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

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!

```
# cars is an example dataset included in the tidyverse package
dataset <- cars
```

Provide a short summary of the dataset(s)

```
summary(cars)
```

```
##      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

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

Variable 2

2.3 Visualize distributions and relationships

Visualize variable 1

Visualize variable 2

Visualize relationship between two variables

2.4 Analysis

Analyze the relationship between two variables

Part 3 - Report

3.1 Discuss your findings

3.2 Provide a description of the input of each project member

Part 4 - Reproducibility

4.1 Github repository link

Provide the link here: ...

4.2 Reference list