Final Project

Veena Advani, Annie Chen, Robert Tung

Due 4/28/2017

# Introduction

This analysis will focus on the Teaching and Learning International Survey (TALIS) Dataset. This data set was 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. Another goal of the TALIS survey was to provide a more global view of education systems, and to produce metrics that could be used to compare education systems from different countries.

TALIS was administered in 2008 and 2013. We will be using the more recent international dataset, from 2013. It contains answers from teachers from more than 10,000 schools in 36 different countries. Not all of the countries were English speaking. The only eligibility requirement was that countries must be industrialized. The TALIS dataset contains several different files. The data is split by school type (primary school, secondary school etc). Although the TALIS dataset contains information from a variety of school types (primary, secondary, etc), the focus of the survey was on lower-secondary schools, and that is the largest data file. As a result, we are going to focus our analysis on lower-secondary schools. At each school that was selected to participate, the principal (or head administrator) and a random selection of up to 22 teachers were chosen to complete voluntary online questionnaires. Our analysis below consists of only the teacher responses, not the responses of the principals, as they were given slightly different questionnaires.

The first question we hope to explore in our analysis, is: (1) 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 countries development and functioning. It's important that a country can educate its youth, so that they can participate in society when they are adults. However, there is a fair amount of diversity across countries with regard to how well the teaching profession is respected, and how well education systems are funded. As a result, teacher working conditions vary quite a bit. However, in order to make sure that there are enough teachers, and that they are able to teach their students well, it is important to make sure that teacher satisfaction is at a high enough level. If teacher satisfaction is too low, it becomes difficult to retain a large enough teaching force, and to attract talented people to the field. Low teacher satisfaction can also indicate that they are having trouble doing their job well, possibly due to environmental factors out of their control. For many reasons such as these, teacher satisfaction is a useful metric to observe in any education system. It's also helpful to understand what kinds of environments keep teachers more satisfied, so that educational systems can encourage and build these types of environments. Therefore, understanding 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, could be helpful information.

Our second research question is: (2) How do teacher satisfaction, as well as some of the significant predictors found from question 1, vary by country?

#### Main Research Questions:

1. Which variables and classroom factors are the best predictors of teacher job satisfaction?
2. How do teacher satisfaction, as well as some of the significant predictors found from question 1, vary by country?

#### Motivation of Study:

* Education is important to any country's basic development and functioning
* There is a fair amount of diversity across countries with regard to how well the teaching profession is respected, as well as how much countries fund their education systems.
* As a result, teacher working conditions vary quite a bit
* However, in order to make sure that there are enough teachers, and that they are able to teach their students well, it is important that teacher satisfaction is at a high enough level. If teacher satisfaction is too low, it becomes difficult to retain a large enough teaching force
* Therefore, understanding 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, could be helpful information in order to make sure that countries have enough teachers, and that teachers have enough resources to do their jobs well.

# Data

Each row of the dataset corresponds to single teacher's response to the questionnaire (There are 117876 rows). Each column in the dataset corresponds to either a particular question in the questionnaire, or a summary variable for a section of the questionnaire.

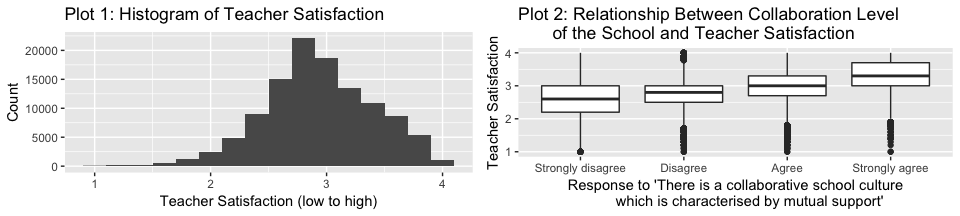
The dataset contains a column for each individual question on the questionnaire, and summary columns for overall sections. There are columns concerning employment status, personal information/background, time spent, professional development, efficacy and amount of feedback, outlook about the school, school climate, overall job satisfaction, and much more. For example, there are individual columns for questions such as “The advantages of this profession clearly outweigh the disadvantages” and “I would recommend my school as a good place to work”, but also an overall column summing up all the scores for the teacher satisfaction questions. The specific variables we are using are listed below. Almost all of the questions in the questionnaire were multiple choice or numerical input questions.

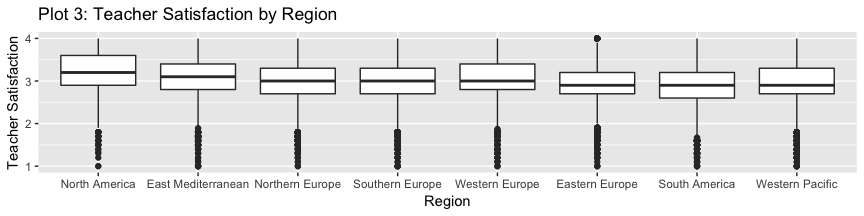
The variables we are using are:

* country (CNTRY): Country of teacher (factor variable, 36 levels)
* region: Region the teacher's country is in (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)
* train\_stat (TT2G11): Training level of teacher (factor variable)
* prep\_A/B/C (TT2G13A/B/C): How prepared the teacher felt (factor variable, 4 levels)
* total\_time (TT2G16): Total time spent on teaching related activities in hours per week (numeric variable)
* feedback\_salary (TT2G30G): Whether performance feedback influences salary (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 (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 (factor variables, 4 levels)
* sat\_avg: An average of the ten questions related to teacher job satisfaction (numeric variable)

#### Initial Plots Describing Data

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





**Plot 1** provides a full histogram of teacher satisfaction. 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 the data.

**Plot 2** provides a box plot of collab vs sat\_avg. This plot is an example of one we would find in regard to our first research question, in which we assess which variables are particularly predictive of overall teacher satisfaction. In looking at this particular plot, we can actually 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, as well as more rigorous analysis to be described below.

**Plot 3** provides a box plot of geographic region vs sat\_avg. This is relevant to our second research question, which revolves around variations in sat\_avg and other variables across different 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). We will explore additional plots like this one, but with other variables, in our full analysis.

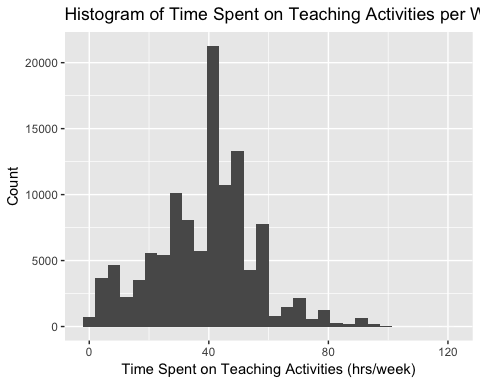
# Analysis

### Removal of Strange Values

Before we analyze the data any further, we should clean up the data slightly more. In previous assignments, 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. We have decided to set these values to NA, since they are likely unreliable.

The total time spent on teaching related activities per week:

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



We can see from this histogram, that most teachers seem to have a 40 hour work week, which is a pretty typical time comittment. There is also another pretty clear dropoff after 60, which also seems like a reasonable addition to the 40 hour workweek for teachers that spend a fair amount of time on class planning and grading in the evenings. We can also see that there are some teachers claiming to have spent over 90 hours per week on teaching related activities. Even if a teacher worked from 6 am to 9 pm Monday through Friday (15 hours a day) due to 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 a physically possible time commitment, it seems highly unlikely. It is possible that edge cases such as teachers at boarding schools could claim this kind of time comittment, but other than unusual situations like that, more than 90 hours a week seems highly improbable. Given this fact, let's see how many teachers reported spending 90 hours a week or more on teaching related activities.

length(teacher\_data[teacher\_data$total\_time>=90 & !is.na(teacher\_data$total\_time),]$total\_time)

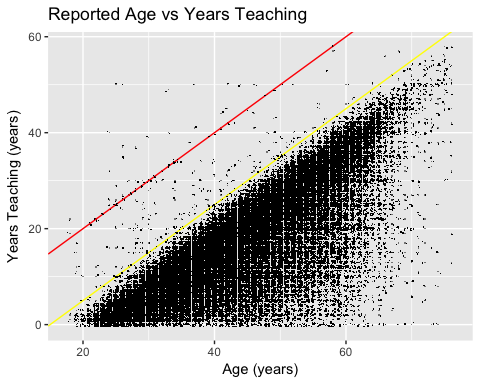
## [1] 795

length(teacher\_data[teacher\_data$total\_time>90 & !is.na(teacher\_data$total\_time),]$total\_time)

## [1] 210

We can see that there are quite a few values of exactly 90, given the fact that the length of the vector where total\_time is greater than or equal to 90 is so much longer than the vector that is strictly greater than 90. Given this fact, we will 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

We can now take a look at reported age vs years teaching. 

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 remove those values.

age\_start <- teacher\_data$age - teacher\_data$years\_teacher  
nrow(teacher\_data[age\_start < 15 & !is.na(age\_start),])

## [1] 232

teacher\_data[age\_start < 15 & !is.na(age\_start), "years\_teacher"] <- NA

### Analysis for Research Question 1

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 attempt to use several techniques.

#### Linear Regression Analysis

Although our earlier plot of satisfaction against the degree of collaboration at the school shows a clear linear trend, let's look at a matrix of plots of the data, before diving into any linear regression models:

Dataset for Linear Regression Model: 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 resposne 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:

##   
## 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) 2.771e+00 1.310e-02 211.484 < 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\_A.L -2.644e-02 1.267e-02 -2.087 0.036903 \*   
## prep\_A.Q 3.576e-02 9.522e-03 3.755 0.000173 \*\*\*  
## prep\_A.C -1.916e-02 6.061e-03 -3.161 0.001574 \*\*   
## prep\_B.L 8.505e-02 1.247e-02 6.818 9.27e-12 \*\*\*  
## prep\_B.Q 4.432e-02 9.019e-03 4.914 8.94e-07 \*\*\*  
## prep\_B.C -2.596e-02 5.154e-03 -5.037 4.73e-07 \*\*\*  
## prep\_C.L 4.432e-02 1.045e-02 4.239 2.24e-05 \*\*\*  
## prep\_C.Q 6.545e-02 7.714e-03 8.484 < 2e-16 \*\*\*  
## prep\_C.C -2.446e-02 4.806e-03 -5.090 3.58e-07 \*\*\*  
## total\_time -4.725e-04 8.231e-05 -5.741 9.46e-09 \*\*\*  
## collab.L 5.599e-01 4.995e-03 112.087 < 2e-16 \*\*\*  
## collab.Q 2.100e-02 4.039e-03 5.198 2.02e-07 \*\*\*  
## collab.C 1.028e-02 2.832e-03 3.629 0.000284 \*\*\*  
## 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

Linear model using region:

##   
## 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) 2.8255085 0.0109682 257.609 < 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\_A.L -0.0971208 0.0128434 -7.562 4.00e-14 \*\*\*  
## prep\_A.Q 0.0721405 0.0096714 7.459 8.78e-14 \*\*\*  
## prep\_A.C -0.0275630 0.0061781 -4.461 8.15e-06 \*\*\*  
## prep\_B.L 0.0835187 0.0128041 6.523 6.93e-11 \*\*\*  
## prep\_B.Q 0.0392924 0.0092765 4.236 2.28e-05 \*\*\*  
## prep\_B.C -0.0241980 0.0053036 -4.563 5.06e-06 \*\*\*  
## prep\_C.L 0.0650401 0.0107319 6.060 1.36e-09 \*\*\*  
## prep\_C.Q 0.0697690 0.0079397 8.787 < 2e-16 \*\*\*  
## prep\_C.C -0.0269999 0.0049410 -5.465 4.65e-08 \*\*\*  
## total\_time -0.0006643 0.0000818 -8.121 4.69e-16 \*\*\*  
## collab.L 0.5594228 0.0051113 109.449 < 2e-16 \*\*\*  
## collab.Q 0.0272365 0.0041516 6.561 5.39e-11 \*\*\*  
## collab.C 0.0129032 0.0029143 4.428 9.54e-06 \*\*\*  
## 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

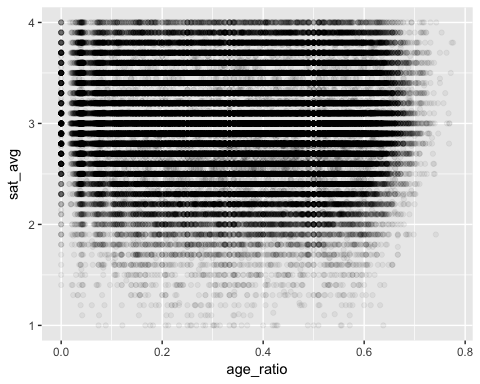
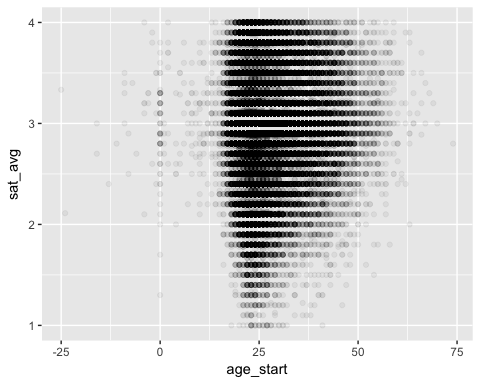
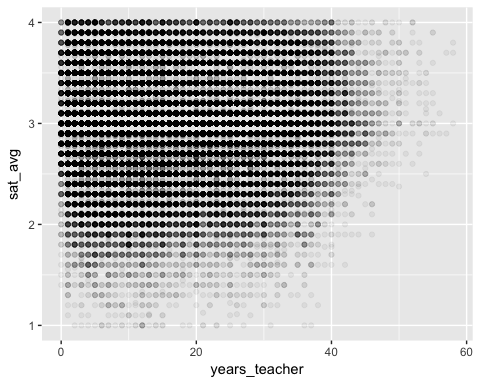
We can see from the two summaries, that a large amount of the country categories are significant, and all of the region categories are also significant. However, the R-squared value for the model using country data was 0.2449, and the R-squared value for the model with the region predictor was 0.1981. This means that removing some geographical information by condensing the countries into regions caused a drop in R-squared by roughly 19%, which is quite a bit. As a result, we are going to 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 also 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. We can also see that they seem quite correlated from the plot earlier of age against years\_teacher. As a result, we probably don't need both predictors in our linear model. Let's check the correlation:

## [1] NA

We can see above that 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. Two options are, the age at which the teacher started teaching, and the ratio of years\_teacher to age. Perhaps the time in your life when you start teaching may affect how satisfied you are. Of course, if we interpret this variable directly as the age teachers start teaching, we are making an implicit assumption that they never took a break/time off from teaching. It's also possible that the fraction of your life that you have spent teaching, affects how satisfied you are.

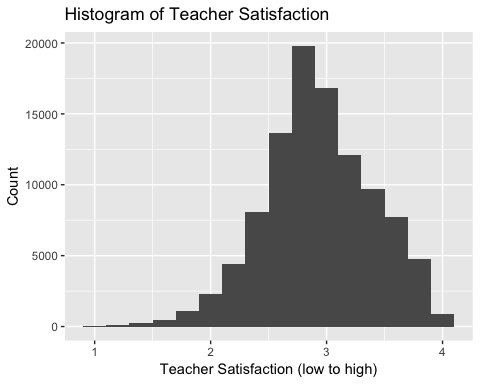


Let's also make sure that the prep\_A, prep\_B, and prep\_C predictors are not too highly correlated.

## 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

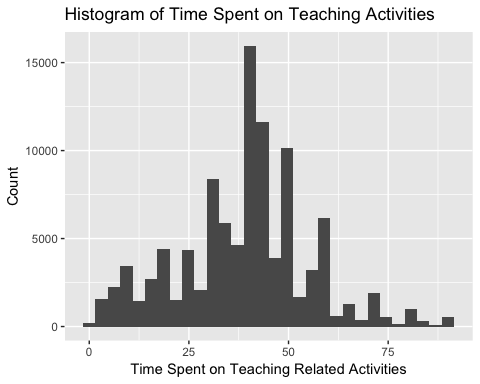
We can see from the correlation matrix above that all three prep questions are pretty correlated. As a result, let's try creating a single predictor that's the average of all three, in order to indicate their overall feeling of how prepared they were.

Now let's see if all of the variables we have in our dataset are normally distributed, or if some transformations might be helpful:

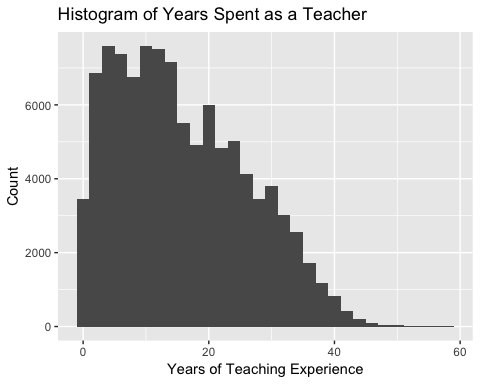


We can see from the histogram of sat\_avg above, that although it isn't perfectly normal, it is close enough that it probably doesn't warrant any transformation of the variable. Additionally, after we have a final model, we can check the qqnorm plot to make sure none of the normality assumptions are drastically violated.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

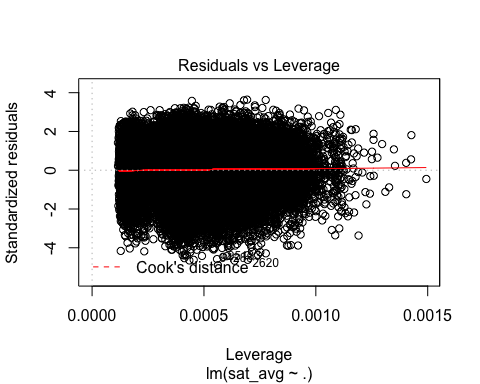
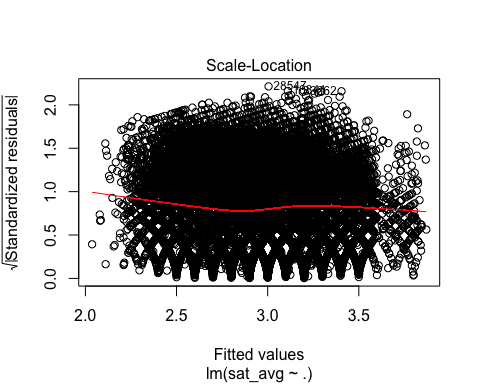
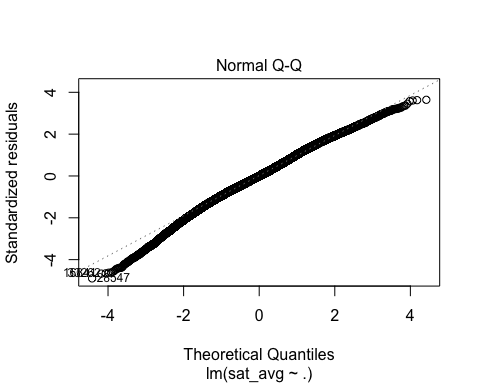
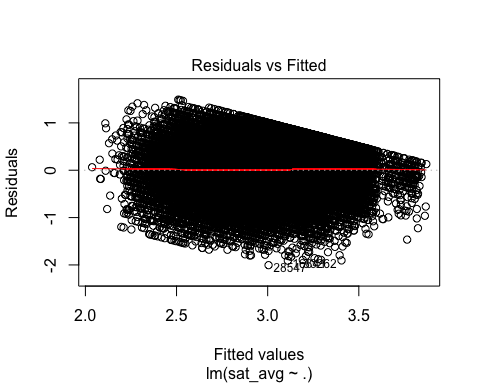


## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 Both of these histograms, although not perfectly normal, also don't have a huge amount of skew. In fact, the histogram for time spent on reaching related activities has very little skew. As a result, we probably don't need to transform either of these variables either.

Model without train\_stat:

##   
## Call:  
## lm(formula = sat\_avg ~ ., data = x.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0048 -0.2573 -0.0035 0.2726 1.4907   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.578e+00 1.688e-02 152.688 < 2e-16 \*\*\*  
## countryAUS 1.805e-01 1.329e-02 13.580 < 2e-16 \*\*\*  
## countryBFL 2.396e-01 1.190e-02 20.138 < 2e-16 \*\*\*  
## countryBGR -3.761e-02 1.250e-02 -3.009 0.002625 \*\*   
## countryBRA -2.783e-02 1.000e-02 -2.782 0.005400 \*\*   
## countryCAB 1.677e-01 1.366e-02 12.282 < 2e-16 \*\*\*  
## countryCHL 9.920e-02 1.482e-02 6.694 2.18e-11 \*\*\*  
## countryCSH -1.589e-01 1.130e-02 -14.066 < 2e-16 \*\*\*  
## countryCZE -1.002e-01 1.177e-02 -8.513 < 2e-16 \*\*\*  
## countryDNK 1.616e-01 1.404e-02 11.516 < 2e-16 \*\*\*  
## countryENG 7.562e-02 1.263e-02 5.989 2.12e-09 \*\*\*  
## countryESP 1.254e-01 1.170e-02 10.718 < 2e-16 \*\*\*  
## countryEST -1.609e-01 1.200e-02 -13.412 < 2e-16 \*\*\*  
## countryFIN 2.443e-01 1.239e-02 19.709 < 2e-16 \*\*\*  
## countryFRA 4.531e-02 1.227e-02 3.692 0.000222 \*\*\*  
## countryGEO -3.226e-02 1.432e-02 -2.253 0.024268 \*   
## countryHRV -2.143e-02 1.202e-02 -1.782 0.074672 .   
## countryISR 1.341e-01 1.177e-02 11.402 < 2e-16 \*\*\*  
## countryITA 3.531e-02 1.182e-02 2.987 0.002822 \*\*   
## countryJPN -8.026e-02 1.196e-02 -6.712 1.93e-11 \*\*\*  
## countryKOR -3.855e-02 1.219e-02 -3.163 0.001563 \*\*   
## countryLVA -1.190e-01 1.311e-02 -9.073 < 2e-16 \*\*\*  
## countryMEX 4.722e-01 1.262e-02 37.426 < 2e-16 \*\*\*  
## countryMYS 2.206e-01 1.203e-02 18.335 < 2e-16 \*\*\*  
## countryNLD 1.933e-01 1.346e-02 14.365 < 2e-16 \*\*\*  
## countryNOR 4.808e-02 1.225e-02 3.925 8.67e-05 \*\*\*  
## countryNZL 1.739e-01 1.248e-02 13.937 < 2e-16 \*\*\*  
## countryPOL -2.771e-02 1.143e-02 -2.425 0.015307 \*   
## countryPRT 8.642e-03 1.159e-02 0.746 0.455746   
## countryROU -5.969e-02 1.166e-02 -5.119 3.07e-07 \*\*\*  
## countryRUS 2.332e-02 1.155e-02 2.019 0.043483 \*   
## countrySGP -3.000e-03 1.198e-02 -0.251 0.802170   
## countrySRB 1.441e-02 1.178e-02 1.223 0.221457   
## countrySVK -1.962e-01 1.162e-02 -16.886 < 2e-16 \*\*\*  
## countrySWE -1.074e-01 1.184e-02 -9.065 < 2e-16 \*\*\*  
## countryUSA 1.583e-01 1.332e-02 11.883 < 2e-16 \*\*\*  
## genderMale -2.481e-02 2.855e-03 -8.690 < 2e-16 \*\*\*  
## years\_teacher 1.052e-02 5.164e-04 20.379 < 2e-16 \*\*\*  
## total\_time -4.054e-04 8.207e-05 -4.939 7.86e-07 \*\*\*  
## collab.L 5.581e-01 4.987e-03 111.923 < 2e-16 \*\*\*  
## collab.Q 2.289e-02 4.027e-03 5.685 1.31e-08 \*\*\*  
## collab.C 9.909e-03 2.828e-03 3.504 0.000458 \*\*\*  
## num\_subjects 3.865e-03 9.375e-04 4.122 3.75e-05 \*\*\*  
## age\_start -1.827e-03 2.964e-04 -6.162 7.20e-10 \*\*\*  
## age\_ratio -6.501e-01 3.376e-02 -19.258 < 2e-16 \*\*\*  
## avg\_prep 1.185e-01 2.419e-03 48.973 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4095 on 102060 degrees of freedom  
## Multiple R-squared: 0.2432, Adjusted R-squared: 0.2428   
## F-statistic: 728.8 on 45 and 102060 DF, p-value: < 2.2e-16

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

Now let's analyze the coefficients of the model. We can see from the summary of m\_final, that a lot of the coefficients have statistically significant p-values. However, this is somewhat expected given the large size of our dataset. In order to decide which predictors seem to be particularly helpful in predicting teacher satisfaction, let's look at the magnitude of the coefficients, and the confidence intervals of the coefficients and see how far away they are from zero.

## 2.5 % 97.5 %  
## (Intercept) 2.5446056168 2.6107828595  
## countryAUS 0.1544370581 0.2065364708  
## countryBFL 0.2162376679 0.2628675113  
## countryBGR -0.0621106314 -0.0131086034  
## countryBRA -0.0474309898 -0.0082239677  
## countryCAB 0.1409550028 0.1944857340  
## countryCHL 0.0701517371 0.1282405622  
## countryCSH -0.1810932988 -0.1367971009  
## countryCZE -0.1232271894 -0.0771033303  
## countryDNK 0.1341349901 0.1891609122  
## countryENG 0.0508734407 0.1003670766  
## countryESP 0.1025080614 0.1483901157  
## countryEST -0.1844189950 -0.1373907152  
## countryFIN 0.2199670606 0.2685473192  
## countryFRA 0.0212606155 0.0693666494  
## countryGEO -0.0603265110 -0.0041942933  
## countryHRV -0.0449988132 0.0021341728  
## countryISR 0.1110843462 0.1572040694  
## countryITA 0.0121367095 0.0584820025  
## countryJPN -0.1036941950 -0.0568233549  
## countryKOR -0.0624439988 -0.0146609729  
## countryLVA -0.1446573144 -0.0932597505  
## countryMEX 0.4474544275 0.4969105657  
## countryMYS 0.1970061160 0.2441672070  
## countryNLD 0.1669267408 0.2196742207  
## countryNOR 0.0240739587 0.0720921230  
## countryNZL 0.1494539038 0.1983683793  
## countryPOL -0.0501024535 -0.0053138615  
## countryPRT -0.0140669278 0.0313505037  
## countryROU -0.0825500065 -0.0368395290  
## countryRUS 0.0006822195 0.0459481754  
## countrySGP -0.0264718920 0.0204712591  
## countrySRB -0.0086880754 0.0375024609  
## countrySVK -0.2189259078 -0.1733899778  
## countrySWE -0.1305827272 -0.0841552035  
## countryUSA 0.1321660983 0.1843769961  
## genderMale -0.0304068529 -0.0192147587  
## years\_teacher 0.0095109852 0.0115351730  
## total\_time -0.0005662157 -0.0002444959  
## collab.L 0.5483659183 0.5679140548  
## collab.Q 0.0150014540 0.0307881765  
## collab.C 0.0043668816 0.0154510269  
## num\_subjects 0.0020271657 0.0057020340  
## age\_start -0.0024075014 -0.0012455937  
## age\_ratio -0.7162731539 -0.5839425065  
## avg\_prep 0.1137333579 0.1232165858

Using stepwise regression

Forward, backwards, and both

We can see from the results above, that with AIC as the criterion, all three stepwise models resulted in using the same 9 predictors, which were: collab, country, prep\_A, prep\_B, prep\_C, gender, age, total\_time, and years\_teacher.

Using the BIC criterion, all three models ended up using all of the same predictors as the AIC stepwise models, except without prep\_A.

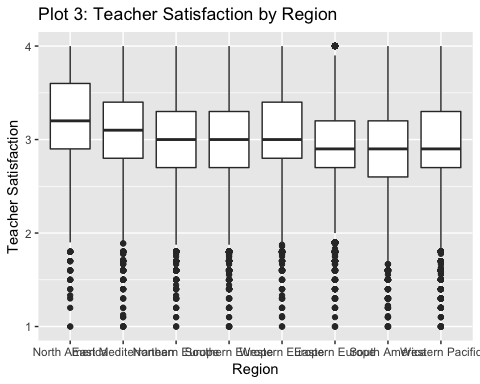
NOTE: need conclusion to this section!!!

### Analysis for Research Question 2

To address our second research question, and try to identify the differences across different regions for teacher job satisfaction and other classroom factors, we are going to try several techniques.

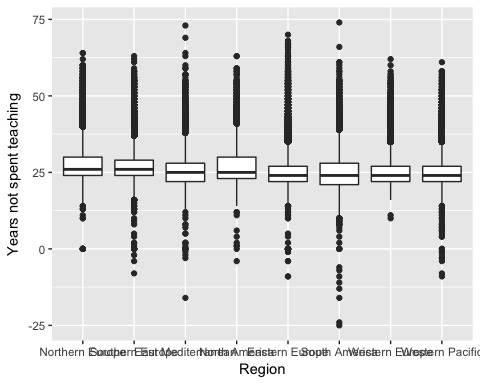
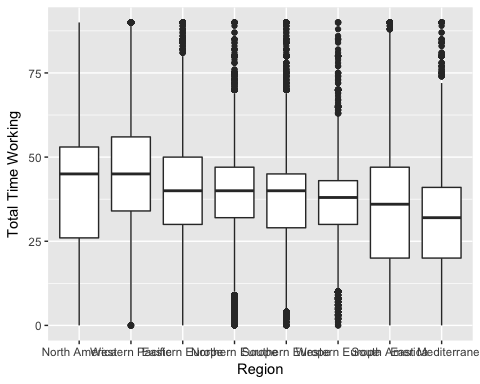
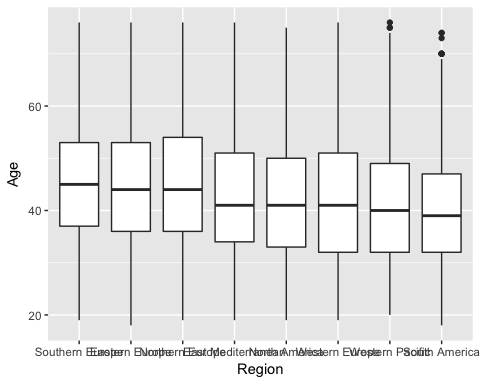
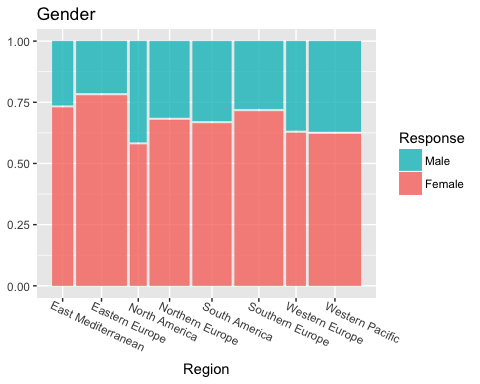
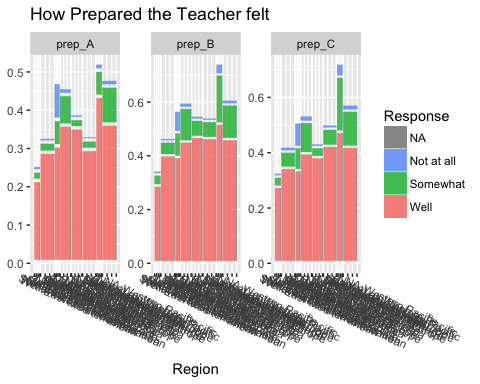
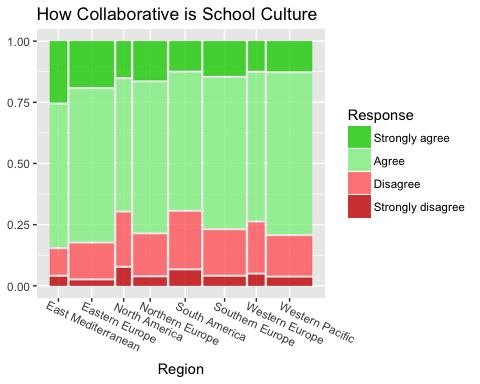
#### Plots

1. We are first going to plot sat\_avg vs region, using the region variable as a factor. This will visually tell us whether there might be a significant difference in satisfaction across regions.



As we stated above, 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). However, we would classify each region's average as "satisfied", as it falls above 2.5

1. We are going to plot each of the predictive factors that were found to be significant in Research Question 1 against region, to see if perhaps the region is affecting these variables and thus affecting satisfaction indirectly. These variables are collab, country, prep\_A, prep\_B, prep\_C, gender, age, total\_time, and years\_teacher. We will not, however, talk about country because of the obvious correlation between that and region.



For collaboration we can see that there is a median response rate of 3 (agree) for all regions, but that they differ in the distribution of quartiles. East Mediterranean's teachers had a quartile of responses in Strongly Agree, while North America, South America, and Western Europe had a quartile of responses in Disagree, and Eastern Europe, Northern Europe, Southern Europe, and Western Pacific had 50% of responses in Agree.

For prep\_a, the question that asked teachers whether they feel prepared for the “content of the subjects that I teach”, we see that East Mediterranean and Western Europe are the slight outliers, with 75% of teachers feeling very well prepared in the former, and more than half the teachers feeling less than very well prepared in the later.

For prep\_b, the question that asked teachers whether they feel prepared for the “pedagogy of the subjects I teach”, we see that East Mediterranean and Eastern Europe are the two regions that have "Very Well" rather than "Well" as the median.

For prep\_c, the question that asked teachers whether they feel prepared for the “classroom practice in the subjects that I teach”, the regions are split equally, with East Mediterranean, Eastern Europe, South America, and Southern Europe respondents median "Very Well", and the other 4 regions median "Well"

The median ages across regions range from mid 40s to high 30s, but the middle 50% of the data overlaps across all the regions. The same is true for total time working, which ranges from mid 30s to mid 40s, and years working as teacher, which ranges from mid to late teens.

1. We will then look at which variables from part (2) are significant when predicted with region. If any are statistically significant, we will graph sat\_avg as a function of that variable for each region.

## Analysis of Variance Table  
##   
## Response: as.numeric(collab)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 916 130.871 274.64 < 2.2e-16 \*\*\*  
## Residuals 112823 53763 0.477   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table  
##   
## Response: as.numeric(prep\_A)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 1917 273.824 709.62 < 2.2e-16 \*\*\*  
## Residuals 115316 44498 0.386   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table  
##   
## Response: as.numeric(prep\_B)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 2443 348.97 776.49 < 2.2e-16 \*\*\*  
## Residuals 114181 51315 0.45   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table  
##   
## Response: as.numeric(prep\_C)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 2743 391.87 826.62 < 2.2e-16 \*\*\*  
## Residuals 114722 54385 0.47   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

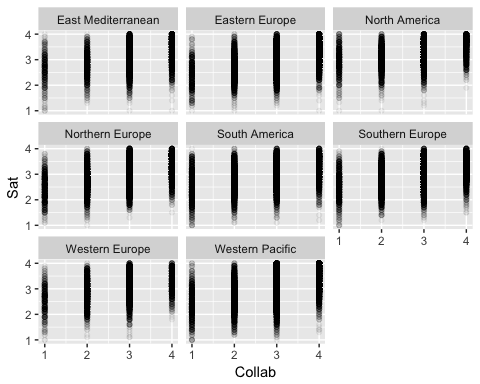
## Analysis of Variance Table  
##   
## Response: as.numeric(age)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 454885 64984 591.13 < 2.2e-16 \*\*\*  
## Residuals 117729 12941978 110   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table  
##   
## Response: as.numeric(total\_time)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 1319802 188543 697.46 < 2.2e-16 \*\*\*  
## Residuals 114376 30919267 270   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

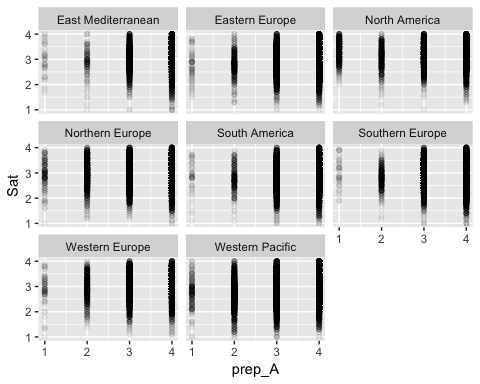
## Analysis of Variance Table  
##   
## Response: as.numeric(years\_teacher)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 265675 37954 352.71 < 2.2e-16 \*\*\*  
## Residuals 109539 11787034 108   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

It looks like all of the variables vary across regions.

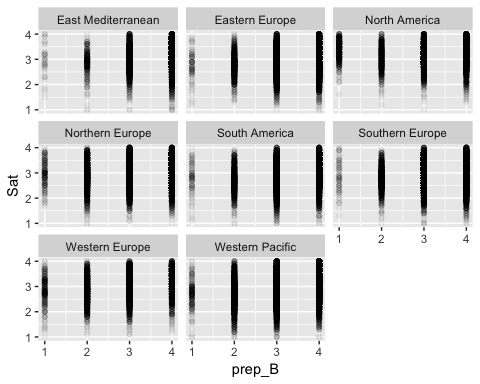
##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(collab), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.31076 -0.28038 -0.00401 0.29599 1.60876   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 2.305058 0.021275 108.348  
## regionEastern Europe -0.222979 0.025546 -8.729  
## regionNorth America 0.336349 0.028702 11.718  
## regionNorthern Europe -0.128803 0.026108 -4.933  
## regionSouth America -0.021319 0.025237 -0.845  
## regionSouthern Europe -0.044542 0.025110 -1.774  
## regionWestern Europe 0.031676 0.029608 1.070  
## regionWestern Pacific -0.210205 0.025300 -8.309  
## as.numeric(collab) 0.251426 0.006735 37.329  
## regionEastern Europe:as.numeric(collab) 0.031849 0.008157 3.904  
## regionNorth America:as.numeric(collab) -0.047453 0.009479 -5.006  
## regionNorthern Europe:as.numeric(collab) 0.024216 0.008413 2.879  
## regionSouth America:as.numeric(collab) -0.026088 0.008242 -3.165  
## regionSouthern Europe:as.numeric(collab) -0.003595 0.008100 -0.444  
## regionWestern Europe:as.numeric(collab) 0.001884 0.009774 0.193  
## regionWestern Pacific:as.numeric(collab) 0.044957 0.008167 5.505  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## regionEastern Europe < 2e-16 \*\*\*  
## regionNorth America < 2e-16 \*\*\*  
## regionNorthern Europe 8.09e-07 \*\*\*  
## regionSouth America 0.39825   
## regionSouthern Europe 0.07609 .   
## regionWestern Europe 0.28468   
## regionWestern Pacific < 2e-16 \*\*\*  
## as.numeric(collab) < 2e-16 \*\*\*  
## regionEastern Europe:as.numeric(collab) 9.45e-05 \*\*\*  
## regionNorth America:as.numeric(collab) 5.57e-07 \*\*\*  
## regionNorthern Europe:as.numeric(collab) 0.00400 \*\*   
## regionSouth America:as.numeric(collab) 0.00155 \*\*   
## regionSouthern Europe:as.numeric(collab) 0.65713   
## regionWestern Europe:as.numeric(collab) 0.84717   
## regionWestern Pacific:as.numeric(collab) 3.70e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4289 on 112562 degrees of freedom  
## (5298 observations deleted due to missingness)  
## Multiple R-squared: 0.1701, Adjusted R-squared: 0.17   
## F-statistic: 1538 on 15 and 112562 DF, p-value: < 2.2e-16



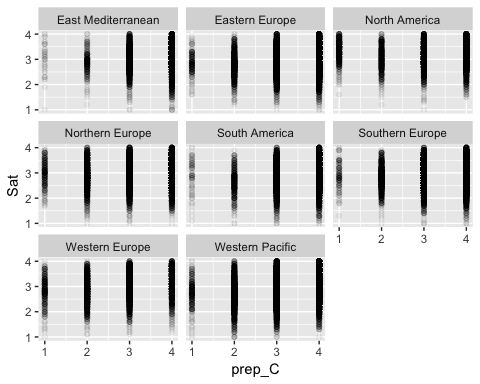
##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(prep\_A), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.21204 -0.28160 -0.00298 0.31332 1.45482   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 2.720271 0.038482 70.689  
## regionEastern Europe -0.103469 0.044654 -2.317  
## regionNorth America 0.527881 0.043677 12.086  
## regionNorthern Europe 0.224803 0.043109 5.215  
## regionSouth America -0.127831 0.046220 -2.766  
## regionSouthern Europe -0.028685 0.045148 -0.635  
## regionWestern Europe 0.024295 0.047771 0.509  
## regionWestern Pacific -0.343304 0.041695 -8.234  
## as.numeric(prep\_A) 0.095677 0.010204 9.376  
## regionEastern Europe:as.numeric(prep\_A) -0.009471 0.011897 -0.796  
## regionNorth America:as.numeric(prep\_A) -0.107714 0.011855 -9.086  
## regionNorthern Europe:as.numeric(prep\_A) -0.085275 0.011596 -7.354  
## regionSouth America:as.numeric(prep\_A) -0.008408 0.012429 -0.676  
## regionSouthern Europe:as.numeric(prep\_A) -0.018552 0.012035 -1.541  
## regionWestern Europe:as.numeric(prep\_A) -0.006167 0.013064 -0.472  
## regionWestern Pacific:as.numeric(prep\_A) 0.072536 0.011193 6.480  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## regionEastern Europe 0.02050 \*   
## regionNorth America < 2e-16 \*\*\*  
## regionNorthern Europe 1.84e-07 \*\*\*  
## regionSouth America 0.00568 \*\*   
## regionSouthern Europe 0.52519   
## regionWestern Europe 0.61105   
## regionWestern Pacific < 2e-16 \*\*\*  
## as.numeric(prep\_A) < 2e-16 \*\*\*  
## regionEastern Europe:as.numeric(prep\_A) 0.42596   
## regionNorth America:as.numeric(prep\_A) < 2e-16 \*\*\*  
## regionNorthern Europe:as.numeric(prep\_A) 1.94e-13 \*\*\*  
## regionSouth America:as.numeric(prep\_A) 0.49874   
## regionSouthern Europe:as.numeric(prep\_A) 0.12320   
## regionWestern Europe:as.numeric(prep\_A) 0.63690   
## regionWestern Pacific:as.numeric(prep\_A) 9.20e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4614 on 112518 degrees of freedom  
## (5342 observations deleted due to missingness)  
## Multiple R-squared: 0.04158, Adjusted R-squared: 0.04145   
## F-statistic: 325.4 on 15 and 112518 DF, p-value: < 2.2e-16



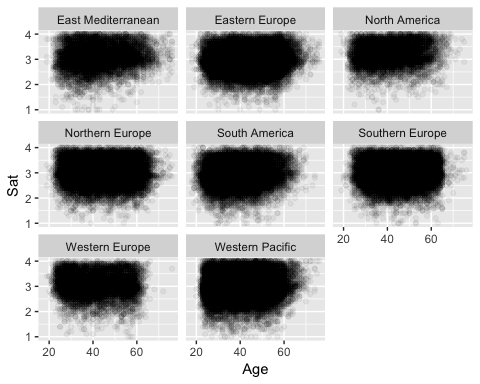
##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(prep\_B), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.20512 -0.28814 -0.00638 0.31032 1.46022   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 2.639253 0.032741 80.610  
## regionEastern Europe -0.091273 0.037784 -2.416  
## regionNorth America 0.539326 0.038980 13.836  
## regionNorthern Europe 0.197644 0.037035 5.337  
## regionSouth America -0.190891 0.039135 -4.878  
## regionSouthern Europe -0.060986 0.037596 -1.622  
## regionWestern Europe -0.055016 0.039281 -1.401  
## regionWestern Pacific -0.282778 0.036033 -7.848  
## as.numeric(prep\_B) 0.121042 0.008938 13.543  
## regionEastern Europe:as.numeric(prep\_B) -0.011002 0.010402 -1.058  
## regionNorth America:as.numeric(prep\_B) -0.112195 0.010980 -10.218  
## regionNorthern Europe:as.numeric(prep\_B) -0.076779 0.010331 -7.432  
## regionSouth America:as.numeric(prep\_B) 0.016084 0.010980 1.465  
## regionSouthern Europe:as.numeric(prep\_B) -0.004576 0.010419 -0.439  
## regionWestern Europe:as.numeric(prep\_B) 0.033675 0.011349 2.967  
## regionWestern Pacific:as.numeric(prep\_B) 0.062260 0.010018 6.215  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## regionEastern Europe 0.01571 \*   
## regionNorth America < 2e-16 \*\*\*  
## regionNorthern Europe 9.48e-08 \*\*\*  
## regionSouth America 1.07e-06 \*\*\*  
## regionSouthern Europe 0.10478   
## regionWestern Europe 0.16134   
## regionWestern Pacific 4.27e-15 \*\*\*  
## as.numeric(prep\_B) < 2e-16 \*\*\*  
## regionEastern Europe:as.numeric(prep\_B) 0.29023   
## regionNorth America:as.numeric(prep\_B) < 2e-16 \*\*\*  
## regionNorthern Europe:as.numeric(prep\_B) 1.08e-13 \*\*\*  
## regionSouth America:as.numeric(prep\_B) 0.14296   
## regionSouthern Europe:as.numeric(prep\_B) 0.66054   
## regionWestern Europe:as.numeric(prep\_B) 0.00301 \*\*   
## regionWestern Pacific:as.numeric(prep\_B) 5.16e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4579 on 111444 degrees of freedom  
## (6416 observations deleted due to missingness)  
## Multiple R-squared: 0.05586, Adjusted R-squared: 0.05573   
## F-statistic: 439.5 on 15 and 111444 DF, p-value: < 2.2e-16



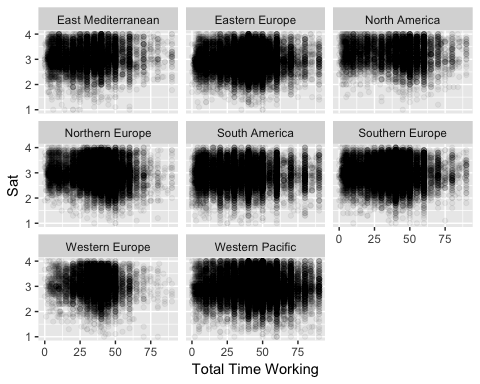
##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(prep\_C), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2032 -0.2808 -0.0072 0.3143 1.4659   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 2.678712 0.033777 79.306  
## regionEastern Europe -0.100252 0.038296 -2.618  
## regionNorth America 0.480853 0.039608 12.140  
## regionNorthern Europe 0.126972 0.037908 3.349  
## regionSouth America -0.245045 0.041041 -5.971  
## regionSouthern Europe -0.074309 0.038378 -1.936  
## regionWestern Europe -0.069686 0.039495 -1.764  
## regionWestern Pacific -0.328525 0.036802 -8.927  
## as.numeric(prep\_C) 0.109496 0.009166 11.945  
## regionEastern Europe:as.numeric(prep\_C) -0.008917 0.010462 -0.852  
## regionNorth America:as.numeric(prep\_C) -0.094951 0.011025 -8.612  
## regionNorthern Europe:as.numeric(prep\_C) -0.056835 0.010462 -5.432  
## regionSouth America:as.numeric(prep\_C) 0.025142 0.011281 2.229  
## regionSouthern Europe:as.numeric(prep\_C) -0.001740 0.010548 -0.165  
## regionWestern Europe:as.numeric(prep\_C) 0.037648 0.011293 3.334  
## regionWestern Pacific:as.numeric(prep\_C) 0.074378 0.010147 7.330  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## regionEastern Europe 0.008851 \*\*   
## regionNorth America < 2e-16 \*\*\*  
## regionNorthern Europe 0.000810 \*\*\*  
## regionSouth America 2.37e-09 \*\*\*  
## regionSouthern Europe 0.052841 .   
## regionWestern Europe 0.077662 .   
## regionWestern Pacific < 2e-16 \*\*\*  
## as.numeric(prep\_C) < 2e-16 \*\*\*  
## regionEastern Europe:as.numeric(prep\_C) 0.394068   
## regionNorth America:as.numeric(prep\_C) < 2e-16 \*\*\*  
## regionNorthern Europe:as.numeric(prep\_C) 5.57e-08 \*\*\*  
## regionSouth America:as.numeric(prep\_C) 0.025835 \*   
## regionSouthern Europe:as.numeric(prep\_C) 0.868972   
## regionWestern Europe:as.numeric(prep\_C) 0.000857 \*\*\*  
## regionWestern Pacific:as.numeric(prep\_C) 2.32e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4579 on 111962 degrees of freedom  
## (5898 observations deleted due to missingness)  
## Multiple R-squared: 0.0557, Adjusted R-squared: 0.05558   
## F-statistic: 440.3 on 15 and 111962 DF, p-value: < 2.2e-16



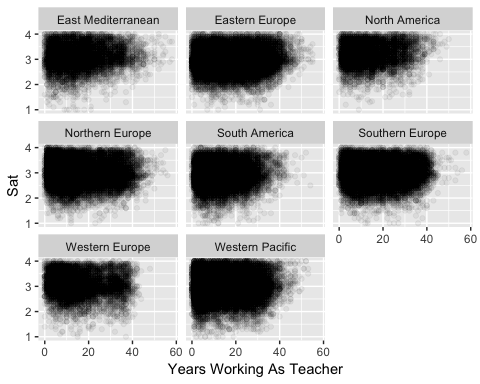
##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(age), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.18443 -0.28116 -0.01673 0.32175 1.18078   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 2.8828651 0.0207816 138.722  
## regionEastern Europe -0.0300498 0.0249326 -1.205  
## regionNorth America 0.1507126 0.0312355 4.825  
## regionNorthern Europe 0.1250463 0.0258885 4.830  
## regionSouth America -0.1441969 0.0261786 -5.508  
## regionSouthern Europe 0.0590989 0.0258715 2.284  
## regionWestern Europe 0.2539550 0.0295529 8.593  
## regionWestern Pacific -0.0265292 0.0244597 -1.085  
## as.numeric(age) 0.0045099 0.0004681 9.634  
## regionEastern Europe:as.numeric(age) -0.0026959 0.0005585 -4.827  
## regionNorth America:as.numeric(age) -0.0003197 0.0007165 -0.446  
## regionNorthern Europe:as.numeric(age) -0.0051077 0.0005764 -8.861  
## regionSouth America:as.numeric(age) -0.0002703 0.0006089 -0.444  
## regionSouthern Europe:as.numeric(age) -0.0037842 0.0005750 -6.581  
## regionWestern Europe:as.numeric(age) -0.0066228 0.0006792 -9.751  
## regionWestern Pacific:as.numeric(age) -0.0021435 0.0005604 -3.825  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## regionEastern Europe 0.228112   
## regionNorth America 1.40e-06 \*\*\*  
## regionNorthern Europe 1.37e-06 \*\*\*  
## regionSouth America 3.63e-08 \*\*\*  
## regionSouthern Europe 0.022354 \*   
## regionWestern Europe < 2e-16 \*\*\*  
## regionWestern Pacific 0.278099   
## as.numeric(age) < 2e-16 \*\*\*  
## regionEastern Europe:as.numeric(age) 1.39e-06 \*\*\*  
## regionNorth America:as.numeric(age) 0.655491   
## regionNorthern Europe:as.numeric(age) < 2e-16 \*\*\*  
## regionSouth America:as.numeric(age) 0.657062   
## regionSouthern Europe:as.numeric(age) 4.68e-11 \*\*\*  
## regionWestern Europe:as.numeric(age) < 2e-16 \*\*\*  
## regionWestern Pacific:as.numeric(age) 0.000131 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4651 on 113859 degrees of freedom  
## (4001 observations deleted due to missingness)  
## Multiple R-squared: 0.02708, Adjusted R-squared: 0.02695   
## F-statistic: 211.3 on 15 and 113859 DF, p-value: < 2.2e-16



##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(total\_time), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.19981 -0.27852 -0.02569 0.32375 1.10924   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 3.0377437 0.0112737 269.454  
## regionEastern Europe -0.1124220 0.0142541 -7.887  
## regionNorth America 0.2310968 0.0174589 13.237  
## regionNorthern Europe 0.0459195 0.0158186 2.903  
## regionSouth America -0.1337803 0.0139587 -9.584  
## regionSouthern Europe -0.0546815 0.0143152 -3.820  
## regionWestern Europe 0.0812770 0.0204217 3.980  
## regionWestern Pacific -0.0515410 0.0141406 -3.645  
## as.numeric(total\_time) 0.0011916 0.0003111 3.830  
## regionEastern Europe:as.numeric(total\_time) -0.0010067 0.0003732 -2.698  
## regionNorth America:as.numeric(total\_time) -0.0027256 0.0004319 -6.310  
## regionNorthern Europe:as.numeric(total\_time) -0.0038341 0.0004116 -9.316  
## regionSouth America:as.numeric(total\_time) -0.0010920 0.0003726 -2.931  
## regionSouthern Europe:as.numeric(total\_time) -0.0014185 0.0003815 -3.718  
## regionWestern Europe:as.numeric(total\_time) -0.0030944 0.0005426 -5.703  
## regionWestern Pacific:as.numeric(total\_time) -0.0019331 0.0003584 -5.393  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## regionEastern Europe 3.12e-15 \*\*\*  
## regionNorth America < 2e-16 \*\*\*  
## regionNorthern Europe 0.003698 \*\*   
## regionSouth America < 2e-16 \*\*\*  
## regionSouthern Europe 0.000134 \*\*\*  
## regionWestern Europe 6.90e-05 \*\*\*  
## regionWestern Pacific 0.000268 \*\*\*  
## as.numeric(total\_time) 0.000128 \*\*\*  
## regionEastern Europe:as.numeric(total\_time) 0.006982 \*\*   
## regionNorth America:as.numeric(total\_time) 2.80e-10 \*\*\*  
## regionNorthern Europe:as.numeric(total\_time) < 2e-16 \*\*\*  
## regionSouth America:as.numeric(total\_time) 0.003382 \*\*   
## regionSouthern Europe:as.numeric(total\_time) 0.000201 \*\*\*  
## regionWestern Europe:as.numeric(total\_time) 1.18e-08 \*\*\*  
## regionWestern Pacific:as.numeric(total\_time) 6.94e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4651 on 111984 degrees of freedom  
## (5876 observations deleted due to missingness)  
## Multiple R-squared: 0.02506, Adjusted R-squared: 0.02493   
## F-statistic: 191.9 on 15 and 111984 DF, p-value: < 2.2e-16



##   
## Call:  
## lm(formula = sat\_avg ~ region \* as.numeric(years\_teacher), data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.16796 -0.27899 -0.01516 0.32420 1.15145   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 3.0115921 0.0101961  
## regionEastern Europe -0.1197846 0.0121146  
## regionNorth America 0.1161870 0.0148371  
## regionNorthern Europe -0.0135535 0.0122295  
## regionSouth America -0.1669378 0.0127656  
## regionSouthern Europe -0.0479019 0.0123454  
## regionWestern Europe 0.0682944 0.0140515  
## regionWestern Pacific -0.0666693 0.0117167  
## as.numeric(years\_teacher) 0.0041929 0.0005123  
## regionEastern Europe:as.numeric(years\_teacher) -0.0020702 0.0005971  
## regionNorth America:as.numeric(years\_teacher) 0.0002722 0.0008043  
## regionNorthern Europe:as.numeric(years\_teacher) -0.0051955 0.0006152  
## regionSouth America:as.numeric(years\_teacher) -0.0002995 0.0006909  
## regionSouthern Europe:as.numeric(years\_teacher) -0.0036030 0.0006120  
## regionWestern Europe:as.numeric(years\_teacher) -0.0061760 0.0007233  
## regionWestern Pacific:as.numeric(years\_teacher) -0.0037246 0.0006025  
## t value Pr(>|t|)   
## (Intercept) 295.367 < 2e-16 \*\*\*  
## regionEastern Europe -9.888 < 2e-16 \*\*\*  
## regionNorth America 7.831 4.89e-15 \*\*\*  
## regionNorthern Europe -1.108 0.267752   
## regionSouth America -13.077 < 2e-16 \*\*\*  
## regionSouthern Europe -3.880 0.000104 \*\*\*  
## regionWestern Europe 4.860 1.17e-06 \*\*\*  
## regionWestern Pacific -5.690 1.27e-08 \*\*\*  
## as.numeric(years\_teacher) 8.184 2.76e-16 \*\*\*  
## regionEastern Europe:as.numeric(years\_teacher) -3.467 0.000526 \*\*\*  
## regionNorth America:as.numeric(years\_teacher) 0.338 0.735063   
## regionNorthern Europe:as.numeric(years\_teacher) -8.445 < 2e-16 \*\*\*  
## regionSouth America:as.numeric(years\_teacher) -0.433 0.664718   
## regionSouthern Europe:as.numeric(years\_teacher) -5.887 3.94e-09 \*\*\*  
## regionWestern Europe:as.numeric(years\_teacher) -8.538 < 2e-16 \*\*\*  
## regionWestern Pacific:as.numeric(years\_teacher) -6.182 6.36e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4655 on 106824 degrees of freedom  
## (11036 observations deleted due to missingness)  
## Multiple R-squared: 0.0246, Adjusted R-squared: 0.02446   
## F-statistic: 179.6 on 15 and 106824 DF, p-value: < 2.2e-16



#### One-way ANOVA Test

After the analysis above is completed, we'll run an ANOVA test to see if there are statistically significant variances in mean sat\_avg across regions.

##   
## Call:  
## lm(formula = sat\_avg ~ region, data = teacher\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2079 -0.2811 -0.0079 0.3229 1.0931   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.077104 0.005168 595.359 < 2e-16 \*\*\*  
## regionEastern Europe -0.144465 0.006105 -23.662 < 2e-16 \*\*\*  
## regionNorth America 0.130792 0.007698 16.991 < 2e-16 \*\*\*  
## regionNorthern Europe -0.096031 0.006374 -15.065 < 2e-16 \*\*\*  
## regionSouth America -0.170238 0.006433 -26.465 < 2e-16 \*\*\*  
## regionSouthern Europe -0.102328 0.006145 -16.651 < 2e-16 \*\*\*  
## regionWestern Europe -0.027556 0.007425 -3.711 0.000206 \*\*\*  
## regionWestern Pacific -0.124727 0.006088 -20.487 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4659 on 113992 degrees of freedom  
## (3876 observations deleted due to missingness)  
## Multiple R-squared: 0.02374, Adjusted R-squared: 0.02368   
## F-statistic: 396 on 7 and 113992 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Response: sat\_avg  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 601.8 85.965 395.98 < 2.2e-16 \*\*\*  
## Residuals 113992 24747.4 0.217   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We will also see if there are statistically significant variances across regions in any of the significant predictors from part (1) that were particularly predictive of teacher satisfaction.

## Analysis of Variance Table  
##   
## Response: sat\_avg  
## Df Sum Sq Mean Sq F value Pr(>F)   
## region 7 601.8 85.965 395.98 < 2.2e-16 \*\*\*  
## Residuals 113992 24747.4 0.217   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Let's see how much of the R-squared in our model from part 1 was due to just the country predictor:

##   
## Call:  
## lm(formula = I(sat\_avg^2) ~ country, data = x.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3654 -1.7517 -0.2096 1.7218 7.9983   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.48445 0.05901 160.725 < 2e-16 \*\*\*  
## countryAUS 0.42479 0.08595 4.942 7.74e-07 \*\*\*  
## countryBFL 0.79886 0.07696 10.380 < 2e-16 \*\*\*  
## countryBGR -0.71253 0.08049 -8.853 < 2e-16 \*\*\*  
## countryBRA -0.99574 0.06469 -15.394 < 2e-16 \*\*\*  
## countryCAB 0.56948 0.08811 6.464 1.03e-10 \*\*\*  
## countryCHL 0.13490 0.09611 1.404 0.160455   
## countryCSH -1.43483 0.07310 -19.630 < 2e-16 \*\*\*  
## countryCZE -0.98612 0.07595 -12.984 < 2e-16 \*\*\*  
## countryDNK 0.57349 0.09037 6.346 2.22e-10 \*\*\*  
## countryENG -0.32528 0.08163 -3.985 6.76e-05 \*\*\*  
## countryESP 0.09667 0.07566 1.278 0.201337   
## countryEST -1.34569 0.07728 -17.414 < 2e-16 \*\*\*  
## countryFIN 0.63684 0.07896 8.065 7.40e-16 \*\*\*  
## countryFRA -0.90825 0.07832 -11.596 < 2e-16 \*\*\*  
## countryGEO -0.19577 0.09216 -2.124 0.033660 \*   
## countryHRV -0.61563 0.07772 -7.921 2.38e-15 \*\*\*  
## countryISR 0.65011 0.07633 8.517 < 2e-16 \*\*\*  
## countryITA -0.39001 0.07566 -5.155 2.54e-07 \*\*\*  
## countryJPN -1.48276 0.07520 -19.718 < 2e-16 \*\*\*  
## countryKOR -1.10748 0.07832 -14.140 < 2e-16 \*\*\*  
## countryLVA -1.04950 0.08448 -12.423 < 2e-16 \*\*\*  
## countryMEX 1.65869 0.08082 20.523 < 2e-16 \*\*\*  
## countryMYS 1.25687 0.07720 16.280 < 2e-16 \*\*\*  
## countryNLD 0.28267 0.08681 3.256 0.001129 \*\*   
## countryNOR 0.13493 0.07855 1.718 0.085842 .   
## countryNZL 0.44180 0.08048 5.490 4.04e-08 \*\*\*  
## countryPOL -0.60515 0.07405 -8.173 3.05e-16 \*\*\*  
## countryPRT -0.64802 0.07469 -8.676 < 2e-16 \*\*\*  
## countryROU -0.23356 0.07567 -3.086 0.002026 \*\*   
## countryRUS -0.09517 0.07419 -1.283 0.199591   
## countrySGP -0.88502 0.07648 -11.573 < 2e-16 \*\*\*  
## countrySRB -0.27500 0.07630 -3.604 0.000313 \*\*\*  
## countrySVK -1.64505 0.07493 -21.955 < 2e-16 \*\*\*  
## countrySWE -1.21307 0.07622 -15.916 < 2e-16 \*\*\*  
## countryUSA 0.28187 0.08614 3.272 0.001068 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.664 on 102070 degrees of freedom  
## Multiple R-squared: 0.0822, Adjusted R-squared: 0.08189   
## F-statistic: 261.2 on 35 and 102070 DF, p-value: < 2.2e-16

0.08257/0.2474 is roughly 33.4%, therefore the country variable alone accounts for 33% of the variability that our final model in Part 1 was able to explain.

# 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.

**insert findings here lol**

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

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 the data particularly 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 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, but are still a concern about the data.

One area that our analysis would have benefited from other data sources would be countries. We created our regions variable using the World Health Organization regions categories, so that it was easier to visualize and interpret our results by group of 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 satistically significant difference.

## Code Appendix

First, here is the data cleaning code from the Data Exploration Assignment

\*note, we need to figure out how many code chunks in final, and then remove the last one here, because otherwise the datacleaning code will be printed twice

knitr::opts\_chunk$set(echo = FALSE, warning=FALSE)  
library("foreign")  
library("dplyr")  
library("tidyr")  
library("countrycode")  
library("rworldmap")  
library(ggplot2)  
library(reshape2)  
library(FNN)  
library(grid)  
library(gridExtra)  
library(leaps)  
library(GGally)  
library(ggmosaic)  
# 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 they 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\_Lang", "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 satisfaction 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 Disagree", "Disagree", "Strongly Agree", and "Agree", are currently factors. Given that these categories do have a natural ordering, we're going to convert them into ordered factors.  
teacher\_data$prep\_A <- factor(teacher\_data$prep\_A, levels=c("Not at all", "Somewhat", "Well", "Very well"), ordered=TRUE)  
teacher\_data$prep\_B <- factor(teacher\_data$prep\_B, levels=c("Not at all", "Somewhat", "Well", "Very well"), ordered=TRUE)  
teacher\_data$prep\_C <- factor(teacher\_data$prep\_C, levels=c("Not at all", "Somewhat", "Well", "Very well"), ordered=TRUE)  
teacher\_data$feedback\_salary <- factor(teacher\_data$feedback\_salary, levels=c("No positive change", "A small change", "A moderate change", "A large change"), ordered=TRUE)  
teacher\_data$collab <- factor(teacher\_data$collab, levels=c("Strongly disagree", "Disagree", "Agree", "Strongly agree"), ordered=TRUE)  
  
# 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 Europe"  
  
# 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", "SGP")] <- "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 high)", 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\_boxplot() +  
 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, median), 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  
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")  
length(teacher\_data[teacher\_data$total\_time>=90 & !is.na(teacher\_data$total\_time),]$total\_time)  
length(teacher\_data[teacher\_data$total\_time>90 & !is.na(teacher\_data$total\_time),]$total\_time)  
teacher\_data[teacher\_data$total\_time > 90 & !is.na(teacher\_data$total\_time), "total\_time"] <- NA  
teacher\_data %>% ggplot(aes(x=age, y=years\_teacher)) + geom\_point(alpha=0.5,shape=".") + geom\_abline(intercept =0, slope=1, color="red") + geom\_abline(intercept = -15, slope=1, color="yellow")+ geom\_jitter(shape=".") + labs(title="Reported Age vs Years Teaching", x="Age (years)", y="Years Teaching (years)")  
age\_start <- teacher\_data$age - teacher\_data$years\_teacher  
nrow(teacher\_data[age\_start < 15 & !is.na(age\_start),])  
teacher\_data[age\_start < 15 & !is.na(age\_start), "years\_teacher"] <- NA  
#ggpairs(teacher\_data)  
# 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\_Lang", "Tech", "Arts", "PE", "Rel\_Ethics", "Practical", "Other")  
subject\_indices <- c(which(colnames(teacher\_data) %in% subject\_names))  
  
# Some columns also have many more NA values, lets make a list of the columns that have NA values 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)  
 }  
}  
#teacher\_data[, -c(sat\_indices, na\_cols, subject\_indices,1, 37)] %>%  
# gather(-sat\_avg, -gender, key = "var", value = "value") %>%  
# ggplot(aes(x = value, y = sat\_avg, color = gender)) +  
# geom\_point(alpha=0.05) +  
# facet\_wrap(~ var, scales = "free") +  
# theme\_bw()  
x <- na.omit(teacher\_data[, -c(sat\_indices, na\_cols, subject\_indices)])  
# Try model with country variable and not region  
m\_country <- lm(sat\_avg ~ ., data=x[,-which(names(x) %in% c("region"))])  
summary(m\_country)  
# Try model with region variable and not country  
m\_region <- lm(sat\_avg ~ ., data=x[,-which(names(x) %in% c("country"))])  
summary(m\_region)  
cor(teacher\_data$age, teacher\_data$years\_teacher)  
teacher\_data$age\_start <- age\_start  
teacher\_data$age\_ratio <- teacher\_data$years\_teacher/teacher\_data$age  
teacher\_data %>% ggplot(aes(x=age, y=sat\_avg)) + geom\_point(alpha=0.05)  
teacher\_data %>% ggplot(aes(x=years\_teacher, y=sat\_avg)) + geom\_point(alpha=0.05)  
teacher\_data %>% ggplot(aes(x=age\_start, y=sat\_avg)) + geom\_point(alpha=0.05)  
teacher\_data %>% ggplot(aes(x=age\_ratio, y=sat\_avg)) + geom\_point(alpha=0.05)  
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)  
prep\_indices <- grep("prep\_", names(teacher\_data))  
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"))])  
x.new %>% ggplot(aes(x=sat\_avg)) + geom\_histogram(binwidth = 0.2) + labs(title="Histogram of Teacher Satisfaction", x="Teacher Satisfaction (low to high)", y="Count")  
#QUESTION: is there a way to do a histogram of all the variables in x?  
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")  
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")  
m\_final <- lm(sat\_avg ~ ., data=x.new)  
summary(m\_final)  
plot(m\_final)  
confint(m\_final)  
# Remove SAT questions, since they obviously explain our predictor variable, and then remove NAs for stepwise regression  
# Also remove subject columns, we have a column that counts the number of subjects each teacher is teaching, to represent this information instead.  
m.full <- lm(sat\_avg ~., data=x[,-which(names(x) %in% c("region"))])  
m.empty <- lm(sat\_avg ~ 1, data=x)  
# With AIC  
m.forward.aic <- step(m.empty, scope=list(upper=m.full),  
 data=x, direction="forward", trace=FALSE)  
m.backward.aic <- step(m.full, scope=list(lower=m.empty),  
 data=x, direction="backward", trace=FALSE)  
m.both.aic <- step(m.empty, scope=list(upper=m.full),  
 data=x, direction="both", trace=FALSE)  
  
# With BIC  
m.forward.bic <- step(m.empty, scope=list(upper=m.full),  
 data=x, direction="forward", trace=FALSE, k=log(nrow(x)))  
m.backward.bic <- step(m.full, scope=list(lower=m.empty),  
 data=x, direction="backward", trace=FALSE, k=log(nrow(x)))  
m.both.bic <- step(m.empty, scope=list(upper=m.full),  
 data=x, direction="both", trace=FALSE, k=log(nrow(x)))  
## p3, used as an example plot earlier in this write-up, is a plot of sat\_avg vs region  
## will fix labels once we figure out how big it can be in the final writeup  
p3  
# collab plot  
teacher\_data %>% filter(!is.na(collab)) %>% ggplot(aes(x = product(collab, region), fill=factor(collab)), na.rm=TRUE) + geom\_mosaic() + scale\_fill\_manual(values=c("red3","indianred1", "lightgreen", "green3")) + theme(axis.text.x=element\_text(angle=-25, hjust= .1)) + labs(x="Region", title='How Collaborative is School Culture') + guides(fill=guide\_legend(title = "Response", reverse = TRUE))  
  
# prep plot  
teacher\_data[, c("prep\_A", "prep\_B", "prep\_C","region")] %>%  
 gather(-region, key = "var", value = "value") %>%  
 mutate(value = factor(value, levels = c("Very Well", "Well", "Somewhat", "Not at all"))) %>%  
 ggplot(aes(x = product(value,region), fill=factor(value))) + geom\_mosaic() +  
 theme(axis.text.x=element\_text(angle=-25, hjust= .1)) + labs(x="Region", title='How Prepared the Teacher felt') + guides(fill=guide\_legend(title = "Response", reverse=TRUE)) +   
 facet\_wrap(~ var, scales = "free")   
  
# gender plot  
teacher\_data %>% filter(!is.na(gender)) %>% ggplot(aes(x = product(gender, region), fill=factor(gender)), na.rm=TRUE) + geom\_mosaic() + theme(axis.text.x=element\_text(angle=-25, hjust= .1)) + labs(x="Region", title='Gender') + guides(fill=guide\_legend(title = "Response", reverse = TRUE))  
  
# age plot  
teacher\_data %>% filter(!is.na(age)) %>% ggplot(aes(x=reorder(region, -as.numeric(age), median), y=as.numeric(age))) + geom\_boxplot() + xlab("Region") + ylab("Age")  
  
# total\_time plot  
teacher\_data %>% filter(!is.na(total\_time)) %>% ggplot(aes(x=reorder(region, -as.numeric(total\_time), median), y=as.numeric(total\_time))) + geom\_boxplot() + xlab("Region") + ylab("Total Time Working")  
  
# age\_start plot  
  
teacher\_data %>% filter(!is.na(age\_start)) %>% ggplot(aes(x=reorder(region, -as.numeric(age\_start), median), y=as.numeric(age\_start))) + geom\_boxplot() + xlab("Region") + ylab("Years not spent teaching")  
  
# Check which variables have a significant correlation with region  
anova(lm(as.numeric(collab) ~ region, data=teacher\_data))  
anova(lm(as.numeric(prep\_A) ~ region, data=teacher\_data))  
anova(lm(as.numeric(prep\_B) ~ region, data=teacher\_data))  
anova(lm(as.numeric(prep\_C) ~ region, data=teacher\_data))  
anova(lm(as.numeric(age) ~ region, data=teacher\_data))  
anova(lm(as.numeric(total\_time) ~ region, data=teacher\_data))  
anova(lm(as.numeric(years\_teacher) ~ region, data=teacher\_data))  
# Collab and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(collab), data=teacher\_data))  
  
# Collab and region plot  
teacher\_data %>% filter(!is.na(as.numeric(collab))) %>% ggplot(aes(x=as.numeric(collab), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("Collab") + ylab("Sat") + facet\_wrap(~region)  
  
# prep\_A and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(prep\_A), data=teacher\_data))  
  
# prep\_A and region plot  
teacher\_data %>% filter(!is.na(as.numeric(prep\_A))) %>% ggplot(aes(x=as.numeric(prep\_A), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("prep\_A") + ylab("Sat") + facet\_wrap(~region)  
  
# prep\_B and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(prep\_B), data=teacher\_data))  
  
# prep\_B and region plot  
teacher\_data %>% filter(!is.na(as.numeric(prep\_B))) %>% ggplot(aes(x=as.numeric(prep\_B), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("prep\_B") + ylab("Sat") + facet\_wrap(~region)  
  
# prep\_C and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(prep\_C), data=teacher\_data))  
  
# prep\_C and region plot  
teacher\_data %>% filter(!is.na(as.numeric(prep\_C))) %>% ggplot(aes(x=as.numeric(prep\_C), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("prep\_C") + ylab("Sat") + facet\_wrap(~region)  
  
# age and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(age), data=teacher\_data))  
  
# age and region plot  
teacher\_data %>% filter(!is.na(as.numeric(age))) %>% ggplot(aes(x=as.numeric(age), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("Age") + ylab("Sat") + facet\_wrap(~region)  
  
# total\_time and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(total\_time), data=teacher\_data))  
  
# total\_time and region plot  
teacher\_data %>% filter(!is.na(as.numeric(total\_time))) %>% ggplot(aes(x=as.numeric(total\_time), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("Total Time Working") + ylab("Sat") + facet\_wrap(~region)  
  
# years\_working and region LM  
# years\_teacher and region LM  
summary(lm(sat\_avg ~ region\*as.numeric(years\_teacher), data=teacher\_data))  
  
# years\_teacher and region plot  
teacher\_data %>% filter(!is.na(as.numeric(years\_teacher))) %>% ggplot(aes(x=as.numeric(years\_teacher), y=as.numeric(sat\_avg))) + geom\_point(alpha=0.05) + xlab("Years Working As Teacher") + ylab("Sat") + facet\_wrap(~region)  
# Regression of sat\_avg vs region  
msatreg <- lm(sat\_avg ~ region, data=teacher\_data)  
summary(msatreg)  
  
# ANOVA of regression  
anova(msatreg)  
# Template code for anova of predictor against region TODO: switch collab with predictive factors after found:  
mcollabreg <- lm(collab ~ region, data=teacher\_data)  
anova(msatreg)  
m\_countryonly <- lm(I(sat\_avg^2) ~ country, data=x.new)  
summary(m\_countryonly)  
nrow(teacher\_data[teacher\_data$sat\_avg >=2.5,])/nrow(teacher\_data)  
# 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.missings=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 related to job satisfaction, which are columns TT2G46A through TT2G46J, as well as a few additional columns which we also think might be useful  
data <- cbind(lower\_second\_data[,cols],lower\_second\_data[,grep("TT2G46", colnames(lower\_second\_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, train\_stat = TT2G11, prep\_A=TT2G13A, prep\_B=TT2G13B, prep\_C=TT2G13C, total\_time=TT2G16, feedback\_salary=TT2G30G, collab=TT2G44E, num\_students=TT2G38, satisfaction=TJOBSATS)  
  
# Columns that start with TT2G15 correspond to several questions asking what subjects teachers 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 for 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 = TT2G15D, For\_Lang = TT2G15E, Gk\_Ln=TT2G15F, Tech=TT2G15G, Arts=TT2G15H, PE=TT2G15I, Rel\_Ethics=TT2G15J, 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 other columns  
  
# All ten columns with responses related to job satisfaction are factors, with 4 levels: "Strongly 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 corresponding 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 so 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 teacher 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 represent No and 1 represent Yes  
data$train\_stat <- relevel(data$train\_stat, "No")