Lab 2: Importing, exploring, managing data using functions

Solutions

DATE

Answer the questions in this lab and submit the compiled WORD or PDF by the deadline.

Importing Data

1. When reading in the `Police Shootings` Excel data set, what do the arguments `sheet=1` and `col_names=TRUE` mean?

   `sheet=1` Takes the data from the first sheet. `col_names` use the first row as column names.

Use the `NCbirths` data set to answer the next set of questions. Read in the data set in the code chunk below.

```
nc <- read.csv("data/NCbirths.csv", header=TRUE)
```

2. How many observations and variables are contained in this data set?

```
dim(nc)
```

```
## [1] 1000   13
```

   There are 1000 observations and 13 variables

- they don't have to have used `dim`. they could have looked in the environment.

3. Calculate the mean age of the mothers (`mage`) in the sample.

```
mean(nc$mage)
```

```
## [1] 27
```

4. Pregnancies last on average 38 weeks. Edit the `weeks` variable to change all records where `weeks` is greater than 38, to equal 38. That is, for all record where `weeks>38`, change the value of `weeks` to `<- 38`.

```
nc$weeks[nc$weeks>38] <- 38
max(nc$weeks, na.rm=TRUE)
```

```
## [1] 38
```

They need to have confirmed that they did the recoding correctly.

5. Use the `summary` function to calculate summary statistics on the fathers age (`fage`). Round to 3 digits using the `digits=` argument. Don't forget that you can look at the bottom of the help for summary (`?summary`) file for examples on how to use this function.

```
summary(nc$fage, digits=3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    14.0    25.0    30.0    30.3    35.0    55.0     171
```

6. What is the distribution of smoking habit (`habit`) amongst the mothers in the sample? i.e. How many are smokers and how many are non-smokers? *Hint: Use the* `table()` function.

```
table(nc$habit)
```

```
##
## nonsmoker    smoker
##       873       126
```

There are 873 non-smokers and 126 smokers in the sample.

7. Use the `ifelse()` function to create a new variable called `missing_fage` to identify if the fathers age is missing. The **logical statement** to identify if something is missing looks like this: `is.na(variable)`.
   - Set this new variable equal to `'MISSING'` if `fage` is missing (the logical statement is TRUE)
   - set this new variable equal to `'OBSERVED'` if `fage` is not missing (the logical statement is FALSE)

```
nc$missing_fage <- ifelse(is.na(nc$fage), 'MISSING', 'OBSERVED')
table(nc$missing_fage)
```

```
##
##  MISSING OBSERVED
##      171      829
```

- They have to make this variable on the nc data set.
- They also have to do something to show that it was created.

8. What class of data is this new variable?

```
class(nc$missing_fage)
```

```
## [1] "character"
```

This new variable is a character variable.

9. What percent of records are missing the fathers age?

```
mean(is.na(nc$fage))*100
```

```
## [1] 17.1
```

17.1% of the data on the fathers age is missing.

- There are many ways they can get this answer.