Week 06 Reproducible data analysis

Open and reproducible science: general reasons and approaches

Homework solutions

Part 1 - Preparation

Step 1

Source: Palmer Station Antarctica LTER and K. Gorman. 2020. URLs: https://doi.org/10.6073/pasta/98b16d7d563f265cb52372c8ca99e60f , https://doi.org/10.6073/pasta/7fca67fb28d56ee2ffa3d9370ebda689 , https://doi.org/10.6073/pasta/c14dfcfada8ea13a17536e73eb6fbe9e

Step 2

```
numericcols <- sapply(colnames(penguins), function(x) is.numeric(penguins[[x]]))
df <- data.frame(Column_Name = names(numericcols)[numericcols],
           Mean = signif(apply((na.omit(penguins[numericcols])),2,mean),4),
           Variance = signif(apply((na.omit(penguins[numericcols])),2,var),4),
           row.names = NULL
)
knitr::kable(df)
```

Column_Name	Mean	Variance
bill_length_mm	43.92	2.981e+01
bill_depth_mm	17.15	3.900e+00
flipper_length_mm	200.90	1.977e+02
body_mass_g	4202.00	6.431e+05
year	2008.00	6.678e-01

Part 2 - Data analysis

Exercise 2.1

```
result_exercise_1 <- dim(penguins)[1]
```

Number of rows: 344

Exercise 2.2

```
result_exercise_2 <- min(penguins$year)
```

First year of records: 2007

Exercise 2.3

```
result_exercise_3 <- sum(penguins$species == "Adelie")
```

Number of Adelie penguins: 152

Exercise 2.4

```
result_exercise_4 <- sum(is.na(penguins))
```

Total number of missing values (NA's): 19

Exercise 2.5

```
result_exercise_5 <- sum(apply(penguins,1, function(x) !any(is.na(x))))
```

Number of complete rows (rows with no missing values, i.e. NA's): 333

Exercise 2.6

```
result_exercise_6 <- unique(penguins$island[penguins$species == "Gentoo"]);
```

Name of islands where the Gentoo penguins were found: Biscoe

Exercise 2.7

```
result_exercise_7 <- sum(penguins$species == "Adelie" & penguins$island == "Dream")/
  sum(penguins$island == "Dream")
```

Proportion of Adelie penguins on Dream island: 0.4516129

Exercise 2.8

```
result_exercise_8 <- quantile(na.omit(penguins$bill_length_mm), 0.93)
```

93 % quantile of `bill_length_mm` : 51.3

Exercise 2.9

```
result_exercise_9 <- abs(coef(lm(bill_depth_mm ~ sex, penguins))[2])
```

Absolute mean difference of `bill_depth_mm` between female and male: 1.4656169

Exercise 2.10

```
result_exercise_10 <- confint(lm(bill_depth_mm ~ sex, penguins),"sexmale" )
```

95% confidence interval of slope of linear regression between `bill_depth_mm` and `sex`: 1.0710254, 1.8602083

Exercise 2.11

```
chins <- na.omit(penguins$species[penguins$flipper_length_mm < 205 & penguins$bill_length_mm > 45])
result_exercise_11 <- sum(chins == "Chinstrap")/length(chins)
```

Proportion of Chinstrap penguins with `flipper_length_mm` smaller than 205 and `bill_length_mm` larger than 45 compared to all penguins with `flipper_length_mm` smaller than 205 and `bill_length_mm` larger than 45: 0.9310345

Exercise 2.12

```
result_exercise_12 <- sum(chins == "Chinstrap")/sum(penguins$species == "Chinstrap")
```

Proportion of Chinstrap penguins with `flipper_length_mm` smaller than 205 and `bill_length_mm` larger than 45 compared to all Chinstrap penguins: 0.7941176