LAY SUMMARY

- **The goals of this analysis are following.**

  1. Give a general description of the work landscape for those with a college degree in the US, as surveyed in 2013
  2. Build a regression model to predict annual salary
  3. Build a regression model to predict job satisfaction
  4. Use our analysis to fact-check news outlets.
  5. Convey our findings in a technical report and in plain terms.

1. Highlights from the Basic Analysis -

This dataset is an extract of two surveys conducted by the National Science Foundation (NSF) in 2013: the National Survey of College Graduates (SESTAT 2013) and Survey of Doctorate Recipients (SESTAT 2013).

2. **Brief description of the population, the sample, and the sampling method for each of the surveys. What biases were discovered?**

Survey 1 - The National Survey of College Graduates (SESTAT 2013)

Population - non-institutionalized individuals living in the US under the age of 76 who had completed at least a bachelor's degree by the date of the Census.

Sample - Stratified Random Selection of Census long-form households in 1990 and 2000, and ACS households in 2010, then taking another stratified into groups of age, race, highest degree type, occupation, and sex.

Sampling Method - A two-stage, stratified systematic sampling method chooses between Census long-form households in 1990 and 2000, and ACS households in 2010, then, from long-form or ACS households, uses a stratification sampling scheme using age, race, highest degree type, occupation, and sex.

Biases - The two biases found were Non response bias, meaning that not all households in this sample were responsive to this survey / and or completed a response, as well as selective bias. For selective bias, In this situation the survey administrators chose who and what kind of people to survey, and that initial screening of who to choose has the possibility to skew the answers.

Survey 2 - The Survey of Doctorate Recipients (SESTAT 2013)

Population - Individuals who have earned a doctorate degree in science, engineering, or health in the United States

Sample - The pool of possible respondents comes from a stratified sample from the Survey of Earned Doctorates

Sampling Method - The sampling method used was stratified sampling

Biases - The two biases found were Non response bias, meaning that not all households in this sample were responsive to this survey / and or completed a response, as well as selective bias. For selective bias, In this situation the survey administrators chose who and what kind of people to survey, and that initial screening of who to choose has the possibility to skew the answers.

3. **Is there an introduction of further biases when analyzing the results of the surveys together**

- We do introduce further biases when analyzing the two surveys together as they will provide results from separate subsets of individuals with different types of academic credentials. National Survey of College Graduates simply surveys individuals in the US under the age of 76 who had completed at least a bachelor's degree, while The Survey of Doctorate Recipients surveyed those who have earned a doctorate degree in science, engineering, or health in the United States. Also, the sample in the The Survey of Doctorate Recipients will potentially overlap in the National Survey of college Graduate. Due to the differences in academic credentials and the overlapping of data, there will be further biases when analyzing the two surveys together, especially when it comes to looking at the results of how college affects the characteristics of one's life.

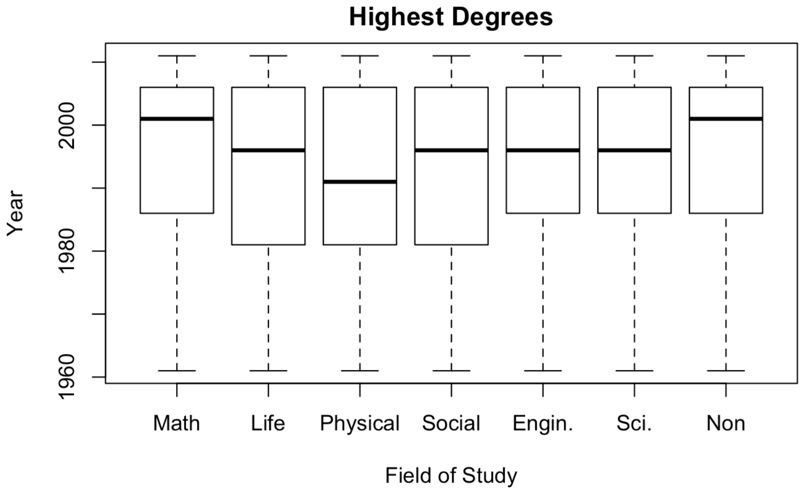
4. **Summary of the demographics of the survey.**

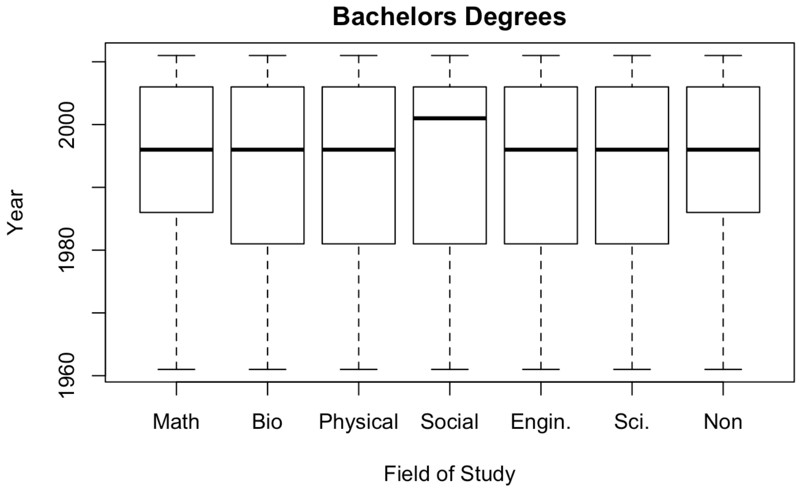
- Proportion of Females - .43350, Proportion of Males - .56649

- Proportion of not Minority - .796112, Proportion of Minority - .203887

- Proportion of "Asian" - .17090454, Proportion of "White" - .62520842, Proportion of "Under Represented Minorities" - .20388704

5. Summary of the distribution of highest degrees and bachelor degrees by field and year obtained obtain.





- In terms of Highest Degrees - Therefore this can show that majors 1,5,6, and 7 have shown more popularity in recent years. Category 1,5,6 and 7 represent computer and mathematic sciences, Engineerings, Science and Engineering Related Fields, and non science fields. On the other hand, 50 percent of the data that is recorded between 1961 and 2011, for majors 2,3 and 4 have a wider quartile range. These majors of category 2,3 and 4 are Biological, agricultural and environmental life sciences, physical and related sciences, and Social and related sciences. Therefore, this shows that these majors have had significant popularity in earlier years (before 1990). Physical and Related Science had a median close to 1990, while computer and mathematic science had a median closer to 2000. This shows that 2,3, and 4 were more populated earlier than 1,5,6,7.

- In terms of all Bachelors Degrees - For the categories 1 and 7, 50 percent of the data that is recorded of highest degree paths in-between 1961 and 2011, are concentrated at later years (1985 Onwards), in terms of quartile range. Therefore this can show that majors 1 and 7 have shown more of their respective popularity in recent years. Category 1 and 7 represent computer and mathematic sciences and non science and engineering fields. On the other hand, 50 percent of the data that is recorded between 1961 and 2011, for majors 2,3 and 4, 5 and 6 have a wider quartile range. Therefore, this shows that these majors have had significant and independent popularity in earlier years in terms of bachelors degrees. These majors of category 2,3 and 4, 5 and 6 are Biological, agricultural and environmental life sciences, physical and related sciences, Social and related sciences, Engineering, and Science and engineering-related fields. 4 has a higher median, therefore that shows that it's popularity has grown over time. Category 4 is Social and related Sciences.

6. **The retention rate of a field is the rate at which people with a bachelor degree in this field would do a higher degree in the same field. Is there a significant difference in retention rates among different field of majors?**

- To compute this, we used a permutation test to compare retention rates of each field. The retention rate was calculated by dividing the number of times an individual had the same bachelors degree as their highest degree, all divided by the total amount in the major. We used this data to find our initial deviation, or D.

- H null : There is no significant difference in retention rates among different field of majors.
- H Alt : There is a significant difference in retention rates among different field of majors.

Upon computation, we shuffled the column of “Highest Degree”, and recalculated the deviation. Upon running this simulation 10**4 times, we see that our final p value is calculated as 0.

This means we were able to reject the null, and accept that There is a significant difference in retention rates among different field of majors.

7. **What does the labor force look like?**

- The percentage of people working is roughly 85.15%

- The percentage of people working part time, where part time is defined as 35 hours or less, is approximately 15.77%

- Approximately 37.75% of the sample works 36-40 hours per week, and another 46.47% work overtime.

- .648% work 1-10 weeks a year, .768% work 11-20 weeks a year, 6.31% work 21-39 weeks a year, and 92.267% work 40-52 weeks a year.

8. Degree relevance

- We are able to see the proportion of job relevance that is "closely related to their principle job" has a percentage of 62.8%. Therefore there is enough data to assume that there is a significant relevance to people's degree with their principle job.

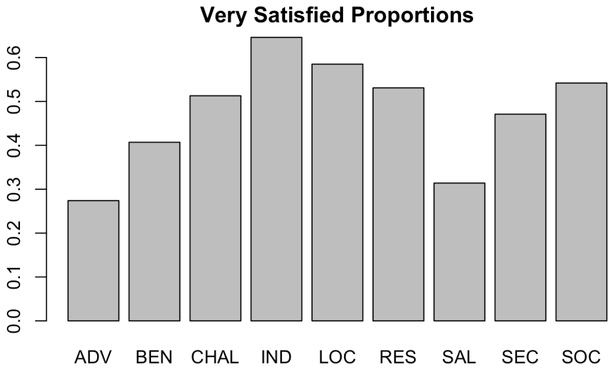
- To test degree relevance, we ran 3 chi-squared tests comparing “Principal job related to highest degree” with Job type, field of major for highest degree, and primary work activity. For each chi-squared test we had a p value of less than 0.05 significance. Therefore,

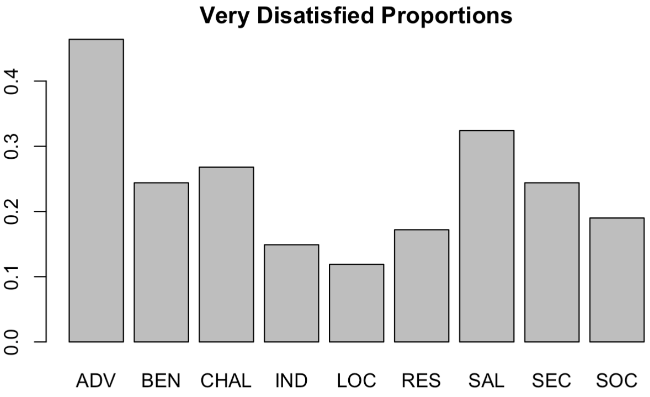
There is a statistical difference in relevance of degree vs job type
There is a statistical difference in relevance of degree vs the degree that they are trained for
There is a statistical difference in relevance of degree vs the type of job that people do

9. Job satisfaction key elements

- The data shows that there is an 89.43% proportion of those who are "Very Satisfied / Somewhat Satisfied" while there is a 10.56% proportion of those who are "Somewhat Dissatisfied / very dissatisfied". Therefore this can be summarized to say that a much larger proportion of respondents believe that they are satisfied with their job.

10. Summary of overall job satisfaction





Those who reported "Somewhat / Very Dissatisfied", they are most dissatisfied with their career advancement, salary, and intellectual challenge.

Those who reported "Somewhat / Very Satisfied", they are most satisfied with Independence, Location, Social Impact.

Highlights from the Salary Model -

1. **Detail how variable selection was done:**

   - To begin, I started by qualitatively removing variables that did not necessarily affect the model. These were variables such as PERSONID, YEAR, CHTOT, and SALARY. Next, I plotted each variable against Salary, to see if there were any obvious non-correlations that could be removed. Then I created a subset where LFSTAT == 1 to get rid of NA values. After this initial cleaning, I started by creating the linear model to predict salary using as many of the variables from the original data left. I used this large model as a control to test the available R^2 as well as the P-values for each coefficient, then started adjusting. I switched all variables to the format of "as.factor()".

   From this point, is when I used backward elimination to slowly move away the coefficients that had a p value higher than 0.05. This type of p value less than the significance level accepts the Null, stating that this coefficient will become 0 during analysis, showing that it will have little effect on the model. To arrive to final_model, all of the coefficients in analysis have a p-value less than 0.05, showing that they will be significant for analysis. At this point, we then use boxplots to plot the remaining values against Salary.

2. **Our final regression model model.lm for salary.**

   ```
   final_model <- lm(SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) +
   as.factor(RACETH) + as.factor(NOCPRMG) + as.factor(NBAMEMG) +
   as.factor(WKSWKGR) + as.factor(HRSWKGR) + as.factor(JOBINS) +
   as.factor(JOBPENS) + as.factor(JOBPROFT) + as.character(JOBVAC) +
   as.factor(FTPRET) + as.factor(OCEDRLP) + as.factor(EMSEC) + as.factor(JOBSATIS) +
   as.factor(SATSAL) + as.factor(SATADV) + as.factor(MGRNAT) + as.factor(MGROTH), data
   = dataset)

   model.lm <- stepAIC(final_model)
   ```

3. **Report the R^2 and adjusted R^2 of your model.**

   Multiple R-squared: 0.5651
   Adjusted R-squared: 0.5648

   Meaning of R^2: Means this model can explain 56.51% of the variance in this data

4. **How do we maximize our SALARY, based on our model?**

- To maximize salary, one should get a professional degree as their highest degree, work in the field of Science and Engineering related occupation for their principle job, obtain a bachelors in engineering, work for 40-52 weeks, work greater than 40 hours per week, have their principal job "somewhat related" to highest degree, work in the employer sector of Business or Industry, and have technical expertise in natural sciences.

Highlights from the Job Satisfaction Analysis -

1. **Detail Variable Selection**

   - The way I conducted variable selection was as such - To begin, I started by qualitatively removing variables that did not necessarily affect the model. These were variables such as PERSONID, YEAR, CHTOT, and SALARY. Next, I plotted each variable against is.satis, which is a simple recode of job satsifaction, to see if there were any obvious non-correlations that could be removed. Then I used the subset where LFSTAT == 1 to not use NA values. After this initial cleaning, I started by creating the logistic model to predict Job Satisfaction using as many of the variables from the data left after cleaning. I used this large model as a control to test the available AUROC as well as the P-values for each coefficient, then started adjusting further from there. I switched all variables to the format of "as.factor()".

   From this point, is when I used backward elimination to slowly move away the coefficients that had a p value higher than 0.05. This type of p value less than the significance level accepts the Null, stating that this coefficient will become 0 during analysis, showing that it will have little effect on the model. To arrive to updated_model, I cleared unnecessary variables to get the highest AUROC value. Once this is done, the logistic model is complete.

2. **Our final regression model model.lm for job-satisfaction**

   - updated_model <- glm(is.satis ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(NDGMEMG) + as.factor(NOCPRMG) + as.factor(RACETH) + as.factor(NBAMEMG) + + as.factor(WKSWKGR) + as.factor(HRSWKGR) + as.factor(JOBINS) + as.factor(JOBPENS) + as.factor(JOBPROFT) + as.factor(JOBVAC) + as.factor(FTPRET) + as.factor(OCEDRLP) + as.factor(EMSEC)+ as.factor(SATADV) + as.factor(SATIND) + as.factor(SATLOC) + as.factor(SATRESP) + as.factor(SATSAL) + as.factor(SATSEC) + as.factor(SATSOC), family = "binomial", data = dataset)

3. **Our model's AUROC curve. Is this model a good fit? Is it easy to interpret?**

   - AUROC - .9268

   - To make this conclusion we use a confusion matrix. After calculation, the optimal accuracy is .926, which is obtained with a .457 cutoff value. Because this value is far above the cutoff, and is in the ~.9 range, therefore because the diagnostic plots pass the conditions, and the accuracy is high, this means that this model is a good fit, as well as clear to interpret.

4. **How do we maximize job satisfaction, based on our model?**

- To maximize job satisfaction, get a professional degree as a highest degree, the field of major for highest degree should be in Science and engineering related fields, have a life and related science bachelors degree, work greater than 40 hours per week, work 40-52 weeks out of the year, be in a science and engineering related occupation, and be in the business or industry employer sector.

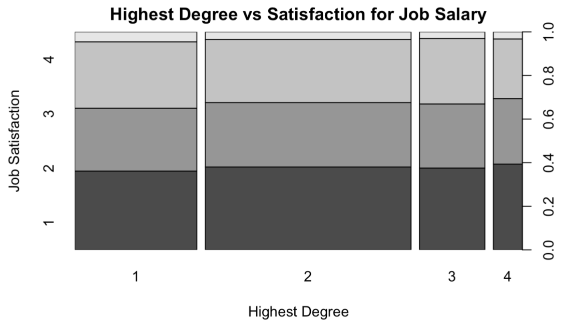
Highlights from the Fact Check Section -

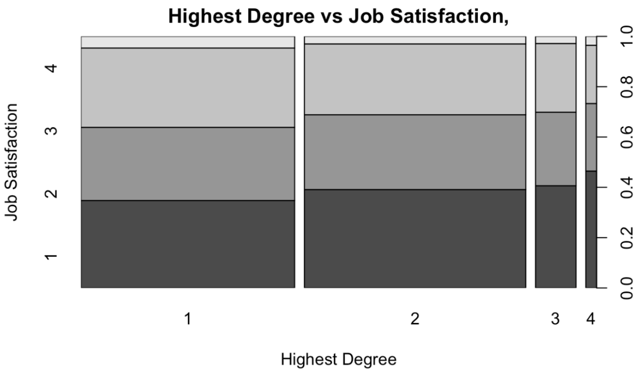
News outlets regularly examine relationships between degrees, job satisfaction and income. Here are various claims from three different outlets.

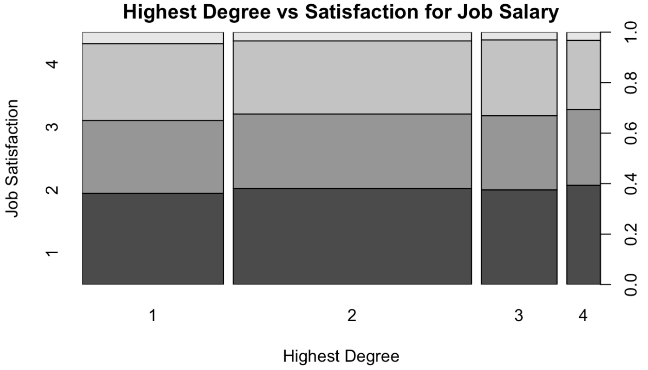
Gallup: Does Higher Learning = Higher Job Satisfaction? https://news.gallup.com/poll/ 6871/ does-higher-learning-higher-job-satisfaction.aspx

This article claims that:

a. **Education level has very little to do with job satisfaction, or satisfaction with income and time flexibility.**







Looking at the plots above, we are able to see the relationship between "Highest Degree" or Highest Education Level, versus job satisfaction, or satisfaction with income and time flexibility.

Each plot maintains the same scale, therefore we can look at each x value, which represents each of the degree types. Within each degree level, we see the different proportions of job satisfaction. We can clearly see that the proportions of job satisfaction, for each type of satisfaction variable, does not change significantly based on the degree type. Therefore this shows that education level has a very low correlation to job satisfaction, or satisfaction with income and time flexibility,

Therefore from this data we can see that education level has very little to do with job satisfaction, or satisfaction with income and time flexibility

b. Having the opportunity to do what you do best is the one factor that correlates most highly with overall job satisfaction is.

Based on the data analysis from question 7, we are able to see what factors correlate to what degree with overall job satisfaction. Looking at that data, we are able to see that "Degree of Independence" has the highest correlation to overall job satisfaction. Therefore it is acceptable to say that having the opportunity to "do what you do best", is indeed, one of the key factors that correlates most highly with overall job satisfaction.

2. Diverse Education: College-educated Americans More Likely Experience Job Satisfaction, Lead Healthier Lives, Study Says https://diverseeducation.com/article/14156/

This article claims that:

a. **Certain race groups earn less than others when they have the same education level.**

- Based on the model created for salary above, each category for RACETH has different coefficients in relation to Salary. Therefore, when you hold education level constant, and solely adjust by race, you will see that certain groups will have less salary than others. For example, based on Linear Regression Model 1, people who classified as white will earn more than those who classified as an Under-represented minority. Hence it is true that certain race groups earn less than others when they have the same education level, looking at the coefficients from linear model 1.

b. **b. STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented, tend to pay more than careers in social sciences.**

- To address that minorities that are underrepresented in STEM, we have created bar plots to show the RACETH distribution for every type of stem related field. In this analysis, we want to look at choice "3", or "Underrepresented Minorities". In each field, choice 3 was the smallest proportion, or close to the smallest proportion, of race types in each of the STEM fields. Therefore it can be fairly concluded that minorities are underrepresented in STEM related fields.

To address the idea that STEM careers pay better than social careers, we take a subset of stem related fields, then a subset of social related fields, then get average salaries for both subsets. Then in comparison of the averages, the Stem Salary Average is 81330.6 while the Social Salary Average is 69880.23. Therefore it can be concluded that STEM careers pay better than social science.

3. PEW: the rising cost of not going to college https://www.pewsocialtrends.org/2014/02/11/ the-rising-cost-of-not-going-to-college/

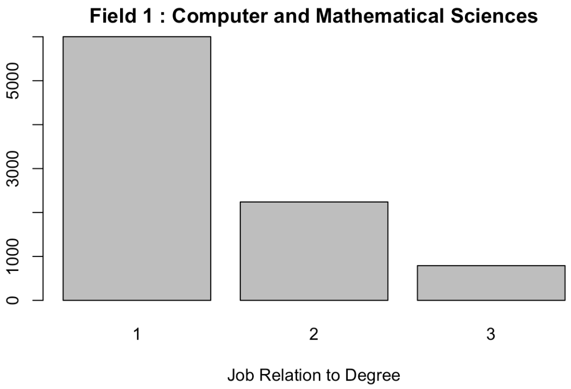
This article claims that:

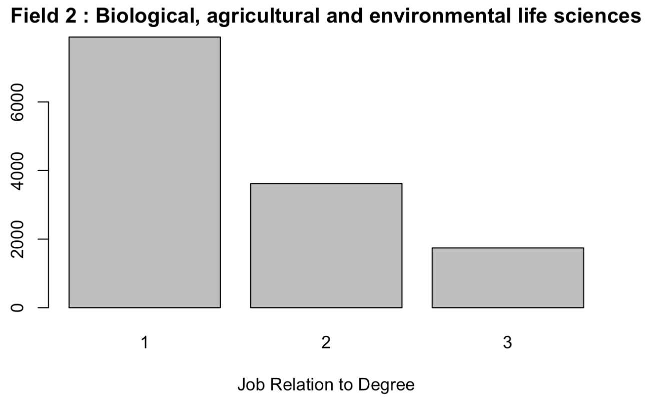
a. **Those who studied science or engineering are the most likely to say that their current job is “very closely” related to their college or graduate field of study.**

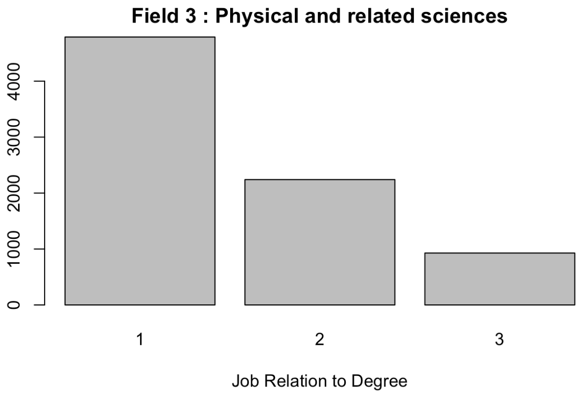
Upon plotting the relationship between each highest field and job relation to highest degree, we are able to see the distribution for relationship to each degree field.

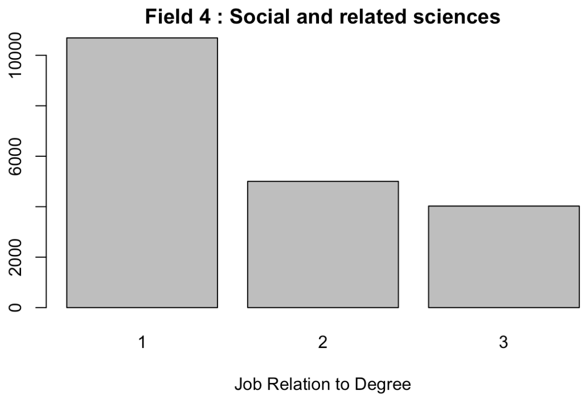
Based on the data from the plots, those who studied scientific majors have the highest amount of individuals choosing choice 1, which means that is the most "closely related" field of study to current job.

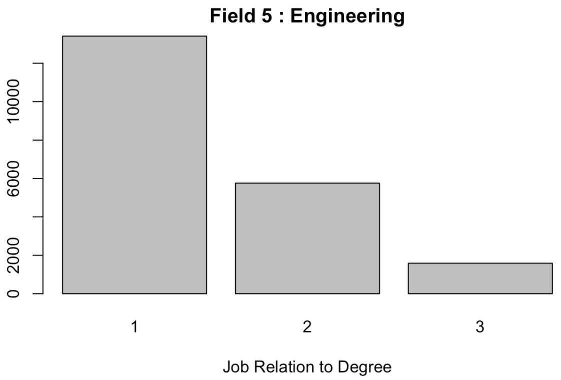
Hence, it is true that those who studied science or engineering are the most likely to say that their current job is “very closely” related to their college or graduate field of study.

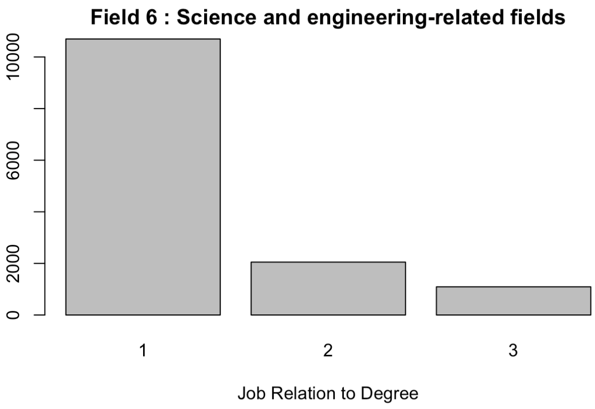

Field 1 : Computer and Mathematical Sciences

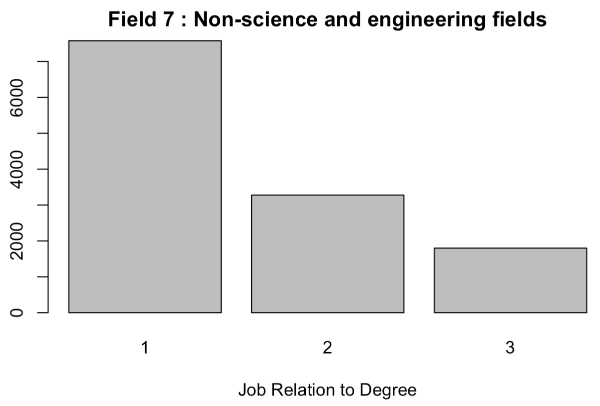

Field 2 : Biological, agricultural and environmental life sciences


Field 3 : Physical and related sciences


Field 4 : Social and related sciences


Field 5 : Engineering


Field 6 : Science and engineering-related fields



Based on this data, there are certain variables that we have found that can greatly alter the value of college. Therefore, based on this data, it is possible to conclude that college is indeed worth it, to not only salary, but job satisfaction.