Data Wrangling

Code

Practical assessment 2

Jason Wang

9/10/2024

Setup

Hide

```
library(tidyr)
library(dplyr)
```

```
Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

Hide

```
library(kableExtra)
```

```
Attaching package: ‘kableExtra’

The following object is masked from ‘package:dplyr’:

    group_rows
```

Hide

```
library(magrittr)
```

```
Attaching package: ‘magrittr’

The following object is masked from ‘package:tidyr’:

    extract
```

Hide

```
library(rmarkdown)
library(openxlsx)
library(stringr)
```

Student names, numbers and percentage of contributions

Group information

Student name	Student number	Percentage of contribution
Jason Wang	s4134626	100%

Executive Summary

- 2 datasets were merged. A crime statistics dataset from the Australian Bureau of Statistics, and an education dataset. They were merged through the common variables of Year and State. The point of this merging was to test if there is a correlation between attaining the proportion of people with non-school qualifications and a lower crime rate.
- Both datasets were initially untidy. Both needed pivot_longer to turn into a tidy format. This tidying process resulted in observations through State and Year, as well as Total Crimes and Proportion of non-school qualifications.
- There are 2 resulted datasets, one with real data and one with extrapolated data. For the actual dataset (2013–2021), only data was used. In the extrapolated dataset, data was missing, a linear model was used to predict missing values, and was subsequently filled in. This proved logical as an extra form of comparison.
- A new categorical variable Crime Severity was created and turned into a factor. This was based on the number of crimes. State and year were also turned into factors.
- Data types were converted. These include Year and State which were converted into factors. Total Crimes was also log transformed to reduce skewness.
- Outliers in Total Crime were identified through boxplots. The log transformation was applied to also normalise the data. Visualisations of histograms and boxplots were made.

Data

- I sourced the crime dataset from the ABS. It shows the amount of crime per year per state beginning from 1993 to 2023. I sourced the education dataset from the ABS. It shows the proportion of people in a state that have a non-school qualification, which includes tradework or university. This comparison is done to see whether or not there is a correlation between higher proportion of people with non-school qualification, and lower crime rate.
- The crime dataset begins with variables of State, and 1993 to 2023 as columns. The education dataset begins with variables of State and 2013 to 2021 as columns. Description is also included to what the percentage means. I eventually merge them to get a resultant dataset called merged_data and work on

that.

- I begin by scraping the crime dataset from the internet and define it. I call it using read.xlsx and allocate the appropriate sheet that I would like to analyse. Start row begins at 6 because there is no data until 6. I remove the bottom 30 rows because they account for no data.
- I then perform data cleaning and transformation because the state names are within the 1993 column. A new column called State is made, and it is filled accordingly to the new column based on the state names found in the column. This results in state names corresponding to a new instance, which is the beginning of becoming tidy. After this, the state names in the 1993 column are now redundant, so i remove them just with the row number they correspond to.
- For values that are not published, I changed it to N/A rather than a “np” string. This is so that we do not have string and numeric clash when summing later on. I did not extrapolate because I believe extrapolating more values would not represent a true representation, so later on i extrapolate totals rather than individual years.
- Afterwards, I use pivot longer to make it tidy, converting 1993-2023 to a column called Year, and the corresponding values to a new column called Count. I also strip the year column since it had letters in it, ensuring only numbers are in the column. I then turned Year and Count into a numeric.
- Lastly, i sum the total crimes to the year to condense it so that it is mergeable with my other dataset and so that there are much less observations to work with.
- Similarly in the education dataset, I scrape it from the web. I choose the sheet I would like to use and the start row. I only wanted 8 rows beginning from 50 as it showed proportions, which I had wanted to use. I rename the columns using colnames and I create a new column for the description called Proportion of persons with non-school qualifications (%). This just serves as a placeholder.
- I then ensured all year columns were numeric and then proceeded to pivot longer. I changed the names of the state so that they matched the crime dataset, rather than being abbreviated.
- I use left join because the crime dataset is much larger, it spans over 30 years, whilst the education database only spans for 10 years. So left join is very ideal because the empty years resulted from the mismatch of years will still show. I then extrapolate to fill in the data later.
- This is where there becomes 2 tables, one for filtered real data, and one that is extrapolated. Filtered real data removes the extra years in the crime dataset that the education dataset does not have, and shows only true numbers.
- I also create a factor variable called Crime Severity, denoting how severe a state is in a particular year for crime.

Hide

```
crime.url <- "https://www.abs.gov.au/statistics/people/crime-and-justice/recorded-crime-victims/2023/2.%20Victims%20of%20crime%2C%20states%20and%20territories%20%28Tables%209%20to%2016%29.xlsx"
crime <- read.xlsx(crime.url, sheet = "Table 9", startRow = 6)
crime <- crime[1:(nrow(crime) - 30), ]


crime <- crime %>%
  mutate(State = case_when(
    str_detect(`1993`, "New South Wales") ~ "New South Wales",
    str_detect(`1993`, "Victoria") ~ "Victoria",
    str_detect(`1993`, "Queensland") ~ "Queensland",
    str_detect(`1993`, "South Australia") ~ "South Australia",
    str_detect(`1993`, "Western Australia") ~ "Western Australia",
    str_detect(`1993`, "Tasmania") ~ "Tasmania",
    str_detect(`1993`, "Northern Territory") ~ "Northern Territory",
    str_detect(`1993`, "Australian Capital Territory") ~ "Australian Capital Territory",
    TRUE ~ NA_character_
  )) %>%
  fill(State, .direction = "down")

crime_clean <- crime[-c(1, 18, 35, 52, 69, 86, 103, 120), ]

crime_clean <- crime_clean %>% mutate(across(`1993`:`2023`, ~ na_if(., "np")))

crime_tidy <- crime_clean %>%
  pivot_longer(
    cols = `1993`:`2023`,
    names_to = "Year",
    values_to = "Count"
  ) %>%
  mutate(
    Year = gsub("[^0-9]", "", Year),
    Year = as.numeric(Year),
    Count = as.numeric(str_replace_all(Count, "[^0-9\\.]", ""))
  )

crime_total <- crime_tidy %>% group_by(State, Year) %>% summarise(Total_Crimes = sum(Count, na.rm = TRUE))
```

```
`summarise()` has grouped output by 'State'. You can override using the `.groups` argument.
```

Hide

```
head(crime_total)
```

State <chr>	Year <dbl>	Total_Crimes <dbl>
Australian Capital Territory	1993	7703
Australian Capital Territory	1994	6974
Australian Capital Territory	1995	22027

State <chr>	Year <dbl>	Total_Crimes <dbl>
Australian Capital Territory	1996	23335
Australian Capital Territory	1997	21965
Australian Capital Territory	1998	26601

6 rows

Hide

```
education.url <- "https://www.abs.gov.au/statistics/people/education/education-and-work-australia/may-2021/Education%20and%20work%2C%202021%2C%20Datacube%2013%20%28Table%2025%29.xlsx"

education <- read.xlsx(education.url, sheet = "2013-2021", startRow = 50)
education <- education [1:8, ]
colnames(education) <- c("State", "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021")
education <- education %>%
  mutate(Description = "Proportion of persons with non-school qualifications (%)")

education_cleaned <- education %>%
  mutate(across(`2013`:`2021`, as.numeric))

education_tidy <- education_cleaned %>%
  pivot_longer(
    cols = `2013`:`2021`,
    names_to = "Year",
    values_to = "Proportion"
  )

education_tidy <- education_tidy %>%
  mutate(State = case_when(
    State == "NSW" ~ "New South Wales",
    State == "Vic." ~ "Victoria",
    State == "Qld" ~ "Queensland",
    State == "SA" ~ "South Australia",
    State == "WA" ~ "Western Australia",
    State == "Tas." ~ "Tasmania",
    State == "NT" ~ "Northern Territory",
    State == "ACT" ~ "Australian Capital Territory",
    TRUE ~ State))

education_tidy <- education_tidy %>% mutate(Year = as.numeric(Year))

head(education_tidy)
```

State <chr>	Description <chr>	Y… <dbl>
New South Wales	Proportion of persons with non-school qualifications (%)	2013
New South Wales	Proportion of persons with non-school qualifications (%)	2014

State	Description	Y…
<chr>	<chr>	<dbl>
New South Wales	Proportion of persons with non-school qualifications (%)	2015
New South Wales	Proportion of persons with non-school qualifications (%)	2016
New South Wales	Proportion of persons with non-school qualifications (%)	2017
New South Wales	Proportion of persons with non-school qualifications (%)	2018

6 rows

Hide

```
# MERGING REAL

merged_data_real <- left_join(crime_total, education_tidy, by = c("State", "Year"))
merged_data_real <- merged_data_real %>% mutate(State = as.factor(State), Year = as.factor(Ye
ar))

merged_data_real <- merged_data_real %>%
  select(-Description) %>%
  rename(Proportion_Non_School_Qualifications = Proportion)

filtered_data_real <- merged_data_real %>%
  mutate(Year = as.numeric(as.character(Year))) %>%
  filter(Year >= 2013 & Year <= 2021)

filtered_data_real <- filtered_data_real %>%
  mutate(Crime_Severity = case_when(
    Total_Crimes < 10000 ~ "Low",
    Total_Crimes >= 10000 & Total_Crimes < 25000 ~ "Moderate",
    Total_Crimes >= 25000 ~ "High"
  ))

filtered_data_real <- filtered_data_real %>%
  mutate(Crime_Severity = factor(Crime_Severity, levels = c("Low", "Moderate", "High"), order
ed = TRUE))

head(filtered_data_real)
```

State	Y…	Total_Crimes	Proportion_Non_School_Qualifica
<fctr>	<dbl>	<dbl>	
Australian Capital Territory	2013	16605	
Australian Capital Territory	2014	16433	
Australian Capital Territory	2015	19287	
Australian Capital Territory	2016	17964	
Australian Capital Territory	2017	19229	
Australian Capital Territory	2018	17272	

6 rows

Hide

```
# MERGING FOR EXTRAPOLATION LATER

merged_data <- left_join(crime_total, education_tidy, by = c("State", "Year"))

merged_data <- merged_data %>%  mutate(State = as.factor(State), Year = as.factor(Year))

merged_data <- merged_data %>%
  select(-Description) %>%
  rename(Proportion_Non_School_Qualifications = Proportion)
```

Understand

- When we inspect crime, we can see that all the columns are all characters. After pivoting, I ensured that the Year column was stripped so that there were no letters with the numbers. I then used as.numeric to turn Year into numeric from character. Likewise, Count was also turned into a numeric. So when we inspect crime_tidy, Offence and State are in character form, and State and Year are in numeric.
- In the education dataset, when we inspect it, we can see that 2014 to 2021 are in numeric, but 2013 is still in character. Using class function confirms it. I change it to numeric and is successfully changed when inspecting it afterwards.
- When merging the datasets, I converted both State and Year to factors because they are categories which is important for analyses such as grouping, summarizing, or creating models.

Hide

```
head(crime)
```

	Offence <chr>	**1993** <chr>	**1…** <chr>	**1995** <chr>	**1996** <chr>	**1997** <chr>	**1998** <chr>	**199** <chr>
1	*NA*	New South Wales	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	Homicide and related offences(h)	206	178	167	198	221	239	267
3	Murder	118	108	103	95	110	93	124
4	Attempted murder	85	60	57	85	100	121	132
5	Manslaughter	7	7	8	15	11	26	14
6	Assault(i)	np	np	37863	47828	55995	59219	638

6 rows | 1-10 of 33 columns

Hide

```
summary(crime)
```

```
   Offence             1993               1994               1995               1996
1997               1998             1999(a)              2000               2001
 Length:136         Length:136         Length:136         Length:136         Length:136
Length:136         Length:136         Length:136         Length:136         Length:136
 Class :character   Class :character   Class :character   Class :character   Class :character
Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
     2002               2003               2004               2005               2006
2007(b)            2008(c)           2009(d)(e)         2010(f)(g)            2011
 Length:136         Length:136         Length:136         Length:136         Length:136
Length:136         Length:136         Length:136         Length:136         Length:136
 Class :character   Class :character   Class :character   Class :character   Class :character
Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
     2012               2013               2014               2015               2016
2017               2018               2019               2020               2021
 Length:136         Length:136         Length:136         Length:136         Length:136
Length:136         Length:136         Length:136         Length:136         Length:136
 Class :character   Class :character   Class :character   Class :character   Class :character
Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
     2022               2023              State
 Length:136         Length:136         Length:136
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

Hide

```
crime_tidy <- crime_clean %>%
  pivot_longer(
    cols = `1993`:`2023`,
    names_to = "Year",
    values_to = "Count"
  ) %>%
  mutate(
    Year = gsub("[^0-9]", "", Year),
    Year = as.numeric(Year),
    Count = as.numeric(str_replace_all(Count, "[^0-9\\.]", ""))
  )

str(crime_tidy)
```

```
tibble [3,968 × 4] (S3: tbl_df/tbl/data.frame)
 $ Offence: chr [1:3968] "Homicide and related offences(h)" "Homicide and related offences
(h)" "Homicide and related offences(h)" "Homicide and related offences(h)" ...
 $ State  : chr [1:3968] "New South Wales" "New South Wales" "New South Wales" "New South Wal
es" ...
 $ Year   : num [1:3968] 1993 1994 1995 1996 1997 ...
 $ Count  : num [1:3968] 206 178 167 198 221 239 267 262 313 256 ...
```

Hide

```
head(crime_tidy)
```

Offence <chr>	State <chr>	Year <dbl>	Count <dbl>
Homicide and related offences(h)	New South Wales	1993	206
Homicide and related offences(h)	New South Wales	1994	178
Homicide and related offences(h)	New South Wales	1995	167
Homicide and related offences(h)	New South Wales	1996	198
Homicide and related offences(h)	New South Wales	1997	221
Homicide and related offences(h)	New South Wales	1998	239

6 rows

Hide

```
str(education)
```

```
'data.frame':   8 obs. of  11 variables:
 $ State      : chr  "NSW" "Vic." "Qld" "SA" ...
 $ 2013       : chr  "56.7" "57" "52.8" "52.9" ...
 $ 2014       : num  58 57.9 55.4 53.9 57.2 52.9 59.6 66.4
 $ 2015       : num  59.2 60.1 55.9 57.7 59.3 54.2 58.5 69.3
 $ 2016       : num  60.2 59.2 56.5 55.1 60.2 57.1 59 69.1
 $ 2017       : num  60.7 60.5 56.8 56.5 60.9 57.4 60.7 66
 $ 2018       : num  60.7 61.6 57.4 57.1 60.2 57.8 59.3 69.9
 $ 2019       : num  62.2 62.4 58.9 55.3 59.8 59.9 61 67.3
 $ 2020       : num  62.9 63.7 60.2 56.3 62 57.8 62.9 66.5
 $ 2021       : num  63.7 63.1 59.8 58.3 62.9 59.9 62.7 68.6
 $ Description: chr  "Proportion of persons with non-school qualifications (%)" "Proportion o
f persons with non-school qualifications (%)" "Proportion of persons with non-school qualific
ations (%)" "Proportion of persons with non-school qualifications (%)" ...
```

Hide

```
class(education[["2013"]])
```

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

Hide

```
education_cleaned <- education %>%
  mutate(across(`2013`:`2021`, as.numeric))

str(education_cleaned)
```

```
'data.frame':   8 obs. of  11 variables:
 $ State      : chr  "NSW" "Vic." "Qld" "SA" ...
 $ 2013       : num  56.7 57 52.8 52.9 55.2 52.7 55 63.6
 $ 2014       : num  58 57.9 55.4 53.9 57.2 52.9 59.6 66.4
 $ 2015       : num  59.2 60.1 55.9 57.7 59.3 54.2 58.5 69.3
 $ 2016       : num  60.2 59.2 56.5 55.1 60.2 57.1 59 69.1
 $ 2017       : num  60.7 60.5 56.8 56.5 60.9 57.4 60.7 66
 $ 2018       : num  60.7 61.6 57.4 57.1 60.2 57.8 59.3 69.9
 $ 2019       : num  62.2 62.4 58.9 55.3 59.8 59.9 61 67.3
 $ 2020       : num  62.9 63.7 60.2 56.3 62 57.8 62.9 66.5
 $ 2021       : num  63.7 63.1 59.8 58.3 62.9 59.9 62.7 68.6
 $ Description: chr  "Proportion of persons with non-school qualifications (%)" "Proportion o
f persons with non-school qualifications (%)" "Proportion of persons with non-school qualific
ations (%)" "Proportion of persons with non-school qualifications (%)" ...
```

Hide

```
merged_data <- merged_data %>%  mutate(State = as.factor(State), Year = as.factor(Year))
str(merged_data)
```

```
gropd_df [248 × 4] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ State                              : Factor w/ 8 levels "Australian Capital Territor
y",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year                               : Factor w/ 31 levels "1993","1994",..: 1 2 3 4 5 6 7
8 9 10 ...
 $ Total_Crimes                       : num [1:248] 7703 6974 22027 23335 21965 ...
 $ Proportion_Non_School_Qualifications: num [1:248] NA NA NA NA NA NA NA NA NA NA ...
 - attr(*, "groups")= tibble [8 × 2] (S3: tbl_df/tbl/data.frame)
  ..$ State: Factor w/ 8 levels "Australian Capital Territory",..: 1 2 3 4 5 6 7 8
  ..$ .rows: list<int> [1:8]
  .. ..$ : int [1:31] 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ : int [1:31] 32 33 34 35 36 37 38 39 40 41 ...
  .. ..$ : int [1:31] 63 64 65 66 67 68 69 70 71 72 ...
  .. ..$ : int [1:31] 94 95 96 97 98 99 100 101 102 103 ...
  .. ..$ : int [1:31] 125 126 127 128 129 130 131 132 133 134 ...
  .. ..$ : int [1:31] 156 157 158 159 160 161 162 163 164 165 ...
  .. ..$ : int [1:31] 187 188 189 190 191 192 193 194 195 196 ...
  .. ..$ : int [1:31] 218 219 220 221 222 223 224 225 226 227 ...
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

Hide

```
summary(merged_data)
```

```
                        State         Year      Total_Crimes    Proportion_Non_School_Quali
fications
 Australian Capital Territory:31   1993   :  8   Min.   :  4771   Min.   :52.70
 New South Wales              :31   1994   :  8   1st Qu.: 23450   1st Qu.:57.08
 Northern Territory           :31   1995   :  8   Median :139539   Median :59.45
 Queensland                   :31   1996   :  8   Mean   :158779   Mean   :59.71
 South Australia              :31   1997   :  8   3rd Qu.:234998   3rd Qu.:62.05
 Tasmania                     :31   1998   :  8   Max.   :719298   Max.   :69.90
 (Other)                      :62   (Other):200                    NA's   :176
```

Tidy & Manipulate Data I

- Tidy data needs to be: Each variable forms a column. Each observation forms a row. Each type of observational unit forms a table.
- In the crime dataset, the years from 1993 to 2023 are represented as multiple column names. It has the count of crime underneath each year. This violates the tidy data principle because Year should be a single variable and each crime count associated with that year should be an observation.
- I used pivot longer to convert the crime dataset. Crime_tidy is tidy, but I group the total crimes into totals for the corresponding year so that there are less observations. Crime_total is in tidy format.

Hide

```
head(crime)
```

	Offence <chr>	**1993** <chr>	**1…** <chr>	**1995** <chr>	**1996** <chr>	**1997** <chr>	**1998** <chr>	**199** <chr
1	*NA*	New South Wales	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	Homicide and related offences(h)	206	178	167	198	221	239	267
3	Murder	118	108	103	95	110	93	124
4	Attempted murder	85	60	57	85	100	121	132
5	Manslaughter	7	7	8	15	11	26	14
6	Assault(i)	np	np	37863	47828	55995	59219	638

6 rows | 1-10 of 33 columns

Hide

```
crime_tidy <- crime_clean %>%
  pivot_longer(
    cols = `1993`:`2023`,
    names_to = "Year",
    values_to = "Count"
  ) %>%
  mutate(
    Year = gsub("[^0-9]", "", Year),
    Year = as.numeric(Year),
    Count = as.numeric(str_replace_all(Count, "[^0-9\\.]", ""))
  )

head(crime_tidy)
```

Offence <chr>	State <chr>	Year <dbl>	Count <dbl>
Homicide and related offences(h)	New South Wales	1993	206
Homicide and related offences(h)	New South Wales	1994	178
Homicide and related offences(h)	New South Wales	1995	167
Homicide and related offences(h)	New South Wales	1996	198
Homicide and related offences(h)	New South Wales	1997	221
Homicide and related offences(h)	New South Wales	1998	239

6 rows

Hide

```
crime_total <- crime_tidy %>% group_by(State, Year) %>% summarise(Total_Crimes = sum(Count, n
a.rm = TRUE))
```

```
`summarise()` has grouped output by 'State'. You can override using the `.groups` argument.
```

Hide

```
head(crime_total)
```

State <chr>	Year <dbl>	Total_Crimes <dbl>
Australian Capital Territory	1993	7703
Australian Capital Territory	1994	6974
Australian Capital Territory	1995	22027
Australian Capital Territory	1996	23335
Australian Capital Territory	1997	21965
Australian Capital Territory	1998	26601

6 rows

Hide

```
NA
```

Tidy & Manipulate Data II

- First thing I did here was to create a new column called State. This is because the column 1993 contains the states which were formatted horizontally initially. The code will look for the corresponding state name, and if it finds it in 1993 column, it will assign it to the corresponding State column. Creating a new column allows us to remove these horizontally formatted states, thus then allowing us to delete it afterwards. When we inspect crime, we can see that the new column is present and filled, but the state names in 1993 column are still present, but this was removed right after.
- A new variable of Crime Severity was created. It was simultaneously turned into factor form. It represents crime rate severity in a state in a particular year. It is ordered afterwards.

Hide

```
crime <- crime %>%
  mutate(State = case_when(
    str_detect(`1993`, "New South Wales") ~ "New South Wales",
    str_detect(`1993`, "Victoria") ~ "Victoria",
    str_detect(`1993`, "Queensland") ~ "Queensland",
    str_detect(`1993`, "South Australia") ~ "South Australia",
    str_detect(`1993`, "Western Australia") ~ "Western Australia",
    str_detect(`1993`, "Tasmania") ~ "Tasmania",
    str_detect(`1993`, "Northern Territory") ~ "Northern Territory",
    str_detect(`1993`, "Australian Capital Territory") ~ "Australian Capital Territory",
    TRUE ~ NA_character_
  )) %>%
  fill(State, .direction = "down")

str(crime)
```

```
'data.frame':   136 obs. of  33 variables:
 $ Offence   : chr  NA "Homicide and related offences(h)" "Murder " "Attempted murder" ...
 $ 1993      : chr  "New South Wales" "206" "118" "85" ...
 $ 1994      : chr  NA "178" "108" "60" ...
 $ 1995      : chr  NA "167" "103" "57" ...
 $ 1996      : chr  NA "198" "95" "85" ...
 $ 1997      : chr  NA "221" "110" "100" ...
 $ 1998      : chr  NA "239" "93" "121" ...
 $ 1999(a)   : chr  NA "267" "124" "132" ...
 $ 2000      : chr  NA "262" "98" "149" ...
 $ 2001      : chr  NA "313" "103" "207" ...
 $ 2002      : chr  NA "256" "96" "144" ...
 $ 2003      : chr  NA "233" "100" "120" ...
 $ 2004      : chr  NA "149" "75" "73" ...
 $ 2005      : chr  NA "153" "86" "60" ...
 $ 2006      : chr  NA "176" "105" "69" ...
 $ 2007(b)   : chr  NA "162" "95" "61" ...
 $ 2008(c)   : chr  NA "153" "80" "66" ...
 $ 2009(d)(e): chr  NA "143" "84" "48" ...
 $ 2010(f)(g): chr  NA "133" "73" "45" ...
 $ 2011      : chr  NA "153" "82" "59" ...
 $ 2012      : chr  NA "107" "61" "35" ...
 $ 2013      : chr  NA "131" "85" "48" ...
 $ 2014      : chr  NA "107" "75" "30" ...
 $ 2015      : chr  NA "104" "67" "32" ...
 $ 2016      : chr  NA "100" "65" "27" ...
 $ 2017      : chr  NA "77" "49" "15" ...
 $ 2018      : chr  NA "102" "70" "22" ...
 $ 2019      : chr  NA "116" "76" "26" ...
 $ 2020      : chr  NA "99" "68" "27" ...
 $ 2021      : chr  NA "81" "55" "28" ...
 $ 2022      : chr  NA "79" "59" "10" ...
 $ 2023      : chr  NA "79" "56" "14" ...
 $ State     : chr  "New South Wales" "New South Wales" "New South Wales" "New South Wales"
...
```

Hide

```
filtered_data_real <- filtered_data_real %>%
  mutate(Crime_Severity = case_when(
    Total_Crimes < 10000 ~ "Low",
    Total_Crimes >= 10000 & Total_Crimes < 25000 ~ "Moderate",
    Total_Crimes >= 25000 ~ "High"
  ))

filtered_data_real <- filtered_data_real %>%
  mutate(Crime_Severity = factor(Crime_Severity, levels = c("Low", "Moderate", "High"), order
ed = TRUE))

str(filtered_data_real)
```

```
gropd_df [72 × 5] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ State                              : Factor w/ 8 levels "Australian Capital Territor
y",..: 1 1 1 1 1 1 1 1 2 ...
 $ Year                               : num [1:72] 2013 2014 2015 2016 2017 ...
 $ Total_Crimes                       : num [1:72] 16605 16433 19287 17964 19229 ...
 $ Proportion_Non_School_Qualifications: num [1:72] 63.6 66.4 69.3 69.1 66 69.9 67.3 66.5 68.
6 56.7 ...
 $ Crime_Severity                     : Ord.factor w/ 3 levels "Low"<"Moderate"<..: 2 2 2 2
2 2 2 2 2 3 ...
 - attr(*, "groups")= tibble [8 × 2] (S3: tbl_df/tbl/data.frame)
  ..$ State: Factor w/ 8 levels "Australian Capital Territory",..: 1 2 3 4 5 6 7 8
  ..$ .rows: list<int> [1:8]
  .. ..$ : int [1:9] 1 2 3 4 5 6 7 8 9
  .. ..$ : int [1:9] 10 11 12 13 14 15 16 17 18
  .. ..$ : int [1:9] 19 20 21 22 23 24 25 26 27
  .. ..$ : int [1:9] 28 29 30 31 32 33 34 35 36
  .. ..$ : int [1:9] 37 38 39 40 41 42 43 44 45
  .. ..$ : int [1:9] 46 47 48 49 50 51 52 53 54
  .. ..$ : int [1:9] 55 56 57 58 59 60 61 62 63
  .. ..$ : int [1:9] 64 65 66 67 68 69 70 71 72
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

Scan I

- There are special values called ‘np’. It stands for not published or not public and is just not ideal to have in a dataset. It can either be turned into N/A or 0. I chose to turn it into N/A.
- I check the columns and the sum of missing values (N/A). It returns with 176 missing values for the proportion of non-school qualifications. This makes sense because there are 6 states and 2 territories in Australia that are missing 22 values each because crime dataset has 30 years range, whilst education has 8 years range. I extrapolate the missing data using linear regression model based on the relationship between year and proportion. It seemed appropriate use linear regression because it does not make many assumptions about the data, and is generally useful for extrapolating missing values over time, which in this case is true.
- By checking the str(merged data) before and after, it can be seen that there are no more N/A entries in the Proportion_Non_School_Qualifications column.
- I also check if there are negative values which prove illogical to be in the dataset.

Hide

```
crime_clean <- crime_clean %>% mutate(across(`1993`:`2023`, ~ na_if(., "np")))

head(crime_clean)
```

	Offence <chr>	**1…** <chr>	**1…** <chr>	**1995** <chr>	**1996** <chr>	**1997** <chr>	**1998** <chr>	**1999(a)** <chr>	**2000** <chr>
2	Homicide and related offences(h)	206	178	167	198	221	239	267	262
3	Murder	118	108	103	95	110	93	124	98

	Offence <chr>	1… <chr>	1… <chr>	1995 <chr>	1996 <chr>	1997 <chr>	1998 <chr>	1999(a) <chr>	2000 <chr>
4	Attempted murder	85	60	57	85	100	121	132	149
5	Manslaughter	7	7	8	15	11	26	14	12
6	Assault(i)	*NA*	*NA*	37863	47828	55995	59219	63813	68714
7	Sexual assault	3794	4611	4159	5038	4660	4503	4427	5975

6 rows | 1-10 of 33 columns

Hide

```
missing_values <- colSums(is.na(merged_data))
print(missing_values)
```

```
                                State                                 Year
Total_Crimes Proportion_Non_School_Qualifications
                                    0                                    0
0                                   176
```

Hide

```
str(merged_data)
```

```
gropd_df [248 × 4] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ State                              : Factor w/ 8 levels "Australian Capital Territor
y",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year                               : Factor w/ 31 levels "1993","1994",..: 1 2 3 4 5 6 7
8 9 10 ...
 $ Total_Crimes                       : num [1:248] 7703 6974 22027 23335 21965 ...
 $ Proportion_Non_School_Qualifications: num [1:248] NA NA NA NA NA NA NA NA NA NA ...
 - attr(*, "groups")= tibble [8 × 2] (S3: tbl_df/tbl/data.frame)
  ..$ State: Factor w/ 8 levels "Australian Capital Territory",..: 1 2 3 4 5 6 7 8
  ..$ .rows: list<int> [1:8]
  .. ..$ : int [1:31] 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ : int [1:31] 32 33 34 35 36 37 38 39 40 41 ...
  .. ..$ : int [1:31] 63 64 65 66 67 68 69 70 71 72 ...
  .. ..$ : int [1:31] 94 95 96 97 98 99 100 101 102 103 ...
  .. ..$ : int [1:31] 125 126 127 128 129 130 131 132 133 134 ...
  .. ..$ : int [1:31] 156 157 158 159 160 161 162 163 164 165 ...
  .. ..$ : int [1:31] 187 188 189 190 191 192 193 194 195 196 ...
  .. ..$ : int [1:31] 218 219 220 221 222 223 224 225 226 227 ...
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

Hide

```
merged_data <- merged_data %>%
  group_by(State) %>%
  mutate(Proportion_Non_School_Qualifications = ifelse(is.na(Proportion_Non_School_Qualificat
ions),
  predict(lm(Proportion_Non_School_Qualifications ~ as.numeric(Year), data = .), newdata =
.),
  Proportion_Non_School_Qualifications)) %>%
  ungroup()

str(merged_data)
```

```
tibble [248 × 4] (S3: tbl_df/tbl/data.frame)
 $ State                              : Factor w/ 8 levels "Australian Capital Territor
y",..: 1 1 1 1 1 1 1 1 1 ...
 $ Year                               : Factor w/ 31 levels "1993","1994",..: 1 2 3 4 5 6 7
8 9 10 ...
 $ Total_Crimes                       : num [1:248] 7703 6974 22027 23335 21965 ...
 $ Proportion_Non_School_Qualifications: num [1:248] 42.8 43.5 44.2 44.9 45.6 ...
```

Hide

```
negative_values <- merged_data %>%
  select(Total_Crimes, Proportion_Non_School_Qualifications) %>%
  summarise_all(~ sum(. < 0, na.rm = TRUE))

print(negative_values)
```

Total_Crimes <int>	Proportion_Non_School_Qualifications <int>
0	0

1 row

Hide

```
NA
```

Scan II

- I use code to define Q1, Q3, IQR, lower and upper bound to begin scanning for outliers. When we inspect merged data, It results in merged data having 8 outliers, all resulting from New South Wales and all have greater than 58,000 crimes in a year.
- When we use z score to determine if there are outliers, it determines that there are 6 outliers in crime count compared to the 8 from IQR methodology. Both IQR and z score methodology shows 0 outliers for proportion.
- Using both forms of methodologies proves beneficial. IQR is useful for detecting outliers based on spread, whereas z score detects outliers based on standard deviation from the mean.

Hide

```
detect_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  return(x < lower_bound | x > upper_bound)
}

merged_data <- merged_data %>%
  mutate(
    Total_Crimes_Outliers = detect_outliers(Total_Crimes),
    Proportion_Outliers = detect_outliers(Proportion_Non_School_Qualifications)
  )

total_crimes_outliers <- sum(merged_data$Total_Crimes_Outliers, na.rm = TRUE)
proportion_outliers <- sum(merged_data$Proportion_Outliers, na.rm = TRUE)

print(total_crimes_outliers)
```

```
[1] 8
```

Hide

```
print(proportion_outliers)
```

```
[1] 0
```

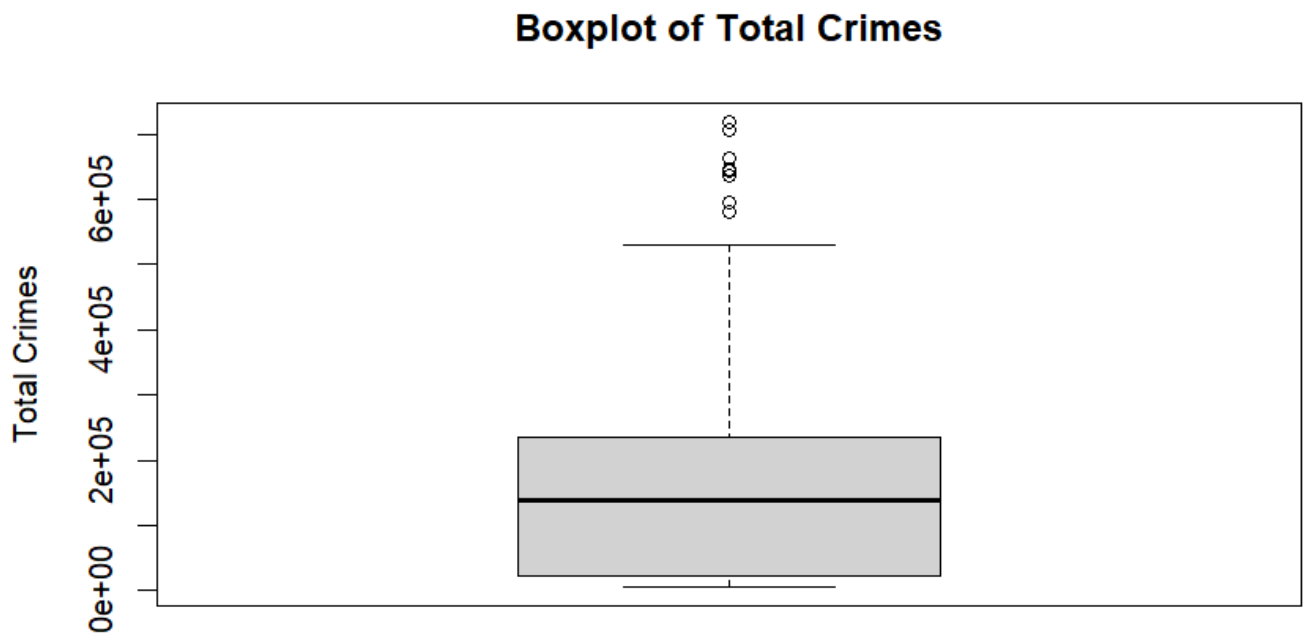
Hide

```
str(merged_data)
```

```
tibble [248 × 10] (S3: tbl_df/tbl/data.frame)
 $ State                              : Factor w/ 8 levels "Australian Capital Territor
y",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year                               : Factor w/ 31 levels "1993","1994",..: 1 2 3 4 5 6 7
8 9 10 ...
 $ Total_Crimes                       : num [1:248] 7703 6974 22027 23335 21965 ...
 $ Proportion_Non_School_Qualifications: num [1:248] 42.8 43.5 44.2 44.9 45.6 ...
 $ Total_Crimes_Outliers              : logi [1:248] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Proportion_Outliers                : logi [1:248] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Total_Crimes_Z_Outliers            : logi [1:248] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Proportion_Z_Outliers              : logi [1:248] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Total_Crimes_Log                   : num [1:248] 8.95 8.85 10 10.06 10 ...
 $ Total_Crimes_Sqrt                  : num [1:248] 87.8 83.5 148.4 152.8 148.2 ...
```

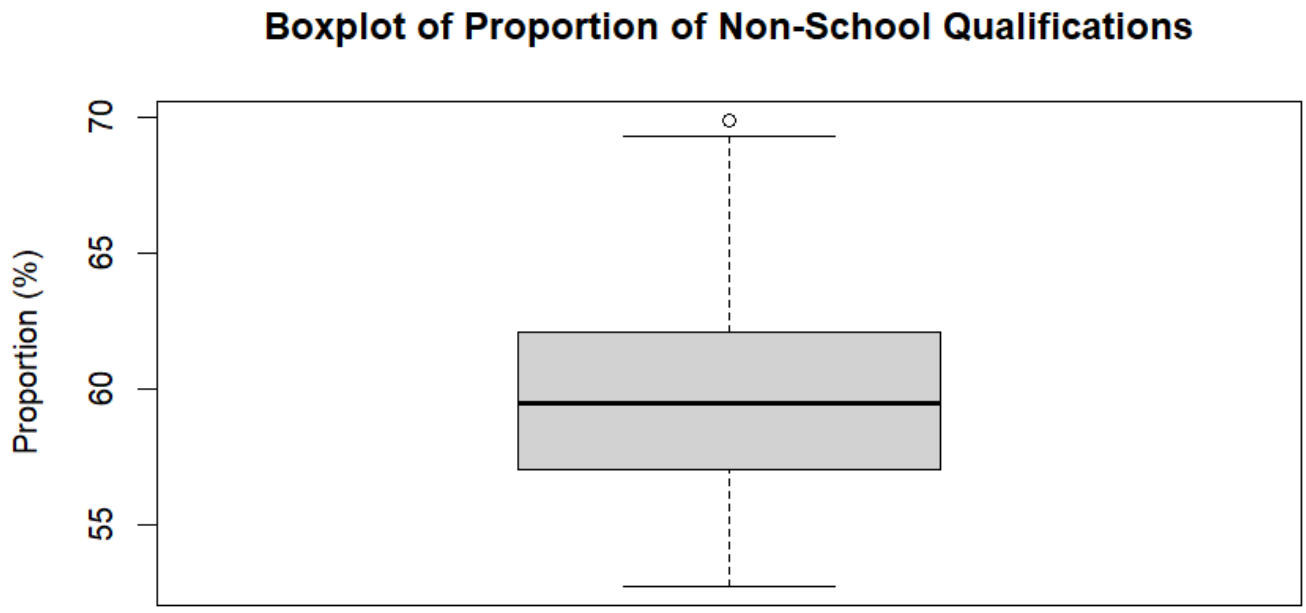
Hide

```
boxplot(merged_data$Total_Crimes, main = "Boxplot of Total Crimes", ylab = "Total Crimes")
```



Hide

```
boxplot(merged_data_real$Proportion_Non_School_Qualifications, main = "Boxplot of Proportion
of Non-School Qualifications", ylab = "Proportion (%)")
```



Hide

```
detect_zscore_outliers <- function(x) {
  z_scores <- (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
  return(abs(z_scores) > 3)
}

merged_data <- merged_data %>%
  mutate(
    Total_Crimes_Z_Outliers = detect_zscore_outliers(Total_Crimes),
    Proportion_Z_Outliers = detect_zscore_outliers(Proportion_Non_School_Qualifications)
  )

summary(merged_data)
```

```
                         State         Year      Total_Crimes    Proportion_Non_School_Quali
fications Total_Crimes_Outliers Proportion_Outliers Total_Crimes_Z_Outliers Proportion_Z_Outl
iers
 Australian Capital Territory:31   1993   :  8   Min.   :  4771   Min.   :42.80
Mode :logical         Mode :logical       Mode :logical           Mode :logical
 New South Wales             :31   1994   :  8   1st Qu.: 23450   1st Qu.:47.74
FALSE:240             FALSE:248           FALSE:242               FALSE:248
 Northern Territory          :31   1995   :  8   Median :139539   Median :53.14
TRUE :8                                   TRUE :6
 Queensland                  :31   1996   :  8   Mean   :158779   Mean   :53.37
 South Australia             :31   1997   :  8   3rd Qu.:234998   3rd Qu.:57.83
 Tasmania                    :31   1998   :  8   Max.   :719298   Max.   :69.90
 (Other)                     :62   (Other):200
 Total_Crimes_Log Total_Crimes_Sqrt
 Min.   : 8.471   Min.   : 69.07
 1st Qu.:10.063   1st Qu.:153.13
 Median :11.846   Median :373.55
 Mean   :11.395   Mean   :351.93
 3rd Qu.:12.367   3rd Qu.:484.77
 Max.   :13.486   Max.   :848.11
```

Transform

- I applied 2 different transformations to see which one suits the dataset more. I use log transformation and square root transformation. Both are great for handling right skew, which is what we have in this dataset. The benefit of square root transformation is that it can handle values of 0. However, log transformation looks best on this dataset. I show the original histogram, as well as the transformed one to compare.

Hide

```
summary(merged_data$Total_Crimes)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4771   23450  139539  158779  234998  719298
```

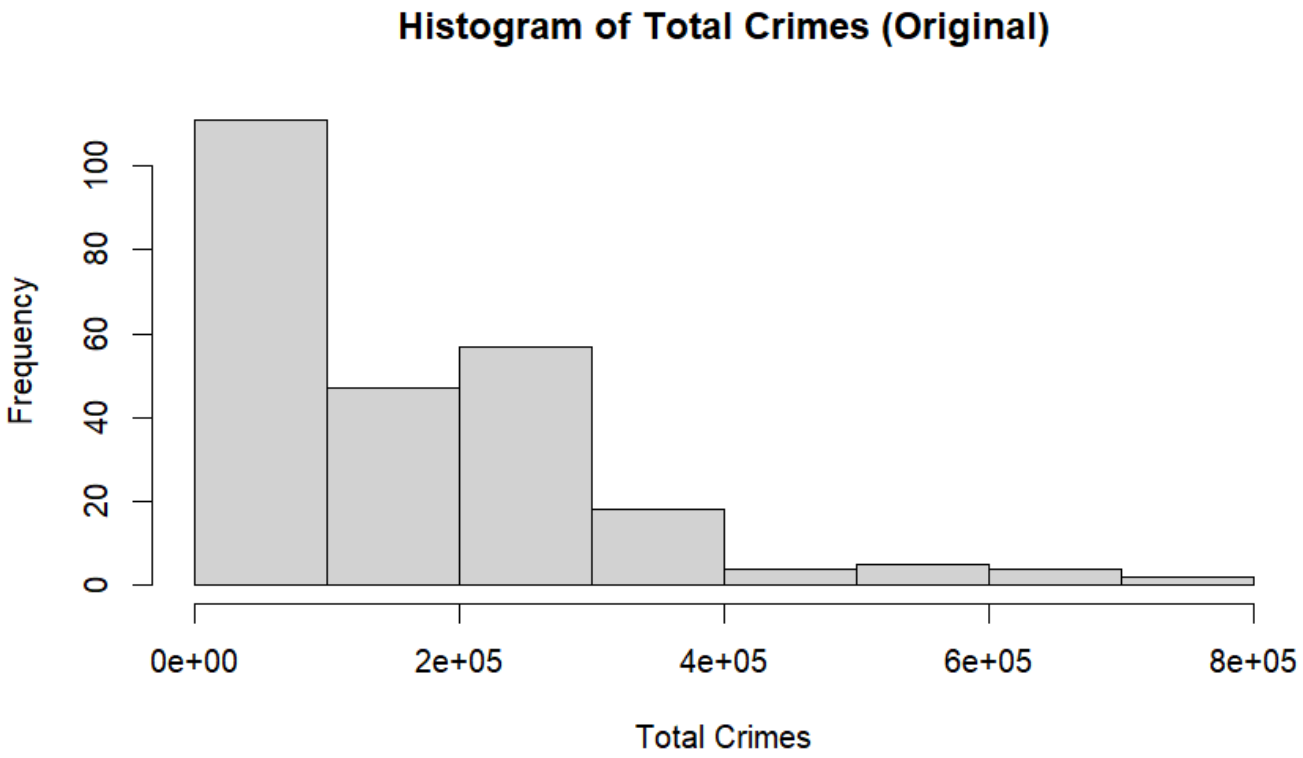
Hide

```
summary(merged_data$Proportion_Non_School_Qualifications)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  42.80   47.74   53.14   53.37   57.83   69.90 
```

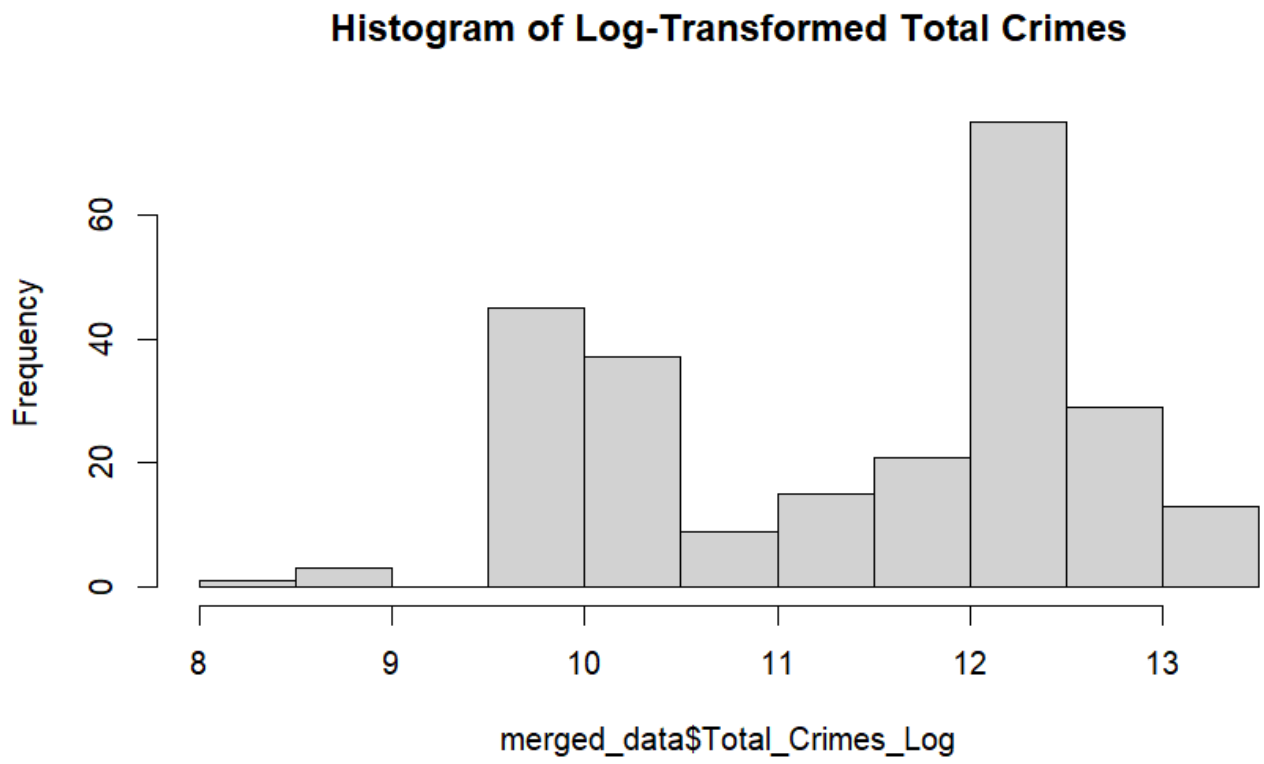
Hide

```
hist(merged_data$Total_Crimes, main = "Histogram of Total Crimes (Original)", xlab = "Total C
rimes", ylab = "Frequency")
```



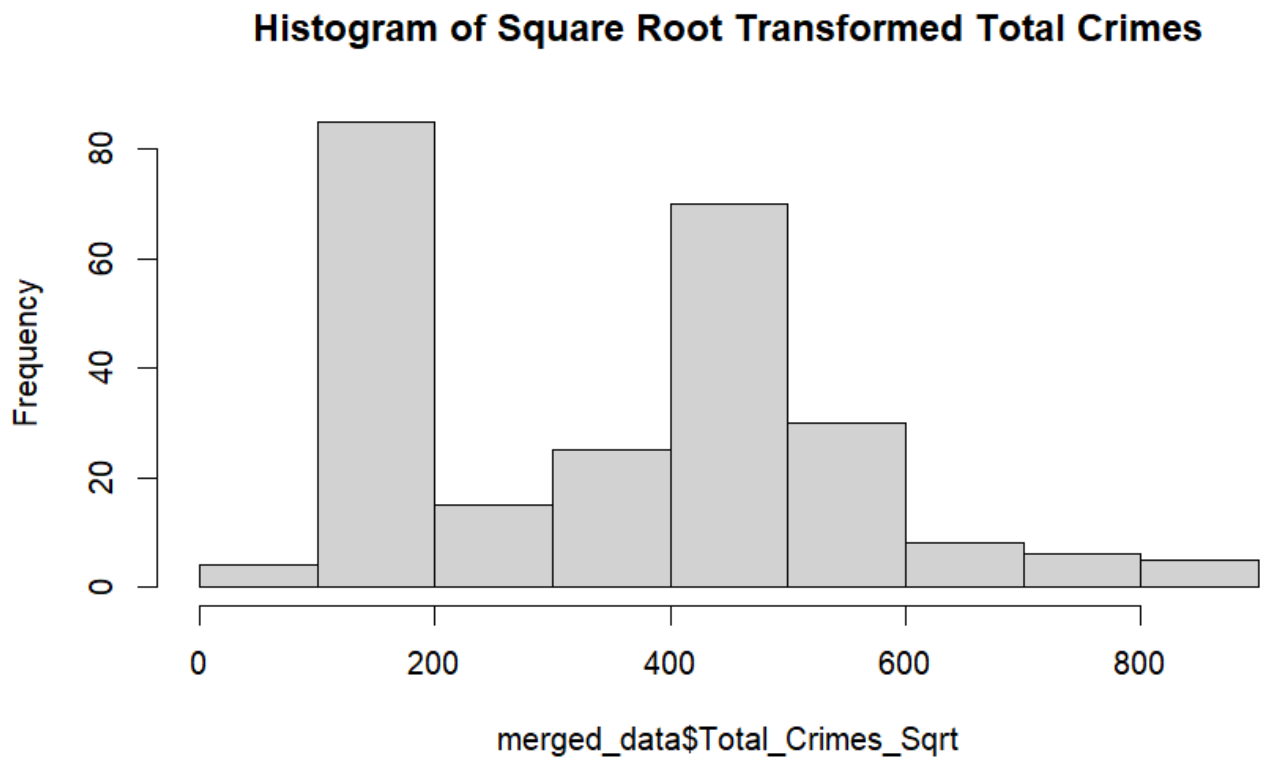
Hide

```
merged_data <- merged_data %>%
  mutate(Total_Crimes_Log = log(Total_Crimes + 1))

hist(merged_data$Total_Crimes_Log, main = "Histogram of Log-Transformed Total Crimes")
```



Hide

```
merged_data <- merged_data %>%
  mutate(Total_Crimes_Sqrt = sqrt(Total_Crimes))

hist(merged_data$Total_Crimes_Sqrt, main = "Histogram of Square Root Transformed Total Crime
s")
```



Hide

```
NA
NA
```

Correlation

Hide

```
correlation_data_real <- filtered_data_real %>%
  select(Total_Crimes, Proportion_Non_School_Qualifications)
```

```
Adding missing grouping variables: `State`
```

Hide

```
correlation_data_real <- na.omit(correlation_data_real)

correlation_result_real <- cor(
  correlation_data_real$Total_Crimes,
  correlation_data_real$Proportion_Non_School_Qualifications,
  method = "pearson"
)

cat("Pearson Correlation for filtered_data_real:", correlation_result_real, "\n")
```

```
Pearson Correlation for filtered_data_real: -0.1315209
```

Hide

```
correlation_data <- merged_data %>%
  select(Total_Crimes, Proportion_Non_School_Qualifications)

correlation_data <- na.omit(correlation_data)

correlation_result <- cor(
  correlation_data$Total_Crimes,
  correlation_data$Proportion_Non_School_Qualifications,
  method = "pearson"
)

cat("Pearson Correlation for merged_data:", correlation_result, "\n")
```

```
Pearson Correlation for merged_data: -0.1123362
```

Presentation

Presentation (https://rmit-arc.instructuremedia.com/embed/e173ff81-b664-44e9-8dea-def7c2dc4386)

References

Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2024). *rmarkdown: Dynamic Documents for R*. R package version 2.27, https://github.com/rstudio/rmarkdown (https://github.com/rstudio/rmarkdown).

Xie Y, Allaire J, Grolemund G (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338, https://bookdown.org/yihui/rmarkdown (https://bookdown.org/yihui/rmarkdown).

Xie Y, Dervieux C, Riederer E (2020). *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9780367563837, https://bookdown.org/yihui/rmarkdown-cookbook (https://bookdown.org/yihui/rmarkdown-cookbook).

Bache S, Wickham H (2022). *magrittr: A Forward-Pipe Operator for R*. R package version 2.0.3, https://CRAN.R-project.org/package=magrittr (https://CRAN.R-project.org/package=magrittr).

Schauberger P, Walker A (2024). *openxlsx: Read, Write and Edit xlsx Files*. R package version 4.2.6.1, https://CRAN.R-project.org/package=openxlsx (https://CRAN.R-project.org/package=openxlsx).

Wickham H (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1, https://CRAN.R-project.org/package=stringr (https://CRAN.R-project.org/package=stringr).

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, https://CRAN.R-project.org/package=dplyr (https://CRAN.R-project.org/package=dplyr).

Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, https://CRAN.R-project.org/package=tidyr (https://CRAN.R-project.org/package=tidyr).

Zhu H (2024). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.4.0, https://CRAN.R-project.org/package=kableExtra (https://CRAN.R-project.org/package=kableExtra).