FinalCapstoneProject

Lindsay Lee

2024-06-10

Final Capstone Project: PISA 2022 Predictors of Well-Being

Introduction

After the COVID-19 pandemic, the factors that influence children and adolescent's well-being has been at the collective forefront of society (UNESCO, 2024). For years, there has been a call to action on how social media has influenced children and teenager's perception of themselves and overall well-being (e.g., cyberbullying, privacy violations, body dissatisfaction, social comparison; Granic et al., 2020; Scully et al., 2023; UNESCO, 2023; UNESCO, 2024; Wong, 2024). Adolescence is a critical period of life that centers around identity development (Erikson, 1968) and other psychosocial self-beliefs continue to develop over the course of time within schools. There are several scholars that argue that psychosocial beliefs likely has influences to well-being that can impact learning environments (Nemiro et al., 2022; Tsang et al., 2011). Beyond the influence of social media and technology usage, there is little research on how psychosocial strength factors predict student overall well-being in adolescents (Nemiro et al., 2022), particularly how I are developing their talent (Subotnik, 2015).

Purpose

The Programme for International Student Assessment (PISA) is developed by The Organization for Economic Cooperation and Development (OECD) and assesses students, principals, and parents on a range of topics (e.g., well-being, creative thinking, financial literacy, ICT [Information, Computers, & Technology]) involved with the education of students internationally (OECD, 2024). For the purpose of this report, I will analyze student demographics (i.e., grade level gender, parental education, and country of origin) and psychosocial variables (i.e., responses on self-report items of belongingness, curiosity, perseverance, stress resistance, problems with self-directed learning, and family support) to predict scores on well-being.

The following questions guided my inquiry: 1. To what extent does student demographics (i.e., grade level, gender, parental education, country of origin) self-beliefs of belongingness, perseverance, curiosity, stress resistance, family support, and problems with self-directed learning predict well-being in teenagers? 2. Of the specific psychosocial and demographic features included, what are the top predictors that highly predict well-being in adolescents?

Methods/Analysis

RStudio was used to conduct analyses (R Core Team, 2023). The following analyses used the PISA 2022 dataset that is freely available online (OECD, 2024). Analyses used both a hierarchical linear regression approach and a random forest algorithm to assess the relationship and predictiveness of the included features. Demographic factor type features include grade level, gender, parental education, country of origin. Psychosocial numeric type features include belongingness, curiosity, perseverance, stress resistance, problems with self-directed learning, and well-being.

Composite variables were already computed and converted to Z-Score metric. Prior to analyses, data cleaning used dplyr (Wickham et al., 2014) and tidyr (Wickem et al., 2024) to rename variables, create an Age variable based on substracting student's birth year from 2022 (testing year), remove erronuous levels (e.g., 96 from International Grade levels) and unneeded variables from my pull of the dataset (i.e., BirthYear, STRATUM). I also converted data from integers and logical vectors to factors (demographics) and numeric (features) variable types to be able to run analyses for my research questions.The countries included in this report are Brazil, Spain, France, Hong Kong (China), Hungry, Ireland, Macao (China), Netherlands, New Zealand, and Slovenia. Parental Education was based on the ISCED indicators used by UNESCO/PISA documentation (see OECD, 2023). All other countries were listwise deleted with na.omit() because they did not include all variables of interest.

*Note - I initially used the haven package () to import spss (.sav) file, but looked for alternatives to create a relative path for this report. I tried to webscrape the dataset off of the PISA website, however webscraping encountered a HTTP 403 error and could not access website even when specifying user agent for my browser. I looked to host the dataset on my github, but the file was too large. I asked ChatGPT the following prompt: "How to automatically download condensedStudentQQQ.xlsx from https://github.com/lindsayelee/Data-Science-Capstone-Project in R". See R Script for output of code that ChatGPT provide to help me set a relative path to automatically download a version of my dataset on my Github repository. This was the only instance of using Open AI ChatGPT.

Regression: Hierarchical Multiple Regression & Random Forest Machine Learning

To assess the first research question, I conducted a series of multiple regressions to assess demographic background to well-being, then the psychosocial variables to well-being. In a final regression model, I included all demographic and psychosocial variables to predict well-being. To assess the predictive power of specific psychosocial and demographic variables on well-being (and specifically the top features), I used a random forest model to assess the predictability of the features to well-being and the importance of all included features. Random forest models are a machine learning algorithm used for classification or regression tasks by creating decision trees based on your training set (Breiman, 2001) Random forest models are useful for continuous outcomes and can use both continuous and categorical features (Ibizarry, 2024). Initially, I used the caret package (Kuhn et al., 2023) and the explored how to tune my model with train() and help determine the mtry value to use for my random forest model that includes all features. To evaluate, Root Mean Square Error (RMSE), MAE, and R-squared was used as performance metrics. Specifically, I sought to find a model with an RMSE lower than 1 (.75 < RMSE <= 1) and Mean Absolute Error (MAE), and highest R-squared values.

Results

Initial descriptive statistics and data visualizations were explored using basic plots (e.g., the ggplot package; Wickham, 2016) to understand the summary statistics and distributions of the dataset in relation to well-being scores.

Demographic Background

Age & Gender

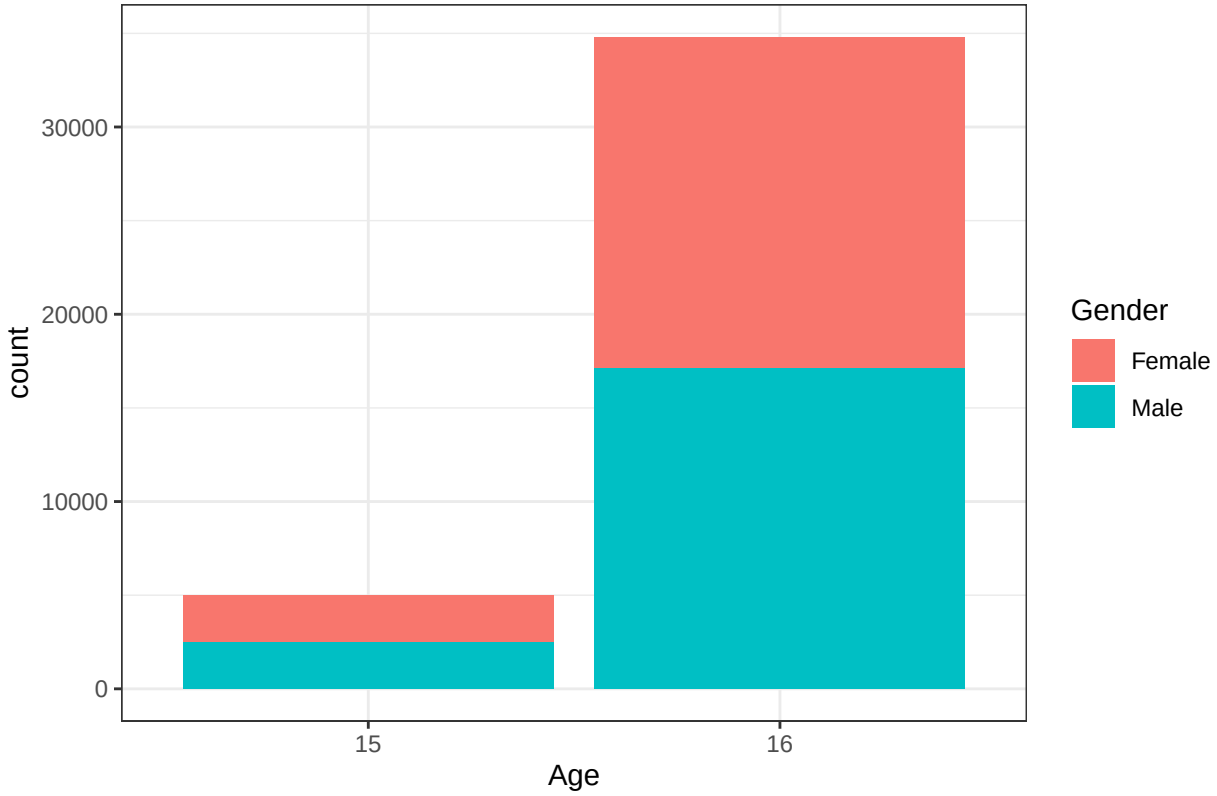
This sample was comprised of students aged 15 & 16 years old, with more students being 16 years old (87.37%) than 15 years old (12.6%). There was also a fairly even split between males (n = 19564; 49.11%) and females (n = 20277;50.89%). The average well-being for 16 year old male students (M = .06, SD = .96) was lower than 15 year old male students (M = .13, SD = .99), however overall both were above the standard average score. However, females overall reported below mean average scores on well-being. The average well-being for 16 year old female students (M = -.21, SD = 1.05) was lower than the 15 year old female students (M = -.07, SD = 1.04). The plots show how there are more 16 year olds overall than 15 year

olds in the dataset. Also, there are several boys and girls who are scoring below the average for well-being. See Table 1 & Figures 1-2.

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

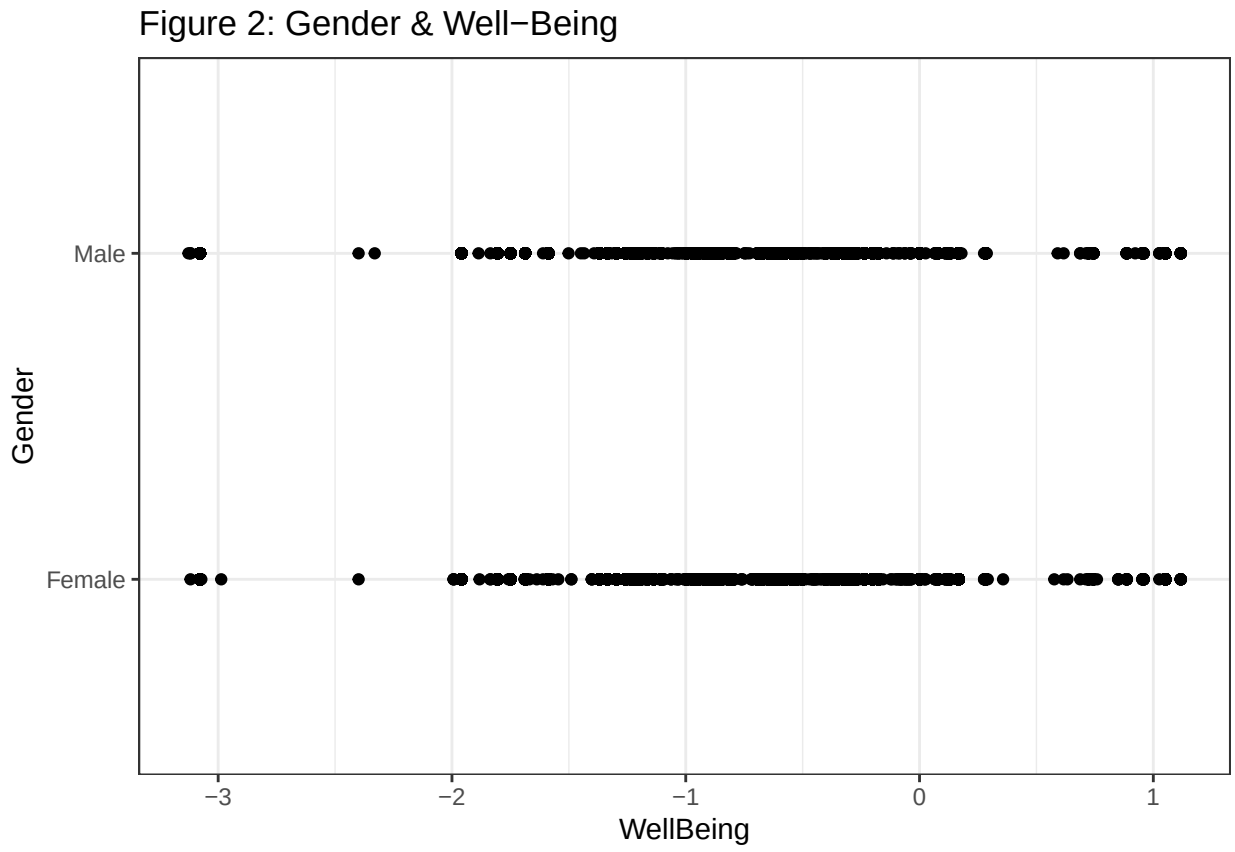
Age	Gender	count	meanWellBeing	sdWellBeing
15	Female	2561	-0.0719860	1.0488119
15	Male	2468	0.1327384	0.9918951
16	Female	17716	-0.2130131	1.0497307
16	Male	17096	0.0591945	0.9578991

Figure 1: Gender & Age




```
##
##        15        16
## 0.1262268 0.8737732
```

```
##
##    Female      Male
## 0.5089481 0.4910519
```



Grade

For grade level, this was tricky to interpret. Across the world, there are different distinctions for grade level. However, the majority of students are in the 10th grade (n = 27673) which is similar to the schooling in the United States and is in line with the 15 to 16 year olds included in the sample. Interestingly, the mean well-being score was highest for students in the 7th grade (M = .11, SD = 1.08) and the 11th grade (M = .01, SD = 1.04), but lowest for students in the 10th grade (M = -.09, SD = 1.01). However, according to Figure 4, you can see that the majority of the grades are scoring below average on the Well Being measure.

Grade	count	meanWellBeing	sdWellBeing
10	27673	-0.0918857	1.008225
9	7168	-0.0141697	1.025951
11	4187	0.0170144	1.044474
8	656	-0.0712848	1.041548
12	91	-0.0714769	1.071555
7	66	0.1187742	1.080813

Figure 3: Grade

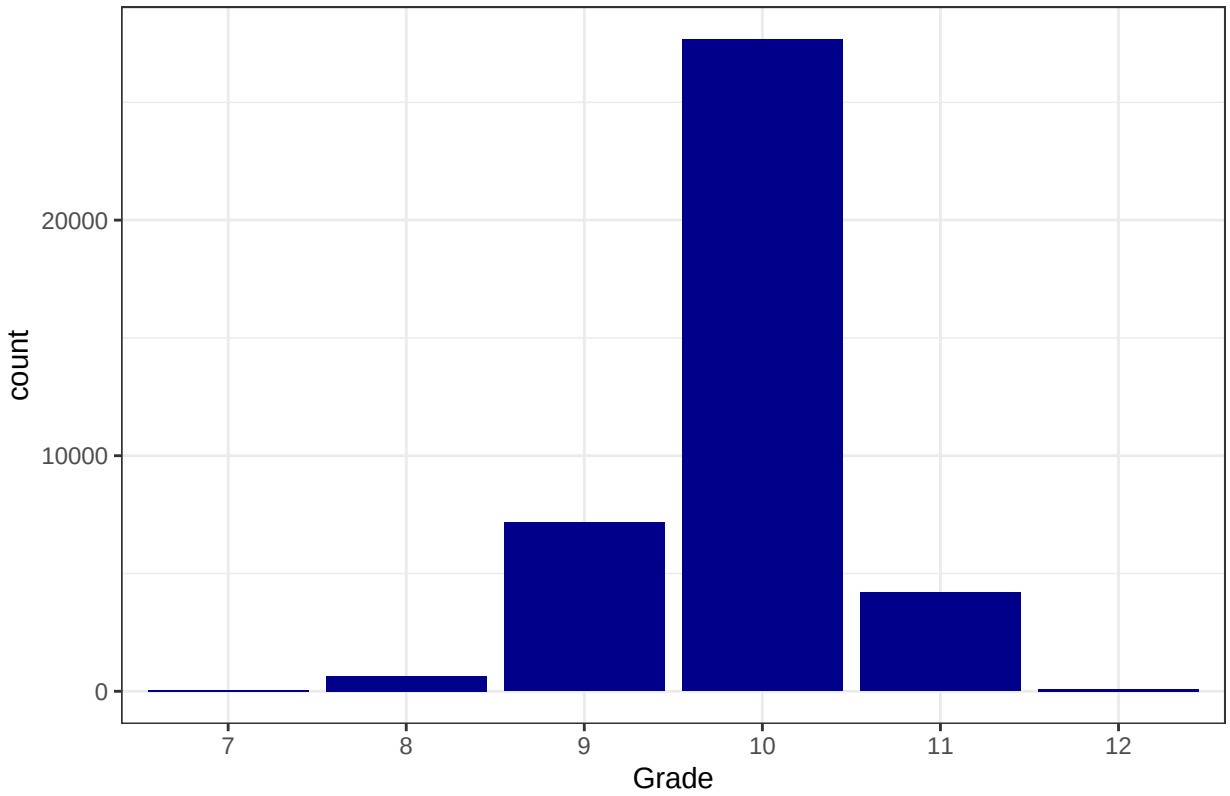
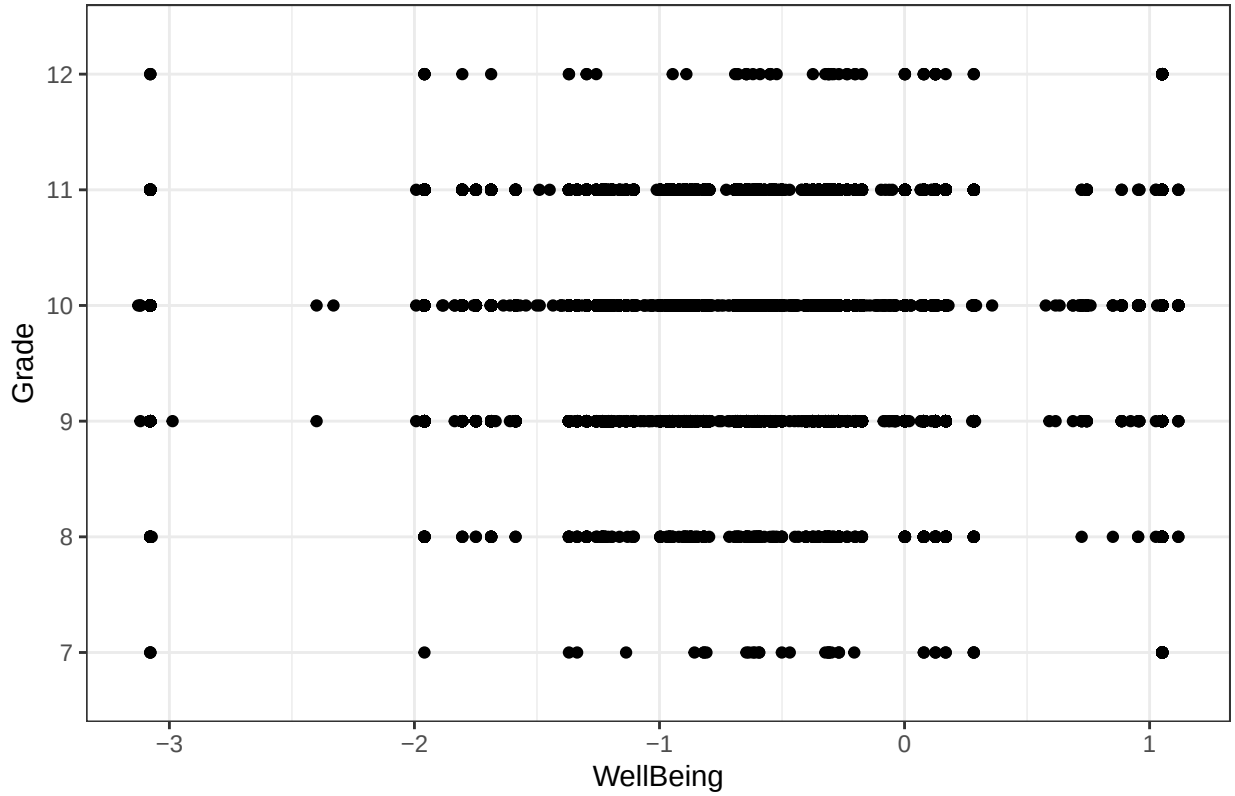


Figure 4: Grade & Well−Being



Parental Education

For parental education, the majority of students in the sample have both parents with a Bachelor's or equivalent (23.65%) or a Master's or equivalent (21.79%). The lowest proportion of students had both parents with no formal education (.06%). See Figure 5. In regard to Well Being, students seem to have lower than average well-being scores regardless of parental education scores.

```
##
##              Bachelor's or equivalent               Doctoral or equivalent
##                             0.236540247                          0.115810346
##                         Lower Secondary               Master's or equivalent
##                             0.072688939                          0.217916217
##                     No Formal Education           Post Secondary non Tertiary
##                             0.006400442                          0.040234934
##                       Primary Education                Two Years of Tertiary
##                             0.017444341                          0.103134961
## Upper Secondary No Access to Tertiary  Upper Secondary w Access to Tertiary
##                             0.044953691                          0.144875882
```

HighestParentalEd	count	meanWellBeing	sdWellBeing
Doctoral or equivalent	4614	-0.0114132	1.0020215
Upper Secondary No Access to Tertiary	1791	-0.0219835	0.9892636
Bachelor's or equivalent	9424	-0.0469990	0.9995827

HighestParentalEd	count	meanWellBeing	sdWellBeing
Two Years of Tertiary	4109	-0.0664917	1.0036594
Upper Secondary w Access to Tertiary	5772	-0.0670828	1.0592292
No Formal Education	255	-0.0860675	1.0563409
Post Secondary non Tertiary	1603	-0.0875082	1.0679325
Master's or equivalent	8682	-0.0880348	0.9939613
Primary Education	695	-0.1122770	1.0572070
Lower Secondary	2896	-0.1445330	1.0642608

Figure 5: Highest Parental Education

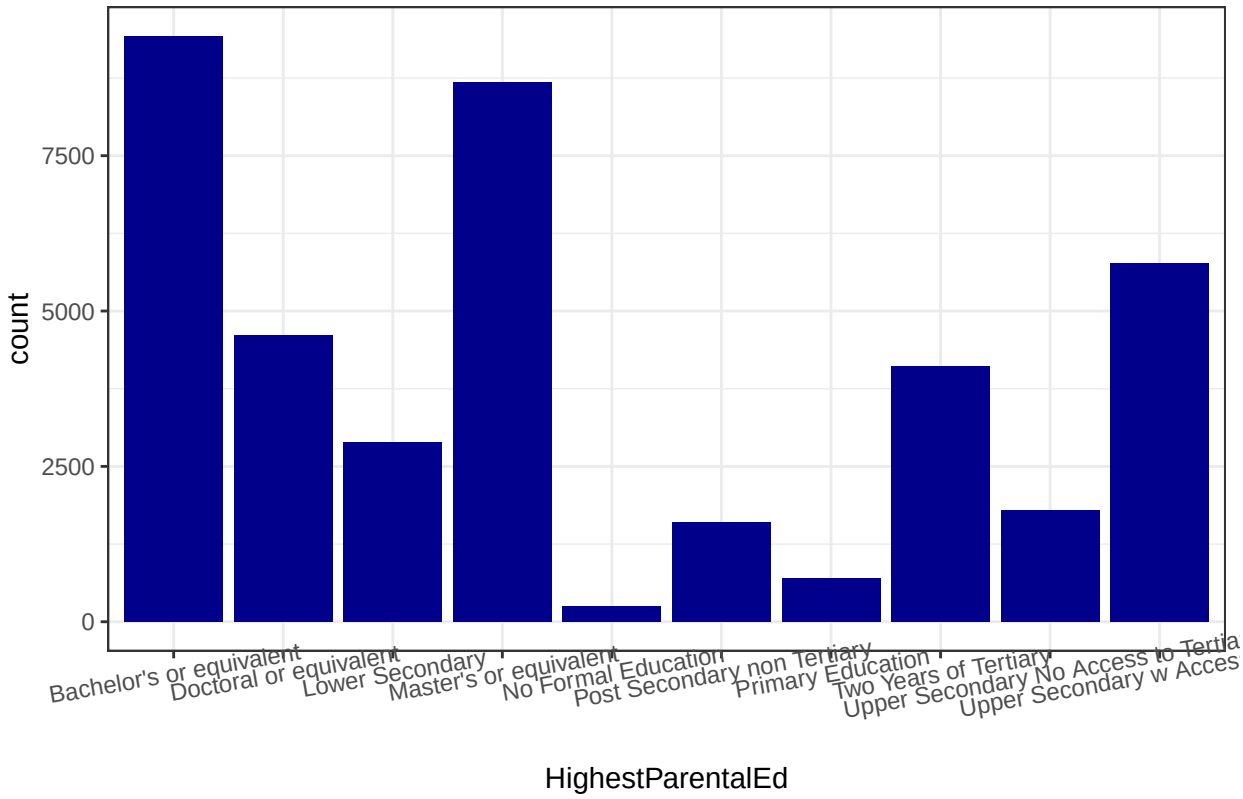
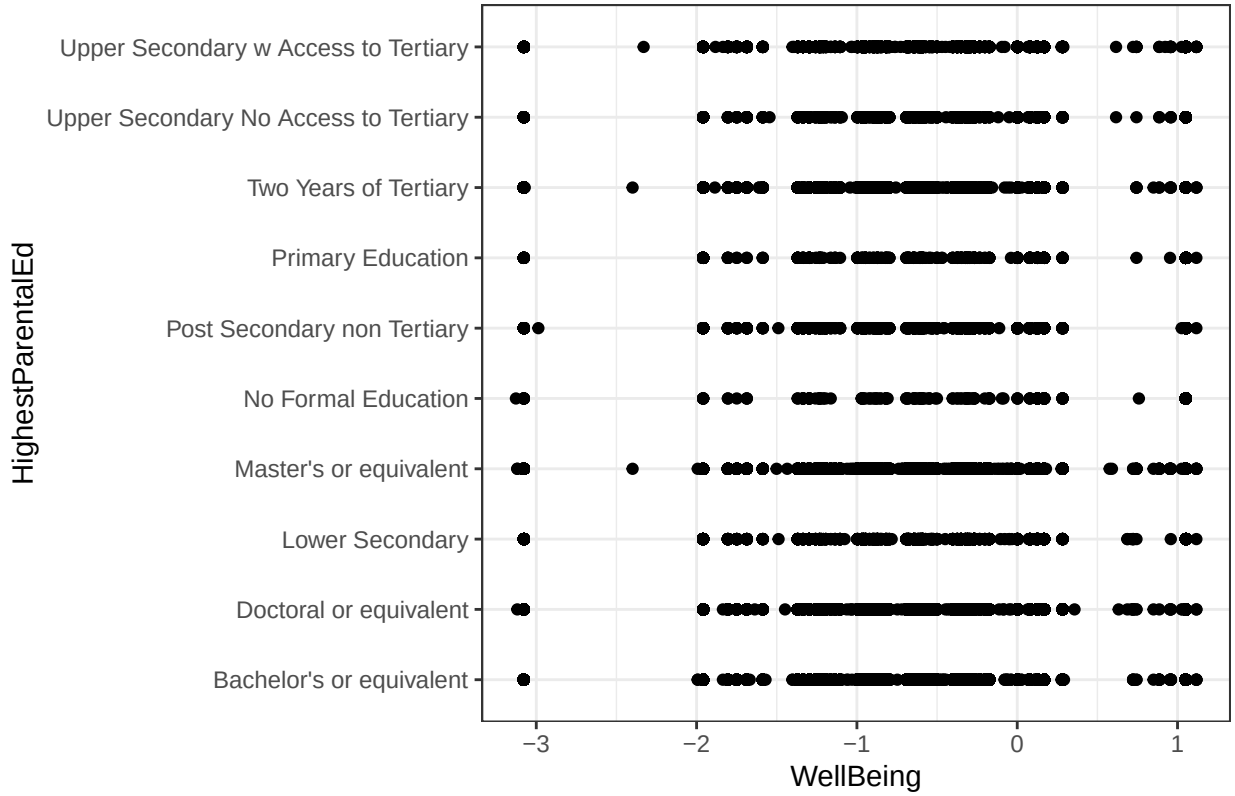


Figure 6: Highest Parental Education & Well−Bein

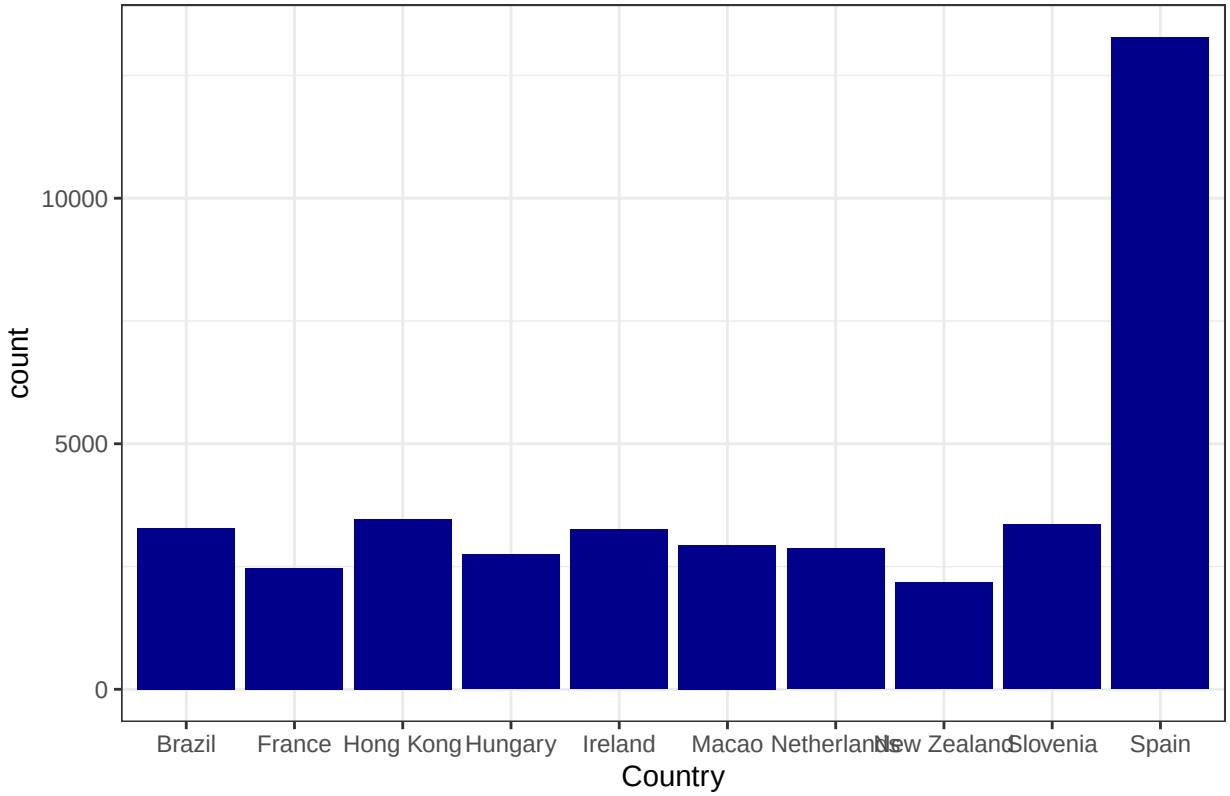


Country of Origin

The majority of the sample was from Spain (33.33%) and a similar amount from the other included countries (Brazil, Frane, Hong Kong, Hungary, Ireland, Macao, Netherlands, New Zealand, and Slovenia) that range from 5% to 8% of the total sample. Due to the large sample from Spain, I used Spain as the reference group in the regressions below. Again, regardless of country of origin, well-being scores were below average. However, the highest reported well-being was found in Brazil (M = .20, SD = .96) and the lowest in Macao (M = -.32, SD = 1.01).

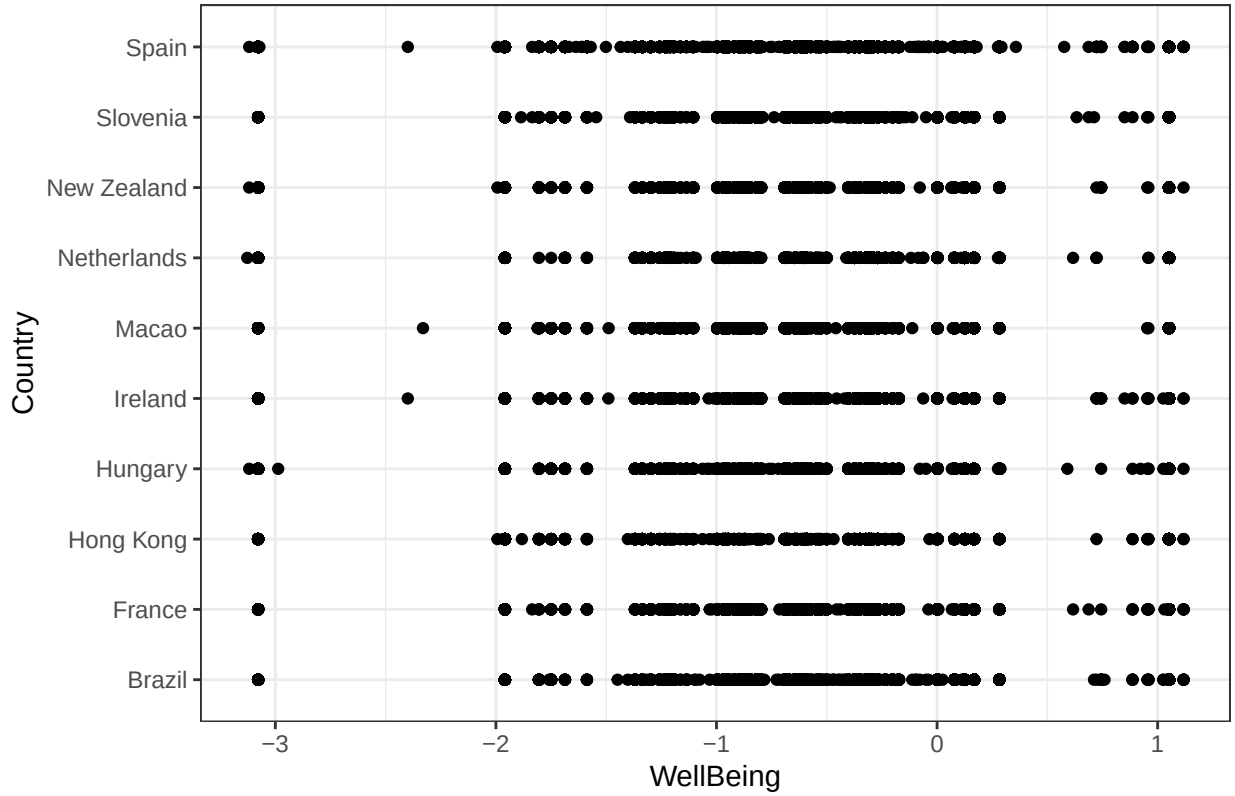
```
##
##      Brazil      France   Hong Kong     Hungary     Ireland       Macao
##  0.08247785  0.06199644  0.08689541  0.06892397  0.08177506  0.07379333
## Netherlands New Zealand    Slovenia       Spain
##  0.07206144  0.05466730  0.08418463  0.33322457
```

Figure 7: Country of Origin



Country	count	meanWellBeing	sdWellBeing
Brazil	3286	0.2065601	0.9598628
Netherlands	2871	0.1375765	0.8806235
Hungary	2746	0.1102630	0.9581029
Hong Kong	3462	0.0118025	1.0880124
Ireland	3258	0.0050479	1.0668032
New Zealand	2178	-0.0286390	1.0706408
Slovenia	3354	-0.0620122	1.0384261
France	2470	-0.1686475	0.9206562
Spain	13276	-0.1812998	0.9876528
Macao	2940	-0.3260127	1.1035605

Figure 8: Country of Origin & Well−Being



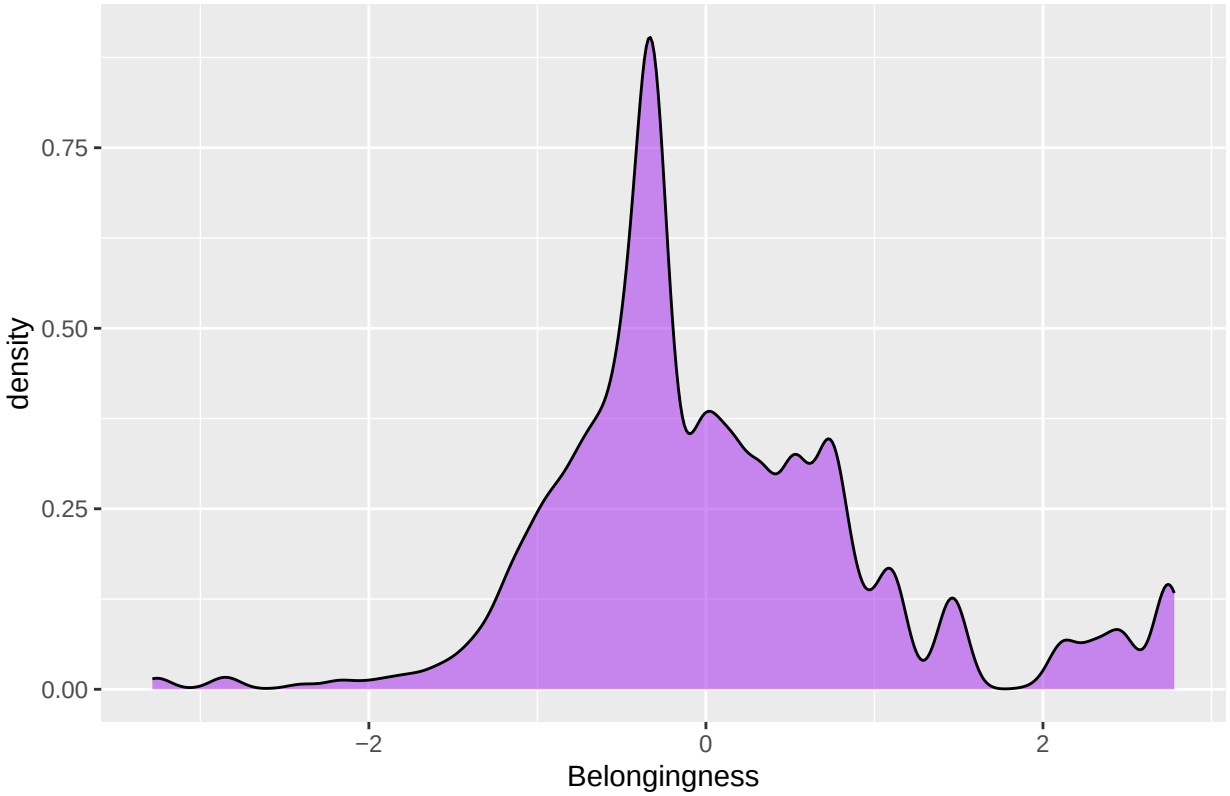
Psychosocial Features

Belongingness

According to Figure 9, the density plot shows belongingness scores across the entire sample are below average. For Figure 10, a generalized additive model ("gam") method was used for a penalized regression. This figure shows with an increase in well-being scores there is an increase on belongingness scores.

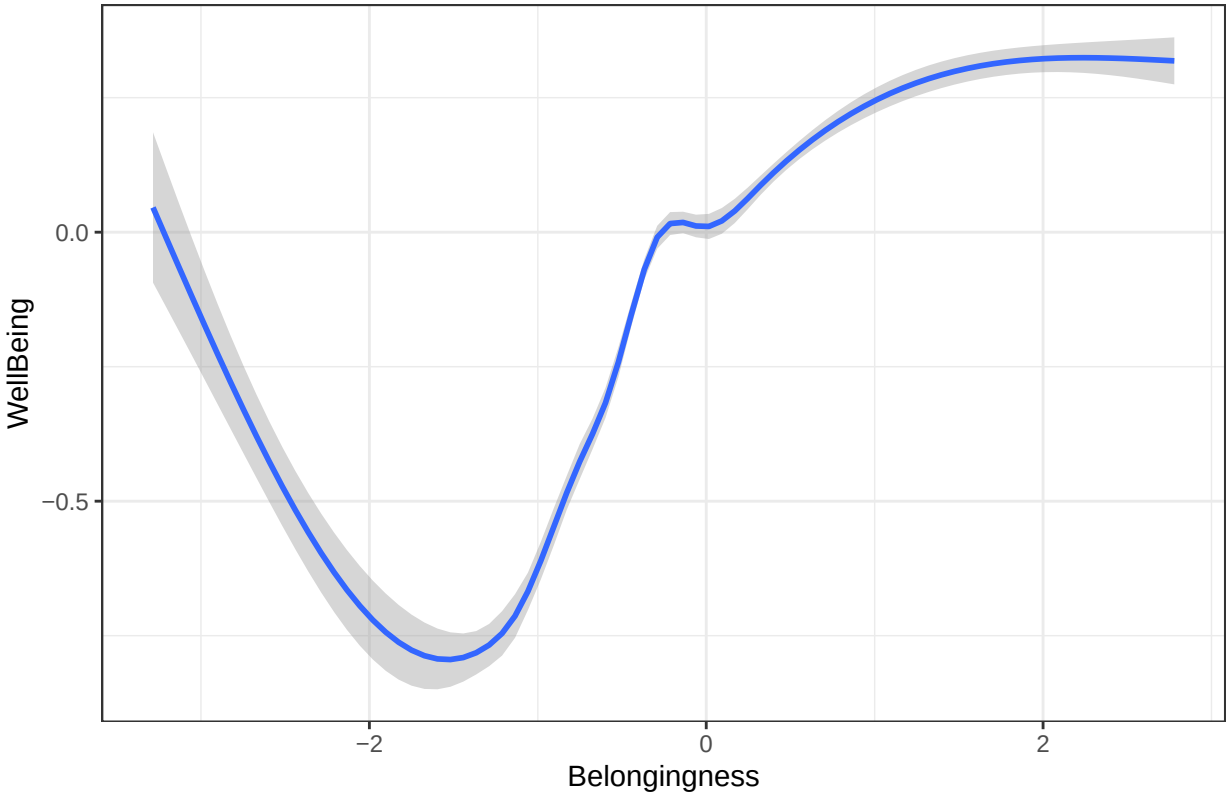
Belongingness
Min. :-3.28470
1st Qu.:-0.49410
Median :-0.19500
Mean : 0.05357
3rd Qu.: 0.54530
Max. : 2.77930

Figure 9: Belongingness




```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

Figure 10: Belongingness & Well−Being

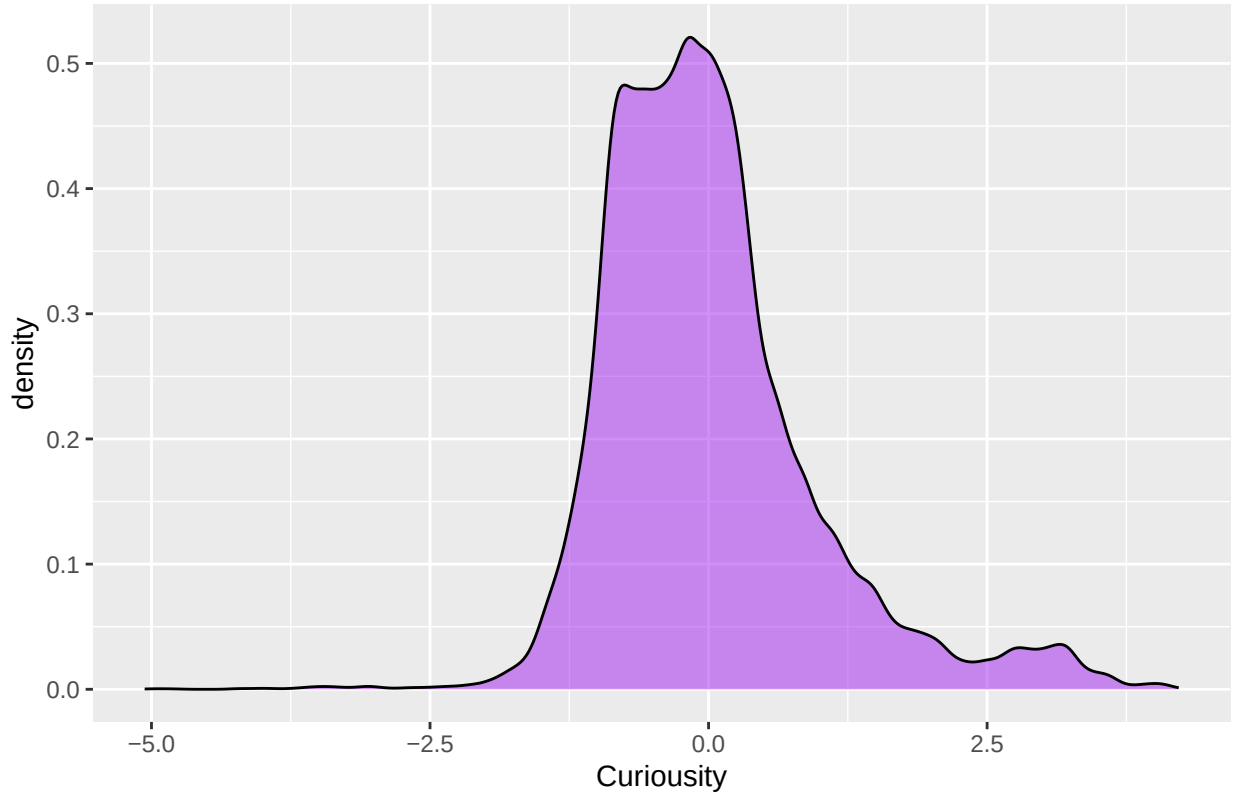


Curiosity

According to Figure 11, the density plot shows curiosity scores across the entire sample are near average. For Figure 12, a generalized additive model ("gam") method was used for a penalized regression. This figure shows with an increase in well-being scores there is an increase on curiosity scores.

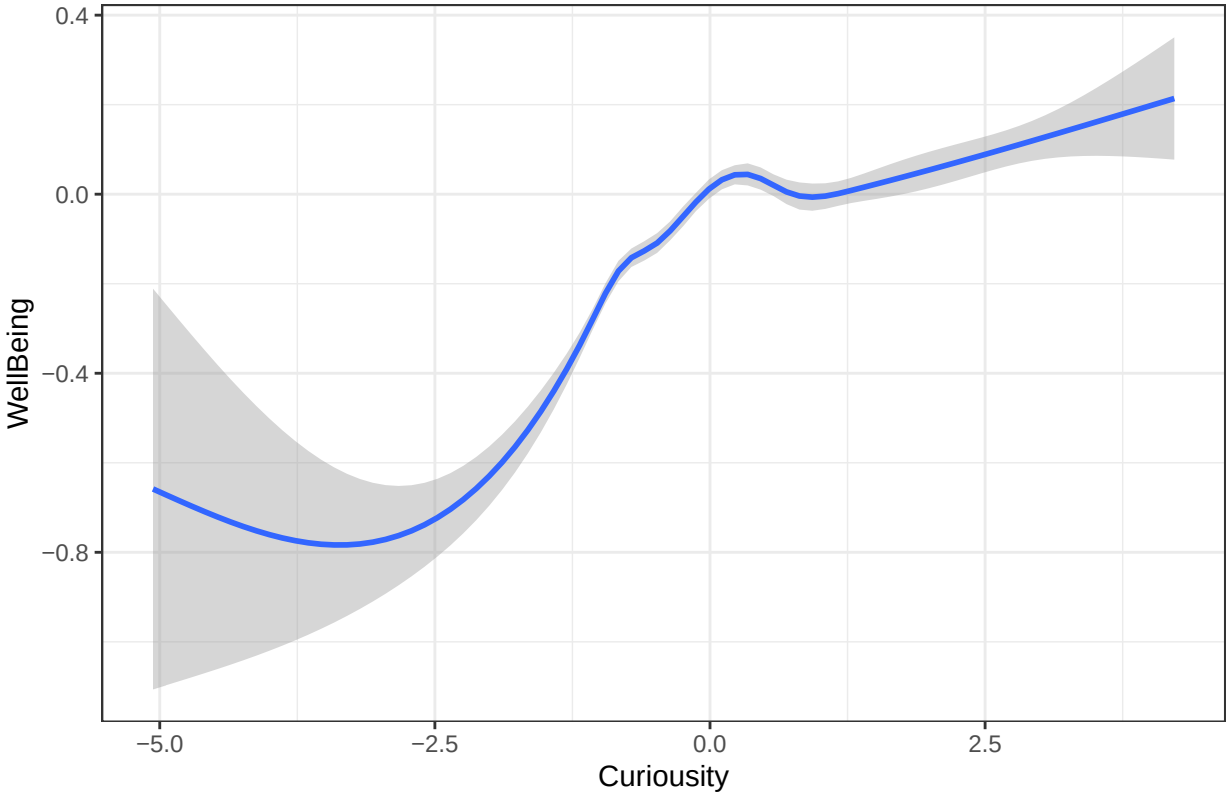
Curiousity
Min. :-5.06130
1st Qu.:-0.63440
Median :-0.13220
Mean : 0.01579
3rd Qu.: 0.40480
Max. : 4.21950

Figure 11: Curiosity




```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

Figure 12: Curiosity & Well−Being

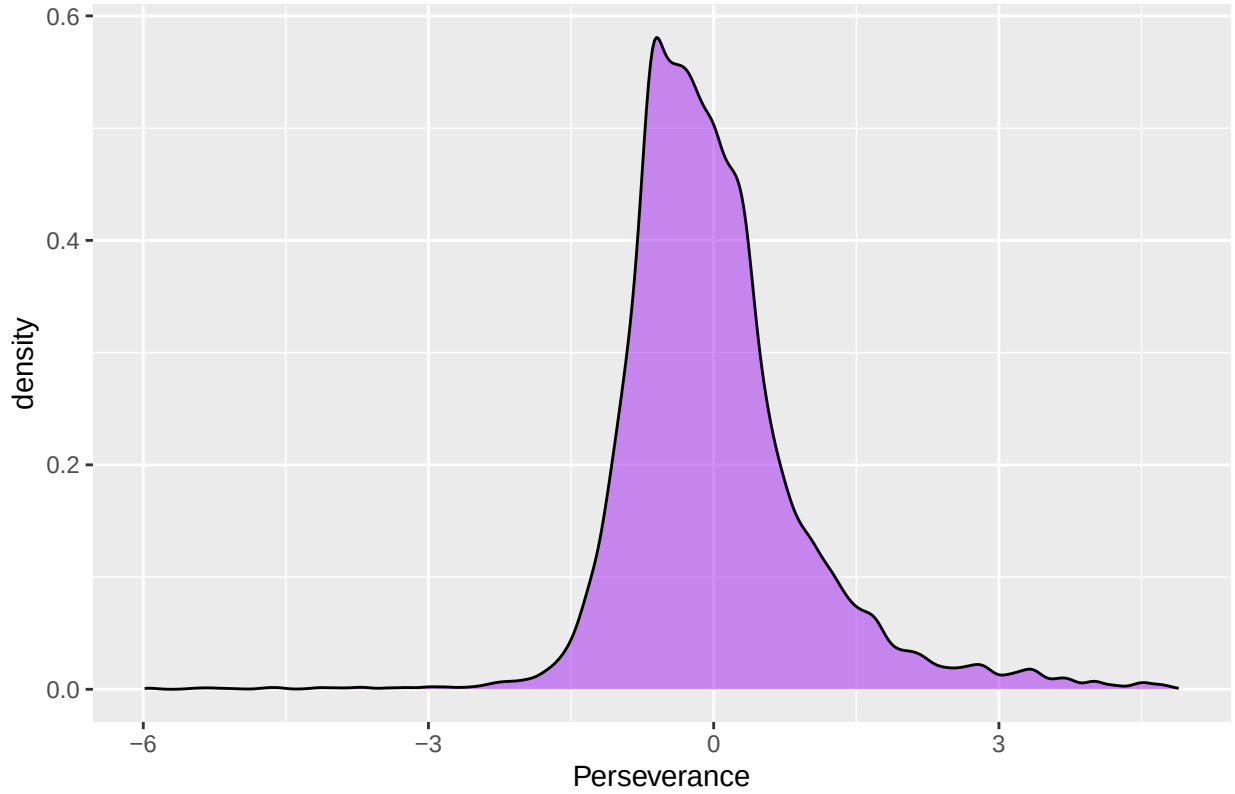


Perseverance

According to Figure 13, the density plot shows perseverance scores across the entire sample are slightly below average. For Figure 14, a generalized additive model ("gam") method was used for a penalized regression. This figure shows with an increase in well-being scores there is an increase on perseverence scores.

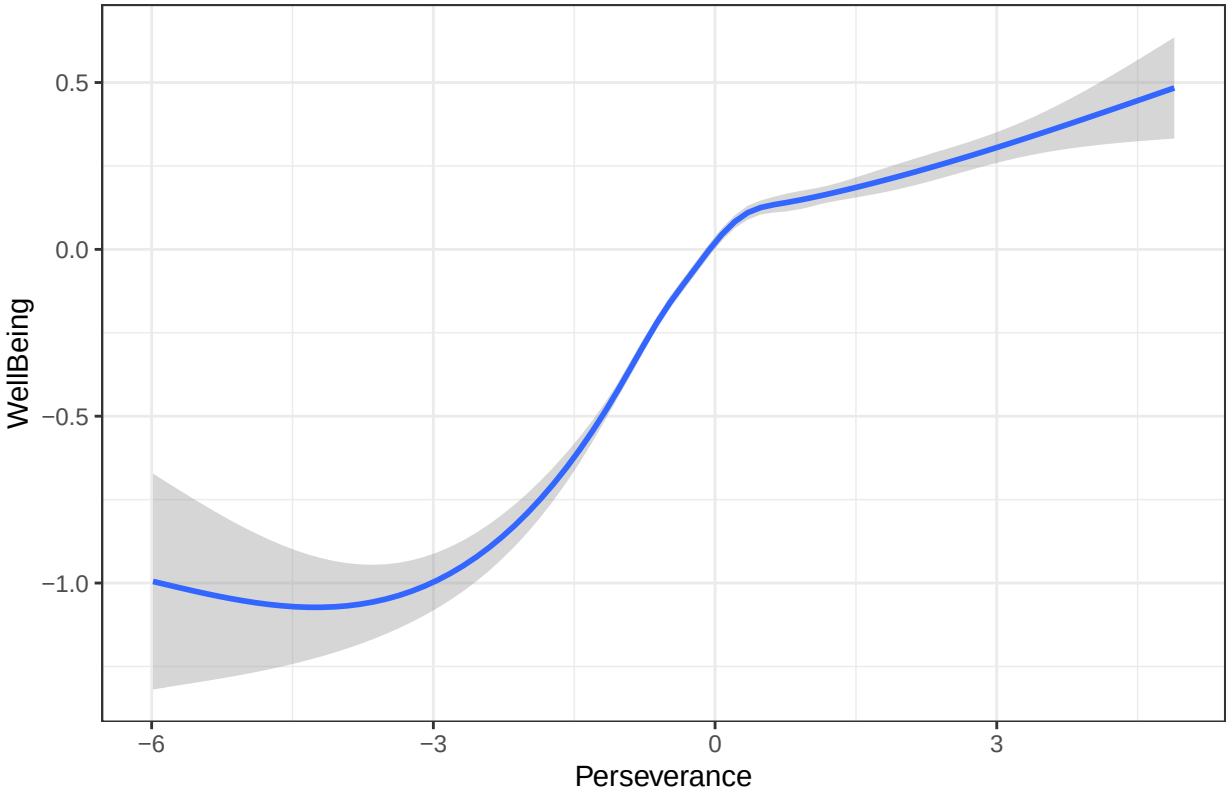
Perseverance
Min. :-5.985400
1st Qu.:-0.594100
Median :-0.140900
Mean : 0.001609
3rd Qu.: 0.381800
Max. : 4.890900

Figure 13: Perseverance




```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

Figure 14: Perseverance & Well−Being

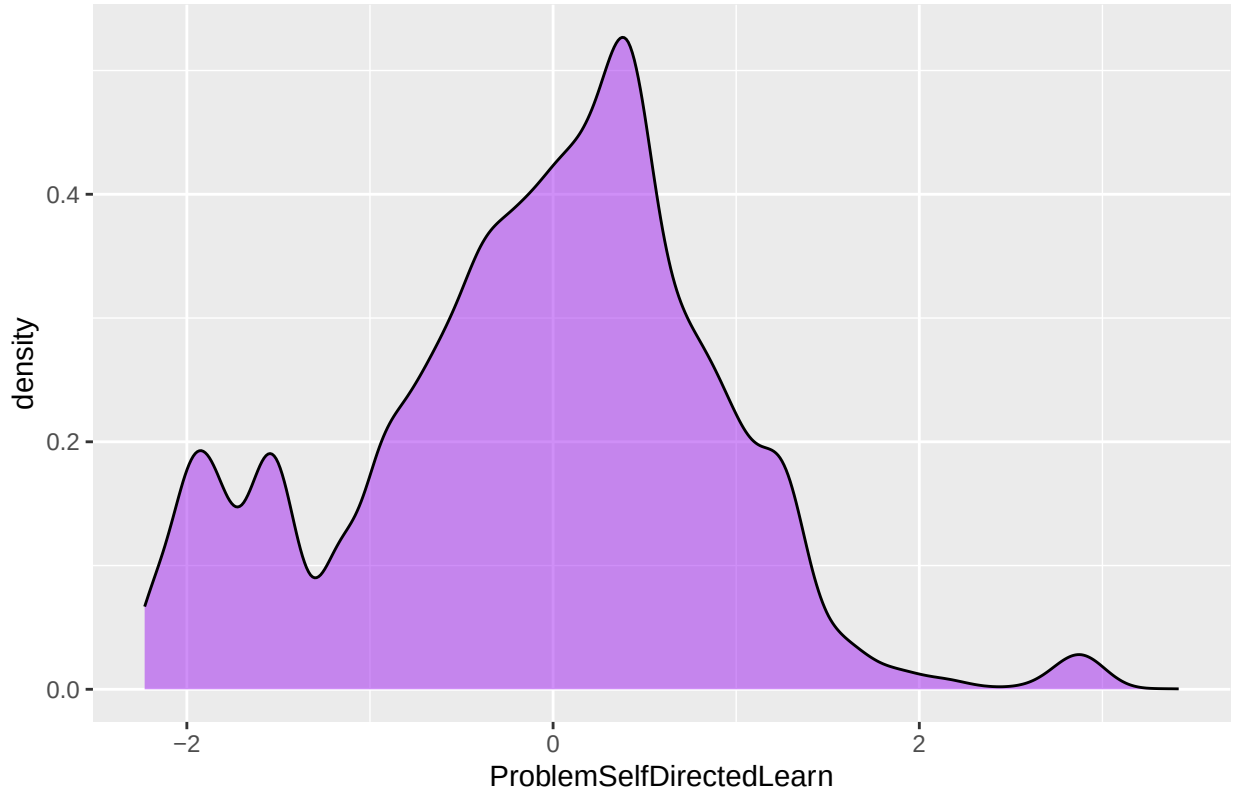


Problems with Self Directed Learning

According to Figure 15, the density plot shows problems with self-direct learning scores show a large proportion of students near the average, however there looks to be a bimodal distribution with an increase in scores around a -2 standard deviation. For Figure 16, a generalized additive model ("gam") method was used for a penalized regression. This figure also shows mixed results, with an increase in well-being scores being associated with both high scores on problems with self-direct learning and low problems with self-direct learning.

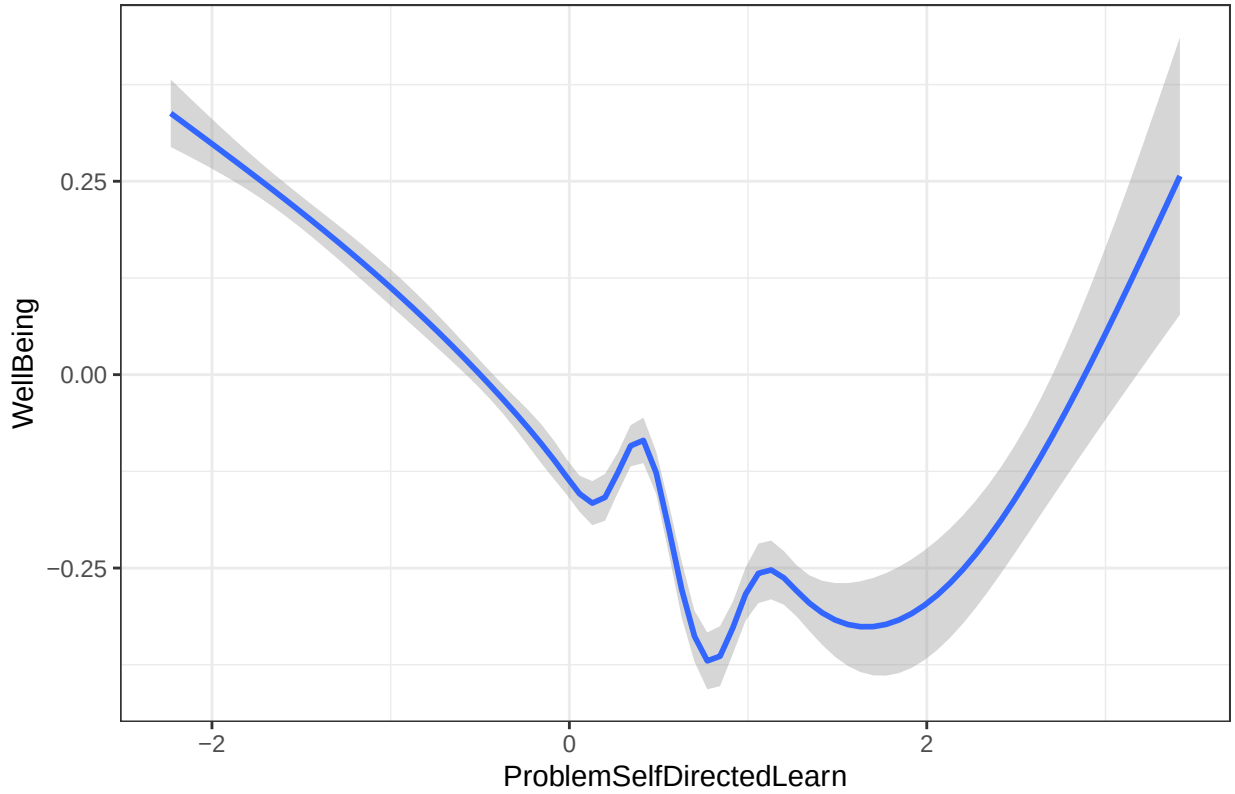
ProblemSelfDirectedLearn
Min. :-2.2306
1st Qu.:-0.6904
Median : 0.0075
Mean :-0.1079
3rd Qu.: 0.5144
Max. : 3.4165

Figure 15: Problems w/ Self Directed Learning




```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

Figure 16: Problems w/ Self Directed Learning & Well−Being

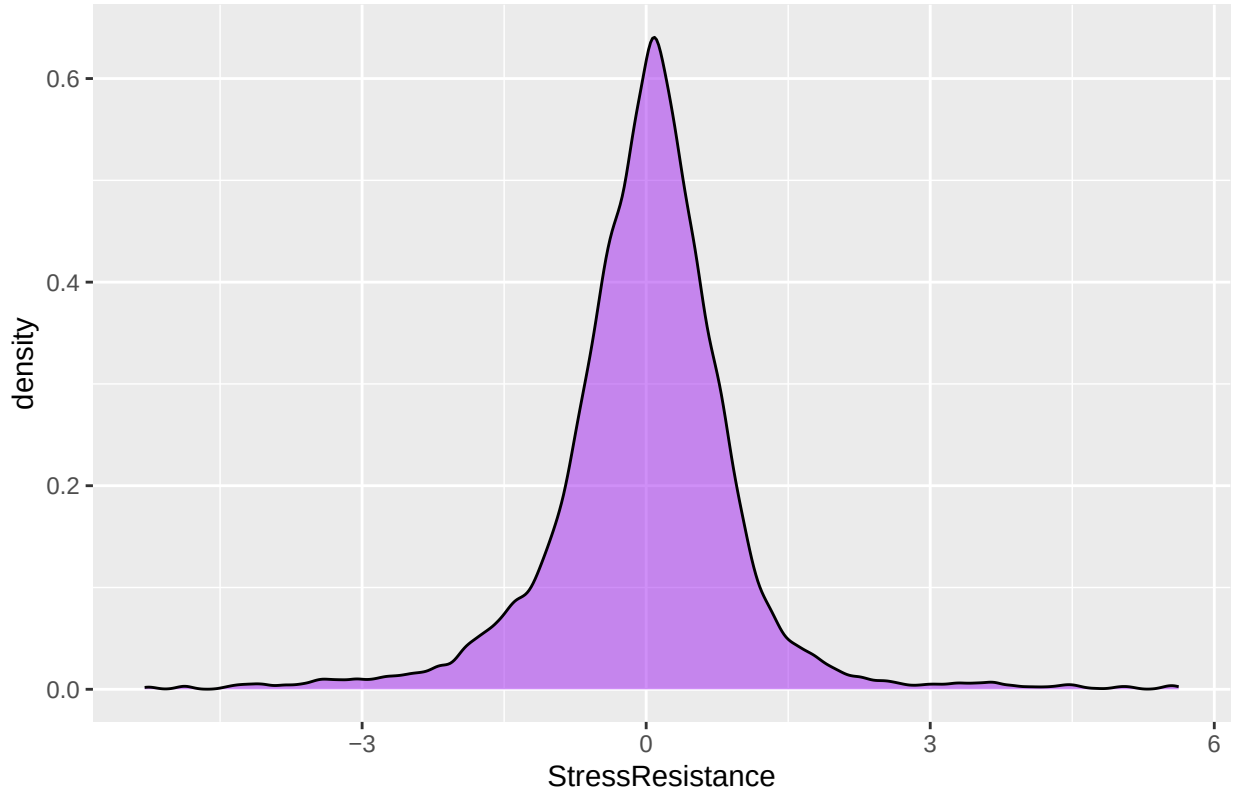


Stress Resistance

According to Figure 17, the density plot shows Stress Resistance scores across the entire sample are normally distributed. For Figure 18, a generalized additive model ("gam") method was used for a penalized regression. This figure shows with an increase in well-being scores there is an increase on stress resistance scores (and vice versa).

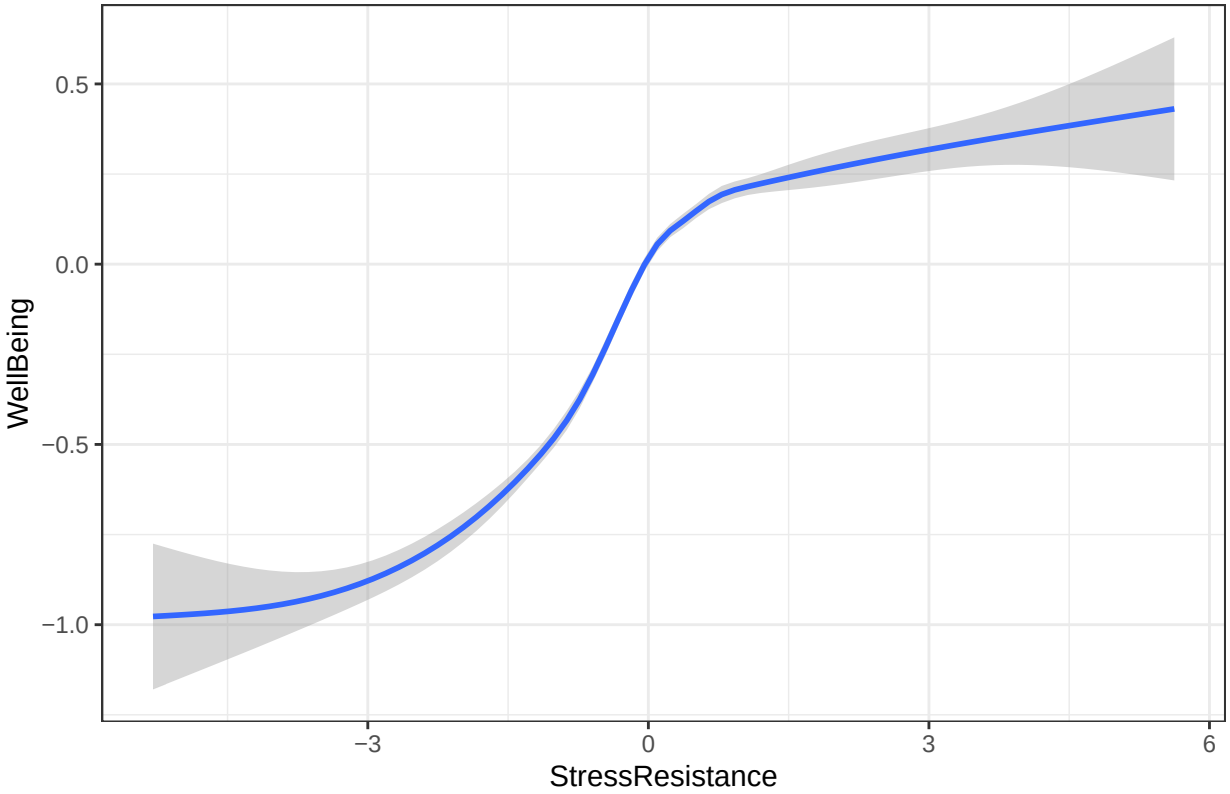
StressResistance
Min. :-5.29690
1st Qu.:-0.44920
Median : 0.03390
Mean :-0.01317
3rd Qu.: 0.46990
Max. : 5.62410

Figure 17: Stress Resistance




```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

Figure 18: Stress Resistance & Well−Being

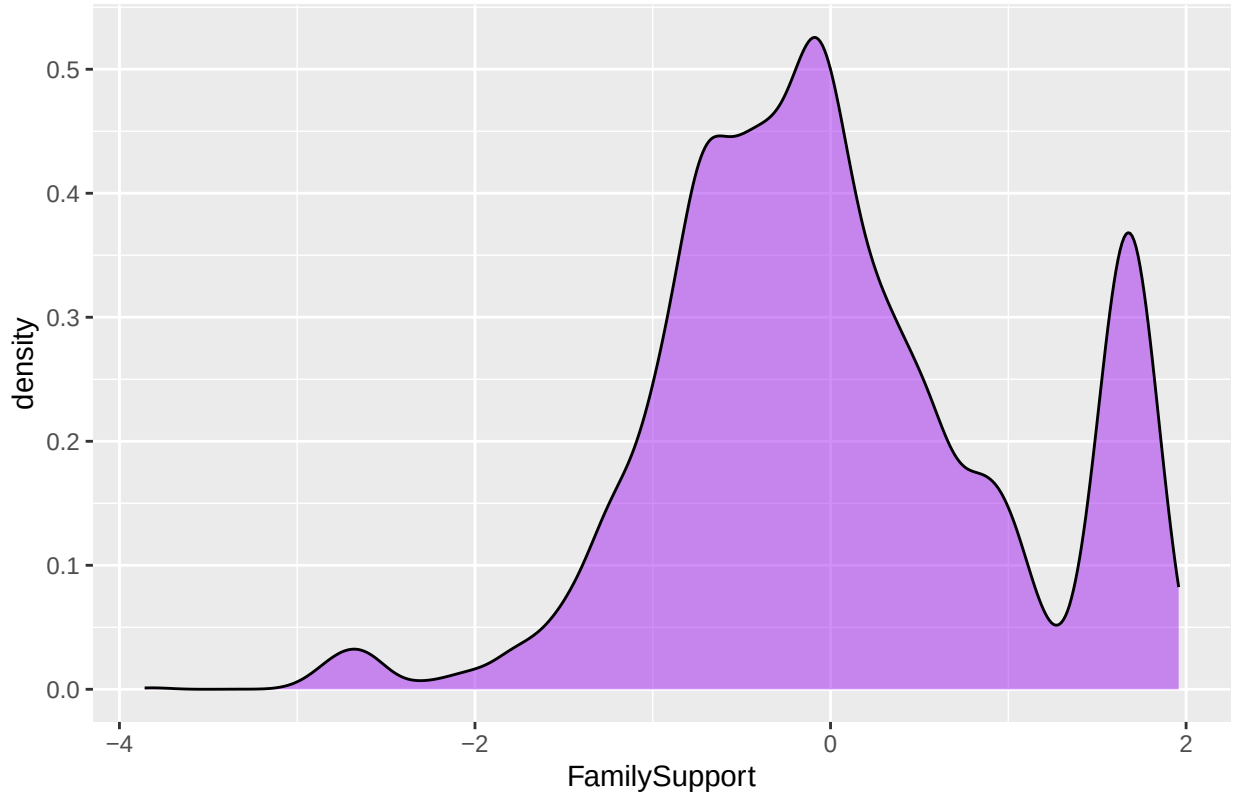


Family Support

According to Figure 19, the density plot shows Family Support scores across the entire sample show a bimodal distribution with a large proportion of scores at the mean, but then another large proportion at the 2 standard deviations above the average. For Figure 20, a generalized additive model ("gam") method was used for a penalized regression. This figure shows with an increase in well-being scores there is an increase on Family Support scores.

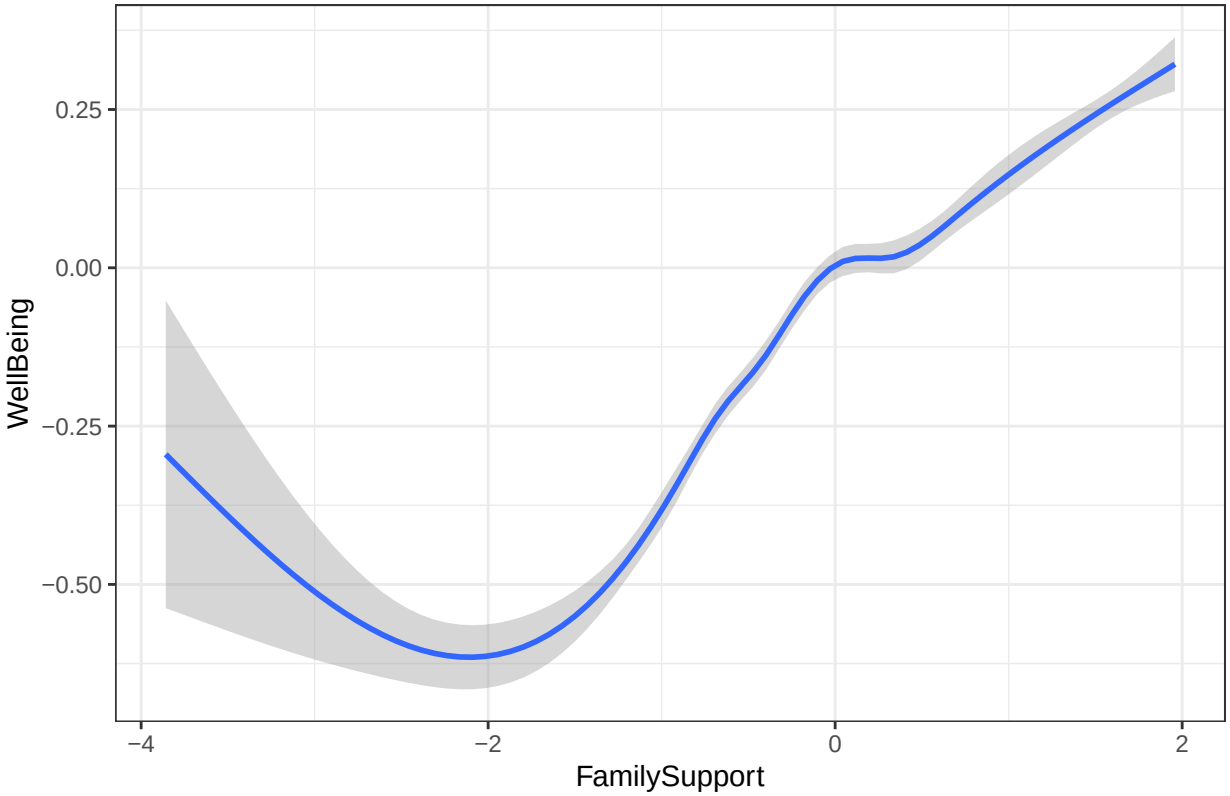
FamilySupport
Min. :-3.85810
1st Qu.:-0.62630
Median :-0.09030
Mean : 0.02691
3rd Qu.: 0.59480
Max. : 1.95830

Figure 19: FamilySupport




```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

Figure 20: FamilySupport & Well−Being



Hierarchical Multiple Linear Regression

A series of three models were assessed using hierarchical multiple linear regression.

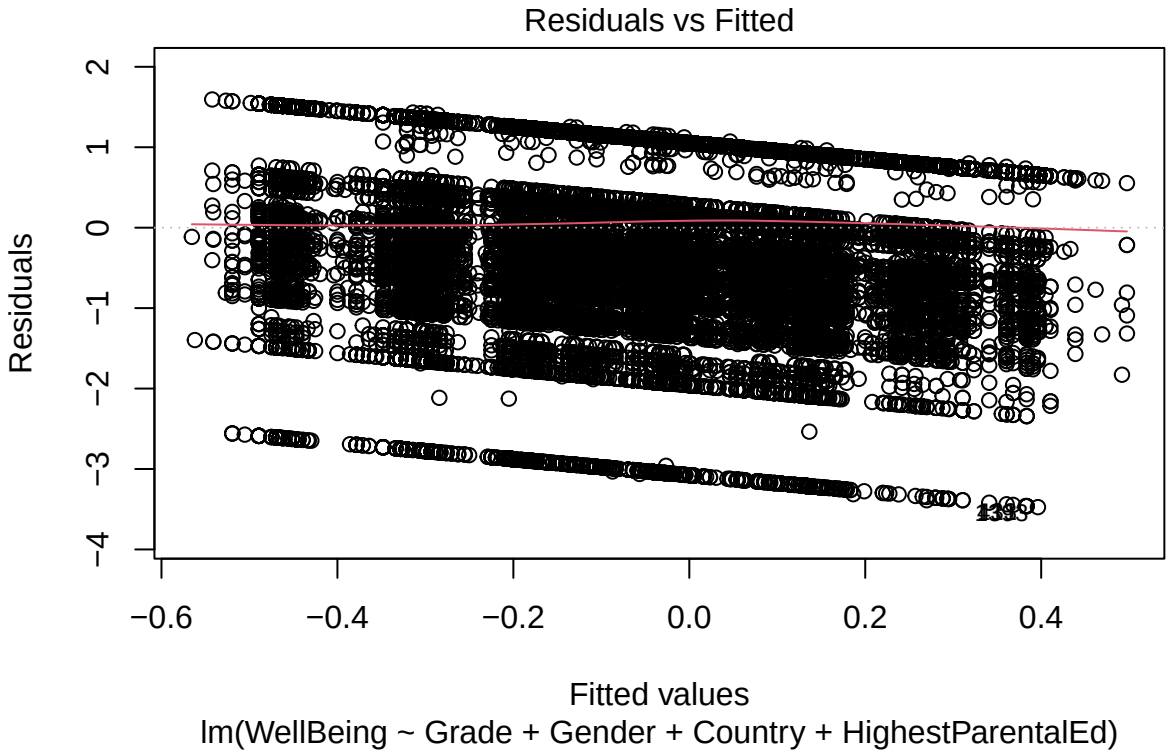
Model 1

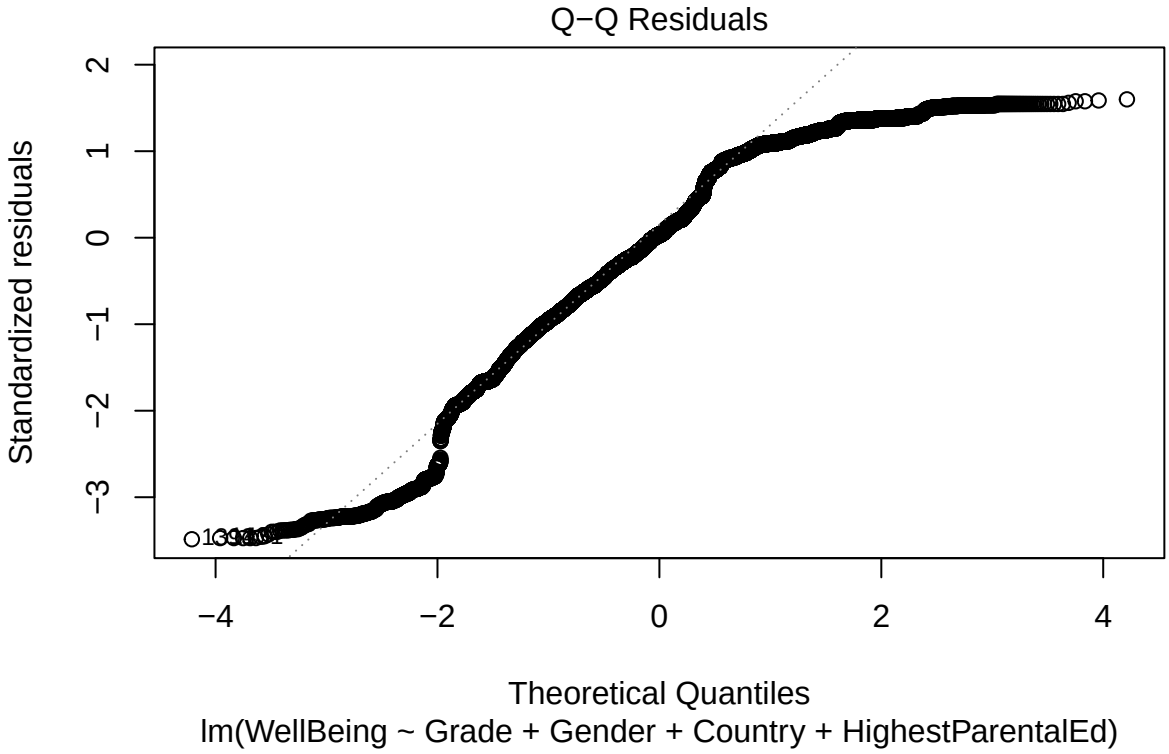
In Model 1, I assessed the contribution of only the demographic predictors/features to the well-being outcome. Overall, the model explained only 3.9% of the variance in scores of well-being, $R\text{-}squared = .039$, $F(24, 39816) = 67.92$, $p < .001$. Caution should be exercised with these results as there were clusters of outliers found. Grade level and highest level of parental education were not significant predictors of adolescent well-being. However, gender and all countries of origin did predict adolescent well-being. For instance, male students, on average, scored higher on well-being compared to females, $b = 0.26$, $p < .001$. See Table below for details by country.

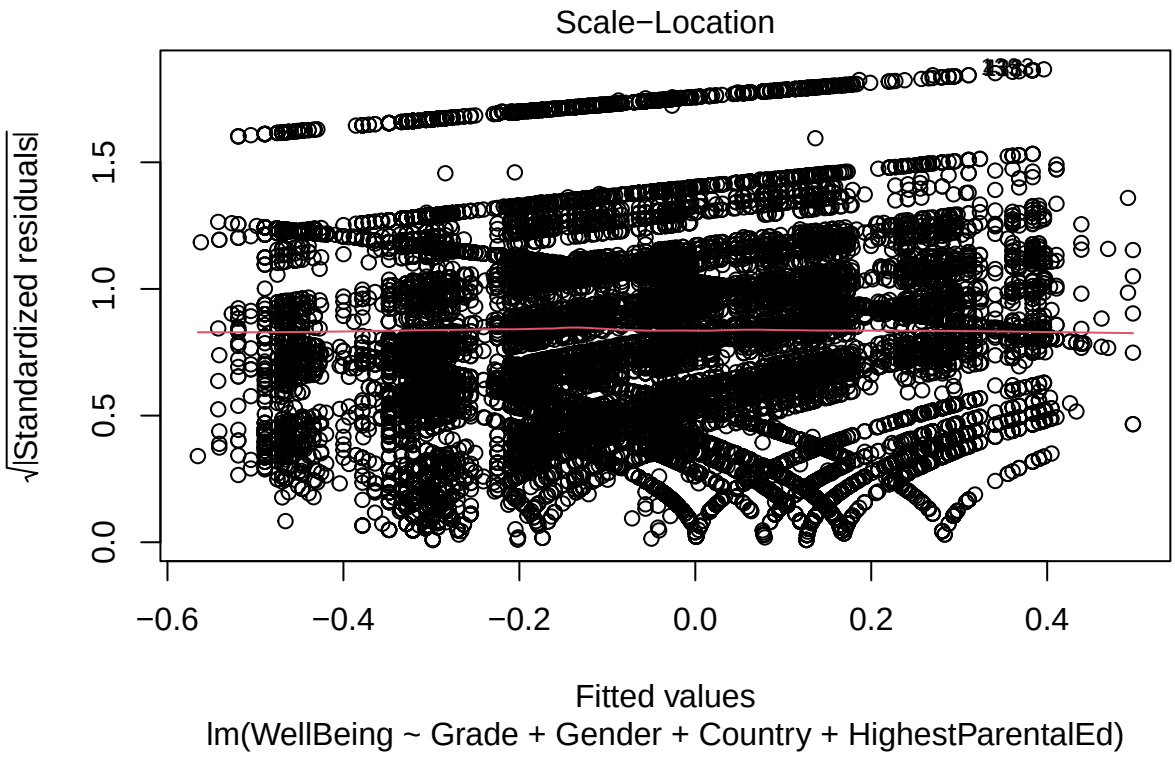
```
##
## Call:
## lm(formula = WellBeing ~ Grade + Gender + Country + HighestParentalEd,
##     data = bcpwb_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4742 -0.6224  0.0345  0.9251  1.5937
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                        0.220233   0.124451
```

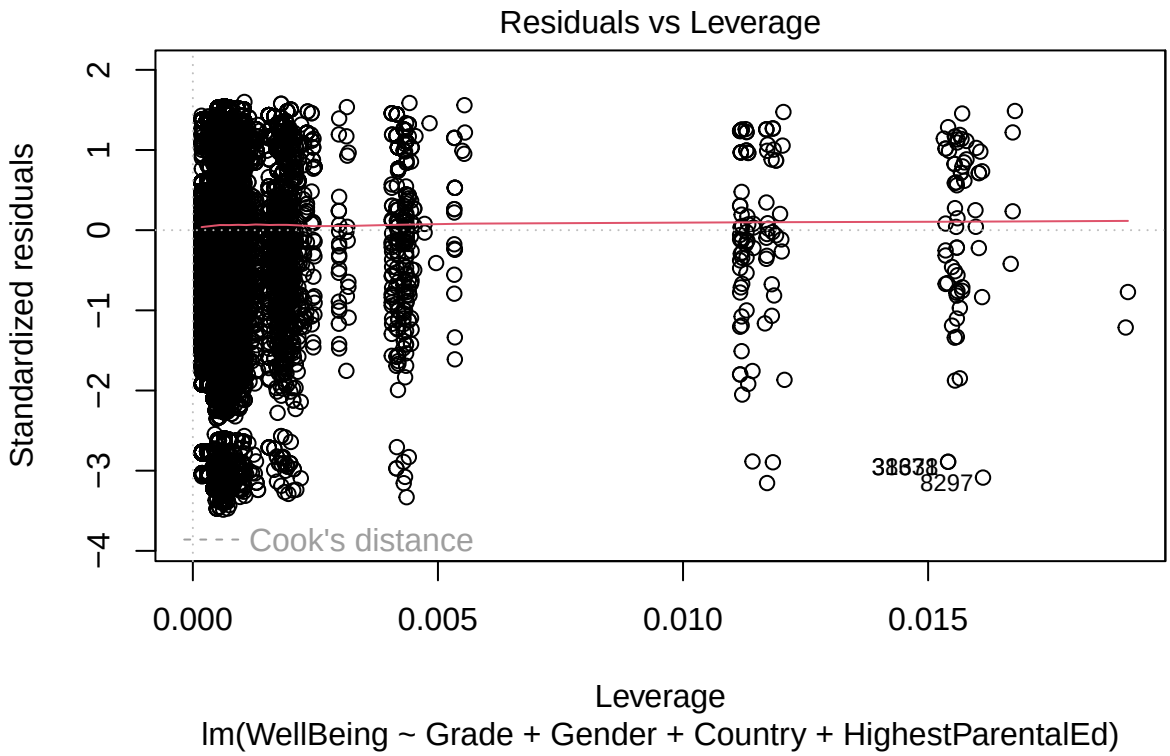
```
## Grade8                                                -0.058634   0.128868
## Grade9                                                -0.086923   0.123505
## Grade10                                               -0.101165   0.123120
## Grade11                                               -0.173915   0.124909
## Grade12                                               -0.204524   0.162777
## GenderMale                                             0.264410   0.010028
## CountryFrance                                         -0.402119   0.028024
## CountryHong Kong                                      -0.227238   0.025999
## CountryHungary                                        -0.140668   0.029067
## CountryIreland                                        -0.231667   0.025688
## CountryMacao                                          -0.565818   0.027132
## CountryNetherlands                                    -0.102784   0.026681
## CountryNew Zealand                                    -0.190940   0.031347
## CountrySlovenia                                       -0.305617   0.026541
## CountrySpain                                          -0.424665   0.021380
## HighestParentalEdDoctoral or equivalent                0.012845   0.018081
## HighestParentalEdLower Secondary                      -0.042539   0.021751
## HighestParentalEdMaster's or equivalent               -0.015375   0.014986
## HighestParentalEdNo Formal Education                  -0.094339   0.063585
## HighestParentalEdPost Secondary non Tertiary          -0.019091   0.027942
## HighestParentalEdPrimary Education                    -0.072743   0.039586
## HighestParentalEdTwo Years of Tertiary                 0.007212   0.018842
## HighestParentalEdUpper Secondary No Access to Tertiary 0.002827   0.026502
## HighestParentalEdUpper Secondary w Access to Tertiary -0.022804   0.017361
##                                                       t value Pr(>|t|)
## (Intercept)                                             1.770 0.076796 .
## Grade8                                                 -0.455 0.649116
## Grade9                                                 -0.704 0.481561
## Grade10                                                -0.822 0.411269
## Grade11                                                -1.392 0.163830
## Grade12                                                -1.256 0.208953
## GenderMale                                             26.368  < 2e-16 ***
## CountryFrance                                         -14.349  < 2e-16 ***
## CountryHong Kong                                       -8.740  < 2e-16 ***
## CountryHungary                                         -4.839 1.31e-06 ***
## CountryIreland                                         -9.019  < 2e-16 ***
## CountryMacao                                          -20.854  < 2e-16 ***
## CountryNetherlands                                     -3.852 0.000117 ***
## CountryNew Zealand                                     -6.091 1.13e-09 ***
## CountrySlovenia                                       -11.515  < 2e-16 ***
## CountrySpain                                          -19.863  < 2e-16 ***
## HighestParentalEdDoctoral or equivalent                 0.710 0.477446
## HighestParentalEdLower Secondary                       -1.956 0.050504 .
## HighestParentalEdMaster's or equivalent                -1.026 0.304934
## HighestParentalEdNo Formal Education                   -1.484 0.137907
## HighestParentalEdPost Secondary non Tertiary           -0.683 0.494460
## HighestParentalEdPrimary Education                     -1.838 0.066129 .
## HighestParentalEdTwo Years of Tertiary                  0.383 0.701906
## HighestParentalEdUpper Secondary No Access to Tertiary  0.107 0.915042
## HighestParentalEdUpper Secondary w Access to Tertiary  -1.314 0.189003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.997 on 39816 degrees of freedom
```

```
## Multiple R-squared:  0.03933,    Adjusted R-squared:  0.03875
## F-statistic: 67.92 on 24 and 39816 DF,  p-value: < 2.2e-16
```




Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
−4
−2
0
2
4
Theoretical Quantiles
lm(WellBeing ~ Grade + Gender + Country + HighestParentalEd)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−0.6
−0.4
−0.2
0.0
0.2
0.4
Fitted values
lm(WellBeing ~ Grade + Gender + Country + HighestParentalEd)

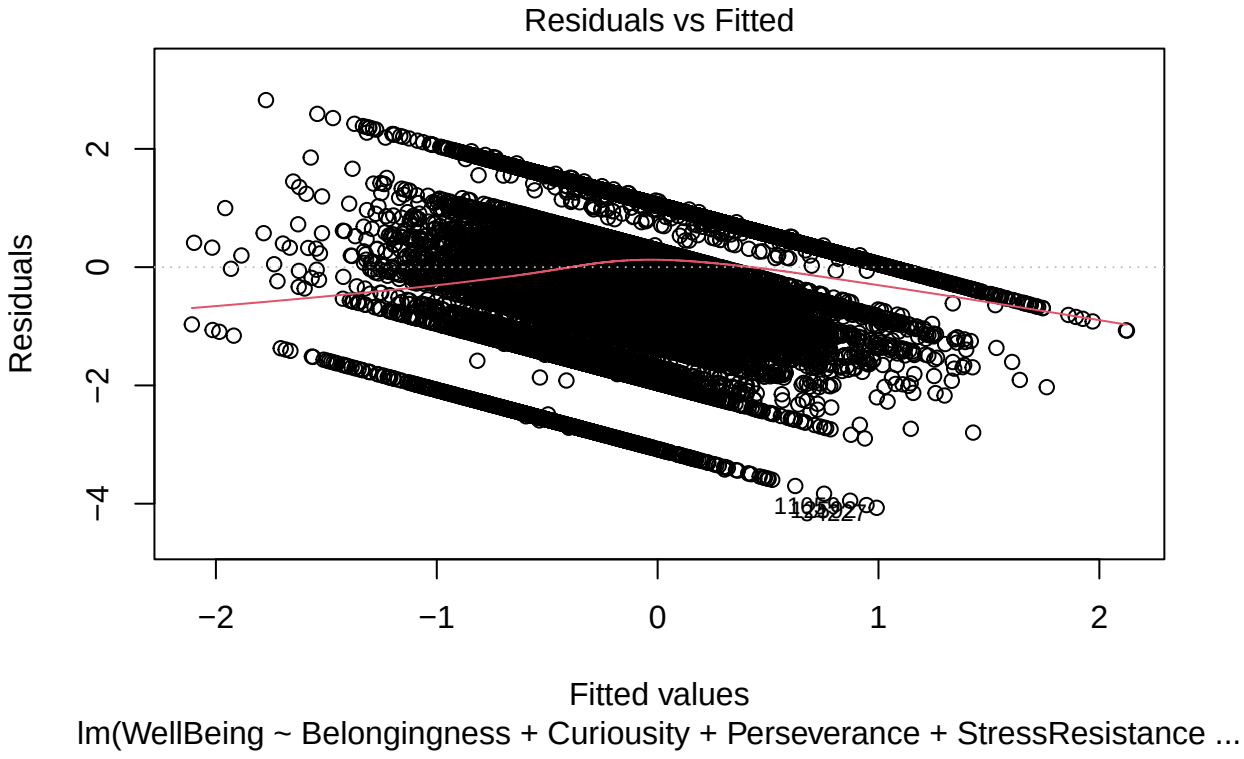


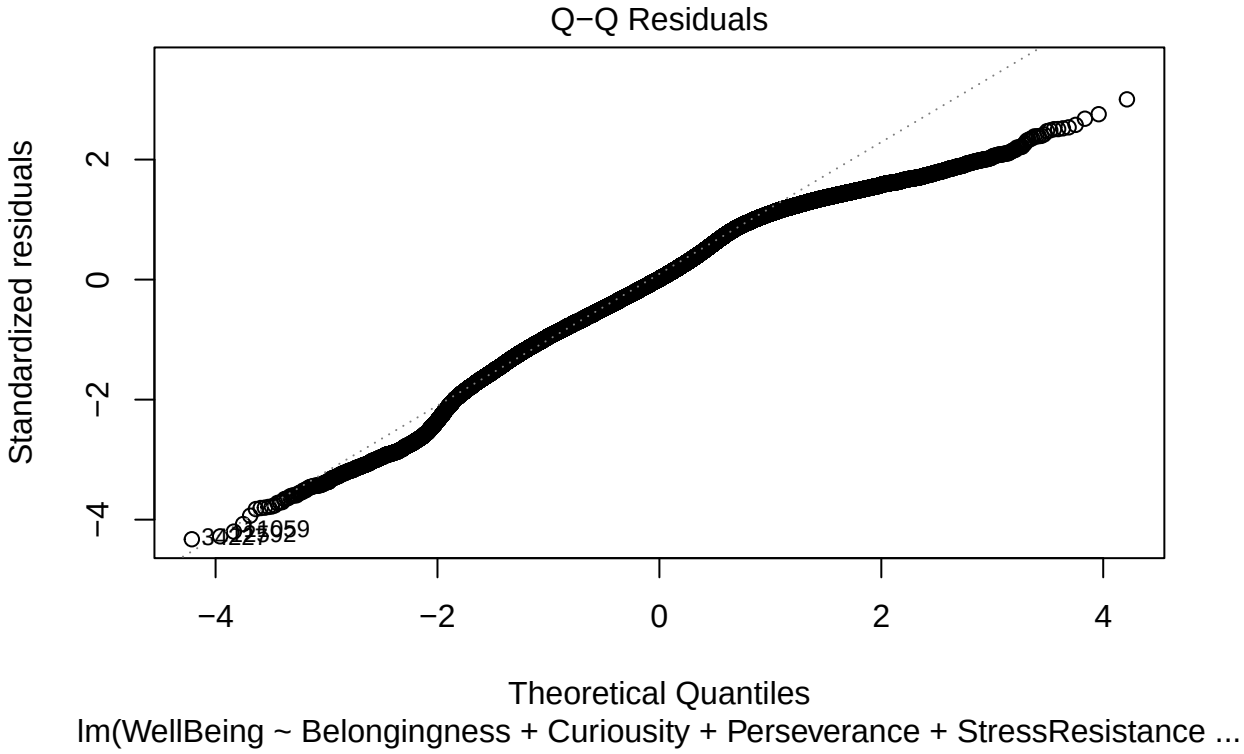
Model 2

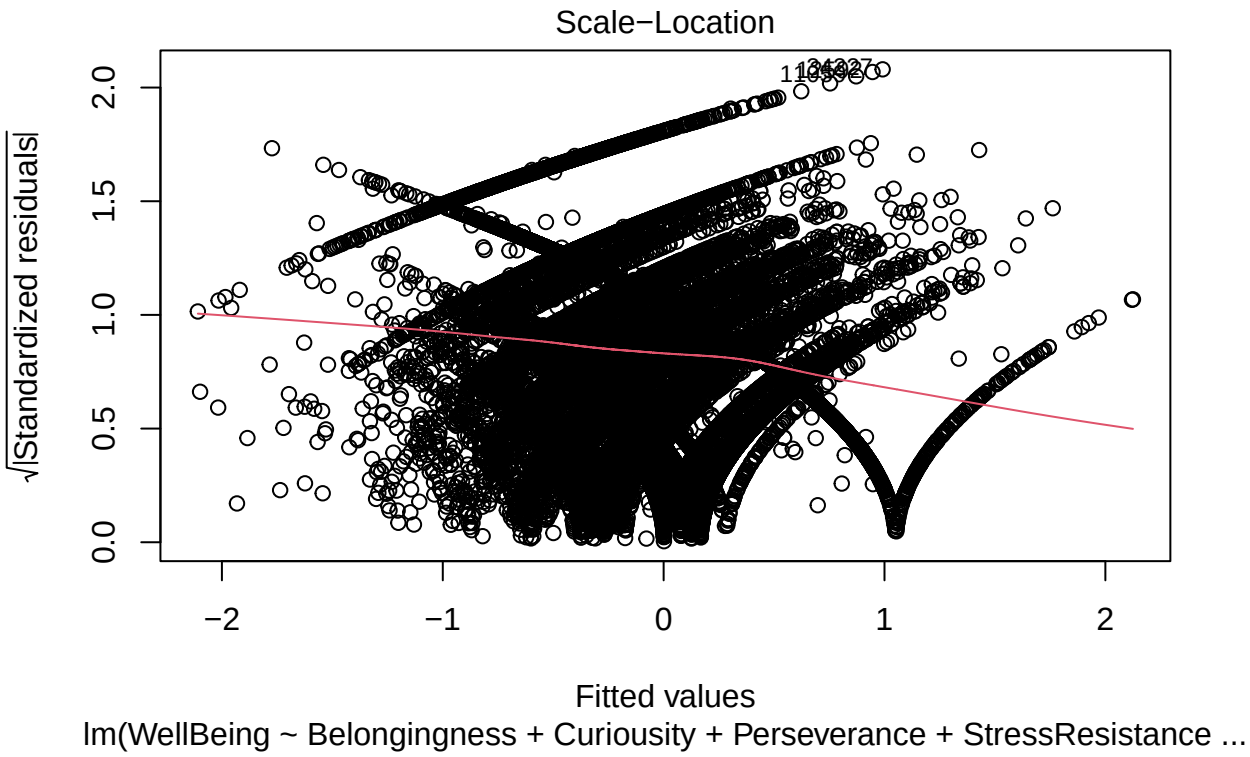
In Model 2, I assessed the psychosocial variables in relation to well-being. Overall, all of the psychosocial predictors of the model explained 14.4% of the variance in scores of well-being, $R\text{-}squared = .14$, $F(6, 39834) = 1118$, $p < .001$. Notably, Stress resistance, $b = .19$, $p < .001$, Belongingness, $b = .14$, $p < .001$, and Family Support, $b = .16$, $p < .001$ were the strongest predictors of increases in adolescent well-being.

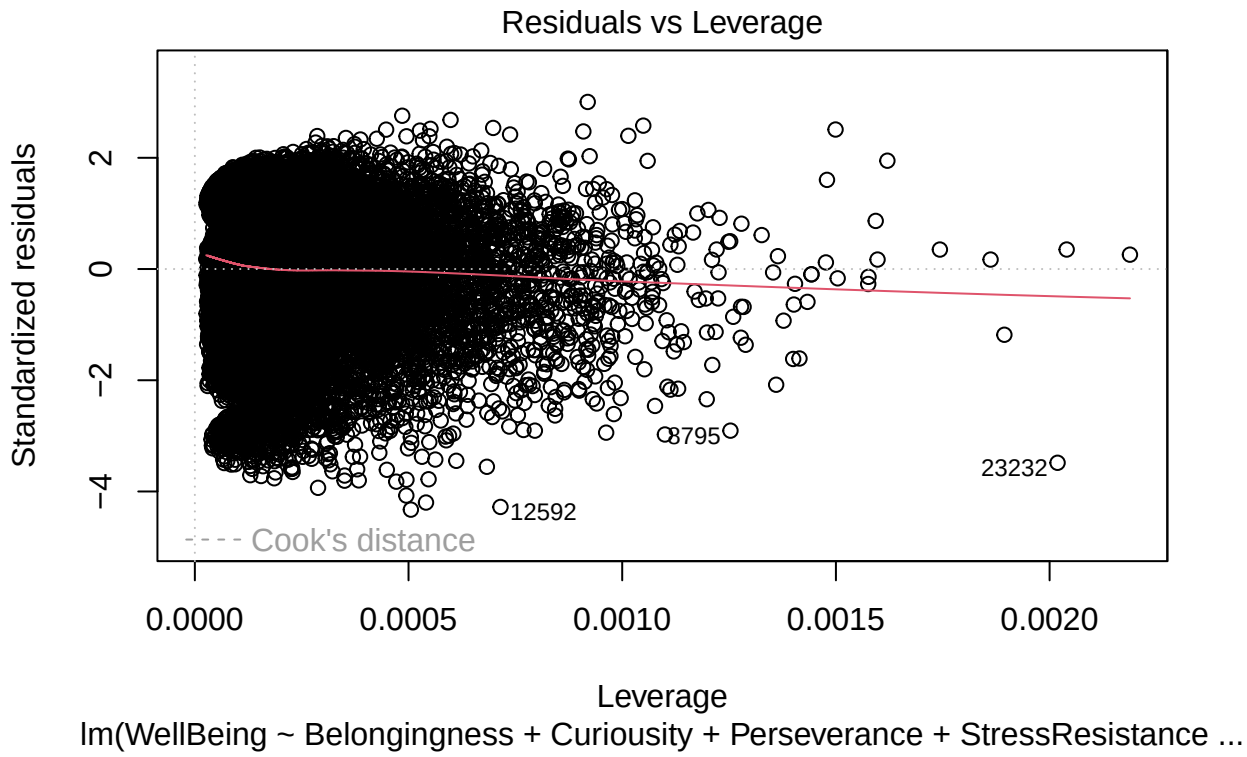
```
##
## Call:
## lm(formula = WellBeing ~ Belongingness + Curiousity + Perseverance +
##     StressResistance + FamilySupport + ProblemSelfDirectedLearn,
##     data = bcpwb_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0695 -0.6061  0.0115  0.7868  2.8247
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.084564   0.004751 -17.800   <2e-16 ***
## Belongingness               0.144102   0.005090  28.312   <2e-16 ***
## Curiousity                  0.012782   0.005310   2.407   0.0161 *
## Perseverance                0.092420   0.005538  16.687   <2e-16 ***
## StressResistance            0.193866   0.005171  37.491   <2e-16 ***
## FamilySupport               0.164015   0.005089  32.227   <2e-16 ***
## ProblemSelfDirectedLearn -0.082565   0.004993 -16.535   <2e-16 ***
## ---
```

```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9408 on 39834 degrees of freedom
## Multiple R-squared:  0.1442, Adjusted R-squared:  0.144
## F-statistic:  1118 on 6 and 39834 DF,  p-value: < 2.2e-16
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(WellBeing ~ Belongingness + Curiousity + Perseverance + StressResistance ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−2
−1
0
1
2
Fitted values
lm(WellBeing ~ Belongingness + Curiousity + Perseverance + StressResistance ...



Model 3

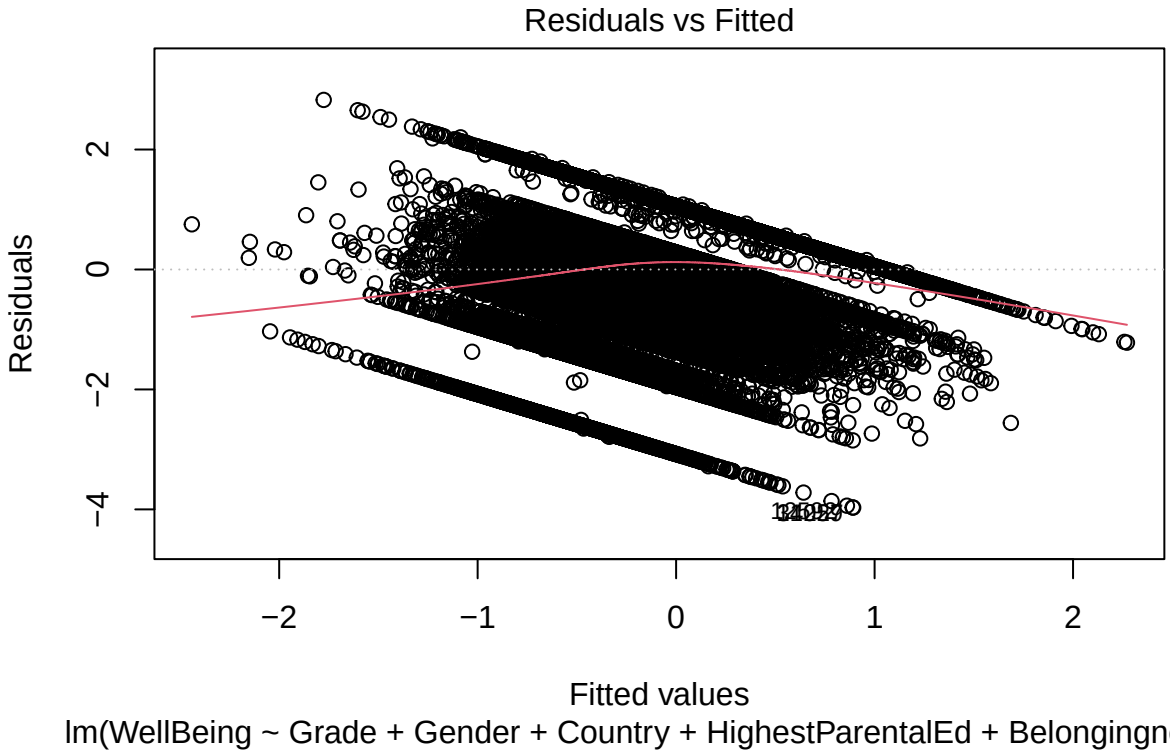
In Model 3, I assessed the demographic and psychosocial variables in relation to well-being. There were changes to demographic predictors to well-being when the psychosocial variables were included. Overall, all of the psychosocial and demographic predictors of the model explained 18.8% of the variance in scores of well-being, *R-squared* = .18, $F(30, 39810) = 294.8$, $p < .001$. Again, Stress resistance, $b = .16$, $p < .001$, Belongingness, $b = .18$, $p < .001$, and, Family Support, $b = .17$, $p < .001$ were the strongest predictors of increases in adolescent well-being.

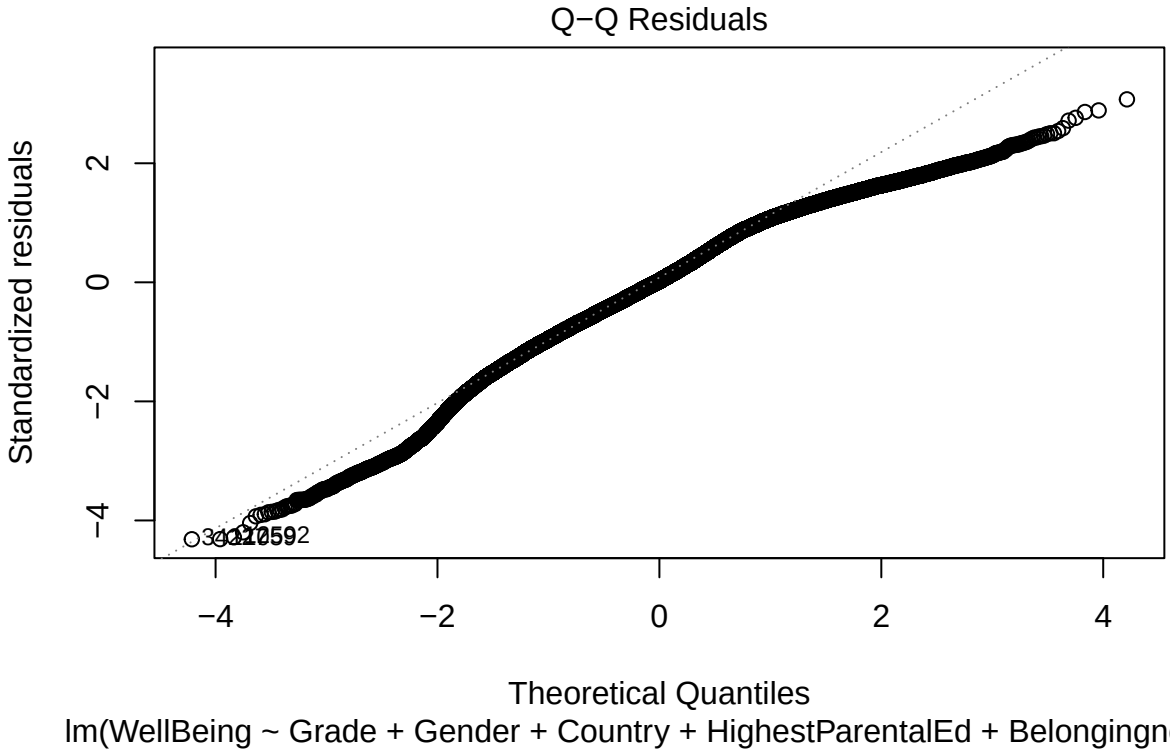
Specific predictors of Highest Parental Education and Grade level did become significant predictors in the model to well-being after inclusion of the psychosocial predictors. Specifically, students with parents with a Lower Secondary education, $b = 0.04$, $p < .05$, Two Years of Tertiary, $b = 0.04$, $p < .05$, Secondary No Access to Tertiary, $b = 0.06$, $p < .01$, and Master's or equivalent $b = -0.03$, $p < .001$ showed contributions to adolescent well-being. Due to multiple comparisons, there would likely need to be a correction to account for potential false discovery rates across all three of these models (e.g., Benjamini & Hochberg, 1995)

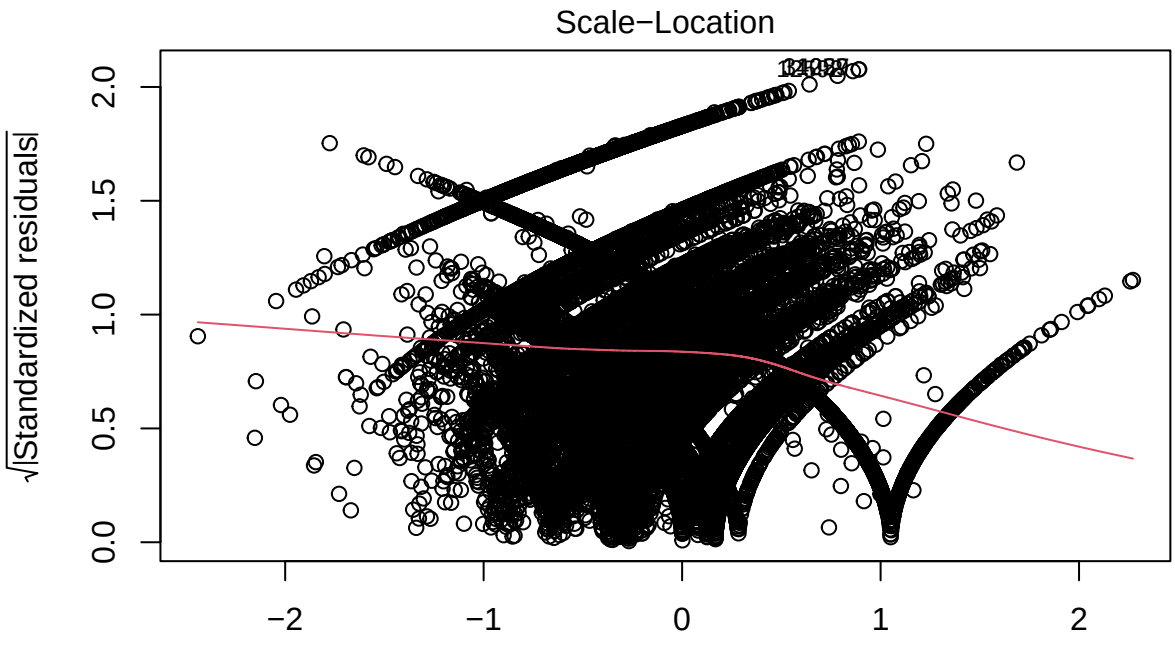
```
##
## Call:
## lm(formula = WellBeing ~ Grade + Gender + Country + HighestParentalEd +
##     Belongingness + Curiousity + Perseverance + FamilySupport +
##     ProblemSelfDirectedLearn + StressResistance, data = bcpwb_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9700 -0.5789  0.0220  0.7278  2.8270
##
## Coefficients:
##                                                    Estimate Std. Error
```

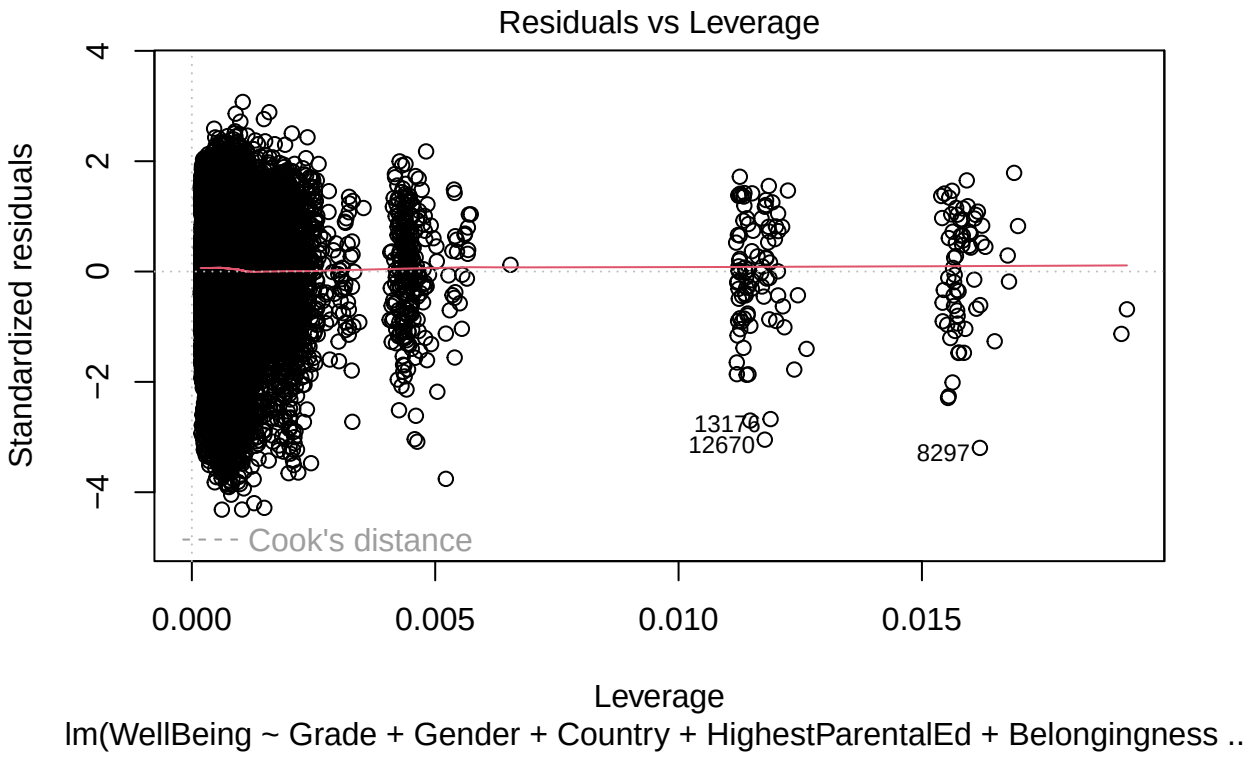
```
## (Intercept)                                          0.406574   0.114892
## Grade8                                              -0.081261   0.118943
## Grade9                                              -0.152694   0.113998
## Grade10                                             -0.188839   0.113648
## Grade11                                             -0.271523   0.115302
## Grade12                                             -0.330760   0.150254
## GenderMale                                           0.117711   0.009883
## CountryFrance                                       -0.451929   0.025975
## CountryHong Kong                                    -0.161103   0.024137
## CountryHungary                                      -0.335049   0.027028
## CountryIreland                                      -0.268838   0.023748
## CountryMacao                                        -0.502831   0.025247
## CountryNetherlands                                  -0.204250   0.024930
## CountryNew Zealand                                  -0.163730   0.028977
## CountrySlovenia                                     -0.329851   0.024642
## CountrySpain                                        -0.592648   0.019916
## HighestParentalEdDoctoral or equivalent              0.008334   0.016691
## HighestParentalEdLower Secondary                     0.041529   0.020112
## HighestParentalEdMaster's or equivalent             -0.028099   0.013834
## HighestParentalEdNo Formal Education                 0.031055   0.058721
## HighestParentalEdPost Secondary non Tertiary         0.014442   0.025796
## HighestParentalEdPrimary Education                   0.035218   0.036569
## HighestParentalEdTwo Years of Tertiary               0.036590   0.017395
## HighestParentalEdUpper Secondary No Access to Tertiary  0.064027   0.024477
## HighestParentalEdUpper Secondary w Access to Tertiary   0.029019   0.016043
## Belongingness                                        0.175014   0.005119
## Curiousity                                           0.016002   0.005242
## Perseverance                                         0.104808   0.005440
## FamilySupport                                        0.177157   0.005073
## ProblemSelfDirectedLearn                            -0.090156   0.004945
## StressResistance                                     0.160587   0.005382
##                                                     t value Pr(>|t|)
## (Intercept)                                           3.539 0.000402 ***
## Grade8                                               -0.683 0.494488
## Grade9                                               -1.339 0.180434
## Grade10                                              -1.662 0.096599 .
## Grade11                                              -2.355 0.018533 *
## Grade12                                              -2.201 0.027718 *
## GenderMale                                           11.910  < 2e-16 ***
## CountryFrance                                       -17.398  < 2e-16 ***
## CountryHong Kong                                     -6.675 2.51e-11 ***
## CountryHungary                                      -12.396  < 2e-16 ***
## CountryIreland                                      -11.321  < 2e-16 ***
## CountryMacao                                        -19.916  < 2e-16 ***
## CountryNetherlands                                   -8.193 2.62e-16 ***
## CountryNew Zealand                                   -5.650 1.61e-08 ***
## CountrySlovenia                                     -13.386  < 2e-16 ***
## CountrySpain                                        -29.758  < 2e-16 ***
## HighestParentalEdDoctoral or equivalent               0.499 0.617582
## HighestParentalEdLower Secondary                      2.065 0.038937 *
## HighestParentalEdMaster's or equivalent              -2.031 0.042249 *
## HighestParentalEdNo Formal Education                  0.529 0.596903
## HighestParentalEdPost Secondary non Tertiary          0.560 0.575589
## HighestParentalEdPrimary Education                    0.963 0.335527
```

```
## HighestParentalEdTwo Years of Tertiary                    2.104 0.035426 *
## HighestParentalEdUpper Secondary No Access to Tertiary    2.616 0.008906 **
## HighestParentalEdUpper Secondary w Access to Tertiary     1.809 0.070479 .
## Belongingness                                            34.186  < 2e-16 ***
## Curiousity                                                3.053 0.002269 **
## Perseverance                                             19.266  < 2e-16 ***
## FamilySupport                                            34.922  < 2e-16 ***
## ProblemSelfDirectedLearn                                -18.232  < 2e-16 ***
## StressResistance                                         29.839  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9202 on 39810 degrees of freedom
## Multiple R-squared:  0.1818, Adjusted R-squared:  0.1812
## F-statistic: 294.8 on 30 and 39810 DF,  p-value: < 2.2e-16
```




Q−Q Residuals
Standardized residuals
−4
−2
0
2
−4
−2
0
2
4
Theoretical Quantiles
lm(WellBeing ~ Grade + Gender + Country + HighestParentalEd + Belongingness ..


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−2
−1
0
1
2
Fitted values
lm(WellBeing ~ Grade + Gender + Country + HighestParentalEd + Belongingness ..



Random Forest Algorithm

Prior to running our random forest algorithm, I need to partition our dataset into a training and test set. The training set is 10% of our entire dataset. I set the seed to 2024 (the current year).

Prior to running the random forest, I need to tune our random forest algorithm with our training set to determine the the optimal number of variables that are randomly sampled at each split (i.e., mtry). I used all of the features in the dataset to run our model to the outcome of Well-Being (similar to our Model 3 regression). The code below was partially used from the edX Harvard Course (Irizarry, 2024) and the Introduction to Data Science textbook for the course (Irizarry, 2024).

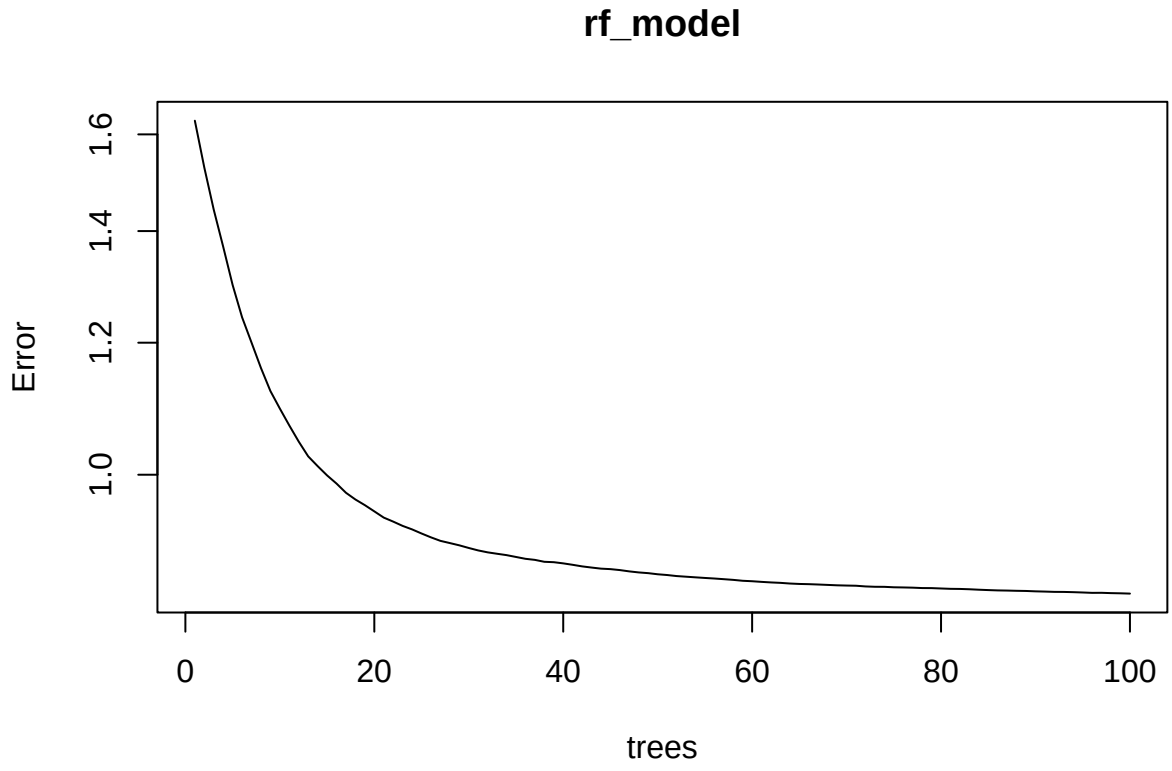
Due to time, I set the ntree to 100, which is the number of branches that will "grow" from each split. I used RMSE as our metric for evaluation. I set a control of cv, which seperates the model into k-folds a specified number of times (I set to 3). The result of our tuning algorithm, I found the optimal mtry was 16, with the lowest RMSE = .915, R-Squared = 19, and MAE at .73.

```
## Random Forest
##
## 35855 samples
##    11 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 23903, 23903, 23904
## Resampling results across tuning parameters:
##
```

```
##   mtry  RMSE       Rsquared   MAE
##    2    0.9282316  0.2034095  0.7521654
##   16    0.9158821  0.1915191  0.7389269
##   31    0.9208402  0.1843264  0.7426414
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 16.
```

Finally, I performed a random forest algorithm model with the training set with all of features with 100 ntrees and mtry set to 16. I also set Importance to true to be able to assess the importance of the included features (to answer our second research question).

```
##
## Call:
##  randomForest(formula = WellBeing ~ ., data = train, Importance = TRUE,      ntree = 100, mtry = 16)
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 11
##
##           Mean of squared residuals: 0.8486277
##                     % Var explained: 18.11
```



The algorithm found the RMSE to be at .92, which is lower than our threshold at 1. I also found the variance explained for algorithm to be at 18%.

```
## [1] 100
```

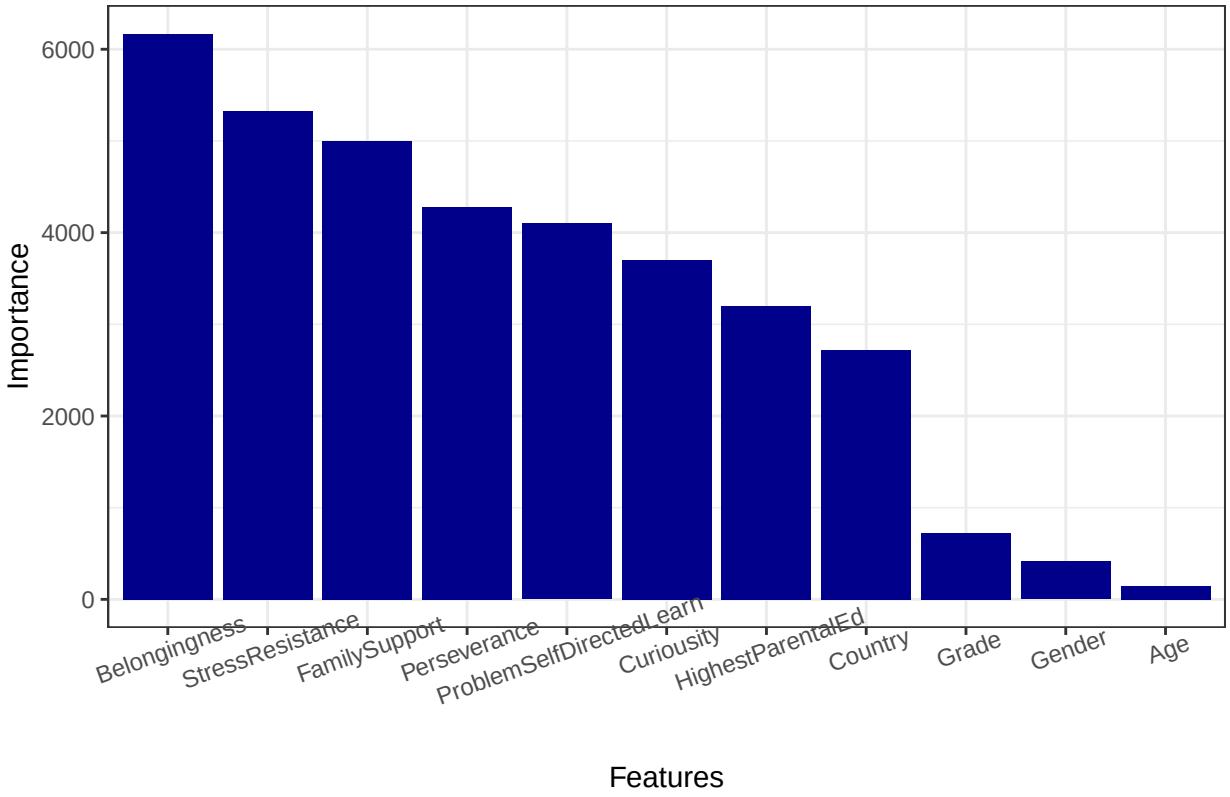
```
## [1] 0.9212099
```

To assess model importance, I isolated importance from the random forest model. I found that Belongingness, Stress Resistance, Family Support, Perseverance, and Problem with Self Directed Learning were the most important predictors of adolescent well-being. See Figure 21 for sorted features by importance.

```
##                                       Feature Importance
## Gender                                 Gender   411.7498
## Grade                                   Grade   721.0858
## HighestParentalEd               HighestParentalEd  3199.4755
## StressResistance                 StressResistance  5323.3518
## FamilySupport                       FamilySupport  5001.3819
## ProblemSelfDirectedLearn ProblemSelfDirectedLearn  4098.0843
## Belongingness                       Belongingness  6169.6799
## Curiousity                             Curiousity  3699.9088
## Perseverance                         Perseverance  4275.6962
## Age                                           Age   143.2716
## Country                                   Country  2717.1654
```

```
## [1] "Belongingness"            "StressResistance"
## [3] "FamilySupport"            "Perseverance"
## [5] "ProblemSelfDirectedLearn"
```

Figure 21: Feature Importance



Finally, I set the seed to 2024 (again) and trained a new random forest model on the reduced dataset. This resulted in a test RMSE of .93, which was higher than what I found in our initial random forest model.

```
## [1] "Test RMSE: 0.940191905677889"
```

Conclusion

The sample had a large proportion of 10th grade students and 16 year olds. A large proportion of students were from Spain as their country of origin. Results from Model 1 show that gender and country origin did significantly predict adolescent well-being. Model 2, show that all of the psychosocial predictors contribute to adolescent well-being.

However, after running our random forest algorithm, I determined the most important features.In this order, Belongingness, Stress Resistance, Family Support, Perseverance, and Problems with Self-directed Learning all were the top features to predict well-being, more so than Curiosity and all the demographic features included.

The potential impact from this finding is that students who felt supported by family, more belongingness (or lack there of), more perseverent, and have higher problems with self-directed learning all related to their well-being. Principals and educators need to consider the learning environments that students are in that encourage a sense of belongingness, parental involvement in a student's life, and the internal factors that can predict well-being in students. This is helpful information for educational practitioners and researchers interested in continuing to find relationships with psychosocial strength variables in relation to well-being in school environments.

Limitations & Future Directions

This report was cross-sectional in nature, and does not report longitudinal trends over time. Likewise, using Spain as the reference group makes us interpret all of the included countries to students in Spain. More research should look at the important features found in relation to other countries as the reference group. One of the most notable limitations is the usage of multiple linear regression with a large dataset. The models were overpowered by a large sample size, and there were notable violations of linearity and outliers that would need to be assessed. Luckily, our predictors did not have several levels and they were able to run. I also need to perform a correction for multiple comparisons for the regression models. The higher RMSE found in the reduced test set shows that I may be overfitting our solution, more replications should be conducted to verify.

Additional more advanced multivariate modeling is needed to account for nested structures. For example, there is teacher level variance that could be accounted for. Future research should also combine datasets by year and examine similar variables over time to analyze specific trends and impacts of the Pandemic. I'd also like to look at the the ICT items related to technology usage and well-being. Additionally, PISA provides item level data for each of their measures. I would like to conduct a machine learning algorithm using an unsupervised machine learning model like PCA to evaluate the item level variance and overall factor structure of each of the subtests, and how these compare to the original measurements included in the student questionnaires. I'd also like to examine potential measurement invariance by country and/or other demographic variables (e.g., gender or socioeconomic status). Other features I was interested in was specifically creativity items. I waited for the release of new variables in June 2024 which were items related to creativity and creative learning environments. Future research should look at creativity in relation to well-being and the other psychosocial features included.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300. http://www.jstor.org/stable/2346101

Breiman, L. (2001). Random forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2022). randomForest: Breiman and Cutler's random forests for classification and regression. R package version 4.7-1.1. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Erikson, E.H. (1968). Identity: youth and crisis. Norton & Co..

Granic, I., Morita, H., & Scholten, H. (2020). Beyond Screen Time: Identity Development in the Digital Age. Psychological Inquiry, 31(3), 195–223. https://doi.org/10.1080/1047840X.2020.1820214

Irizarry, R. A. (2024). Introduction to data science. https://rafalab.dfci.harvard.edu/dsbook/

Kuhn, M. et al. (2023). caret: Classification and regression training. R package version 6.0-94. https://cran.r-project.org/web/packages/caret/index.html

Nemiro, A., Hijazi, Z., O'Connell, R. Coetzee, A.,Snider, L. (2022) Mental health and psychosocial wellbeing in education: The case to integrate core actions and interventions into learning environments. Intervention 20(1), 36-45. https://doi.org/10.4103/intv.intv_20_21

OECD.(2023). PISA 2022 Results (Volume II): Learning during- and from - Disruption. Organisation for Economic Co-operation and Development (OECD). https://www.oecd.org/publications/pisa-2022-results-volume-ii-a97db61c-en.htm

OECD.(2024). PISA 2022 Database. Programme for International Student Assessment (PISA). https://www.oecd.org/pisa/data/2022database/

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Scully, M., Swords, L., & Nixon, E. (2023). Social comparisons on social media: online appearance-related activity and body dissatisfaction in adolescent girls. Irish Journal of Psychological Medicine, 40(1), 31–42. https://doi.org/10.1017/ipm.2020.93

Subotnik, R. F. (2015). Psychosocial Strength Training: The Missing Piece in Talent Development. Gifted Child Today, 38(1), 41-48. https://doi.org/10.1177/1076217514556530

Tsang, K.L.V., Wong, P.Y.H. and Lo, S.K. (2012), Assessing psychosocial well-being of adolescents: a systematic review of measuring instruments. Child: Care, Health and Development, 38, 629-646. https://doi.org/10.1111/j.1365-2214.2011.01355.x

UNESCO. (2023). What you need to know about education for health and well-being. https://www.unesco.org/en/health-education/need-know?hub=79846

UNESCO. (2024, April). UNESCO report spotlights harmful effects of social media on young girls.https://news.un.org/en/story/2024/04/1149021

Wickham, H., Vaughan, D., & Girlich, M. (2024). tidyr: Tidy messy data. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org.

Wickham, H., Miller, E.,Smith, D., & Posit Software (2023). haven: Import and export 'SPSS', 'Stata', and 'SAS' files. R package version 2.5.4. https://haven.tidyverse.org/reference/read_spss.html

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham et al. (2014). dplyr. https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html

Wong, J. (2024, April). Social media hurts girls' mental health and education potential, says UNESCO report. CBC News. https://www.cbc.ca/news/unesco-gem-technology-social-media-1.7184717