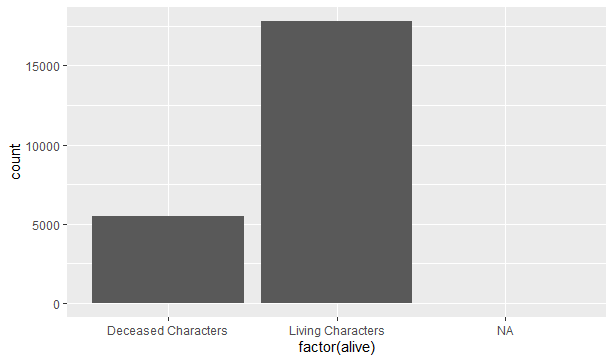
For variable gender



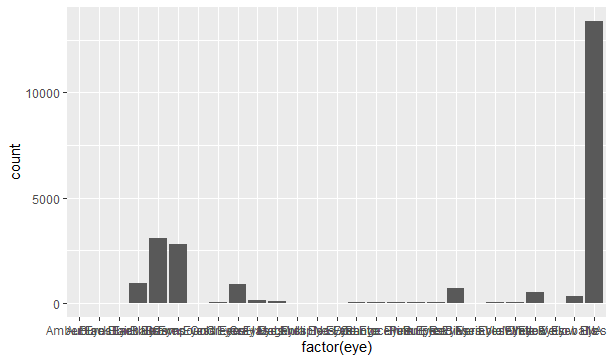
For variable gsm- it has more than 90% NA values, thus dropped this column from data



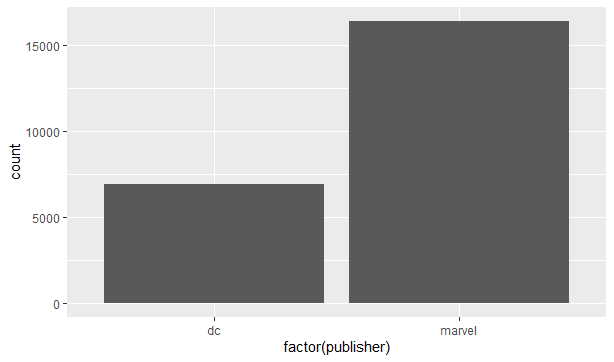
For variable alive



For variable eye- it has more null values thus dropped this variable from data



For variable publisher- did not have any null values



Distribution of id and align variables:

