MATH2349 Semester 2, 2018

Code

Assignment 2

Phalgun Haribabu Chintal, s3702107 and Syed Junaid Ahmed, s3731300

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)
```

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")
```

```
Parsed with column specification:
cols(
  .default = col_integer(),
  country = col_character(),
  iso2 = col_character(),
  iso3 = col_character()
)
See spec(...) for full column specifications.
```

Tidy Task 1:

Hide

```
format_1 <- WHO %>% gather(new_sp_m014:new_rel_f65, key = "Code", value = "value")
format_1
```

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	Code <chr>	value <int>
Afghanistan	AF	AFG	1980	new_sp_m014	*NA*
Afghanistan	AF	AFG	1981	new_sp_m014	*NA*

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	Code <chr>	value <int>
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*
Afghanistan	AF	AFG	1986	new_sp_m014	*NA*
Afghanistan	AF	AFG	1987	new_sp_m014	*NA*
Afghanistan	AF	AFG	1988	new_sp_m014	*NA*
Afghanistan	AF	AFG	1989	new_sp_m014	*NA*

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

Tidy Task 2:

Hide

```
# This is an R chunk for tidy task 2. Provide your R codes here:
format_2 <- format_1 %>% separate(Code, into = c("new", "var", "sex_age"), sep = "_")
format_3 <- format_2 %>% separate(sex_age, into = c("sex", "age"), sep = 1)
format_3
```

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	var <chr>	sex <chr>	age <chr>	value <int>
Afghanistan	AF	AFG	1980	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1981	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1982	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1983	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1984	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1985	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1986	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1987	new	sp	m	014	*NA*
Afghanistan	AF	AFG	1988	new	sp	m	014	*NA*
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

```
format_4 <- format_3 %>% spread(key = var, value = value)
format_4
```

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	sex <chr>	age <chr>	ep <int>	rel <int>	sn <int>
Afghanistan	AF	AFG	1980	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1981	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1982	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1983	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1984	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1985	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1986	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1987	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1988	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1989	new	m	014	NA	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:
format_5 <- format_4 %>% mutate(sex = factor(sex,
                                  levels = c("m","f"),
                                  labels = c("Male","Female")),
                        age = factor(age,
                                  levels = c("014", "1524", "2534", "3544", "4554", "5564"
, "65"),
                                  labels = c("<15", "15-24", "25-34", "35-44", "45-54", "5
5-64", "65>="),
                                  ordered = TRUE))
format_5
```

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	sex <fctr>	age <ord>	ep <int>	rel <int>	sn <int>
Afghanistan	AF	AFG	1980	new	Male	<15	NA	NA	NA
Afghanistan	AF	AFG	1981	new	Male	<15	NA	NA	NA
Afghanistan	AF	AFG	1982	new	Male	<15	NA	NA	NA
Afghanistan	AF	AFG	1983	new	Male	<15	NA	NA	NA
Afghanistan	AF	AFG	1984	new	Male	<15	NA	NA	NA
Afghanistan	AF	AFG	1985	new	Male	<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	Male	<15	*NA*	*NA*	*NA*
Afghanistan	AF	AFG	1987	new	Male	<15	*NA*	*NA*	*NA*
Afghanistan	AF	AFG	1988	new	Male	<15	*NA*	*NA*	*NA*
Afghanistan	AF	AFG	1989	new	Male	<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 <- format_5 %>% select(-c(iso2,new)) %>% filter(country == "Czech Republic" | coun
try == "Switzerland" | country == "United Arab Emirates")
WHO_subset
```

country <chr>	iso3 <chr>	year <int>	sex <fctr>	age <ord>	ep <int>	rel <int>	sn <int>	sp <int>
Czech Republic	CZE	1980	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1981	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1982	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1983	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1984	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1985	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1986	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1987	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1988	Male	<15	*NA*	*NA*	*NA*	*NA*
Czech Republic	CZE	1989	Male	<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

Hide

```
species <- read_csv("species.csv")
```

```
Parsed with column specification:
cols(
  species_id = col_character(),
  genus = col_character(),
  species = col_character(),
  taxa = col_character()
)
```

Hide

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

```
Parsed with column specification:
cols(
  record_id = col_integer(),
  month = col_integer(),
  day = col_integer(),
  year = col_integer(),
  species_id = col_character(),
  sex = col_character(),
  hindfoot_length = col_integer(),
  weight = col_integer()
)
```

Task 6: Join

Hide

```
surveys_combined <- surveys %>% full_join(species, by = "species_id")
surveys_combined
```

record_id <int>	mo… <int>	… <int>	y… <int>	species_id <chr>	… <chr>	hindfoot_length <int>	wei… <int>	genus <chr>	species <chr>
1	7	16	1977	NL	M	32	*NA*	Neotoma	albigula
2	7	16	1977	NL	M	33	*NA*	Neotoma	albigula
3	7	16	1977	DM	F	37	*NA*	Dipodomys	merriami
4	7	16	1977	DM	M	36	*NA*	Dipodomys	merriami
5	7	16	1977	DM	M	35	*NA*	Dipodomys	merriami
6	7	16	1977	PF	M	14	*NA*	Perognathus	flavus
7	7	16	1977	PE	F	*NA*	*NA*	Peromyscus	eremicus
8	7	16	1977	DM	M	37	*NA*	Dipodomys	merriami
9	7	16	1977	DM	F	34	*NA*	Dipodomys	merriami
10	7	16	1977	PF	F	20	*NA*	Perognathus	flavus

1-10 of 35,555 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Task 7: Calculate

Hide

```
# This is a chunk for Task 7. Provide your R codes here:
avg <- surveys_combined %>% filter(species_id == "NL") %>% group_by(month) %>% summarise(mean
_weight = mean(weight, na.rm = TRUE), mean_hindfoot = mean(hindfoot_length, na.rm = TRUE))
avg
```

month <int>	mean_weight <dbl>	mean_hindfoot <dbl>
1	179.3443	32.54098
2	181.3818	32.82353
3	177.4516	32.75862
4	153.0690	32.02439
5	142.7536	31.60000
6	143.7879	32.18889
7	141.7415	32.35398
8	152.5100	32.07143
9	164.9920	32.50427
10	169.1364	32.43119

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

Task 8: Missing Values

Hide

```
surveys_combined_year <- surveys_combined %>% filter(year == '1995')
surveys_combined_year %>% group_by(species) %>% summarise(val = sum(is.na(weight)))
```

species <chr>	val <int>
albigula	2
baileyi	0
bilineata	4
brunneicapillus	2
chlorurus	1
eremicus	1
flavus	3
fuscus	1

species <chr>	val <int>
harrisi	36
hispidus	0

1-10 of 23 rows Previous **1** 2 3 Next

Hide

```
surveys_weight_imputed <- surveys_combined_year %>% group_by(species) %>% mutate(weight = imp
ute(weight, fun = mean))
surveys_weight_imputed
```

record_id <int>	mo… <int>	… <int>	y… <int>	species_id <chr>	… <chr>	hindfoot_length <S3: impute>	weight <dbl>	genus <chr>	s <c
21993	1	11	1995	PF	F	16.00000	7.000000	Perognathus	fl
21994	1	11	1995	DO	M	36.00000	47.000000	Dipodomys	o
21995	1	11	1995	DO	M	36.00000	51.000000	Dipodomys	o
21996	1	11	1995	PF	F	14.00000	7.000000	Perognathus	fl
21997	1	11	1995	RM	M	15.00000	10.000000	Reithrodontomys	n
21998	1	11	1995	DM	M	38.00000	46.000000	Dipodomys	n
21999	1	11	1995	PF	F	15.00000	8.000000	Perognathus	fl
22000	1	11	1995	DM	F	37.00000	45.000000	Dipodomys	n
22001	1	11	1995	DO	M	36.00000	41.000000	Dipodomys	o
22002	1	11	1995	PF	F	16.00000	8.000000	Perognathus	fl

1-10 of 1,222 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Task 9: Inconsistencies or Special Values

Hide

```
surveys_weight_imputed$weight %>% is.nan() %>% sum()
```

```
[1] 98
```

Hide

```
surveys_weight_imputed$weight %>% is.infinite() %>% sum()
```

```
[1] 0
```

Hide

```
which(is.nan(surveys_weight_imputed$weight))
```

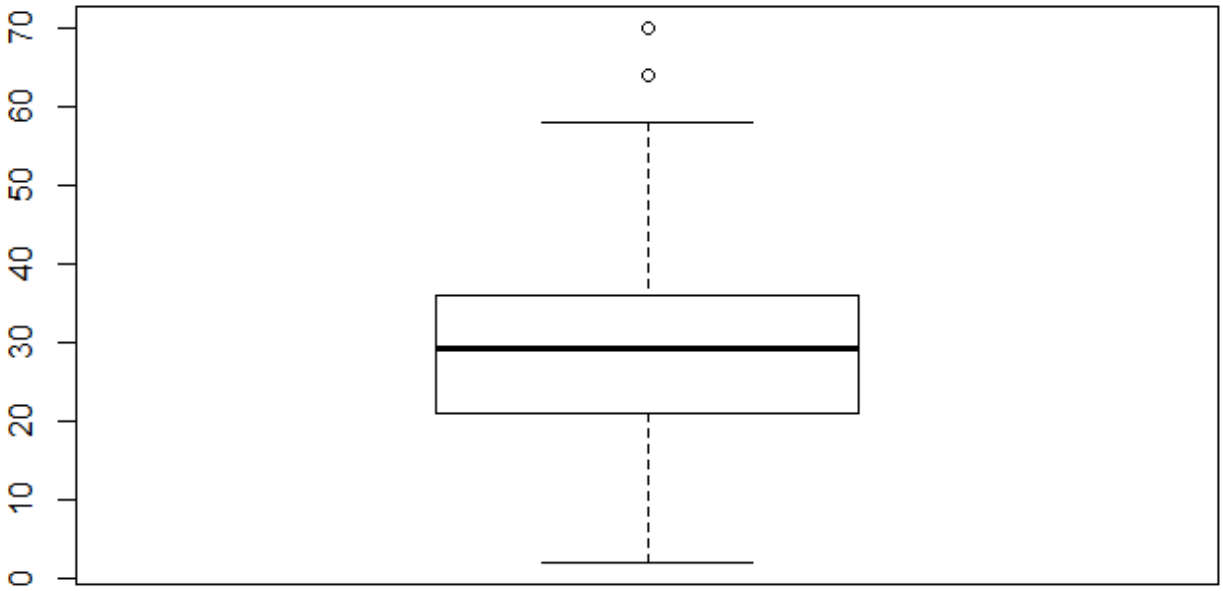
```
 [1]   20   21   22   32   38   39   40   48   49   53   62   67   72   83   85   88   91  10
3  107  111
[21]  112  119  136  137  138  139  140  156  198  199  200  201  230  235  241  244  255  25
9  262  266
[41]  272  276  283  301  302  303  324  325  326  327  338  350  353  356  359  370  399  40
0  407  428
[61]  431  441  485  490  515  516  546  592  597  608  617  652  655  678  706  712  718  74
4  779  820
[81]  838  893  901  948  998  999 1101 1103 1119 1121 1122 1130 1173 1180 1206 1220 1221 122
2
```

In the above task, we have to determine the number of inconsistencies or special values of the weight column in the survey_weight_imputed dataset. It is calculated from is.nan() and is. infinite() functions along with sum(). As per the results, there are 98 NaN and 0 infinite. The location of NaN (not a number) is identified by using which(). The reason for getting NaN on weight column is because of the invalid number when they were replaced by the mean in task 8. So, it resulted in NaN value.

Task 10: Outliers

Hide

```
surveys_combined$hindfoot_length <- with(surveys_combined, impute(hindfoot_length, mean))
boxplot(as.numeric(surveys_combined$hindfoot_length))
```



Hide

```
score <- surveys_combined$hindfoot_length %>% scores(type = "z")
length(which(abs(score)>3))
```

```
[1] 7
```

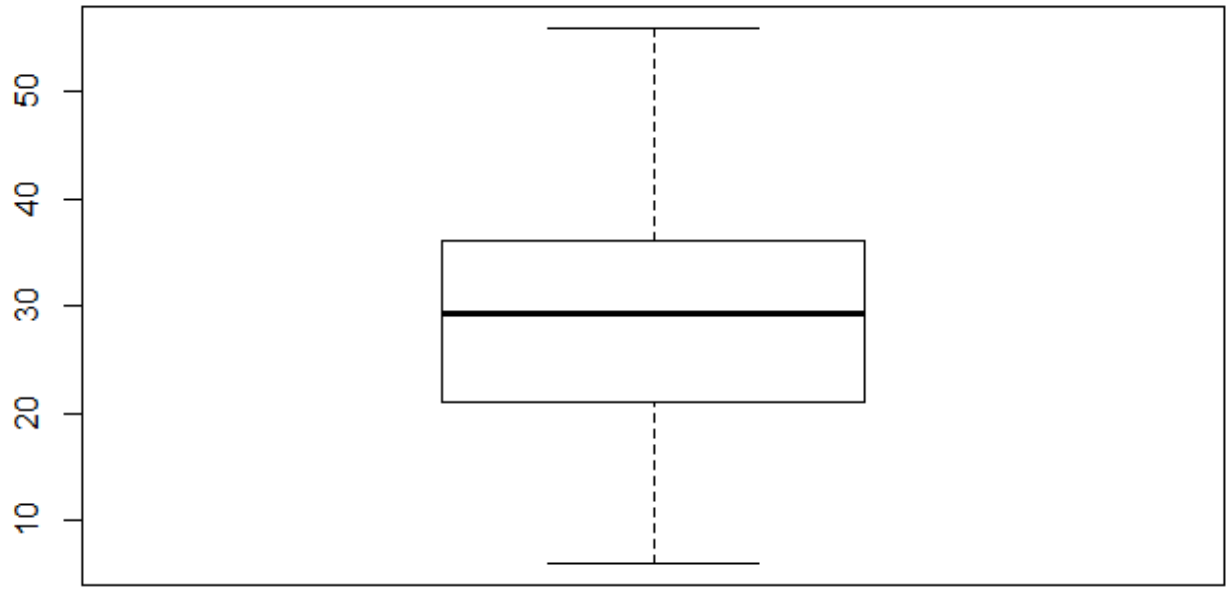
Hide

```
which(abs(score)>3)
```

```
[1]  1694  3784  4449 10574 22049 30425 31400
```

Hide

```
hindfoot_length <- surveys_combined$hindfoot_length[-which(abs(score)>3)]
boxplot(as.numeric(hindfoot_length))
```



Firstly, outliers are found by using boxplot. Then, the z.scores are calculated. The number of outliers are found from length function. Finally, the outliers are dealt.