# 01_Cleaning_and_ETL

Tamara

2025-06-15

## Introduction

This notebook cleans and merges attendance, volatility, and school Equity Index (EQI) data for attendance risk analysis.

**Goals:** - Predict attendance risk using volatility (attendance SD) and EQI - Compare regions and analyze pre/post-COVID patterns - Assess explanatory power of equity and volatility metrics

---

## 1. Load Required Packages

```
library(readxl)
library(dplyr)
library(writexl)
library(here)
library(readr)
```

## 2. Load raw attendance and volatility data

```
# File paths
attendance_file <- here("data_clean", "Regular-attendance-data-cleaned.xlsx")
volatility_file <- here("data_clean", "data_with_volatility.xlsx")

# Load datasets
attendance_df <- read_excel(attendance_file)
volatility_df <- read_excel(volatility_file)

# Preview the structure of the dataframes
glimpse(attendance_df)
```

```
## Rows: 3,144
## Columns: 6
## $ year             <dbl> 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024,~
## $ term             <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,~
```

```
## $ education_region <chr> "Tai Tokerau", "Tai Tokerau", "Tai Tokerau", "Tai Tok~
## $ category         <chr> "total students", "students attending more than 90per~
## $ count            <dbl> 30315, 13815, 7280, 3954, 5266, NA, NA, NA, 101427, 6~
## $ percent          <dbl> NA, 45.6, 24.0, 13.0, 17.4, 83.9, 7.2, 8.9, NA, 62.8,~
```

```
glimpse(volatility_df)
```

```
## Rows: 14
## Columns: 5
## $ education_region   <chr> "All", "Auckland", "Bay of Plenty, Waiariki", "Cant~
## $ volatility_present <dbl> 2.267119, 2.698955, 2.374588, 1.994094, 2.178783, 2~
## $ avg_present        <dbl> 88.52187, 89.54615, 87.60625, 89.58125, 87.30000, 8~
## $ n_obs              <dbl> 32, 13, 32, 32, 32, 32, 32, 32, 20, 20, 20, 32, 32,~
## $ coef_variation     <dbl> 0.02561083, 0.03014038, 0.02710524, 0.02226017, 0.0~
```

## Merge attendance with volatility data

```
# Merge datasets on education_region
merged_df <- attendance_df %>%
  left_join(volatility_df, by = "education_region")

# Preview merged dataset
head(merged_df)
```

```
## # A tibble: 6 x 10
##    year  term education_region category         count percent volatility_present
##   <dbl> <dbl> <chr>            <chr>            <dbl>   <dbl>              <dbl>
## 1  2024     4 Tai Tokerau      total students   30315    NA                 2.78
## 2  2024     4 Tai Tokerau      students attend~ 13815    45.6               2.78
## 3  2024     4 Tai Tokerau      students attend~  7280    24                 2.78
## 4  2024     4 Tai Tokerau      students attend~  3954    13                 2.78
## 5  2024     4 Tai Tokerau      students attend~  5266    17.4               2.78
## 6  2024     4 Tai Tokerau      present half da~    NA    83.9               2.78
## # i 3 more variables: avg_present <dbl>, n_obs <dbl>, coef_variation <dbl>
```

## Save merged attendance-volatility data

```
output_file <- here("data_clean", "attendance_with_volatility_merged.xlsx")
write_xlsx(merged_df, output_file)
```

## Load and prepare EQI data

```
# File paths for EQI and school directory data
eqi_file <- here("data_raw", "School-EQI-numbers-2023-2025.xlsx")
```

```r
school_directory_file <- here("data_raw", "directory.csv")

# Load EQI data (skip first row for headers)
eqi_df <- read_excel(eqi_file, skip = 1)

# Load school directory data (skip initial rows for headers)
directory_df <- read_csv(school_directory_file, skip = 16)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```r
# Rename columns in EQI data for consistency
eqi_df <- eqi_df %>%
  rename(
    School_ID = `School Number`,
    School_Name = `School Name`,
    EQI_2023 = `2023 - School Equity Index Number`,
    EQI_2024 = `2024 - School Equity Index Number`,
    EQI_2025 = `2025 - School Equity Index Number`
  ) %>%
  mutate(School_Name_clean = trimws(tolower(School_Name)))

# Clean school names in directory data
directory_df <- directory_df %>%
  mutate(School_Name_clean = trimws(tolower(`School Name`)))

# Map regional councils to education regions (hardcoding these)
region_map <- c(
  "Northland Region" = "Tai Tokerau",
  "Auckland Region" = "Tāmaki Makaurau",
  "Bay of Plenty Region" = "Bay of Plenty, Waiariki",
  "Waikato Region" = "Waikato",
  "Manawatū-Whanganui Region" = "Taranaki, Whanganui, Manawatū",
  "Hawke's Bay Region" = "Hawke's Bay, Tairāwhiti",
  "Taranaki Region" = "Taranaki, Whanganui, Manawatū",
  "Whanganui Region" = "Taranaki, Whanganui, Manawatū",
  "Canterbury Region" = "Canterbury, Chatham Islands",
  "Chatham Islands" = "Canterbury, Chatham Islands",
  "Otago Region" = "Otago, Southland",
  "Southland Region" = "Otago, Southland",
  "Gisborne Region" = "Hawke's Bay, Tairāwhiti",
  "Marlborough Region" = "Nelson, Marlborough, West Coast",
  "Tasman Region" = "Nelson, Marlborough, West Coast",
  "Nelson Region" = "Nelson, Marlborough, West Coast",
  "West Coast Region" = "Nelson, Marlborough, West Coast",
  "Wellington Region" = "Wellington"
)

directory_df <- directory_df %>%
  mutate(education_region = region_map[`Regional Council`])
```

```r
# Join EQI data with directory on cleaned school names
merged_eqi_df <- eqi_df %>%
  inner_join(directory_df, by = "School_Name_clean") %>%
  select(School_ID, School_Name, education_region, EQI_2023, EQI_2024, EQI_2025)

# Preview cleaned EQI data
head(merged_eqi_df)
```

```
## # A tibble: 6 x 6
##   School_ID School_Name           education_region EQI_2023 EQI_2024 EQI_2025
##       <dbl> <chr>                 <chr>               <dbl>    <dbl>    <dbl>
## 1         1 Te Kura o Te Kao      Tai Tokerau           521      527      532
## 2         2 Taipa Area School     Tai Tokerau           534      532      532
## 3         3 Kaitaia College       Tai Tokerau           519      519      521
## 4         4 Whangaroa College     Tai Tokerau           539      538      538
## 5         5 Kerikeri High School  Tai Tokerau           457      459      461
## 6         6 Broadwood Area School Tai Tokerau           555      561      562
```

```r
# Save cleaned school-region EQI data
write_xlsx(merged_eqi_df, here("data_clean", "school_region_EQI.xlsx"))
```

## Merge attendance and volatility with regional EQI

```r
# Load merged attendance-volatility dataset

attendance_vol_df <- read_excel(here("data_clean", "attendance_with_volatility_merged.xlsx"))


# Load regional mean EQI data

eqi_region_df <- read_excel(here("data_clean", "eqi_by_region.xlsx"))


# Merge on education_region
final_merged_df <- attendance_vol_df %>%
  left_join(eqi_region_df, by = "education_region")

# Preview final dataset
head(final_merged_df)
```

```
## # A tibble: 6 x 13
##    year  term education_region category         count percent volatility_present
##   <dbl> <dbl> <chr>            <chr>            <dbl>   <dbl>              <dbl>
## 1  2024     4 Tai Tokerau      total students   30315    NA                 2.78
## 2  2024     4 Tai Tokerau      students attend~ 13815    45.6               2.78
## 3  2024     4 Tai Tokerau      students attend~  7280    24                 2.78
## 4  2024     4 Tai Tokerau      students attend~  3954    13                 2.78
## 5  2024     4 Tai Tokerau      students attend~  5266    17.4               2.78
## 6  2024     4 Tai Tokerau      present half da~    NA    83.9               2.78
## # i 6 more variables: avg_present <dbl>, n_obs <dbl>, coef_variation <dbl>,
## #   eqi_mean <dbl>, eqi_median <dbl>, schools_in_region <dbl>
```

```r
# Save final merged dataset for analysis
write_xlsx(final_merged_df, here("data_clean", "merged_attendance_eqi.xlsx"))
```

## Files Created

- **attendance_with_volatility_merged.xlsx** - Attendance + volatility data
- **school_region_EQI.xlsx** - School-level EQI with regions

- **eqi_by_region.xlsx** - Regional EQI averages
- **merged_attendance_eqi.xlsx** - Final dataset for analysis