Code ▾

MATH2349 Semester 2, 2018

Assignment 2

Vikas Virani - 3715555, Vishwa Gandhi - 3714805, Jigar Mangukiya - 3715807

Setup

Install and load the necessary packages to reproduce the report here:

Hide

```
# This is a chunk where you can load the necessary packages required to reproduce the report.
 Here are some example packages, you may add others if you require
library(readr)
library(tidyr)
library(dplyr)
library(Hmisc)
library(outliers)
library(magrittr)
```

Read WHO Data

Read the WHO data using an appropriate function.

Hide

```
# This is an R chunk for reading the WHO data. Provide your R codes here:
WHO <- read_csv("WHO.csv")
```

Hide

```
WHO
```

country <chr>	is… <chr>	is… <chr>	y… <int>	new_sp_m… <int>	new_sp_m1… <int>	new_sp_m2… <int>	new_sp_m3… <int>	new_s
Afghanistan	AF	AFG	1980	NA	NA	NA	NA	
Afghanistan	AF	AFG	1981	NA	NA	NA	NA	
Afghanistan	AF	AFG	1982	NA	NA	NA	NA	
Afghanistan	AF	AFG	1983	NA	NA	NA	NA	
Afghanistan	AF	AFG	1984	NA	NA	NA	NA	
Afghanistan	AF	AFG	1985	NA	NA	NA	NA	
Afghanistan	AF	AFG	1986	NA	NA	NA	NA	
Afghanistan	AF	AFG	1987	NA	NA	NA	NA	
Afghanistan	AF	AFG	1988	NA	NA	NA	NA	
Afghanistan	AF	AFG	1989	NA	NA	NA	NA	

1-10 of 7,240 rows | 1-9 of 60 columns Previous **1** 2 3 4 5 6 ... 100 Next

Tidy Task 1:

Hide

```
# This is an R chunk for tidy task 1. Provide your R codes here:
WHO <- WHO %>% gather(new_sp_m014:new_rel_f65,key='code',value='value' )
WHO %>% head()
```

country	iso2	iso3	year	code	value
<chr>	<chr>	<chr>	<int>	<chr>	<int>
Afghanistan	AF	AFG	1980	new_sp_m014	NA
Afghanistan	AF	AFG	1981	new_sp_m014	NA
Afghanistan	AF	AFG	1982	new_sp_m014	NA
Afghanistan	AF	AFG	1983	new_sp_m014	NA
Afghanistan	AF	AFG	1984	new_sp_m014	NA
Afghanistan	AF	AFG	1985	new_sp_m014	NA

6 rows

Tidy Task 2:

Hide

```
# This is an R chunk for tidy task 2. Provide your R codes here:
WHO <- WHO %>% separate(col = code, into = c("new","var", "sex_age"),sep = "_")
WHO <-  WHO %>% separate(col = "sex_age",into = c("sex","age"),sep = 1)
WHO
```

	country	iso2	iso3	year	new	var	sex	age	value
	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<int>
1	Afghanistan	AF	AFG	1980	new	sp	m	014	NA
2	Afghanistan	AF	AFG	1981	new	sp	m	014	NA
3	Afghanistan	AF	AFG	1982	new	sp	m	014	NA
4	Afghanistan	AF	AFG	1983	new	sp	m	014	NA
5	Afghanistan	AF	AFG	1984	new	sp	m	014	NA
6	Afghanistan	AF	AFG	1985	new	sp	m	014	NA
7	Afghanistan	AF	AFG	1986	new	sp	m	014	NA
8	Afghanistan	AF	AFG	1987	new	sp	m	014	NA
9	Afghanistan	AF	AFG	1988	new	sp	m	014	NA
10	Afghanistan	AF	AFG	1989	new	sp	m	014	NA

1-10 of 405,440 rows Previous **1** 2 3 4 5 6 ... 100 Next

Tidy Task 3:

Hide

```
# This is an R chunk for tidy task 3. Provide your R codes here:
WHO <- WHO %>% spread(key = var,value = value)
WHO
```

	country	iso2	iso3	year	new	sex	age	ep	rel
	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<int>	<int>
1	Afghanistan	AF	AFG	1980	new	m	014	NA	NA
2	Afghanistan	AF	AFG	1981	new	m	014	NA	NA
3	Afghanistan	AF	AFG	1982	new	m	014	NA	NA
4	Afghanistan	AF	AFG	1983	new	m	014	NA	NA
5	Afghanistan	AF	AFG	1984	new	m	014	NA	NA
6	Afghanistan	AF	AFG	1985	new	m	014	NA	NA
7	Afghanistan	AF	AFG	1986	new	m	014	NA	NA
8	Afghanistan	AF	AFG	1987	new	m	014	NA	NA
9	Afghanistan	AF	AFG	1988	new	m	014	NA	NA
10	Afghanistan	AF	AFG	1989	new	m	014	NA	NA

1-10 of 101,360 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 ... 100 Next

Tidy Task 4:

Hide

```
# This is a chunk for Task 4. Provide your R codes here:
WHO <- WHO %>% mutate(sex= factor(sex,levels = c("m","f"),labels = c("m","f")),age= factor(age,levels = c("014","1524","2534","3544","4554","5564","65"),labels = c("<15","15-24","25-34",
"3544","4554","5564","65"),ordered = TRUE))
WHO
```

country	iso2	iso3	year	new	sex	age	ep	rel	sn
<chr>	<chr>	<chr>	<int>	<chr>	<fctr>	<ord>	<int>	<int>	<int>
Afghanistan	AF	AFG	1980	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1981	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1982	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1983	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1984	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1985	new	m	<15	NA	NA	NA

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	sex <fctr>	age <ord>	ep <int>	rel <int>	sn <int> ▶
Afghanistan	AF	AFG	1986	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1987	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1988	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1989	new	m	<15	NA	NA	NA

1-10 of 101,360 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 ... 100 Next

Task 5: Filter & Select

Hide

```
# This is a chunk for Task 5. Provide your R codes here:
WHO_subset <- select(WHO, -iso2, - new) %>% filter(country=="India" | country=="Mexico" | cou
ntry=="Australia")
WHO_subset
```

country <chr>	iso3 <chr>	year <int>	sex <fctr>	age <ord>	ep <int>	rel <int>	sn <int>	sp <int>
Australia	AUS	1980	m	<15	NA	NA	NA	NA
Australia	AUS	1981	m	<15	NA	NA	NA	NA
Australia	AUS	1982	m	<15	NA	NA	NA	NA
Australia	AUS	1983	m	<15	NA	NA	NA	NA
Australia	AUS	1984	m	<15	NA	NA	NA	NA
Australia	AUS	1985	m	<15	NA	NA	NA	NA
Australia	AUS	1986	m	<15	NA	NA	NA	NA
Australia	AUS	1987	m	<15	NA	NA	NA	NA
Australia	AUS	1988	m	<15	NA	NA	NA	NA
Australia	AUS	1989	m	<15	NA	NA	NA	NA

1-10 of 1,428 rows Previous **1** 2 3 4 5 6 ... 100 Next

Read Species and Surveys data sets

Read the Species and Surveys data sets using an appropriate function. Name these data frames as `species` and `surveys` , respectively.

Hide

```
# This is an R chunk for reading the Species and Surveys data sets. Provide your R codes her
e:
species <- read.csv("species.csv")
species
```

species_id <fctr>	genus <fctr>	species <fctr>	taxa <fctr>
AB	Amphispiza	bilineata	Bird
AH	Ammospermophilus	harrisi	Rodent
AS	Ammodramus	savannarum	Bird
BA	Baiomys	taylori	Rodent
CB	Campylorhynchus	brunneicapillus	Bird
CM	Calamospiza	melanocorys	Bird
CQ	Callipepla	squamata	Bird
CS	Crotalus	scutalatus	Reptile
CT	Cnemidophorus	tigris	Reptile
CU	Cnemidophorus	uniparens	Reptile

1-10 of 54 rows Previous **1** 2 3 4 5 6 Next

Hide

```
surveys <- read.csv("surveys.csv")
surveys
```

record_id <int>	month <int>	day <int>	year <int>	species_id <fctr>	sex <fctr>	hindfoot_length <int>	weight <int>
1	7	16	1977	NL	M	32	NA
2	7	16	1977	NL	M	33	NA
3	7	16	1977	DM	F	37	NA
4	7	16	1977	DM	M	36	NA
5	7	16	1977	DM	M	35	NA
6	7	16	1977	PF	M	14	NA
7	7	16	1977	PE	F	NA	NA
8	7	16	1977	DM	M	37	NA
9	7	16	1977	DM	F	34	NA
10	7	16	1977	PF	F	20	NA

1-10 of 35,549 rows Previous **1** 2 3 4 5 6 ... 100 Next

Task 6: Join

Hide

```
# This is a chunk for Task 6. Provide your R codes here:
surveys_combined <- surveys %>% left_join(species,by="species_id")
```

```
joining factors with different levels, coercing to character vector
```

Hide

```
head(surveys_combined)
```

	record_id	mo…	…	y…	species_id	sex	hindfoot_length	weight	genus	▸
	<int>	<int>	<int>	<int>	<chr>	<fctr>	<int>	<int>	<fctr>	
1	1	7	16	1977	NL	M	32	NA	Neotoma	
2	2	7	16	1977	NL	M	33	NA	Neotoma	
3	3	7	16	1977	DM	F	37	NA	Dipodomys	
4	4	7	16	1977	DM	M	36	NA	Dipodomys	
5	5	7	16	1977	DM	M	35	NA	Dipodomys	
6	6	7	16	1977	PF	M	14	NA	Perognathus	

6 rows | 1-10 of 11 columns

Task 7: Calculate

Hide

```
# This is a chunk for Task 7. Provide your R codes here:
surveys_combined_filtered <- surveys_combined %>% filter(species_id=='DM') %>% group_by(mont
h)
 summarise(surveys_combined_filtered,
 average_hindfoot = mean(hindfoot_length,na.rm = TRUE),
 average_weight = mean(weight,na.rm = TRUE))
```

month	average_hindfoot	average_weight
<int>	<dbl>	<dbl>
1	36.09476	42.93697
2	36.18777	43.95270
3	36.11765	45.19864
4	36.20646	44.75049
5	35.81557	43.18730
6	35.97699	41.52889
7	35.71283	41.93692
8	35.79850	41.84119
9	35.84908	43.35076
10	35.94261	42.50429

1-10 of 12 rows Previous **1** 2 Next

Task 8: Missing Values

Hide

```
# This is a chunk for Task 8. Provide your R codes here:
surveys_combined_year <- surveys_combined %>% filter(year=='1977')
surveys_combined_year %>% group_by(species_id) %>% summarise(count=sum(is.na(weight)))
```

species_id	count
<chr>	<int>
	16
DM	80
DO	0
DS	66
NL	31
OL	7
OT	15
OX	4
PE	4
PF	8

1-10 of 14 rows Previous **1** 2 Next

Hide

```
surveys_combined_temp <- surveys_combined_year %>% group_by(species_id) %>% summarise(mean =
mean(weight,na.rm = TRUE))

surveys_combined_year <- surveys_combined_year %>% left_join(surveys_combined_temp,by="specie
s_id")
surveys_combined_year$weight[is.na(surveys_combined_year$weight)]  <- surveys_combined_year$m
ean[is.na(surveys_combined_year$weight)]
surveys_weight_imputed <- surveys_combined_year %>% select(-mean)
surveys_weight_imputed
```

record_id	mo…	…	y…	species_id	sex	hindfoot_length	weight	genus	spe
<int>	<int>	<int>	<int>	<chr>	<fctr>	<int>	<dbl>	<fctr>	<fct
1	7	16	1977	NL	M	32	NaN	Neotoma	albi
2	7	16	1977	NL	M	33	NaN	Neotoma	albi
3	7	16	1977	DM	F	37	41.141304	Dipodomys	mer
4	7	16	1977	DM	M	36	41.141304	Dipodomys	mer
5	7	16	1977	DM	M	35	41.141304	Dipodomys	mer
6	7	16	1977	PF	M	14	7.173913	Perognathus	flavu
7	7	16	1977	PE	F	NA	19.500000	Peromyscus	erer

record_id	mo…	…	y…	species_id	sex	hindfoot_length	weight	genus	spe
<int>	<int>	<int>	<int>	<chr>	<fctr>	<int>	<dbl>	<fctr>	<fct
8	7	16	1977	DM	M	37	41.141304	Dipodomys	mer
9	7	16	1977	DM	F	34	41.141304	Dipodomys	mer
10	7	16	1977	PF	F	20	7.173913	Perognathus	flavu

1-10 of 503 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 ... 51 Next

Task 9: Inconsistencies or Special Values

Inspecting the variable 'Surveys_weight_imputed' for special values, it becomes clear that out of 503 total observations, 454 are finite values, whereas 49 are "NaN" values. There are no positive or negative infinite values present in the data.

Further investigating the source of the "NaN" type of special values, We found out that the special values were induced in the table when we replaced "NA" values of weight field of some of the species in variable "surveys_combined_year". After checking out the observations for species having "NaN" values in the "surveys_combined_year", it became apparant that those species did not have any values for weight attributes at all. so when we try to take mean value of such species to replace the NAs, keeping in mind the "na.rm = TRUE" attribute of the mean() function, the formula results in 0/0 = "NaN". Hence the NaN values are induced in the dataset while performing task 8.

Hide

```
# This is a chunk for Task 9. Provide your R codes here:
cat("Finite Observations : ",sum(is.finite(surveys_weight_imputed$weight)),"\n")
```

```
Finite Observations :  454
```

Hide

```
cat("Infinite Observations : ",sum(is.infinite(surveys_weight_imputed$weight)),"\n")
```

```
Infinite Observations :  0
```

Hide

```
cat("NaN Observations : ",sum(is.nan(surveys_weight_imputed$weight)),"\n\n")
```

```
NaN Observations :  49
```

Hide

```
cat("Total Observations : ",count(surveys_weight_imputed)[[1]])
```

```
Total Observations :  503
```

Hide

Task 10: Outliers

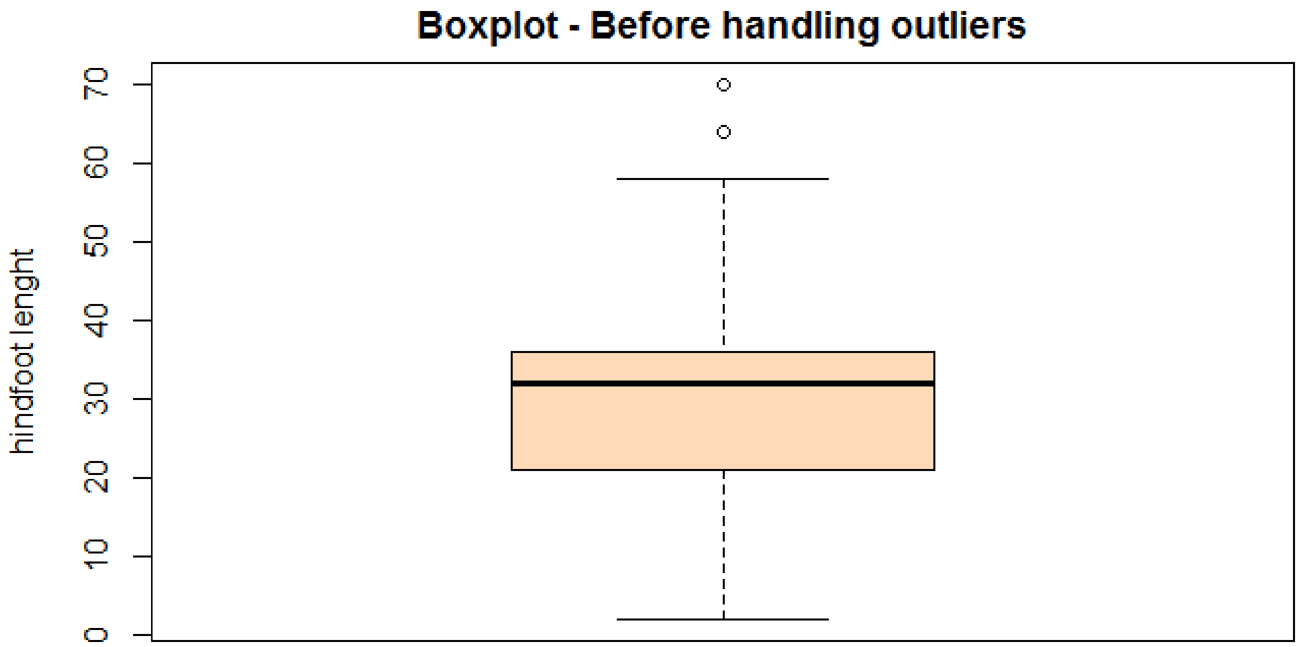
Using the `surveys_combined` data frame, inspect the variable hindfoot length for possible univariate outliers. If you detect any outliers use any of the methods outlined in the Module 6 notes to deal with them. Explain briefly the actions that you take to handle outliers.

To detect the outliers, first we replaced NA values present in data with the mean values. It had to be done so that the quantile function can deal calculate the quantiles properly without being affected by the NAs.

After replacing the NAs with the mean, we replaced outliers with the median values. Plotting the boxplot before and after the replacement operations makes it clear that there are no more outliers present in the data anymore.

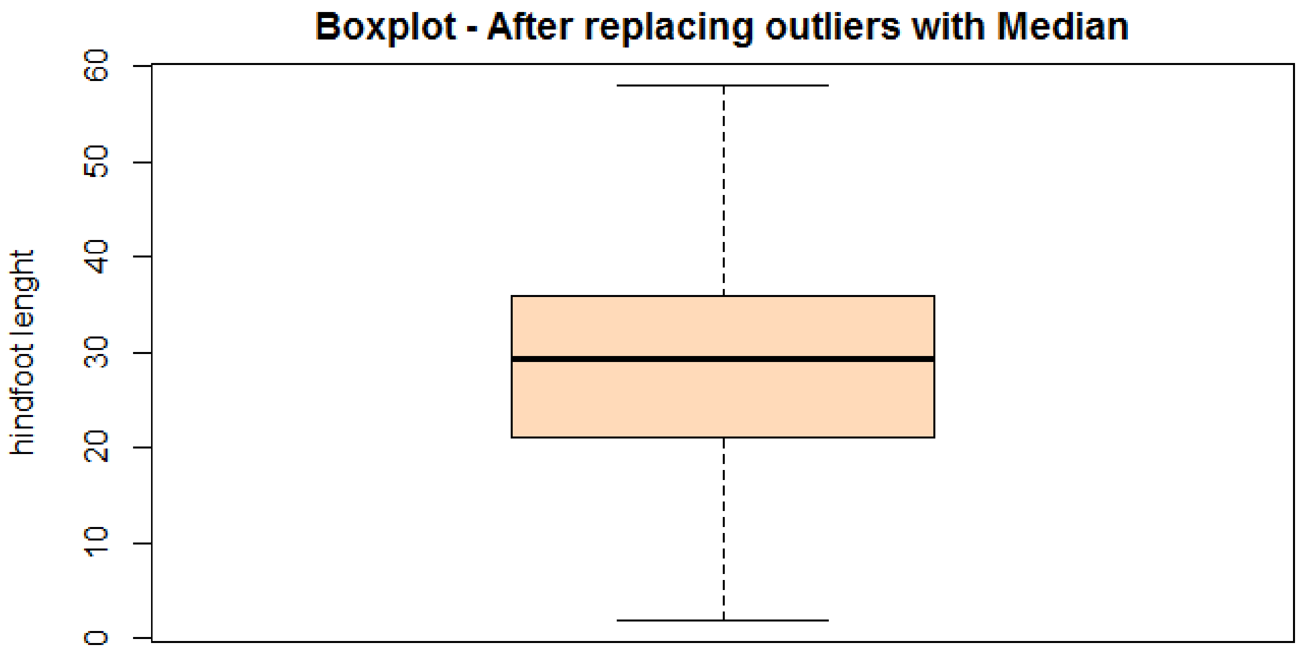
Hide

```
# This is a chunk for Task 10. Provide your R codes here:
surveys_combined$hindfoot_length %>% boxplot(main="Boxplot - Before handling outliers",col =
"peachpuff", border = TRUE, ylab="hindfoot lenght")
```



Hide

```
replace_median <- function(x){
    quantiles <- quantile( x,c(0.25, 0.5, 0.75))
    x[ x < quantiles[1] - 1.5*IQR(x) ] <- quantiles[2]
    x[ x > quantiles[3] + 1.5*IQR(x) ] <- quantiles[2]
    x
}
# Recode missing data with Mean
surveys_combined$hindfoot_length[is.na(surveys_combined$hindfoot_length)] <- mean(surveys_combined$hindfoot_length,na.rm = TRUE)
# Replace Outliers with Median values
surveys_combined$hindfoot_length %>% replace_median() %>% boxplot(main="Boxplot - After replacing outliers with Median",col = "peachpuff", border = TRUE, ylab="hindfoot lenght")
```


Boxplot - After replacing outliers with Median
hindfoot lenght