STAT 230 Final Project

Veena Advani, Annie Chen, Robert Tung

Introduction

This analysis will focus on the Teaching and Learning International Survey (TALIS) Dataset, coordinated by the Organization for Economic Cooperation and Development (OECD), an intergovernmental organization of industrialized countries. The TALIS dataset comes from a survey of the teaching workforce that aims to collect information about teaching as a profession, the working conditions of teachers, and the learning environments of schools. It also aimed to provide a more global view of education systems, and metrics that could be used to compare education systems from different countries. TALIS was administered in 2008 and 2013. We are using the dataset from 2013, which contains answers from teachers from more than 10,000 schools in 36 different industrialized countries. Not all of the countries were English speaking. The TALIS dataset contains several different files, which are split by school type. Although the dataset contains information from a variety of school types (primary, secondary, etc), most of the data was on lower-secondary schools, so we only used that file. At each school that was selected to participate, the principal and a random selection of up to 22 teachers were chosen to complete voluntary online questionnaires. We only analyzed teacher responses.

1. Our 1st question is: Which variables and classroom factors are the best predictors of teacher job satisfaction?

There are several reasons motivating this question. Education is incredibly important to any country's development and functioning. However, countries vary in how much the teaching profession is respected, and how well education systems are funded. As a result, teacher working conditions differ across countries. It's helpful to understand what kinds of environments keep teachers more satisfied, so that educational systems can encourage and build these types of environments. If teacher satisfaction is too low, it becomes difficult to retain a large enough teaching force, and to attract talented people to the field. Therefore, we want to understand how teacher satisfaction varies across the 36 countries in the data set, and which aspects of a school and environment correlate most with teacher satisfaction.

2. Our 2nd question is: How do teacher satisfaction, and some of the significant predictors found from question 1, vary by country?

Because the TALIS dataset allows cross country comparison, we wanted to look at how some of the variables in the dataset varied by country. Looking at how teacher satisfaction varies by region could be particularly helpful to educators and policy makers. There are a lot of environmental factors that can affect teacher satisfaction, and they might not all be addressed in the TALIS questionnaire. However, if we can identify which countries have particularly satisfied teachers, then perhaps people who are looking to increase teacher satisfaction can further research how schools work in those countries, past the information covered in the questionnaire. Additionally, if we are able to identify any variables that are particularly helpful in predicting teacher satisfaction (from research question 1), we want to see if the countries that have a higher average teacher satisfaction, also have good values for those significant variables. Or in other words, after identifying qualities that seem to correlate with high teacher satisfaction, do the more satisfied countries seem to have those qualitites? Or does it seem like other factors, that are possibly not in the data set, are contributing to the higher levels of teacher satisfaction in those countries?

Data

Each row of the dataset corresponds to single teacher's response to the questionnaire (There are 117876 rows). The original dataset had 532 columns, which corresponded to either a particular question in the questionnaire, or a summary variable for a section of the questionnaire. There are columns concerning employment status, personal information/background, time spent, professional development, efficacy and amount of feedback, outlook about the school, school climate, overall job satisfaction, and much more. Almost all of the questions in the questionnaire were multiple choice or numerical input questions.

After picking a subset of those columns and creating some of our own, we were left with 40 columns, resulting in a dataset that has 117876 rows and 40 columns. The variables we are using are:

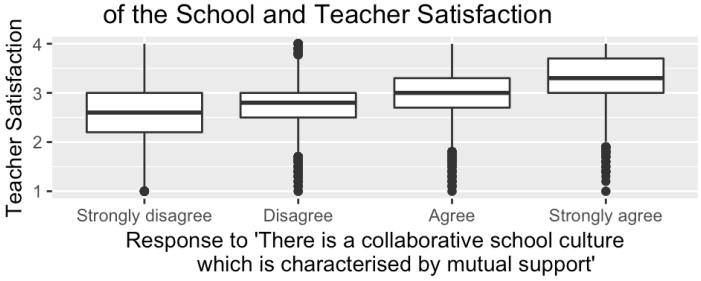
- `country (CNTRY)`: Country of teacher (factor variable, 36 levels)
- `region`: Region the teacher's country is in, based on World Health Organization regions (factor variable, 8 levels)
- `gender (TT2G01)`: Teacher Gender (factor variable, 2 levels)
- `age (TT2G02)`: Teacher Age (numeric variable)
- `years_teacher (TT2G05B)`: Years of teaching experience (numeric variable)
- `age_ratio`: years_teacher divided by age, or in other words, fraction of life spent teaching (numeric variable)
- `train_stat (TT2G11)`: Training level of teacher (factor variable)
- `prep_A/B/C (TT2G13A/B/C)`: How prepared the teacher felt with respect to 3 different areas of teaching, which were content, pedagogy and classroom practice respectively (ordered factor variable, 4 levels)
- `avg_prep`: The average of each teachers responses for prep_A/B/C (numeric variable between 1 and 4)
- `total_time (TT2G16)`: Total time spent on teaching related activities in hours per week (numeric variable)
- `feedback_salary (TT2G30G)`: Whether performance feedback influences salary (ordered factor variable, 4 levels)
- `collab (TT2G44E)`: How collaborative is the school culture? (factor variable, 4 levels)
- `num_students (TT2G38)`: Number of students in the teacher's class (numeric variable)
- `satisfaction (TJOBSATS)`: Self rated teacher job satisfaction (ordered factor variable, 4 levels)
- `Various Subject Names (TT2G15[A - L])`: 12 columns representing which subjects the teacher teaches (boolean numeric variable)
- `num_subjects`: The number of subjects the teacher teaches (numeric variable)
- `sat1-sat10 (TT2G46[A - J])`: Ten questions related to teacher job satisfaction (ordered factor variables, 4 levels)
- `sat_avg`: An average of the ten questions related to teacher job satisfaction (numeric variable). This is our repsonse variable for research question 1.

Initial Plots Describing Data

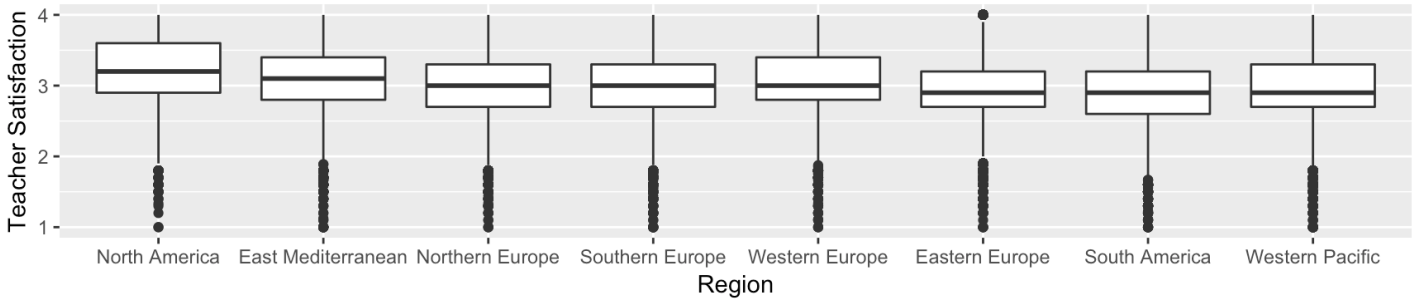
Below are three plots that show our data and hopefully motivate our research questions and the analysis.

Plot 1: Histogram of Teacher Satisfaction

Plot 2: Relationship Between Collaboration Level of the School and Teacher Satisfaction



Plot 3: Teacher Satisfaction by Region



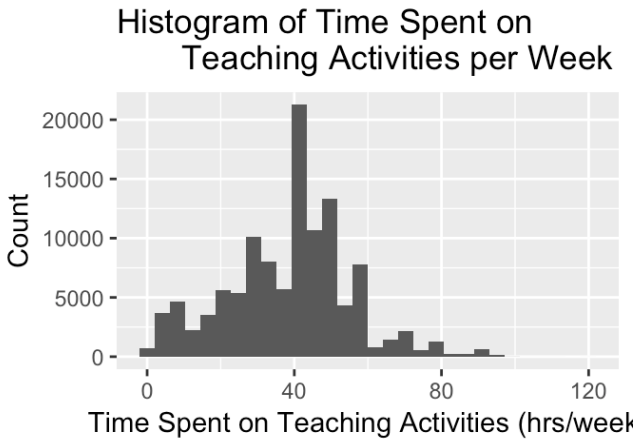
Plot 1 As we are later analyzing the effect of different variables on teacher satisfaction, and the variation between different regions, it is important to note the overall distribution of our response variable.

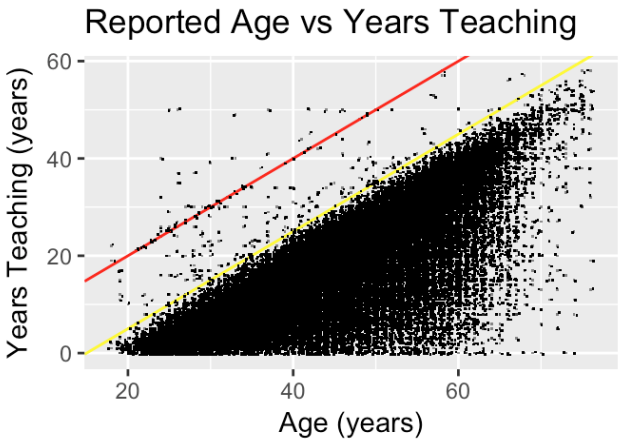
Plot 2 provides a box plot of collab vs sat_avg. This plot helps answer our first research question, in which we assess which variables are particularly predictive of overall teacher satisfaction. We can see a strong visual trend between how collaborative a teacher viewed his/her school environment to be, and how satisfied that teacher was. We will provide similar plots for other variables in the analysis of our first research question.

Plot 3 provides a box plot of geographic region vs sat_avg. This is relevant to our 2nd question, which looks at variations in sat_avg and other variables across geographical regions. Visually, while all median values for sat_avg hover around the range of 2.5 to 3.5, there is still a noticeable difference in some of the distributions (e.g. the median and both quartiles of the satisfaction value for North America is noticeably above that of Eastern Europe).

Removal of Unusual Values

In a previous submission, we cleaned the dataset with respect to missing values and obvious errors. However, when considering the context of some of our variables, some of the entries are still unusual. Let's look at 2 plots that illustrate this





We can see from the histogram on the left, that most teachers seem to have a 40 hour work week, which is a pretty typical time comittment. There is another dropoff after 60 hours/week, which is a reasonable addition to the 40 hour workweek for teachers that spend a fair amount of time class planning and grading in the evenings. Some teachers, however, are claiming to have spent over 90 hours per week on teaching related activities. If a teacher worked from 6 am to 9 pm Monday to Friday (15 hours a day) from time at school and grading in the evening, and another 15 hours of related teaching activities on the weekend such as class planning and grading, they would be at 90 hours in a week. Although this is physically possible, it seems highly unlikely except for in unusual situations like boarding school. Let's see how many teachers reported 90 hours a week or more.

```
nrow(teacher_data[teacher_data$total_time==90 & !is.na(teacher_data$total_time),])
```

```
## [1] 585
```

```
nrow(teacher_data[teacher_data$total_time>90 & !is.na(teacher_data$total_time),])
```

```
## [1] 210
```

We can see that there are many more values of exactly 90 than greater than 90. So we keep the values of 90, but set all values that are higher than 90 hours a week to NA.

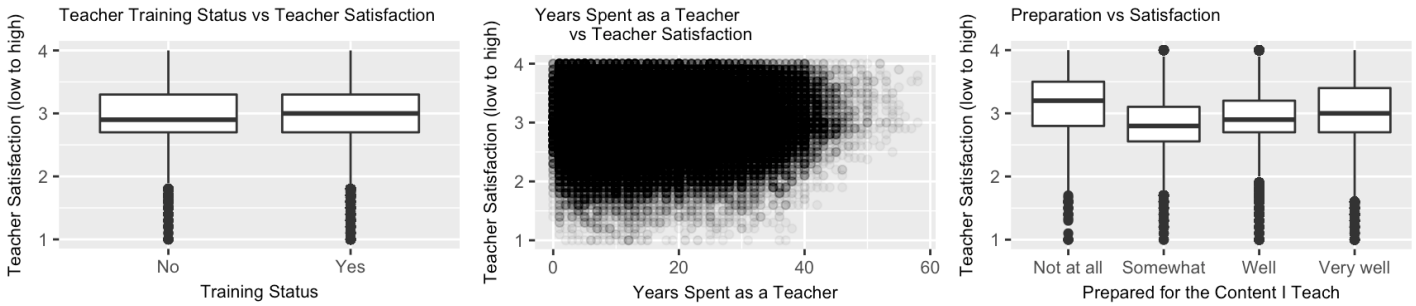
```
teacher_data[teacher_data$total_time > 90 & !is.na(teacher_data$total_time), "total_time"] <-
 NA
```

Now, looking at the plot on the right, we can see that there are points above the red line, which are teachers who reported they had been teaching longer than they had been alive. Indeed, let's see how many responses reported years that would indicate that they had started teaching before age 15 (points above the yellow line), and set those values to NA.

```
age_start <- teacher_data$age - teacher_data$years_teacher
teacher_data[age_start < 15 & !is.na(age_start), "years_teacher"] <- NA
```

Analysis for Research Question 1 - Linear Regression

To address our first research question, and to try to identify which variables are the best indicators of teacher job satisfaction, we are going to create a linear model. Although our earlier plot of satisfaction against the degree of collaboration at the school shows a clear linear trend, let's look at several more plots of the data against our response variable, before diving into any linear regression models:



From these plots we can see that teacher's training status doesn't seem to have a huge relationship with teacher satisfaction, however the median of people who did receive some type of training is slightly higher than those who didn't have any training. There also appears to be some linear relationship between years spent teaching and teacher satisfaction in the center graph. And in the graph on the right, there does seem to be differences in teacher satisfaction between different levels of preparedness. Consequently, it does seem to be worth exploring a linear model, where we try to use all of our predictors to predict teacher satisfaction.

Dataset for Linear Regression Model: To start, for our linear regression model, we are going to remove columns with a large number of NAs, the subject data columns (we're only interested in the number of subjects each teacher is teaching, not the specific subjects), and the 10 satisfaction related questions, as they were used to create our response variable.

One of the things we are interested in is how geography impacts teacher satisfaction, however we have two different geography variables. We can't include both in the model, as they are linearly dependent, and R will only find a set of coefficients for one of the two variables if we put both in the model. Because of this, let's try both and compare the results:

Linear model using country (summary has been left out for space):

```
m_country <- lm(sat_avg ~ ., data=x[,-which(names(x) %in% c("region"))])
summary(m_country)$r.squared
```

```
## [1] 0.2449442
```

Linear model using region (summary has been left out for space):

```
m_region <- lm(sat_avg ~ ., data=x[,-which(names(x) %in% c("country"))])
summary(m_region)$r.squared
```

```
## [1] 0.1982519
```

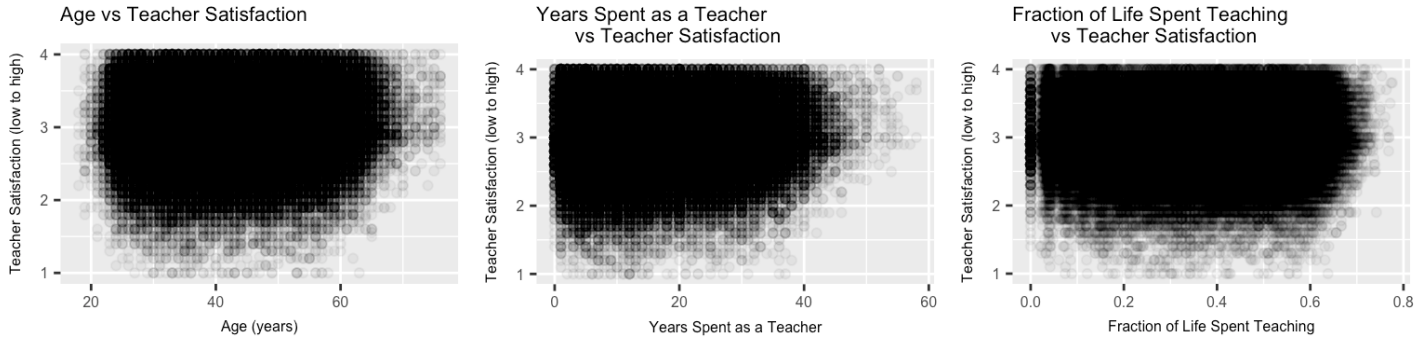
If one looks at the two summaries (Summaries 1 and 2 in the Appendix), one can see that a large amount of the country categories have coefficients with significant p-values in the first model, and all of the region categories have coefficients with significant p-values in the second model. However, the R-squared value for the model using country data was 0.2449, and the R-squared value for the model with the region predictor was 0.1983. This means R-squared dropped by roughly 19%, which is quite a bit. Thus, we use the country data in the linear model. We can also see that train_stat was not significant in either of the models, so let's remove that from future models.

Lastly, it seems very possible that age and years_teacher are correlated, since teachers that have been teaching for longer are likely to be older. As a result, we probably shouldn't use both predictors in our linear model. Let's check the correlation:

```
## [1] 0.8462086
```

Age and years_teacher are pretty highly correlated (correlation of 0.846). As a result, let's see if there's another way we can use our age and years_teacher data, and combine it into one metric. One option is the ratio of years_teacher to age. It's possible that the fraction of your life that you have spent teaching, affects how satisfied you are.

Let's see whether these new predictors, or one of our older predictors, seems to have the clearest linear relationship with our response variable, sat_avg:



From the plots above, it looks like years_teacher has the clearest linear relationship with sat_avg, so we're going to use that predictor in our linear model.

Let's also make sure that the prep_A, prep_B, and prep_C predictors are not too highly correlated.

```
##           prep_A    prep_B    prep_C
## prep_A 1.0000000 0.6463396 0.6254094
## prep_B 0.6463396 1.0000000 0.7381657
## prep_C 0.6254094 0.7381657 1.0000000
```

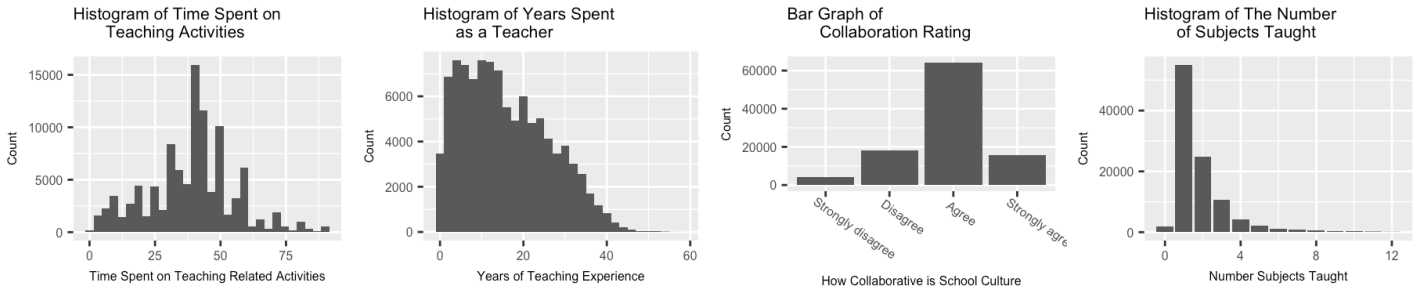
We can see from the correlation matrix above that all three prep questions are pretty correlated, and therefore we shouldn't put all three in our linear model. Let's try creating a single predictor that's the average of all three, in order to indicate teachers' overall sense of preparedness.

```
teacher_data$avg_prep <- apply(data, 1, mean)
```

Now let's make a new dataset with all of the variable selections we just made. So we need to remove region, train_stat, age, age_ratio, and all three prep predictors.

```
x.new <- teacher_data[, -c(sat_indices, na_cols, subject_indices, prep_indices)]
x.new <- na.omit(x.new[, -which(names(x.new) %in% c("train_stat", "region", "age", "age_star
t", "age_ratio"))])
```

Now let's see if all of the variables that we've decided to use for the linear model, are normally distributed, or if some transformations might be helpful:We can see from the histogram of sat_avg (Plot 1, displayed in the Dataset section), that although it isn't perfectly normal, it is close enough that it probably doesn't warrant any transformation of the variable.



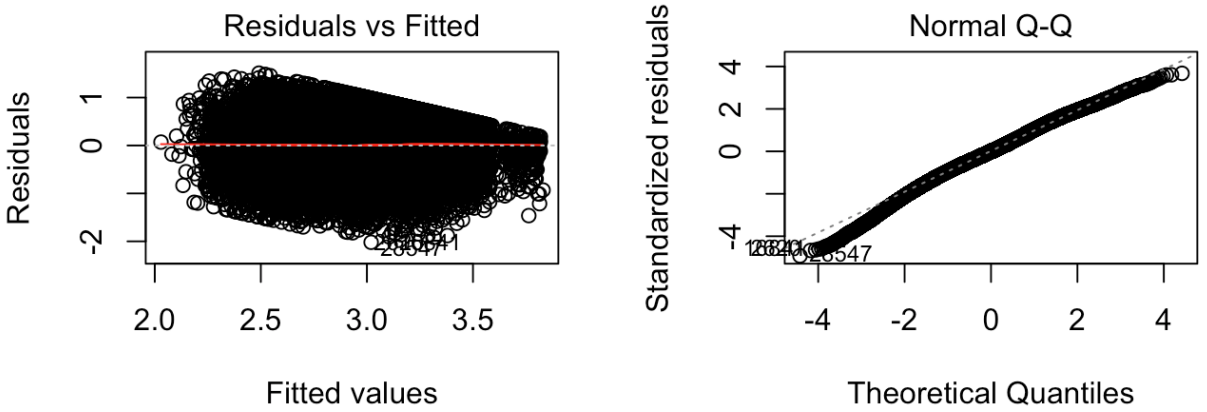
Three of the plots above, although not perfectly normal, don't have a huge amount of skew. We don't need to transform any of these variables. The number of subjects histogram does have a lot of skew. However, we tried several different transformations, and none of them seemed to help much. After the creating the model, we will look at the normal Q-Q plot to make sure the normality assumption wasn't violated too significantly.

Now that we have a new data set with all of our variable selections, let's make a final model (summary in Appendix):

```
m_final <- lm(sat_avg ~ ., data=x.new)
summary(m_final)$r.squared
```

```
## [1] 0.2400472
```

Before analyzing this model, let's make sure that the assumptions for linear regression hold.



We can see from the two plots above that there are no trends in the residuals, and the Normal Q-Q plot is close to a straight line, so our assumptions hold.

General Analysis of the Final Linear Model:

Our final model used the dataset x.new, which had 102186 observations, and 8 columns (7 predictors and the response variable). That means that x.new had roughly 87% of our original observations.

Looking at the model summary (Summary 3 in Appendix), our final linear model has an R-squared value of 0.24. Many of the countries have coefficients with significant p-values, and all of the other predictors have coefficients with significant p-values at a level of 0.001. However, this is not that surprising given how large our dataset is. We might be able to get more insight by analyzing the values of the coefficients.

To decide which predictors seem to be particularly helpful in predicting teacher satisfaction, we look at the magnitude of the coefficients in the model summary (summary 3 in Appendix) and the confidence intervals of the coefficients (Summary 4 in Appendix) and see how far away they are from zero.

The largest coefficients in the model in terms of magnitude were: collab.Sronglyagree, countryMEX, collabAgree, countryFIN, countryBFL, collabDisagree, countryMYS, and countryNLD. All the other coefficients had magnitudes that were less than 0.2.

The confidence interval for the coefficient of num_subjects is close to zero (it has a range of 0.0017 to 0.0054). The confidence interval for the coefficient gender is also pretty close to zero, with a range of -0.027 to -0.016. Lastly, two of the confidence intervals closest to zero are for the coefficient of years_teacher (0.00045, 0.00095) and the coefficient for total_time (-0.00063,-0.00031). Given the small magnitudes of the coefficients compared to the scale of teacher satisfaction, and how close the confidence intervals of these four predictors are to zero, they do not seem that important to predicting teacher satisfaction.

On the other hand, the coefficients for collab.Stronglyagree, collab.Agree, collab.Disagree and avg_prep had confidence intervals of roughly (0.743, 0.771), (0.460, 0.487), (0.221, 0.250) and (0.11, 0.12) respectively, all of which are intervals that are a decent amount away from 0. Given that teacher satisfaction was on a scale of 1 to 4, a coefficient for one of the categories of the collab variable with a magnitude of around 0.7 or 0.5 seems pretty influential, since an intercept shift of 0.7 is almost an entire point in teacher satisfaction. A coefficient of a little over 0.1 also seems significant relative to how small many of the other coefficients were. Consequently, it looks like both collab and avg_prep were useful predictors for predicting teacher satisfaction.

Moreover, we can see that several of the countries have coefficients with pretty large magnitudes, and confidence intervals for those coefficients that are pretty far from zero. For example, the confidence interval for the coefficient for countryMEX is (0.4494836350, 0.4990107641) and for countryFIN is (0.2306808115, 0.2792257572). In fact, 15 countries have coefficients with a magnitude of at least 0.1, and confidence intervals for those coefficients that are also at least 0.1 away from 0. This suggests that country is also significant in predicting teacher satisfaction.

Lastly, let's look at the sign of the coefficients in the model, to see the direction of the relationships between the predictors and sat_avg. Of course we should keep in mind the fact that there are so many variables in this model and thus the coefficients may not completely explain the relationship of the variables.

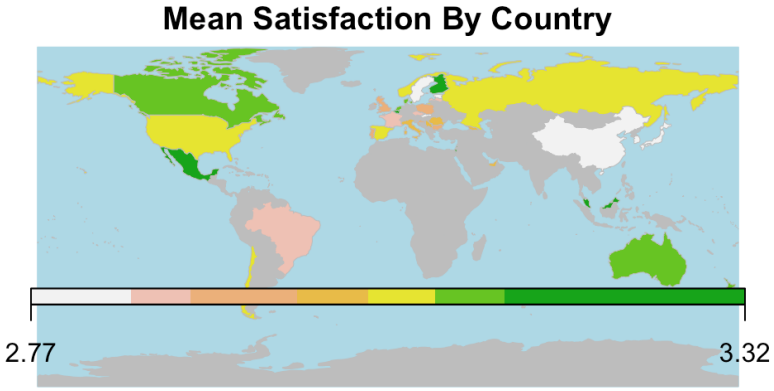
years_teacher has a very small positive coefficient (0.0006979) with sat_avg. A positive correlation may make sense, as those who are satisfied with being a teacher are more likely to stay for longer. However, this coefficient is very small. total_time has a very small negative coefficient (-0.0004706). This may be because of the duality between those who are overworked being unsatisfied, while also addressing that those who are satisfied with their jobs may be more inclined to work more. From this coefficient alone it seems that the first phenomenon may be more prominant. The collab coefficients actually get larger and larger with higher levels of belief that the work environment is collaborative. This makes a lot of sense, as a more collaborative work environment is generally thought of as a positive aspect and thus may be positively correlated with satisfaction. It seems that there is also a positive coefficient for the number of subjects taught and a negative coefficient for gender being male, which do not have any obvious intuition either for or against, but are interesting. One potential explanation for number of subjects taught could be that a satisfied teacher is willing to put in more work and teach more than one subject. Finally, there is a positive coefficient for the average amount of preparation, which may make sense as those who are satisfied may be more inclined to prepare more and those who feel prepared may be more satisfied.

Conclusion: This linear model suggests that teachers' perceptions of how collaborative the school culture is, as well as how prepared the teachers feel, are significant predictors for how satisfied teachers are. Interestingly, although how prepared teachers felt was significant in predicting teacher satisfaction, the degree of teacher training was not. It is

possible that this is because it was a “Yes” or “No” question on the survey, and therefore we didn’t have more information on the quantity or quality of training that teachers received, whereas teachers were able to rate how prepared they felt in several areas on a scale of “Not at all”, “Somewhat”, “Well”, and “Very well”. Lastly, country was very useful in predicting teacher satisfaction. This could be due to a variety of reasons, including cultural values and attitudes at work, school legislation and restrictions, funding of schools, the ratio of public to private schools, and the respect that the teaching profession receives, all of which vary between countries, and are environmental factors at the country level, rather than the kind of information available in a teacher survey.

Analysis for Research Question 2

Given how useful country was in predicting teacher_satisfaction, it would be interesting to see whether there are any significant regional differences in some of the other significant predictors from our linear model above. For example, in the regions with higher teacher satisfaction, are collaboration ratings also higher? Therefore, let’s now address our second research question, and try to identify the differences across different regions for teacher job satisfaction and other classroom factors. First, lets look at a map of average teacher satisfaction by country.



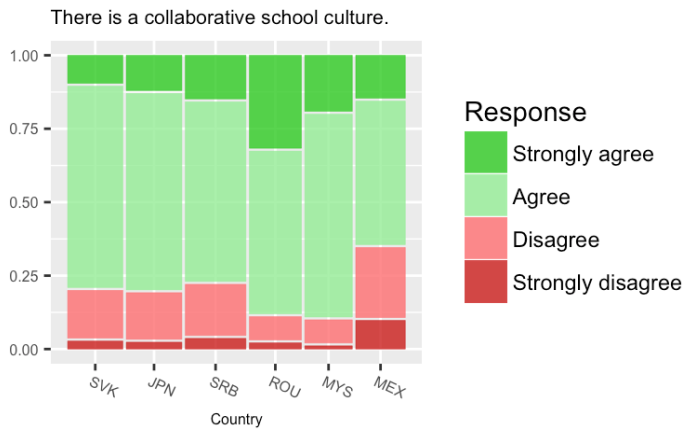
Below is also an anova test showing that there is a significant difference between the mean satisfaction of at least two of the countries.

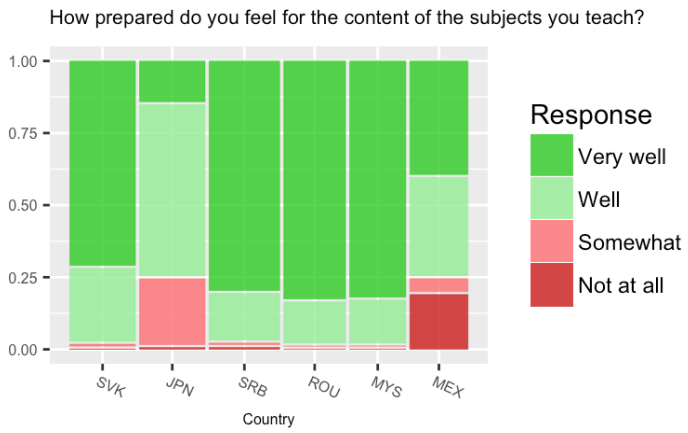
```
## Analysis of Variance Table
##
## Response: sat_avg
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## country       35  1962.7  56.077  273.27 < 2.2e-16 ***
## Residuals 113964 23386.5   0.205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we first began analyzing factors by country, we grouped the countries into 8 regions using the World Health Organization geographical categories. The difference in mean satisfaction was still statistically significant (p-value < 0.05) across all region pairs other than Western Europe vs Eastern Europe, as shown in the pairwise.t.test in the appendix.

However, we felt that these geographical groups didn’t capture all the factors that might influence teacher satisfaction by country (for instance, China and Australia were grouped together geographically, but have vastly different governments). Thus, we took the 2 most satisfied countries (Mexico and Malaysia), 2 least satisfied countries (Slovak Republic and Japan) , and two countries in the middle of the list (Serbia and Romania), and decided to analyze that subset of countries, which well give us unsight into how these factors change with more and less satisfied countries. (See “Selected Countries” in Appendix)

Based on the predictors affected job satisfaction in Part 1, we can see which countries stand out, either positively or negatively, in those factors.



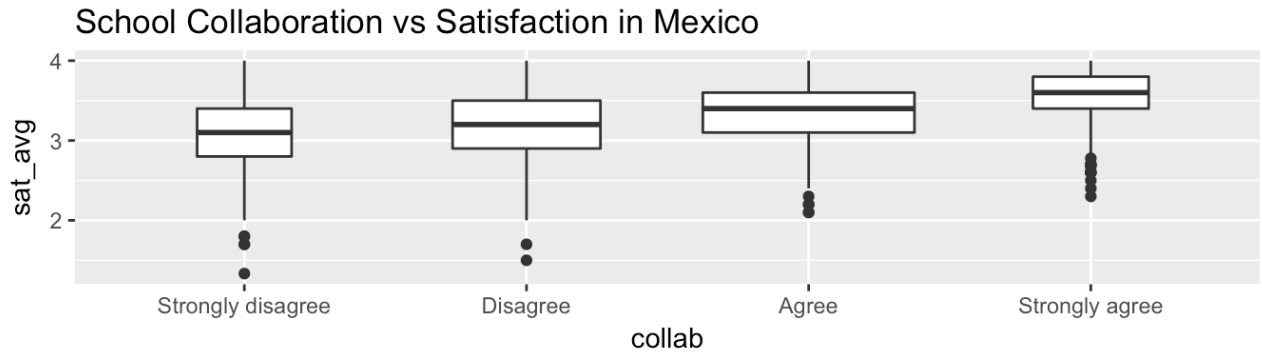


Indeed, from an anova we can see that collab and prep_A are themselves significant in variation by country.

```
## Analysis of Variance Table
##
## Response: as.numeric(collab)
##               Df Sum Sq Mean Sq F value    Pr(>F)
## country       35   2207  63.057  135.55 < 2.2e-16 ***
## Residuals 112795  52472   0.465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: as.numeric(avg_prep)
##               Df Sum Sq Mean Sq F value    Pr(>F)
## country       35   6973 199.238  676.72 < 2.2e-16 ***
## Residuals 113598  33445   0.294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the first plot, there isn't really any trend, but Mexico jumps out as an outlier. It's the country with the highest teacher satisfaction, but almost 30% of teachers say their school isn't collaborative. Does this mean that collaboration doesn't matter as much to satisfaction in Mexico?



Teacher satisfaction is still correlated positively with teacher satisfaction in Mexico, so there must be some other factor that makes the average satisfaction in Mexico higher, other than having a higher level of collaboration.

Interestingly, Mexico is also an outlier in the second plot, where 30% of teachers did not feel prepared to teach their subject well.

The takeaway is that although there are predictors that correlate positively with higher teacher satisfaction, like collaborative schools, and countries that have significantly higher teacher satisfaction than others, like Mexico, the countries with more satisfied teachers don't necessarily have the most collaborative schools. Teacher satisfaction is complex, and we have explored a very small part of the picture.

Conclusion:

Overall, our models didn't completely explain the variation in the data. We expected real datasets to be messier than the ones we had dealt with in problem sets, and so we were pleased with the observations we did find.

The first question we were interested in for this project was how satisfaction was related to the different factors, and as mentioned previously it seems that the most significant factors by linear model coefficient were collab, avg_prep, and country. In fact, we noted that many of the country coefficients were very large, implying that satisfaction varies quite a bit by country.

This brings us to our second research question, which addressed how various factors changed across countries. We first confirmed the above finding with a graph of satisfaction for each country, and supported this with an anova. Then, we also noted that some of the factors that were important to satisfaction, in particular level of collaboration and amount of preparation were themselves also significantly different across countries.

Before we started looking into the data, we expected there to be a lot more unsatisfied teachers than there actually were. Indeed, if we take `sat_avg` scores of above 2.5 to mean that the respondent is generally satisfied (if on average, they responded Agree or Strongly Agree more than Disagree or Strongly Disgree on the teacher satisfaction questions), then

```
nrow(teacher_data[teacher_data$sat_avg >=2.5,])/nrow(teacher_data)
```

```
## [1] 0.8844549
```

roughly 88% of our respondents were satisfied. Of course, this is a very low baseline (it just means the teacher didn't respond "Agree" or "Strongly Agree" to a question like "I regret that I decided to become a teacher.") If we were to collect data specifically to understand how various factors affect teacher satisfication, we would want to get a more precise sense of how satisfied or unsatisfied teachers are about specific aspects of teaching. We might ask respondents to our survey to rate, on a scale from 1 - 10, their satisfaction with things like working conditions, pay, sense of impact, and more.

As another problem, we don't know how many teachers were sampled comparatively by country - we know it was no more than 22 per school, but were unable to analyze how representative our sample was of a country as a whole. In addition, there were some country codes that were just cities (e.g. all of our teachers from China came from Shanghai).

As far as concerns about the data, when we did our data cleaning, we did find some suspicious responses. There were things we could say were impossible, such as working longer than one has been alive, but we could have other incorrectly reported responses that we weren't able to immediately mark as suspicious. Because all of the data was a self reported survey, inaccurate/falsely reported answers were pretty inevitable, and are still a concern about the data.

One area where our analysis would have benefited from other data sources would be in gathering more information about the countries. We created our `regions` variable using the World Health Organization regions categories, and ended up not using it because we wanted to capture the differences in individual countries. However, there are many other ways we could have tried to categorize the countries - for instance, because more collaborative school environments correlated with higher teacher satisfaction, we could have tried to group the countries by their rankings on the individualism vs collectivism scale, to see whether national attitudes made a statistically significant difference.

Code Appendix

Code Outputs

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  teacher_data$sat_avg and teacher_data$region
##
##                East Mediterranean Eastern Europe North America
## Eastern Europe  < 2e-16           -               -
## North America   < 2e-16           < 2e-16         -
## Northern Europe < 2e-16           < 2e-16         < 2e-16
## South America   < 2e-16           8.1e-06         < 2e-16
## Southern Europe < 2e-16           < 2e-16         < 2e-16
## Western Europe  0.00578           < 2e-16         < 2e-16
## Western Pacific < 2e-16           0.00045         < 2e-16
##                Northern Europe South America Southern Europe
## Eastern Europe  -                -             -
## North America   -                -             -
## Northern Europe -                -             -
## South America   < 2e-16          -             -
## Southern Europe 1.00000          < 2e-16       -
## Western Europe  < 2e-16          < 2e-16       < 2e-16
## Western Pacific 1.6e-07          < 2e-16       3.6e-05
##                Western Europe
## Eastern Europe  -
## North America   -
## Northern Europe -
## South America   -
## Southern Europe -
## Western Europe  -
## Western Pacific < 2e-16
##
## P value adjustment method: bonferroni
```

Summary 1

```
##
## Call:
## lm(formula = sat_avg ~ ., data = x[, -which(names(x) %in% c("region"))])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01995 -0.25710 -0.00271  0.27279  1.47375
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.423e+00  2.000e-02 121.121  < 2e-16 ***
## countryAUS            1.968e-01  1.330e-02  14.800  < 2e-16 ***
## countryBFL            2.536e-01  1.194e-02  21.245  < 2e-16 ***
## countryBGR           -2.543e-02  1.253e-02  -2.029 0.042440 *
## countryBRA           -2.131e-02  1.005e-02  -2.119 0.034096 *
## countryCAB            1.807e-01  1.367e-02  13.216  < 2e-16 ***
## countryCHL            1.156e-01  1.485e-02   7.787 6.92e-15 ***
## countryCSH           -1.535e-01  1.134e-02 -13.540  < 2e-16 ***
## countryCZE           -8.030e-02  1.180e-02  -6.807 1.01e-11 ***
## countryDNK            1.846e-01  1.405e-02  13.138  < 2e-16 ***
## countryENG            8.971e-02  1.263e-02   7.103 1.22e-12 ***
## countryESP            1.468e-01  1.177e-02  12.479  < 2e-16 ***
## countryEST           -1.431e-01  1.201e-02 -11.915  < 2e-16 ***
## countryFIN            2.626e-01  1.243e-02  21.120  < 2e-16 ***
## countryFRA            6.877e-02  1.238e-02   5.554 2.80e-08 ***
## countryGEO           -8.114e-03  1.435e-02  -0.566 0.571701
## countryHRV            5.347e-03  1.205e-02   0.444 0.657327
## countryISR            1.434e-01  1.178e-02  12.172  < 2e-16 ***
## countryITA            5.573e-02  1.186e-02   4.700 2.60e-06 ***
## countryJPN           -5.805e-02  1.203e-02  -4.824 1.41e-06 ***
## countryKOR           -2.747e-02  1.221e-02  -2.250 0.024480 *
## countryLVA           -1.042e-01  1.314e-02  -7.928 2.26e-15 ***
## countryMEX            4.347e-01  1.283e-02  33.888  < 2e-16 ***
## countryMYS            2.322e-01  1.204e-02  19.282  < 2e-16 ***
## countryNLD            2.124e-01  1.347e-02  15.764  < 2e-16 ***
## countryNOR            6.962e-02  1.228e-02   5.671 1.42e-08 ***
## countryNZL            1.908e-01  1.249e-02  15.278  < 2e-16 ***
## countryPOL           -2.336e-02  1.145e-02  -2.040 0.041370 *
## countryPRT            1.578e-02  1.165e-02   1.355 0.175498
## countryROU           -4.693e-02  1.167e-02  -4.021 5.80e-05 ***
## countryRUS            3.804e-02  1.158e-02   3.284 0.001023 **
## countrySGP            3.038e-02  1.199e-02   2.534 0.011266 *
## countrySRB            2.710e-02  1.189e-02   2.279 0.022661 *
## countrySVK           -1.794e-01  1.163e-02 -15.431  < 2e-16 ***
## countrySWE           -9.084e-02  1.185e-02  -7.663 1.83e-14 ***
## countryUSA            1.733e-01  1.333e-02  13.004  < 2e-16 ***
## genderMale           -2.198e-02  2.859e-03  -7.686 1.53e-14 ***
## age                   1.550e-03  2.398e-04   6.462 1.04e-10 ***
## years_teacher        -8.778e-04  2.419e-04  -3.629 0.000285 ***
## train_statYes        -4.837e-03  4.491e-03  -1.077 0.281492
## prep_ASomewhat       -6.471e-02  1.866e-02  -3.467 0.000526 ***
## prep_AWell           -5.083e-02  1.830e-02  -2.778 0.005475 **
## prep_AVery well      -4.404e-02  1.841e-02  -2.393 0.016736 *
## prep_BSomewhat       -2.950e-02  1.757e-02  -1.679 0.093157 .
## prep_BWell            4.336e-02  1.783e-02   2.432 0.015016 *
## prep_BVery well       1.025e-01  1.810e-02   5.663 1.49e-08 ***
```

```
## prep_CSomewhat       -6.751e-02  1.487e-02  -4.540 5.64e-06 ***
## prep_CWell           -1.487e-02  1.488e-02  -0.999 0.317687
## prep_CVery well       4.852e-02  1.512e-02   3.208 0.001336 **
## total_time           -4.725e-04  8.231e-05  -5.741 9.46e-09 ***
## collabDisagree        2.386e-01  7.177e-03  33.242  < 2e-16 ***
## collabAgree           4.752e-01  6.731e-03  70.599  < 2e-16 ***
## collabStrongly agree  7.557e-01  7.331e-03 103.084  < 2e-16 ***
## num_subjects          3.207e-03  9.396e-04   3.413 0.000643 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.409 on 101511 degrees of freedom
## Multiple R-squared:  0.2449, Adjusted R-squared:  0.2445
## F-statistic: 621.3 on 53 and 101511 DF,  p-value: < 2.2e-16
```

Summary 2

```
##
## Call:
## lm(formula = sat_avg ~ ., data = x[, -which(names(x) %in% c("country"))])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08944 -0.26399 -0.00771  0.28638  1.52074
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.5346760  0.0186100 136.200  < 2e-16 ***
## genderMale             -0.0160205  0.0028997  -5.525 3.30e-08 ***
## age                     0.0027291  0.0002414  11.305  < 2e-16 ***
## years_teacher          -0.0023157  0.0002440  -9.492  < 2e-16 ***
## train_statYes          -0.0004790  0.0045126  -0.106 0.915470
## prep_ASomewhat         -0.1402274  0.0189587  -7.396 1.41e-13 ***
## prep_AWell             -0.1466814  0.0185819  -7.894 2.96e-15 ***
## prep_AVery well        -0.1426278  0.0186637  -7.642 2.16e-14 ***
## prep_BSomewhat         -0.0235850  0.0180771  -1.305 0.192001
## prep_BWell              0.0462307  0.0183216   2.523 0.011628 *
## prep_BVery well         0.1012304  0.0185891   5.446 5.17e-08 ***
## prep_CSomewhat         -0.0648316  0.0153081  -4.235 2.29e-05 ***
## prep_CWell              0.0004794  0.0153054   0.031 0.975015
## prep_CVery well         0.0751858  0.0155372   4.839 1.31e-06 ***
## total_time             -0.0006643  0.0000818  -8.121 4.69e-16 ***
## collabDisagree          0.2344860  0.0073885  31.737  < 2e-16 ***
## collabAgree             0.4673560  0.0069133  67.603  < 2e-16 ***
## collabStrongly agree    0.7563150  0.0075131 100.666  < 2e-16 ***
## num_subjects            0.0076145  0.0009260   8.223  < 2e-16 ***
## regionEastern Europe   -0.1149916  0.0061300 -18.759  < 2e-16 ***
## regionNorth America     0.2050444  0.0078319  26.181  < 2e-16 ***
## regionNorthern Europe  -0.0249957  0.0064271  -3.889 0.000101 ***
## regionSouth America    -0.0719078  0.0067231 -10.696  < 2e-16 ***
## regionSouthern Europe  -0.0290054  0.0062328  -4.654 3.26e-06 ***
## regionWestern Europe    0.1154316  0.0073789  15.644  < 2e-16 ***
## regionWestern Pacific  -0.0291111  0.0062595  -4.651 3.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4214 on 101539 degrees of freedom
## Multiple R-squared:  0.1983, Adjusted R-squared:  0.1981
## F-statistic:  1004 on 25 and 101539 DF,  p-value: < 2.2e-16
```

Summary 3

```
##
## Call:
## lm(formula = sat_avg ~ ., data = x.new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01987 -0.25819 -0.00272  0.27384  1.50881
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.101e+00  1.461e-02 143.779  < 2e-16 ***
## countryAUS            1.931e-01  1.329e-02  14.533  < 2e-16 ***
## countryBFL            2.414e-01  1.191e-02  20.270  < 2e-16 ***
## countryBGR           -2.938e-02  1.251e-02  -2.350 0.018794 *
## countryBRA           -2.955e-02  1.001e-02  -2.951 0.003165 **
## countryCAB            1.778e-01  1.366e-02  13.020  < 2e-16 ***
## countryCHL            1.133e-01  1.482e-02   7.648 2.06e-14 ***
## countryCSH           -1.672e-01  1.130e-02 -14.795  < 2e-16 ***
## countryCZE           -9.050e-02  1.177e-02  -7.688 1.50e-14 ***
## countryDNK            1.767e-01  1.403e-02  12.602  < 2e-16 ***
## countryENG            8.578e-02  1.263e-02   6.793 1.11e-11 ***
## countryESP            1.318e-01  1.171e-02  11.260  < 2e-16 ***
## countryEST           -1.470e-01  1.198e-02 -12.270  < 2e-16 ***
## countryFIN            2.550e-01  1.238e-02  20.587  < 2e-16 ***
## countryFRA            4.580e-02  1.229e-02   3.726 0.000194 ***
## countryGEO           -1.493e-02  1.430e-02  -1.044 0.296421
## countryHRV           -4.438e-03  1.201e-02  -0.370 0.711701
## countryISR            1.410e-01  1.178e-02  11.971  < 2e-16 ***
## countryITA            5.030e-02  1.178e-02   4.268 1.98e-05 ***
## countryJPN           -7.552e-02  1.197e-02  -6.309 2.81e-10 ***
## countryKOR           -3.231e-02  1.220e-02  -2.648 0.008107 **
## countryLVA           -1.141e-01  1.313e-02  -8.688  < 2e-16 ***
## countryMEX            4.742e-01  1.263e-02  37.536  < 2e-16 ***
## countryMYS            2.250e-01  1.205e-02  18.675  < 2e-16 ***
## countryNLD            2.042e-01  1.346e-02  15.174  < 2e-16 ***
## countryNOR            6.453e-02  1.223e-02   5.277 1.32e-07 ***
## countryNZL            1.884e-01  1.245e-02  15.138  < 2e-16 ***
## countryPOL           -2.935e-02  1.144e-02  -2.565 0.010316 *
## countryPRT            3.777e-03  1.160e-02   0.326 0.744646
## countryROU           -5.231e-02  1.167e-02  -4.482 7.42e-06 ***
## countryRUS            3.166e-02  1.155e-02   2.740 0.006138 **
## countrySGP            1.645e-02  1.195e-02   1.376 0.168808
## countrySRB            2.864e-02  1.177e-02   2.434 0.014936 *
## countrySVK           -1.862e-01  1.162e-02 -16.023  < 2e-16 ***
## countrySWE           -9.280e-02  1.181e-02  -7.860 3.87e-15 ***
## countryUSA            1.713e-01  1.329e-02  12.889  < 2e-16 ***
## genderMale           -2.148e-02  2.853e-03  -7.528 5.18e-14 ***
## years_teacher         6.980e-04  1.284e-04   5.437 5.42e-08 ***
## total_time           -4.706e-04  8.211e-05  -5.731 1.00e-08 ***
## collabDisagree        2.355e-01  7.170e-03  32.838  < 2e-16 ***
## collabAgree           4.735e-01  6.722e-03  70.448  < 2e-16 ***
## collabStrongly agree  7.569e-01  7.328e-03 103.287  < 2e-16 ***
## num_subjects          3.571e-03  9.389e-04   3.804 0.000143 ***
## avg_prep              1.151e-01  2.418e-03  47.595  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4104 on 102142 degrees of freedom
## Multiple R-squared:   0.24,  Adjusted R-squared:  0.2397
## F-statistic: 750.3 on 43 and 102142 DF,  p-value: < 2.2e-16
```

Summary 4

```
##                              2.5 %        97.5 %
## (Intercept)           2.0718977968  2.1291664292
## countryAUS            0.1671017243  0.2191982708
## countryBFL            0.2180608930  0.2647446693
## countryBGR           -0.0538969688 -0.0048729942
## countryBRA           -0.0491803769 -0.0099267590
## countryCAB            0.1510620572  0.2046046494
## countryCHL            0.0842709971  0.1423484344
## countryCSH           -0.1893090175 -0.1450183893
## countryCZE           -0.1135712088 -0.0674291217
## countryDNK            0.1492583814  0.2042377021
## countryENG            0.0610273057  0.1105288540
## countryESP            0.1088670541  0.1547525172
## countryEST           -0.1705123357 -0.1235392035
## countryFIN            0.2306808115  0.2792257572
## countryFRA            0.0217088368  0.0698893204
## countryGEO           -0.0429452324  0.0130923462
## countryHRV           -0.0279757893  0.0190993456
## countryISR            0.1179064500  0.1640745421
## countryITA            0.0271980845  0.0733943515
## countryJPN           -0.0989857835 -0.0520627956
## countryKOR           -0.0562273641 -0.0083916154
## countryLVA           -0.1397995238 -0.0883318262
## countryMEX            0.4494836350  0.4990107641
## countryMYS            0.2013556169  0.2485775811
## countryNLD            0.1778618104  0.2306266470
## countryNOR            0.0405604252  0.0884971717
## countryNZL            0.1640289325  0.2128210410
## countryPOL           -0.0517789334 -0.0069240896
## countryPRT           -0.0189540002  0.0265089813
## countryROU           -0.0751897414 -0.0294330413
## countryRUS            0.0090171858  0.0543119368
## countrySGP           -0.0069788569  0.0398705791
## countrySRB            0.0055768722  0.0516989676
## countrySVK           -0.2090032614 -0.1634454339
## countrySWE           -0.1159457544 -0.0696628901
## countryUSA            0.1452273307  0.1973166893
## genderMale           -0.0270671971 -0.0158848861
## years_teacher         0.0004463662  0.0009495282
## total_time           -0.0006314828 -0.0003096251
## collabDisagree        0.2214015033  0.2495088025
## collabAgree           0.4603686288  0.4867182162
## collabStrongly agree  0.7425126836  0.7712377626
## num_subjects          0.0017310041  0.0054113949
## avg_prep              0.1103231565  0.1197998034
```

Selected Countries

```
## # A tibble: 6 × 2
##   country mean_satisfaction
##    <fctr>             <dbl>
## 1     SVK          2.774664
## 2     JPN          2.784752
## 3     SRB          3.004426
## 4     ROU          3.008698
## 5     MYS          3.251498
## 6     MEX          3.321342
```

Code used throughout report:

```
knitr::opts_chunk$set(echo = FALSE, warning=FALSE)
library("foreign")
library("tidyr")
library("countrycode")
library("rworldmap")
library(ggplot2)
library(reshape2)
library(grid)
library(gridExtra)
library(leaps)
library(GGally)
library(ggmosaic)
library(plyr)
library("dplyr")
# To start, let's read in the cleaned datasets from our last submission:
teacher_data <- read.csv("TALIS.csv")

# Change column names of 10 satisfaction questions to make more readable
satcols <- grep("TT2G46", colnames(teacher_data))
indices <- satcols[satcols != which(colnames(teacher_data) == "TT2G46_avg")]
colnames(teacher_data)[indices] <- c("sat1", "sat2", "sat3", "sat4", "sat5", "sat6", "sat7",
"sat8", "sat9", "sat10")
names(teacher_data)[names(teacher_data)=="TT2G46_avg"] <- "sat_avg"

# Column "TT2G15" contains no useful data, since it's all NAs, so we're going to remove it
# Column "ID" is no longer necessary, was used for making the subject dataset
teacher_data <- teacher_data[,-which(names(teacher_data) %in% c("TT2G15","ID"))]

# Change some of the subject names (one of the comments in our dataset submission was that th
ey weren't clear):
names(teacher_data)[names(teacher_data) == "For_Lang"] <- "Modern_Lang"
names(teacher_data)[names(teacher_data) == "Gk_Ln"] <- "Ancient_Lang"

# We also thought it might be useful to have a variable that indicates the number of subjects
 a given teacher teaches, so let's add it below:
subject_names <- c("R_WR_Lit", "Math", "Science", "Soc_Studies", "Modern_Lang", "Ancient_Lan
g", "Tech", "Arts", "PE", "Rel_Ethics", "Practical", "Other")
subject_indices <- c(which(colnames(teacher_data) %in% subject_names))

teacher_data$num_subjects <- rowSums(teacher_data[,subject_indices], na.rm=TRUE)

# We discussed in our dataset submission why we weren't going to use their teacher satisfacti
on variable, and we are instead going to use the average of the 10 satisfaction questions, as
 a result, let's remove the teacher satisfaction variable:
teacher_data <- teacher_data[, -which(names(teacher_data) %in% c("satisfaction"))]

# A lot of the variables stored as categorical variables with the categories "Strongly Disagr
ee", "Disagree", "Strongly Agree", and "Agree", are currently factors.  Let's change the orde
ring to go from low to high.
teacher_data$prep_A <- factor(teacher_data$prep_A, levels=c("Not at all", "Somewhat", "Well",
 "Very well"))
teacher_data$prep_B <- factor(teacher_data$prep_B, levels=c("Not at all", "Somewhat", "Well",
 "Very well"))
teacher_data$prep_C <- factor(teacher_data$prep_C, levels=c("Not at all", "Somewhat", "Well",
 "Very well"))
teacher_data$feedback_salary <- factor(teacher_data$feedback_salary, levels=c("No positive ch
```

```
ange", "A small change", "A moderate change", "A large change"))
teacher_data$collab <- factor(teacher_data$collab, levels=c("Strongly disagree", "Disagree",
"Agree", "Strongly agree"))

# We'll split the country data into regions using the World Health Organization  categories,
 so that we can try to find regional patterns

# Brazil, Chile
teacher_data$region[teacher_data$country %in% c("BRA", "CHL")] <- "South America"

# Canada, Mexico, United States
teacher_data$region[teacher_data$country %in% c("CAB", "MEX", "USA")] <- "North America"

#  Bulgaria, Czech Republic,  Poland,  Romania, Russia, Slovak Republic
teacher_data$region[teacher_data$country %in%
                      c("BGR", "CZE", "POL", "ROU", "RUS", "SVK")] <- "Eastern Europe"

# Denmark, England, Estonia, Finland,   Latvia,  Norway, Sweden
teacher_data$region[teacher_data$country %in%
                      c("DNK", "ENG",  "EST", "FIN", "LVA", "NOR", "SWE")] <- "Northern Eur
ope"

# Spain, Croatia, Italy, Latvia,  Portugal,  Serbia,
teacher_data$region[teacher_data$country %in%
                      c("ESP", "HRV", "ITA", "LVA", "PRT", "SRB")] <- "Southern Europe"

# Belgium,  France, Netherlands
teacher_data$region[teacher_data$country %in%
                      c("BFL", "FRA","NLD")] <- "Western Europe"

# Abu Dhabi, Georgia, Israel
teacher_data$region[teacher_data$country %in% c("AAD", "GEO", "ISR")] <- "East Mediterranean"

# Australia, China, Japan, South Korea, Malaysia, New Zealand, Singapore
teacher_data$region[teacher_data$country %in% c("AUS", "CSH", "JPN", "KOR", "MYS", "NZL", "SG
P")] <- "Western Pacific"

teacher_data$region <- as.factor(teacher_data$region)
# Histogram of teacher satisfaction
p1 <- teacher_data %>% ggplot(aes(x=sat_avg)) + geom_histogram(binwidth = 0.2) +
labs(title="Plot 1: Histogram of Teacher Satisfaction", x="Teacher Satisfaction (low to hig
h)", y="Count")

# Boxplot of whether they consider their school a collaborative environment to satisfaction
p2 <- teacher_data %>% filter(!is.na(collab)) %>% ggplot(aes(x=collab, y=sat_avg)) + geom_box
plot() +
  labs(title="Plot 2: Relationship Between Collaboration Level
       of the School and Teacher Satisfaction",
       x="Response to 'There is a collaborative school culture
       which is characterised by mutual support'",
       y="Teacher Satisfaction")

# Boxplot of region to satisfaction
p3 <- teacher_data %>% filter(!is.na(sat_avg)) %>% ggplot(aes(x=reorder(region, -sat_avg, med
ian), y=sat_avg)) + geom_boxplot() + labs(title="Plot 3: Teacher Satisfaction by Region",
x="Region", y="Teacher Satisfaction")
grid.arrange(p1, p2, ncol=2)
```

```
p3
plot1 <- teacher_data %>% ggplot(aes(x=total_time)) + geom_histogram() +
  labs(title="Histogram of Time Spent on
       Teaching Activities per Week",
       x="Time Spent on Teaching Activities (hrs/week)",
       y="Count")

plot2 <- teacher_data %>% ggplot(aes(x=age, y=years_teacher)) +
geom_point(alpha=0.5,shape=".") + geom_abline(intercept =0, slope=1, color="red") +  geom_abl
ine(intercept = -15, slope=1, color="yellow")+  geom_jitter(shape=".") +
  labs(title="Reported Age vs Years Teaching",
       x="Age (years)",
       y="Years Teaching (years)")

grid.arrange(plot1, plot2, ncol=2)
nrow(teacher_data[teacher_data$total_time==90 & !is.na(teacher_data$total_time),])
nrow(teacher_data[teacher_data$total_time>90 & !is.na(teacher_data$total_time),])
teacher_data[teacher_data$total_time > 90 & !is.na(teacher_data$total_time), "total_time"] <-
 NA
age_start <- teacher_data$age - teacher_data$years_teacher
teacher_data[age_start < 15 & !is.na(age_start), "years_teacher"] <- NA
plot1 <- teacher_data %>% filter(!is.na(teacher_data$train_stat)) %>%
ggplot(aes(x=train_stat, y=sat_avg)) + geom_boxplot() +
  labs(title="Teacher Training Status vs Teacher Satisfaction",
       x = "Training Status",
       y="Teacher Satisfaction (low to high)") +
  theme(axis.text=element_text(size=8),
        axis.title=element_text(size=8),
        plot.title=element_text(size=8))
plot2 <- teacher_data %>% ggplot(aes(x=years_teacher, y=sat_avg)) + geom_point(alpha=0.05) +
  labs(title="Years Spent as a Teacher
       vs Teacher Satisfaction",
       x="Years Spent as a Teacher", y="Teacher Satisfaction (low to high)") +
  theme(axis.text=element_text(size=8),
        axis.title=element_text(size=8),
        plot.title=element_text(size=8))
plot3 <- teacher_data %>% filter(!is.na(prep_A)) %>% ggplot(aes(x=prep_A, y=sat_avg)) + geom_
boxplot() + labs(title="Preparation vs Satisfaction", x="Prepared for the Content I Teach",
y="Teacher Satisfaction (low to high)") +
  theme(axis.text=element_text(size=8),
        axis.title=element_text(size=8),
        plot.title=element_text(size=8))

grid.arrange(plot1, plot2, plot3, ncol=3)
# In order to have easier control over which predictors to use in our model, below we create
 several lists of indices that we might want to include or exclude as a group

# We don't want to use the 10 satisfaction questions to predict sat_avg, since they were used
 to create the sat_avg predictor, so let's create a list of the indices we need to avoid
satcols <- grep("sat", colnames(teacher_data))
sat_indices <- satcols[satcols != which(colnames(teacher_data) == "sat_avg")]

subject_names <- c("R_WR_Lit", "Math", "Science", "Soc_Studies", "Modern_Lang", "Ancient_Lan
g", "Tech", "Arts", "PE", "Rel_Ethics", "Practical", "Other")
subject_indices <- c(which(colnames(teacher_data) %in% subject_names))

# We might not want to include the three questions about how prepared they felt:
```

```
prep_indices <- grep("prep_", names(teacher_data))

# Some columns also have many more NA values, lets make a list of the columns that have NA va
lues in more than 10% of the data (this ends up being num_students and feedback_salary), that
 way we can choose whether or not to include them in our linear models
na_cols <- c()
for(i in 1:ncol(teacher_data)) {
  if(sum(is.na(teacher_data[,i]))/nrow(teacher_data) > 0.10) {
    na_cols <- c(na_cols,i)
  }
}
x <- na.omit(teacher_data[, -c(sat_indices, na_cols, subject_indices)])
m_country <- lm(sat_avg ~ ., data=x[,-which(names(x) %in% c("region"))])
summary(m_country)$r.squared
m_region <- lm(sat_avg ~ ., data=x[,-which(names(x) %in% c("country"))])
summary(m_region)$r.squared
cor(teacher_data$age, teacher_data$years_teacher, use = "complete.obs")
teacher_data$age_start <- age_start
teacher_data$age_ratio <- teacher_data$years_teacher/teacher_data$age
plot1 <- teacher_data %>% ggplot(aes(x=age, y=sat_avg)) + geom_point(alpha=0.05) +
  labs(title="Age vs Teacher Satisfaction", x="Age (years)", y="Teacher Satisfaction (low to
 high)") +
  theme(axis.text=element_text(size=6),
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))
plot2 <- teacher_data %>% ggplot(aes(x=years_teacher, y=sat_avg)) + geom_point(alpha=0.05) +
  labs(title="Years Spent as a Teacher
       vs Teacher Satisfaction",
       x="Years Spent as a Teacher",
       y="Teacher Satisfaction (low to high)") +
  theme(axis.text=element_text(size=6),
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))
plot3 <- teacher_data %>% ggplot(aes(x=age_ratio, y=sat_avg)) + geom_point(alpha=0.05) +
  labs(title="Fraction of Life Spent Teaching
       vs Teacher Satisfaction",
       x="Fraction of Life Spent Teaching",
       y="Teacher Satisfaction (low to high)") +
  theme(axis.text=element_text(size=6),
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))
grid.arrange(plot1, plot2, plot3, ncol=3)
data <- teacher_data[, grep("prep_", names(teacher_data))]
data <- data.frame(prep_A = as.numeric(teacher_data$prep_A),
                   prep_B = as.numeric(teacher_data$prep_B),
                   prep_C = as.numeric(teacher_data$prep_C))
cor(data, use = "complete.obs")
teacher_data$avg_prep <- apply(data, 1, mean)
x.new <- teacher_data[, -c(sat_indices, na_cols, subject_indices, prep_indices)]
x.new <- na.omit(x.new[, -which(names(x.new) %in% c("train_stat", "region", "age", "age_star
t", "age_ratio"))])
h1 <- x.new %>% ggplot(aes(x=total_time)) + geom_histogram() +
  labs(title="Histogram of Time Spent on
       Teaching Activities",
       x="Time Spent on Teaching Related Activities",
       y="Count") +
  theme(axis.text=element_text(size=6),
```

```
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))

h2 <- x.new %>% ggplot(aes(x=years_teacher)) + geom_histogram() +
  labs(title="Histogram of Years Spent
       as a Teacher",
       x="Years of Teaching Experience",
       y="Count") +
  theme(axis.text=element_text(size=6),
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))

h3 <- x.new %>% ggplot(aes(x=collab)) + geom_bar() +
  labs(title="Bar Graph of
       Collaboration Rating",
       x="How Collaborative is School Culture",
       y="Count") +
  theme(axis.text=element_text(size=6),
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))  + theme(axis.text.x=element_text(angle=-35, hjust=
.1))

h4 <- x.new %>% ggplot(aes(x=num_subjects)) + geom_bar() +
  labs(title="Histogram of The Number
       of Subjects Taught",
       x="Number Subjects Taught",
       y="Count") +
  theme(axis.text=element_text(size=6),
        axis.title=element_text(size=6),
        plot.title=element_text(size=8))

grid.arrange(h1, h2, h3, h4, ncol = 4)
m_final <- lm(sat_avg ~ ., data=x.new)
summary(m_final)$r.squared
par(mfrow=c(1,2))
plot(m_final, which=1)
plot(m_final, which=2)
confint(m_final)
# set to ISO codes
teacher_data$country <- revalue(teacher_data$country, c("AAD"="ARE",
"BFL"="BEL","CAB"="CAN","CSH"="CHN","ENG"="GBR"))
# apparently plyr dependencies will mess things up
detach(package:plyr)
country_avg <- teacher_data %>% group_by(country) %>% summarize(mean_satisfaction = mean(sat_
avg, na.rm=TRUE))
library(plyr); library(dplyr)
#code for map with average teacher satisfaction by country
sPDF1 <- joinCountryData2Map(country_avg, joinCode = "ISO3", nameJoinColumn = "country")
 par(mar=c(0,0,1,0))
mapParams1 <- mapCountryData(sPDF1, nameColumnToPlot = "mean_satisfaction", colourPalette="te
rrain", oceanCol = "lightblue", missingCountryCol="gray", addLegend = FALSE, mapTitle = "Mean
 Satisfaction By Country")
do.call( addMapLegend, c(mapParams1, legendWidth=0.5, legendShrink=0.5))
# Regression of sat_avg vs region
msatreg <- lm(sat_avg ~ country, data=teacher_data)

# ANOVA of regression
```

```
anova(msatreg)
td_country <-teacher_data[teacher_data$country %in% c("SVK", "JPN", "SRB", "ROU", "MYS", "ME
X"),]
# remove unused factors
td_country$country <- factor(td_country$country, levels=c("SVK", "JPN",  "SRB", "ROU", "MYS",
 "MEX"))
# collab plot
p41 <- td_country %>% filter(!is.na(collab)) %>% ggplot(aes(x = product(collab, country), fil
l=factor(collab)), na.rm=TRUE) +  geom_mosaic() + scale_fill_manual(values=c("red3","indianre
d1", "lightgreen", "green3")) +  theme(axis.text.x=element_text(angle=-25, hjust= .1)) +
labs(x="Country", title='There is a collaborative school culture.') + guides(fill=guide_legen
d(title = "Response", reverse = TRUE)) + theme(axis.text=element_text(size=6), axis.title=ele
ment_text(size=6), plot.title=element_text(size=8))

p42 <- td_country %>% filter(!is.na(prep_A)) %>% ggplot(aes(x = product(prep_A, country), fil
l=factor(prep_A)), na.rm=TRUE) +  geom_mosaic() + scale_fill_manual(values=c("red3","indianre
d1", "lightgreen", "green3")) +  theme(axis.text.x=element_text(angle=-25, hjust= .1)) +
labs(x="Country", title='How prepared do you feel for the content of the subjects you
teach?') + guides(fill=guide_legend(title = "Response", reverse = TRUE)) + theme(axis.text=el
ement_text(size=6), axis.title=element_text(size=6), plot.title=element_text(size=8))

grid.arrange(p41,p42, ncol=2)
# Check which variables have a significant correlation with region
anova(lm(as.numeric(collab) ~ country, data=teacher_data))
anova(lm(as.numeric(avg_prep) ~ country, data=teacher_data))
td_country %>%
    filter(country == "MEX") %>% filter(!is.na(collab)) %>%
    ggplot(aes(x=collab, y=sat_avg)) + geom_boxplot(varwidth=TRUE) + labs(title="School Colla
boration vs Satisfaction in Mexico")
nrow(teacher_data[teacher_data$sat_avg >=2.5,])/nrow(teacher_data)
pairwise.t.test(teacher_data$sat_avg, teacher_data$region, p.adj="bonf")
summary(m_country)
summary(m_region)
summary(m_final)
confint(m_final)
country_avg[order(country_avg$mean_satisfaction)[c(1,2,18,19,35,36)],]

# Data Cleaning code from Data Exploration
# Lower-secondary-school teacher data (international file, all countries) is the file titled
 BTGINTT2
lower_second_data <- read.spss("SPSS_International/BTGINTT2.sav", to.data.frame=TRUE, use.mis
sings=TRUE)

cols <- c("CNTRY", "TT2G01", "TT2G02", "TT2G05B", "TT2G11", "TT2G13A", "TT2G13B", "TT2G13C",
"TT2G16", "TT2G30G", "TT2G38", "TT2G44E", "TJOBSATS")

# We are creating a dataset with the columns from our project proposal, the 10 questions rela
ted to job satisfaction, which are columns TT2G46A through TT2G46J, as well as a few addition
al columns which we also think might be useful
data <- cbind(lower_second_data[,cols],lower_second_data[,grep("TT2G46", colnames(lower_secon
d_data))], lower_second_data[,grep("TT2G15", colnames(lower_second_data))])

# Rename some of the columns so that they're easier to work with
data <- data %>% rename(gender=TT2G01, age=TT2G02, country=CNTRY, years_teacher = TT2G05B, tr
ain_stat = TT2G11, prep_A=TT2G13A, prep_B=TT2G13B, prep_C=TT2G13C, total_time=TT2G16, feedbac
k_salary=TT2G30G, collab=TT2G44E, num_students=TT2G38, satisfaction=TJOBSATS)
```

```
# Columns that start with TT2G15 correspond to several questions asking what subjects teacher
s teach:
#Let's manipulate the subject data so that it's in a form that's useful to us

# First make an ID column to merge back after isolating the subject as one column
data$ID <- seq.int(nrow(data))

# Indices of columns dealing with subject taught
indices <- grep("TT2G15", colnames(data))[-1] #-1 to deal with the TT2G15 column that isn't f
or a specific subject

# Convert "Yes" "No" levels to 0 for didn't teach subject and 1 for taught subject
for (i in indices) {
  data[,i] <- 2 - as.numeric(data[,i])
}

# Rename columns by subject
data <- data %>% rename(R_WR_Lit=TT2G15A, Math=TT2G15B, Science=TT2G15C, Soc_Studies = TT2G15
D, For_Lang = TT2G15E, Gk_Ln=TT2G15F, Tech=TT2G15G, Arts=TT2G15H, PE=TT2G15I, Rel_Ethics=TT2G
15J, Practical=TT2G15K, Other=TT2G15L)

# Names of columns
col_names <- colnames(data)[indices]

# Below is just the R code for extracting the subject column
# This is separated for clarity since it is extensive in and of itself

df <- data[25:37]

# Melt to format -- similar to "gather"
temp_by_subj <- melt(df,id="ID")

# Only keep 1's (subjects taught)
temp_by_subj <- temp_by_subj[which(temp_by_subj$value==1),]

# Merge the subject table in with data
data_by_subject <- merge(data,temp_by_subj,all.y=T, by = "ID")

# Clean up unnecessary columns
data_by_subject <- data_by_subject[,-(c(15:24,26:37,39))]
data_by_subject <- data_by_subject %>% rename(Subject=variable)

# Below is the code used to clean up the satisfaction related questions, as well as a few oth
er columns

# All ten columns with responses related to job satisfaction are factors, with 4 levels: "Str
ongly Disagree", "Disagree", "Agree", "Strongly Agree".
# We are going to convert these 10 columns to numerical data, where the factors get converted
 to the integers 1, 2, 3, and 4, with 1 corresponding to Strongly Disagree and 4 correspondin
g to Strongly Agree.
for (i in grep("TT2G46", colnames(data))) {
  data[,i] <- as.numeric(data[,i])
}

# Questions C, D, and F were negatively phrased questions, so we are reversing their scales s
o we can compare them to the positively phrased questions.
data$TT2G46C <- 5 - data$TT2G46C
```

```
data$TT2G46D <- 5 - data$TT2G46D
data$TT2G46F <- 5 - data$TT2G46F

# Questions TT2G46A through J were all related to teacher satisfaction.  Consequently, we are
 going to create a column that averages the results of these 10 questions to get a sense of t
eacher satisfaction
data$TT2G46_avg <- rowMeans(data[,grep("TT2G46", colnames(data))], na.rm=TRUE)

# Clean up country column, which has a lot of whitespace
data$country <- as.factor(trimws(as.character(data$country)))

# Clean up train_stat variable so it's easier to work with
# Make No the first factor, so that if we convert it to numeric, it's easy to have 0 represen
t No and 1 represent Yes
data$train_stat <- relevel(data$train_stat, "No")
```