Template

Studentnames and studentnumbers here

2025-04-25

Set-up your environment

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Part 1 - Identify a Social Problem

Use APA referencing throughout your document.

1.1 Describe the Social Problem

Include the following:

- Why is this relevant?
- ...

1.2 Data Sourcing

Load in the data

Preferably from a URL, but if not, make sure to download the data and store it in a shared location that you can load the data in from. Do not store the data in a folder you include in the Github repository!

```
dataset <- cars
```

cars is an example dataset included in the tidyverse package

Provide a short summary of the dataset(s)

```
summary(dataset)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

In this case we see two variables, speed and distance but we miss information on what units they are in. km/hour? Or meters/second?

These are things that are usually included in the metadata of the dataset. Provide us with the information from your metadata that we need to understand your dataset of choice.

Describe the type of variables included

Think of things like:

- Do the variables contain health information or SES information?
- Have they been measured by interviewing individuals or is the data coming from administrative sources?

Part 2 - Quantifying

2.1 Data cleaning

Say we want to include only larger distances (above 2) in our dataset, we can filter for this.

```
dataset <- dataset %>% filter(dist > 2)
```

Please use a separate 'R block' of code for each type of cleaning. So, e.g. one for missing values, a new one for removing unnecessary variables etc.

2.2 Generate necessary variables

Variable 1

```
dataset$duration <- dataset$speed / dataset$dist
```

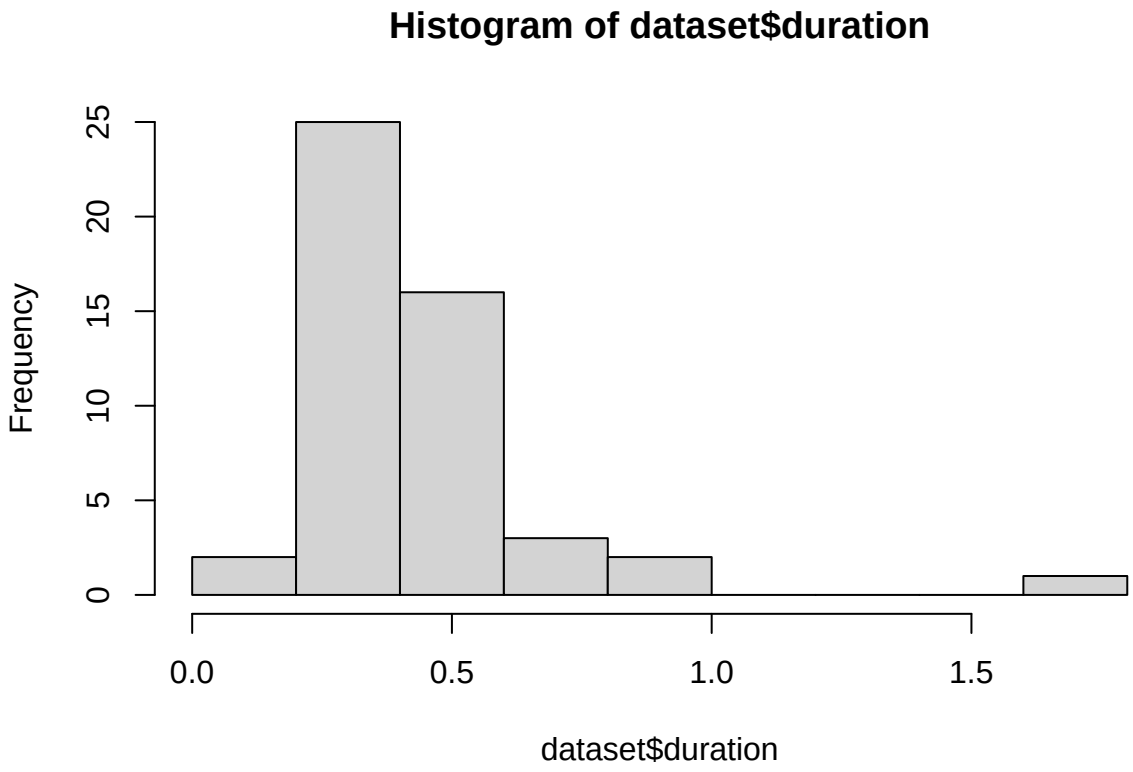
Variable 2

```
dataset$duration <- dataset$speed / dataset$dist
```

2.3 Visualize distributions and relationships

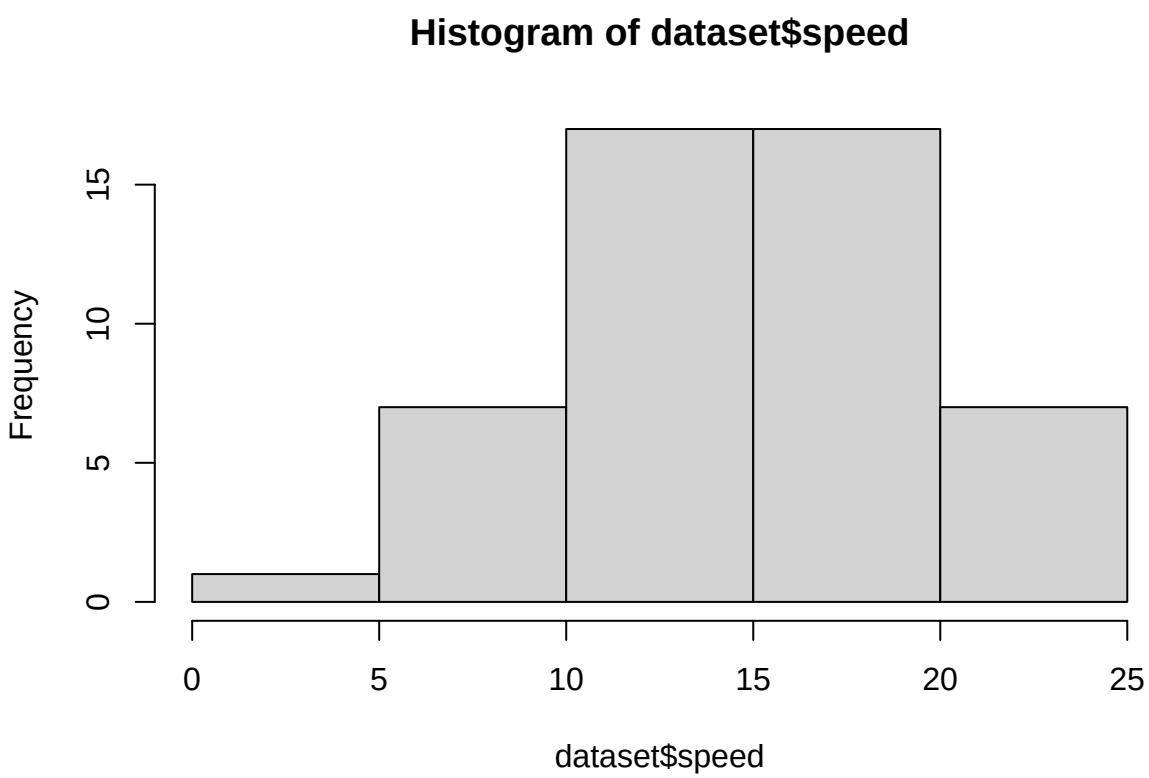
This is way too simple an example, but is just meant for your understanding of the expected markdown structure.

```
hist(dataset$duration)
```



Idem for variable 2.

```
hist(dataset$speed)
```



Visualize relationship between two variables. Again this is way too simple an example, but is just meant for your understanding of the expected markdown structure.

```r
cor(dataset$dist, dataset$speed)
```

```
## [1] 0.795126
```

2.4 Analysis

Analyze the relationship between two variables. Again this is way too simple an example, but is just meant for your understanding of the expected markdown structure.

```r
glm(duration ~ speed, data = dataset) %>%
  summary()
```

```
##
## Call:
## glm(formula = duration ~ speed, data = dataset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.30365  -0.11641  -0.01663   0.06158   1.13055
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.760250   0.105998   7.172 4.48e-09 ***
## speed       -0.020115   0.006455  -3.116  0.00312 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.05156235)
##
##     Null deviance: 2.9241  on 48  degrees of freedom
## Residual deviance: 2.4234  on 47  degrees of freedom
## AIC: -2.2692
##
## Number of Fisher Scoring iterations: 2
```

Part 3 - Report

3.1 Discuss your findings

See assignment hand-out for the guidelines.

3.2 Provide a description of the input of each project member

See assignment hand-out for the guidelines.

Part 4 - Reproducibility

4.1 Github repository link

Provide the link here: ...

4.2 Reference list

Use APA referencing throughout your document.