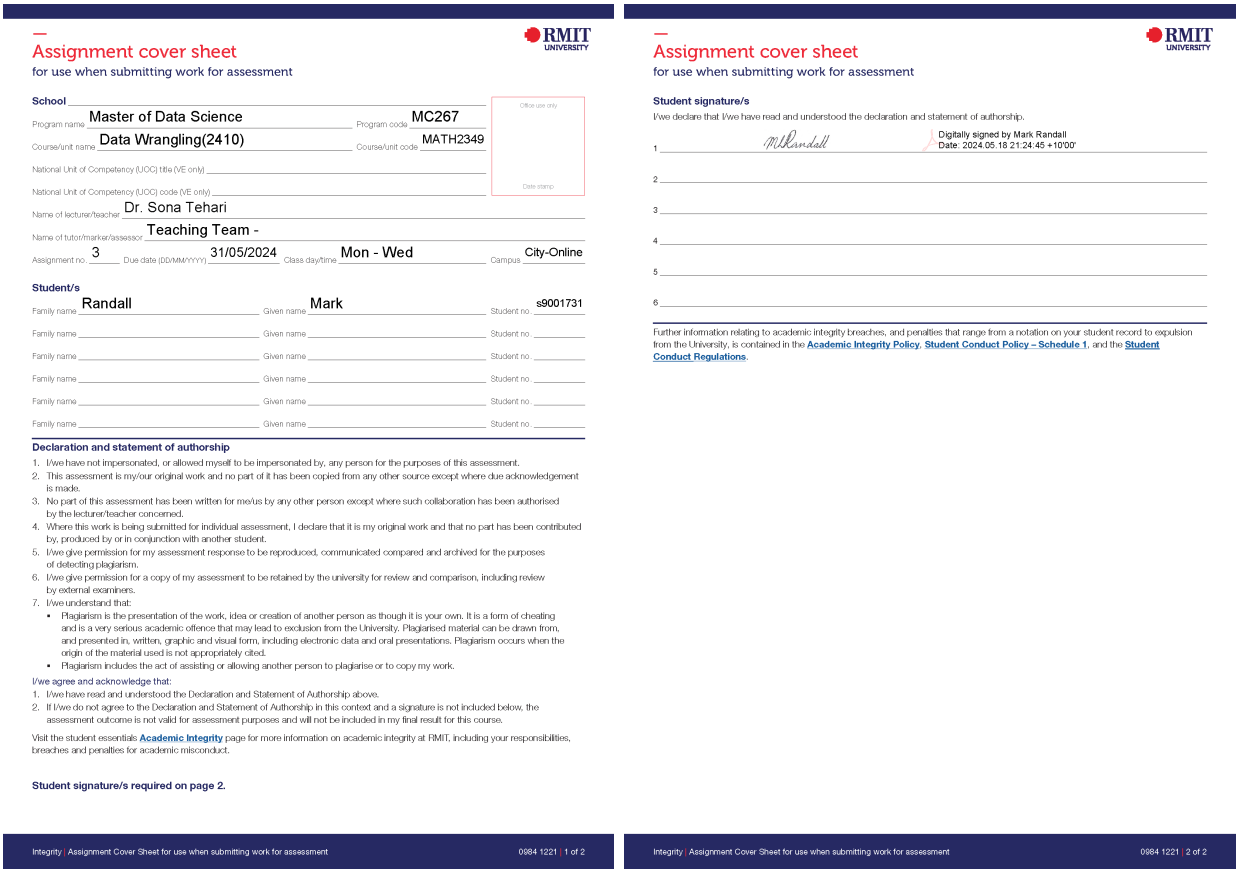
# Data Wrangling (Data Preprocessing)

## Practical Assessment 2

s9001731 Mark randall

2024-05-23



Assignment cover sheet

for use when submitting work for assessment

RMIT UNIVERSITY

School

Program name: Master of Data Science — Program code: MC267

Course/unit name: Data Wrangling(2410) — Course/unit code: MATH2349

National Unit of Competency (UOC) title (VE only)

National Unit of Competency (UOC) code (VE only)

Office use only — Date stamp

Name of lecturer/teacher: Dr. Sona Tehari

Name of tutor/marker/assessor: Teaching Team -

Assignment no.: 3 — Due date (DD/MM/YYYY): 31/05/2024 — Class day/time: Mon - Wed — Campus: City-Online

**Student/s**

Family name: Randall — Given name: Mark — Student no.: s9001731

Family name — Given name — Student no.

Family name — Given name — Student no.

Family name — Given name — Student no.

Family name — Given name — Student no.

Family name — Given name — Student no.

**Declaration and statement of authorship**

1. I/we have not impersonated, or allowed myself to be impersonated by, any person for the purposes of this assessment.
2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made.
3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned.
4. Where this work is being submitted for individual assessment, I declare that it is my original work and that no part has been contributed by, produced by or in conjunction with another student.
5. I/we give permission for my assessment response to be reproduced, communicated compared and archived for the purposes of detecting plagiarism.
6. I/we give permission for a copy of my assessment to be retained by the university for review and comparison, including review by external examiners.
7. I/we understand that:
   - Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.
   - Plagiarism includes the act of assisting or allowing another person to plagiarise or to copy my work.

I/we agree and acknowledge that:

1. I/we have read and understood the Declaration and Statement of Authorship above.
2. If I/we do not agree to the Declaration and Statement of Authorship in this context and a signature is not included below, the assessment outcome is not valid for assessment purposes and will not be included in my final result for this course.

Visit the student essentials **Academic Integrity** page for more information on academic integrity at RMIT, including your responsibilities, breaches and penalties for academic misconduct.

**Student signature/s required on page 2.**

Integrity | Assignment Cover Sheet for use when submitting work for assessment — 0984 1221 | 1 of 2

Assignment cover sheet

for use when submitting work for assessment

RMIT UNIVERSITY

**Student signature/s**

I/we declare that I/we have read and understood the declaration and statement of authorship.

1 MRandall — Digitally signed by Mark Randall Date: 2024.05.18 21:24:45 +10'00'

2

3

4

5

6

Further information relating to academic integrity breaches, and penalties that range from a notation on your student record to expulsion from the University, is contained in the **Academic Integrity Policy**, **Student Conduct Policy – Schedule 1**, and the **Student Conduct Regulations**.

Integrity | Assignment Cover Sheet for use when submitting work for assessment — 0984 1221 | 2 of 2

# Student names, numbers and percentage of contributions

Table 1: Group information

| Student name | Student number | Percentage of contribution |
|---|---|---|
| Mark Randall | s9001731 | 100 |

## Library Load

Package 1 : OpenStreetMap (Fellows & JMapViewer library by Jan Peter Stotz 2023)
Package 2 : tidyterra (Hernangómez 2024)
Package 3 : maptiles (Giraud 2024)
Package 4 : sf (Pebesma 2024)
Package 5 : sn (Azzalini 2023)
Package 6 : stats4 (R Core Team 2024a)
Package 7 : moments (Komsta & Novomestky 2022)
Package 8 : ggnewscale (Campitelli 2024)
Package 9 : Hmisc (Harrell Jr 2024)
Package 10 : validate (van der Loo & de Jonge 2024)
Package 11 : deducorrect (van der Loo, de Jonge & Scholtus 2015)
Package 12 : editrules (de Jonge & van der Loo 2024)
Package 13 : igraph (Csárdi et al. 2024)
Package 14 : deductive (van der Loo & de Jonge 2021)
Package 15 : tidyselect (Henry & Wickham 2024)
Package 16 : rvest (Wickham 2024)
Package 17 : here (Müller 2020)
Package 18 : glue (Hester & Bryan 2024)
Package 19 : magrittr (Bache & Wickham 2022)
Package 20 : lubridate (Spinu, Grolemund & Wickham 2023)
Package 21 : forcats (Wickham 2023a)
Package 22 : stringr (Wickham 2023b)
Package 23 : purrr (Wickham & Henry 2023)
Package 24 : tibble (Müller & Wickham 2023)
Package 25 : ggplot2 (Wickham et al. 2024)
Package 26 : tidyverse (Wickham 2023c)
Package 27 : kableExtra (Zhu 2024)
Package 28 : knitr (Xie 2024)
Package 29 : readxl (Wickham & Bryan 2023)
Package 30 : readr (Wickham, Hester & Bryan 2024)
Package 31 : dplyr (Wickham et al. 2023)
Package 32 : tidyr (Wickham, Vaughan & Girlich 2024)
Package 33 : openxlsx (Schauberger & Walker 2023)
Package 34 : stats (R Core Team 2024b)
Package 35 : graphics (R Core Team 2024c)
Package 36 : grDevices (R Core Team 2024d)
Package 37 : utils (R Core Team 2024e)
Package 38 : datasets (R Core Team 2024f)
Package 39 : methods (R Core Team 2024g)
Package 40 : base (R Core Team 2024h)

## Abstract

***"Most vehicular accidents in Victoria involve a male driver between 18 to 30 years of Age and a high powered car."***

This project will use some empirical data collected by the Victorian State Government(Vic 2024) to examine

this statement.

## Executive Summary

The Victoria Road Crash Data URL(Vic 2024) contains nine comma-separated value (csv) and one geo spatial java script object notation (GeoJSON) file. These were downloaded to a Data folder for examination. The files are:

```
csvFileNames <- list.files("../../Data",pattern = "*.csv", full.names = TRUE)
fullFileNames <- list.files("../../Data", full.names = TRUE)

for (file in fullFileNames) {
  cat(paste("-\t", file,"\n"))
}
```

- ../../Data/ACCIDENT.csv
- ../../Data/ACCIDENT_EVENT.csv
- ../../Data/ACCIDENT_LOCATION.csv
- ../../Data/ATMOSPHERIC_COND.csv
- ../../Data/NODE.csv
- ../../Data/PERSON.csv
- ../../Data/ROAD_SURFACE_COND.csv
- ../../Data/SUB_DCA.csv
- ../../Data/VEHICLE.csv
- ../../Data/VICTORIAN_ROAD_CRASH_DATA.geojson

```
#Create data frames fro csv files
vicRoadsDFList <- sapply(csvFileNames, read.csv)
#Change Key of list to mor HR form
names(vicRoadsDFList) <- c(gsub("../../Data/","", names(vicRoadsDFList)))
```

The URL indicates that the metadata was updated 29 April 2024, data observations updated as at 27 November 2024 and that observations temporal start was 1 January 2012.

```
posnOfDF <- 1
cat("These files have duplicated accident number keys in the data frame:")
```

These files have duplicated accident number keys in the data frame:

```
for (tmpDF in vicRoadsDFList) {
  if (n_distinct(tmpDF$ACCIDENT_NO) < dim(tmpDF)[[1]]) {
    cat(paste("-\t",names(vicRoadsDFList)[[posnOfDF]],"\n"))
  }
  posnOfDF <- posnOfDF + 1
}
```

- ACCIDENT_EVENT.csv
- ATMOSPHERIC_COND.csv
- NODE.csv

- PERSON.csv
- ROAD_SURFACE_COND.csv
- SUB_DCA.csv
- VEHICLE.csv

This would suggest that the ACCIDENT_NO key/attribute is a foreign key in these files. The following is a lst of the attributes/column names by data frame.

```
vecColNamesDF <- list()
posnCN <- 1
for (df in vicRoadsDFList) {
    vecColNamesDF[[names(vicRoadsDFList)[[posnCN]]]] <- colnames(df)

    posnCN <- posnCN + 1
}
# https://stackoverflow.com/questions/60199801/how-to-view-a-list-like-table-style-in-r
max_len <- max(lengths(vecColNamesDF))
df <- purrr::map_df(vecColNamesDF, ~c(., rep("", max_len -
    length(.))))
df[, 1:5] %>%
    kable(caption = "Attributes Files 1-4", longtable = TRUE,
        format = "latex", booktabs = TRUE) %>%
    kable_styling(font_size = 5)
```

Table 2: Attributes Files 1-4

| ACCIDENT.csv | ACCIDENT_EVENT.csv | ACCIDENT_LOCATION.csv | ATMOSPHERIC_COND.csv | NODE.csv |
|---|---|---|---|---|
| ACCIDENT_NO | ACCIDENT_NO | ACCIDENT_NO | ACCIDENT_NO | ACCIDENT_NO |
| ACCIDENT_DATE | EVENT_SEQ_NO | NODE_ID | ATMOSPH_COND | NODE_ID |
| ACCIDENT_TIME | EVENT_TYPE | ROAD_ROUTE_1 | ATMOSPH_COND_SEQ | NODE_TYPE |
| ACCIDENT_TYPE | EVENT_TYPE_DESC | ROAD_NAME | ATMOSPH_COND_DESC | AMG_X |
| ACCIDENT_TYPE_DESC | VEHICLE_1_ID | ROAD_TYPE | | AMG_Y |
| DAY_OF_WEEK | VEHICLE_1_COLL_PT | ROAD_NAME_INT | | LGA_NAME |
| DAY_WEEK_DESC | VEHICLE.1.COLL.PT.DESC | ROAD_TYPE_INT | | DEG_URBAN_NAME |
| DCA_CODE | VEHICLE_2_ID | DISTANCE_LOCATION | | LATITUDE |
| DCA_DESC | VEHICLE_2_COLL_PT | DIRECTION_LOCATION | | LONGITUDE |
| LIGHT_CONDITION | VEHICLE.2.COLL.PT.DESC | | | POSTCODE_CRASH |
| NODE_ID | PERSON_ID | | | |
| NO_OF_VEHICLES | OBJECT_TYPE | | | |
| NO_PERSONS_KILLED | OBJECT_TYPE_DESC | | | |
| NO_PERSONS_INJ_2 | | | | |
| NO_PERSONS_INJ_3 | | | | |
| NO_PERSONS_NOT_INJ | | | | |
| NO_PERSONS | | | | |
| POLICE_ATTEND | | | | |
| ROAD_GEOMETRY | | | | |
| ROAD_GEOMETRY_DESC | | | | |
| SEVERITY | | | | |
| SPEED_ZONE | | | | |
| RMA | | | | |

```
df[, 6:9] %>%
    kable(caption = "Attributes Files 5-9", longtable = TRUE,
        format = "latex", booktabs = TRUE) %>%
    kable_styling(font_size = 5)
```

Table 3: Attributes Files 5-9

| PERSON.csv | ROAD_SURFACE_COND.csv | SUB_DCA.csv | VEHICLE.csv |
|---|---|---|---|
| ACCIDENT_NO | ACCIDENT_NO | ACCIDENT_NO | ACCIDENT_NO |
| PERSON_ID | SURFACE_COND | SUB_DCA_CODE | VEHICLE_ID |
| VEHICLE_ID | SURFACE_COND_DESC | SUB_DCA_SEQ | VEHICLE_YEAR_MANUF |
| SEX | SURFACE_COND_SEQ | SUB_DCA_CODE_DESC | VEHICLE_DCA_CODE |
| AGE_GROUP | | | INITIAL_DIRECTION |
| INJ_LEVEL | | | ROAD_SURFACE_TYPE |
| INJ_LEVEL_DESC | | | ROAD_SURFACE_TYPE_DESC |
| SEATING_POSITION | | | REG_STATE |
| HELMET_BELT_WORN | | | VEHICLE_BODY_STYLE |
| ROAD_USER_TYPE | | | VEHICLE_MAKE |
| ROAD_USER_TYPE_DESC | | | VEHICLE_MODEL |
| LICENCE_STATE | | | VEHICLE_POWER |
| TAKEN_HOSPITAL | | | VEHICLE_TYPE |
| EJECTED_CODE | | | VEHICLE_TYPE_DESC |
| | | | VEHICLE_WEIGHT |
| | | | CONSTRUCTION_TYPE |
| | | | FUEL_TYPE |
| | | | NO_OF_WHEELS |
| | | | NO_OF_CYLINDERS |
| | | | SEATING_CAPACITY |
| | | | TARE_WEIGHT |
| | | | TOTAL_NO_OCCUPANTS |
| | | | CARRY_CAPACITY |
| | | | CUBIC_CAPACITY |
| | | | FINAL_DIRECTION |
| | | | DRIVER_INTENT |
| | | | VEHICLE_MOVEMENT |
| | | | TRAILER_TYPE |
| | | | VEHICLE_COLOUR_1 |
| | | | VEHICLE_COLOUR_2 |
| | | | CAUGHT_FIRE |
| | | | INITIAL_IMPACT |
| | | | LAMPS |
| | | | LEVEL_OF_DAMAGE |
| | | | TOWED_AWAY_FLAG |
| | | | TRAFFIC_CONTROL |
| | | | TRAFFIC_CONTROL_DESC |

The following details those tables with common attribute names and what those names are.

```
namesOfFile <- names(vecColNamesDF)
for (posnOne in 1:length(namesOfFile)) {
    cat(namesOfFile[[posnOne]], "at list no.", posnOne,
        "intersects with the following file:")
    cat("  \n")
    if (posnOne == length(namesOfFile)) {
        break
    }
    for (posnTwo in (posnOne + 1):length(namesOfFile)) {
        cat("-\t", namesOfFile[[posnTwo]], "at list no.",
            posnTwo, " with these attributes:")
        cat("  \n")
        cat("\t\t-\t", intersect(vecColNamesDF[[posnOne]],
            vecColNamesDF[[posnTwo]]))
        cat("  \n")

        posnTwo <- posnTwo + 1
    }
    cat("  \n")
    posnOne <- posnOne + 1

}
```

ACCIDENT.csv at list no. 1 intersects with the following file:
- ACCIDENT_EVENT.csv at list no. 2 with these attributes:

- ACCIDENT_NO
- ACCIDENT_LOCATION.csv at list no. 3 with these attributes:
- ACCIDENT_NO NODE_ID
- ATMOSPHERIC_COND.csv at list no. 4 with these attributes:
- ACCIDENT_NO
- NODE.csv at list no. 5 with these attributes:
- ACCIDENT_NO NODE_ID
- PERSON.csv at list no. 6 with these attributes:
- ACCIDENT_NO
- ROAD_SURFACE_COND.csv at list no. 7 with these attributes:
- ACCIDENT_NO
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

ACCIDENT_EVENT.csv at list no. 2 intersects with the following file:
- ACCIDENT_LOCATION.csv at list no. 3 with these attributes:
- ACCIDENT_NO
- ATMOSPHERIC_COND.csv at list no. 4 with these attributes:
- ACCIDENT_NO
- NODE.csv at list no. 5 with these attributes:
- ACCIDENT_NO
- PERSON.csv at list no. 6 with these attributes:
- ACCIDENT_NO PERSON_ID
- ROAD_SURFACE_COND.csv at list no. 7 with these attributes:
- ACCIDENT_NO
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

ACCIDENT_LOCATION.csv at list no. 3 intersects with the following file:
- ATMOSPHERIC_COND.csv at list no. 4 with these attributes:
- ACCIDENT_NO
- NODE.csv at list no. 5 with these attributes:
- ACCIDENT_NO NODE_ID
- PERSON.csv at list no. 6 with these attributes:
- ACCIDENT_NO
- ROAD_SURFACE_COND.csv at list no. 7 with these attributes:
- ACCIDENT_NO
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

ATMOSPHERIC_COND.csv at list no. 4 intersects with the following file:
- NODE.csv at list no. 5 with these attributes:
- ACCIDENT_NO
- PERSON.csv at list no. 6 with these attributes:
- ACCIDENT_NO
- ROAD_SURFACE_COND.csv at list no. 7 with these attributes:
- ACCIDENT_NO
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

NODE.csv at list no. 5 intersects with the following file:
- PERSON.csv at list no. 6 with these attributes:
- ACCIDENT_NO
- ROAD_SURFACE_COND.csv at list no. 7 with these attributes:
- ACCIDENT_NO
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

PERSON.csv at list no. 6 intersects with the following file:
- ROAD_SURFACE_COND.csv at list no. 7 with these attributes:

- ACCIDENT_NO
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO VEHICLE_ID

ROAD_SURFACE_COND.csv at list no. 7 intersects with the following file:
- SUB_DCA.csv at list no. 8 with these attributes:
- ACCIDENT_NO
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

SUB_DCA.csv at list no. 8 intersects with the following file:
- VEHICLE.csv at list no. 9 with these attributes:
- ACCIDENT_NO

VEHICLE.csv at list no. 9 intersects with the following file:

## Data

Provide explanations here.

```
# Import the data, provide your R codes here.
```

## Understand

```
# This is the R chunk for the Understand Section
```

Provide explanations here.

## Tidy & Manipulate Data I

```
# This is the R chunk for the Tidy & Manipulate Data I
```

Provide explanations here.

## Tidy & Manipulate Data II

```
# This is the R chunk for the Tidy & Manipulate Data II
```

Provide explanations here.

## Scan I

```
# This is the R chunk for the Scan I
```

Provide explanations here.

## Scan II

```
# This is the R chunk for the Scan II
```

Provide explanations here.

## Transform

```
# This is the R chunk for the Transform Section
```

Provide explanations here.

## Bibliography


Azzalini, A 2023, *Sn: The skew-normal and related distributions such as the skew-t and the SUN*,.

Bache, SM & Wickham, H 2022, *Magrittr: A forward-pipe operator for r*,.

Campitelli, E 2024, *Ggnewscale: Multiple fill and colour scales in 'ggplot2'*,.

Csárdi, G, Nepusz, T, Traag, V, Horvát, S, Zanini, F, Noom, D & Müller, K 2024, *Igraph: Network analysis and visualization*,.

de Jonge, E & van der Loo, M 2024, *Editrules: Parsing, applying, and manipulating data cleaning rules*,.

Fellows, I & JMapViewer library by Jan Peter Stotz, using the 2023, *OpenStreetMap: Access to open street map raster images*,.

Giraud, T 2024, *Maptiles: Download and display map tiles*,.

Harrell Jr, FE 2024, *Hmisc: Harrell miscellaneous*,.

Henry, L & Wickham, H 2024, *Tidyselect: Select from a set of strings*,.

Hernangómez, D 2024, *Tidyterra: 'Tidyverse' methods and 'ggplot2' helpers for 'terra' objects*,.

Hester, J & Bryan, J 2024, *Glue: Interpreted string literals*,.

Komsta, L & Novomestky, F 2022, *Moments: Moments, cumulants, skewness, kurtosis and related tests*,.

Müller, K 2020, *Here: A simpler way to find your files*,.

Müller, K & Wickham, H 2023, *Tibble: Simple data frames*,.

Pebesma, E 2024, *Sf: Simple features for r*,.

R Core Team 2024a, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2024b, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2024c, *R: A language and environment for statistical computing*, R Foundation

for Statistical Computing, Vienna, Austria.

R Core Team 2024d, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2024e, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2024f, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2024g, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2024h, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Schauberger, P & Walker, A 2023, *Openxlsx: Read, write and edit xlsx files*,.

Spinu, V, Grolemund, G & Wickham, H 2023, *Lubridate: Make dealing with dates a little easier*,.

van der Loo, M & de Jonge, E 2021, *Deductive: Data correction and imputation using deductive methods*,.

van der Loo, M & de Jonge, E 2024, *Validate: Data validation infrastructure*,.

van der Loo, M, de Jonge, E & Scholtus, S 2015, *Deducorrect: Deductive correction, deductive imputation, and deterministic correction*,.

Vic, D 2024, 'Victoria road crash data', vol. 2024.

Wickham, H 2023a, *Forcats: Tools for working with categorical variables (factors)*,.

Wickham, H 2023b, *Stringr: Simple, consistent wrappers for common string operations*,.

Wickham, H 2023c, *Tidyverse: Easily install and load the 'tidyverse'*,.

Wickham, H 2024, *Rvest: Easily harvest (scrape) web pages*,.

Wickham, H & Bryan, J 2023, *Readxl: Read excel files*,.

Wickham, H, Chang, W, Henry, L, Pedersen, TL, Takahashi, K, Wilke, C, Woo, K, Yutani, H, Dunnington, D & van den Brand, T 2024, *ggplot2: Create elegant data visualisations using the grammar of graphics*,.

Wickham, H, François, R, Henry, L, Müller, K & Vaughan, D 2023, *Dplyr: A grammar of data manipulation*,.

Wickham, H & Henry, L 2023, *Purrr: Functional programming tools*,.

Wickham, H, Hester, J & Bryan, J 2024, *Readr: Read rectangular text data*,.

Wickham, H, Vaughan, D & Girlich, M 2024, *Tidyr: Tidy messy data*,.

Xie, Y 2024, *Knitr: A general-purpose package for dynamic report generation in r*,.

Zhu, H 2024, *kableExtra: Construct complex table with 'kable' and pipe syntax*,.