Data cleansing in R

Dr Akhter Raza

Review 1

Difference between parametric and non-parametric statistics?

Review 2

Difference between descriptive and inferential statistics?

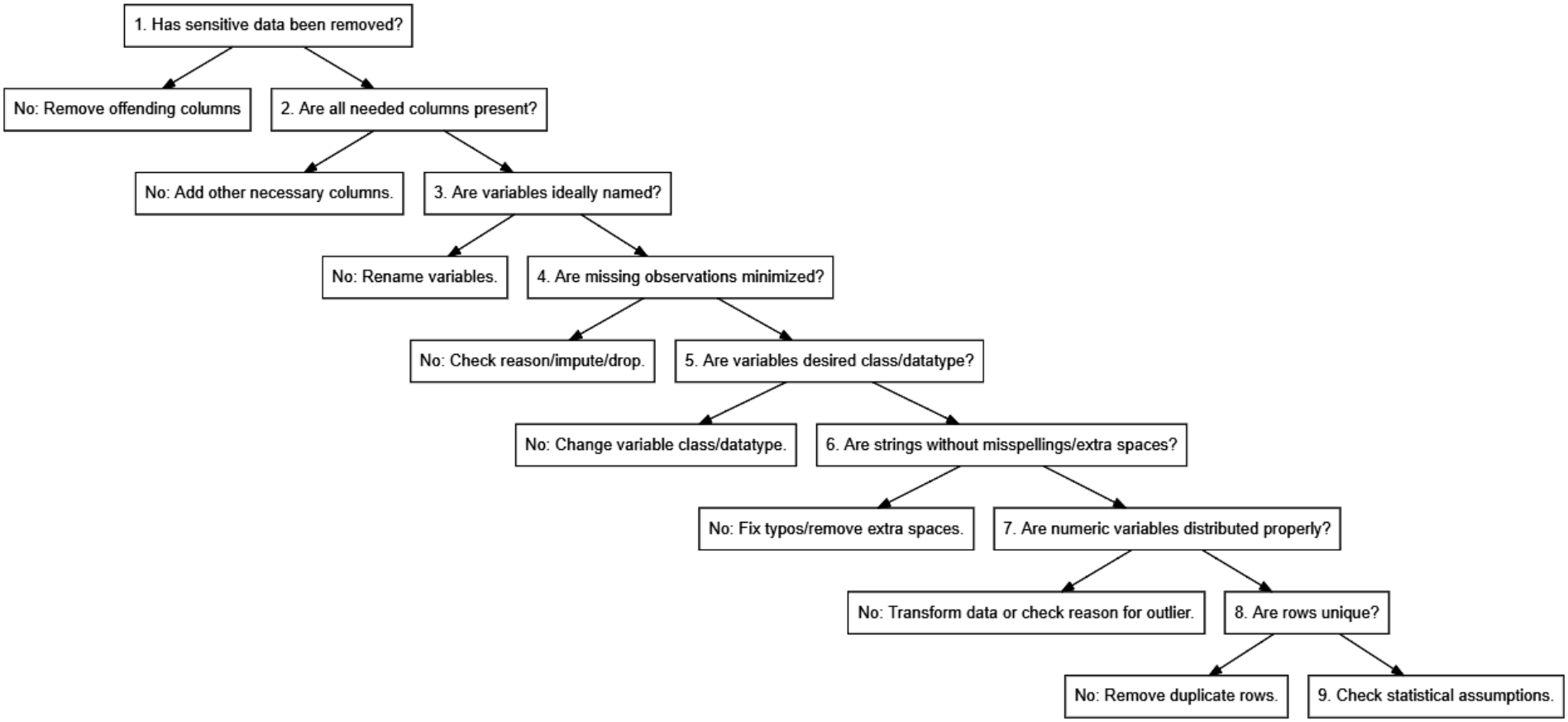
Review 3

Difference between Parameter and statistiC?

Steps in data preparation

- Check for sensitive data
- Check for missing columns
- Check variables names
- Check missing observations
- Check variable classification
- Check misspellings/extra spaces
- Check numeric data distribution
- Check duplicate rows
- Check statistical assumptions

Steps in data preparation



Missing cases

One of the big issue in data is

i) NA

ii) NaN

NA's are the missing casses

NaN are not a number

Function to be used in cleansing

head()

tail()

is.na()

any(is.na())

colSums(is.na())

na.omit()

complete.cases()

Exploring and handling NA's

The airquality data set is used for this purpose. This set is found in Base R

```
df <- airquality
str(df)
```

this data contains 153 observations of 6 variables

```
is.na(df)
```

Exploring and handling NA's

Now we are deliberately creating NA's in data.

Add new column and a row full of NA's

```
df[,7]      <-    c(NA)
df[154,]    <-    c(NA)
any(is.na(df))
is.na(df)
```

Exploring and handling NA's

Removing column number 7 because it is full of NA's

```
df   <-   df[,-7]
str(df)
```

Removing last row

```
df   <-   df[-154,]
str(df)
```

Exploring and handling NA's

```
any(is.na(df))
```

How many total NA's are there

```
sum(is.na(df))
```

Now we check each column for na's

```
sum(is.na(df$Solar.R))
```

Exploring and handling NA's

instead of checking columns 1 by 1 for NA's we can use colSums function

```
colSums(is.na(df))
```

This shows that majority of NA's are in first column which is 37 and there are 7 missing cases in column 2 rest of the columns are full and doesn't have NA's

Exploring and handling NA's

na.omit function can be used to remove all missing cases

```
df.clean <- na.omit(df)
```

Most na's are in first column which are 37 if this column does not plays any important role in data analysis then we can omit this column

Exploring and handling NA's

we will remove na's this will enhance our sample size

```
df.clean2 <- na.omit(df[,-1])
nrow(df.clean2)
```

df.clean contains 111 rows

df.clean2 contain 146 rows

Exploring and handling NA's

We can implement a rule of keeping all those columns in which NA's are less than 10

```
df.clean3 <- df[, colSums(is.na(df))<10]
```

```
nrow(df.clean3)
```

Exploring and handling NA's

mean, median and standard deviation results in NA if variable having NA

```
mean(airquality$Solar.R)
median(airquality$Solar.R)
sd(airquality$Solar.R)
```

All three results are NA's

Exploring and handling NA's

To find mean and sd of remaining values we use following

```
mean(!is.na(airquality$Solar.R))
sd(!is.na(airquality$Solar.R))
```

Imputing NA's

instead of deleting missing rows we can impute them by mean or by median

```
df.meanImputed   <- df
df.medianImputed <- df
```

Imputing NA's

All NA's are replaced by mean of the rest of data

```
df.meanImputed$Solar.R[is.na(df.meanImputed$Solar.R)] <- mean(!is.na(df.meanImputed$Solar.R))
```

Imputing NA's

All NA's are replaced by median

```
df.medianImputed$Solar.R[is.na(df.medianImputed$Solar.R)] <- median(!is.na(df.medianImputed$Solar.R))
```

Imputing NA's

now we check is there any na in solar.r of the two data frames

```
any(is.na(df.meanImputed$Solar.R))
any(is.na(df.medianImputed$Solar.R))
```

Removing outliers

```
str(df.clean2)
boxplot(df.clean2$Temp)
```

No outlier in Temp variable

```
boxplot(df.clean2$Wind)
```

There are three outliers in the Wind variable

```
summary(df.clean2$Wind)
```

Questions?