04_risk_scoring_typology

Tamara

2025-06-14

Loading the data (merged)

```
## Rows: 14
## Columns: 12
## $ education_region   <chr> "Tai Tokerau", "Tamaki Herenga Tangata", "Tamaki He~
## $ volatility_present <dbl> 2.780397, 2.235338, 2.183786, 3.361719, 2.555947, 2~
## $ avg_present        <dbl> 85.16875, 89.19000, 89.60500, 84.37000, 87.40625, 8~
## $ n_obs              <dbl> 32, 20, 20, 20, 32, 32, 32, 32, 32, 32, 32, 32, 32,~
## $ coef_variation     <dbl> 0.03264574, 0.02506265, 0.02437125, 0.03984496, 0.0~
## $ reg_below_70       <dbl> 13.553125, 8.055000, 7.865000, 16.600000, 10.090625~
## $ reg_80_90          <dbl> 26.29688, 22.21500, 20.99500, 23.00500, 24.30625, 2~
## $ reg_90             <dbl> 47.64688, 61.19000, 63.12500, 48.13500, 55.69375, 5~
## $ reg_70_80          <dbl> 12.512500, 8.530000, 8.010000, 12.275000, 9.915625,~
## $ eqi_mean           <dbl> 506.2925, 428.1176, 419.6805, 485.1792, 476.1022, 4~
## $ eqi_median         <dbl> 514, 422, 409, 493, 473, 489, 478, 493, 446, 453, 4~
## $ schools_in_region  <dbl> 148, 189, 172, 175, 275, 189, 231, 174, 281, 123, 2~
```

STEP 1: Standardise predictors

Standardise EQI and Volatility to z-scores

STEP 2: Assigning weighted risk score based on variance explained in 03_modeling_risk_factors (~63% EQI, ~37% volatility)

STEP 3: Quick check of risk score distribution

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -1.21225 -0.83644 -0.10750 -0.01922  0.48716  1.44948        2
```

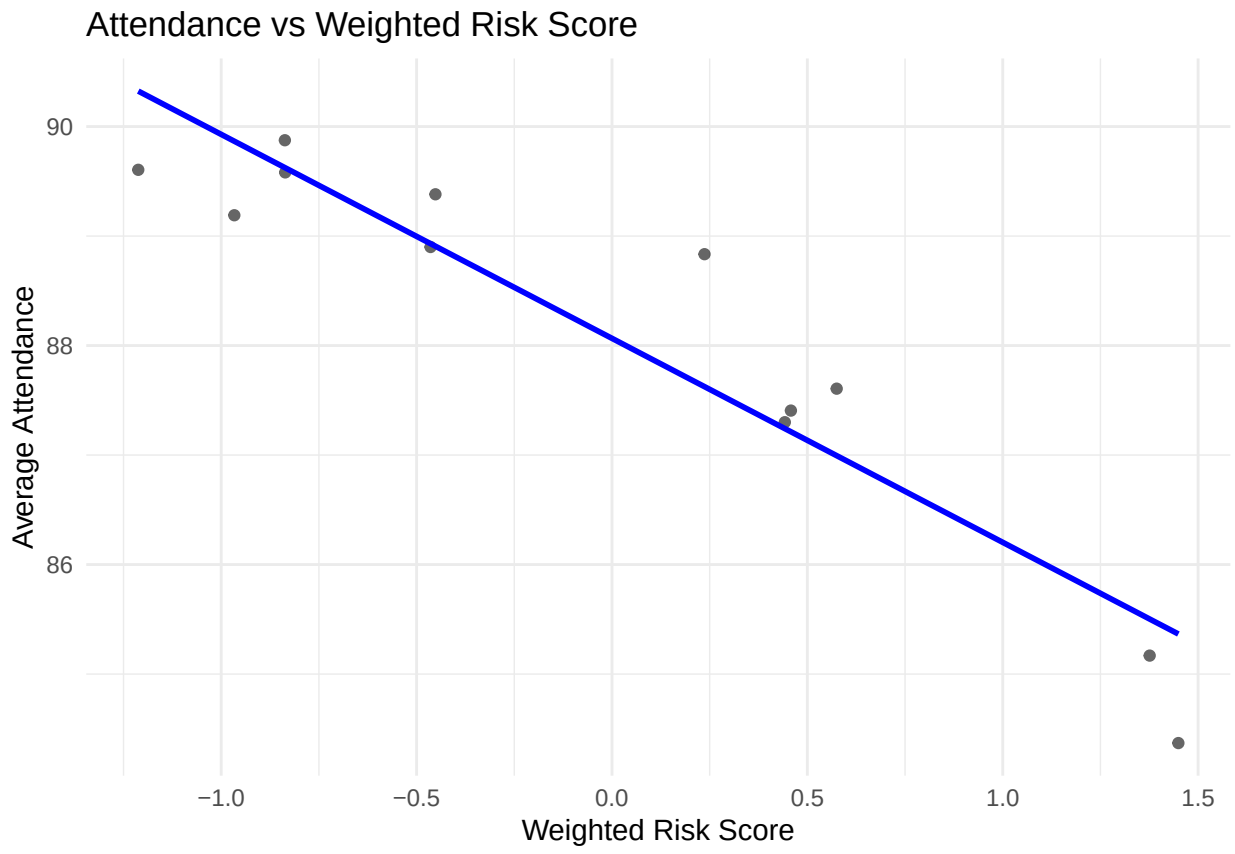
STEP 4: Validating the association between risk score and attendance using a linear model

```
##
## Call:
## lm(formula = avg_present ~ risk_score, data = df_clean)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99584 -0.41937  0.01383  0.30604  1.20967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  88.0657     0.1881 468.130  < 2e-16 ***
## risk_score   -1.8627     0.2171  -8.579 6.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6515 on 10 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8804, Adjusted R-squared:  0.8684
## F-statistic: 73.61 on 1 and 10 DF,  p-value: 6.349e-06
```

STEP 5: Visualise relationship

```
## `geom_smooth()` using formula = 'y ~ x'
```



STEP 6: Building the typology classification.

Thresholds were chosen as ±0.5 standard deviations from the mean, capturing regions meaningfully above or below average. This results in four interpretable categories:

High risk (structural + instability): High EQI disadvantage and high attendance volatility.

Structural risk: Disadvantaged but attendance is relatively stable.

Instability risk: Less disadvantaged, but engagement fluctuates.

Lower risk: Relatively advantaged and stable attendance.

STEP 7: Summary Table for README (NA values are Auckland and All)

```
## # A tibble: 4 x 3
##   risk_typology    mean_attendance count
##   <fct>                      <dbl> <int>
## 1 Structural Risk             89.4     5
## 2 Instability Risk            85.6     3
## 3 Lower Risk                  88.3     4
## 4 <NA>                        89       2
```

STEP 8: Visualization of the typology from step 6