R PLOT DRAFT

1. Factors that impact the investments for learning at institutions/organizations:

1.1

a) Does the amount of money spent on learning how to code vary with age?

b) Is there a relationship between programming experience and money spent on learning how to code?

The data provided can be used to analyze relationships between various variables that can possibly impact the investments made on learning how to code from various institutions/organizations. A good measure of impact would be the amount of money spent in USD by people who are learning how to code, best described by "*MoneyForLearning*" variable. This variable becomes the dependent variable. Pearson’s correlation test would reveal correlation between money spent for learning how to code and multiple independent variables. Pearson’s correlation function provides a p-value. The existence of correlation between two variables is proved if the p-value is less than the standard significance level (0.05).

|  |  |  |
| --- | --- | --- |
| Predictor Variables | Outcome Variables | Result: r , p-value |
| Age | MoneyForLearning | r = 0.1013 | p = 0.00017 |
| MonthsProgramming | MoneyForLearning | r = 0.0543 | p = 0.04427 |

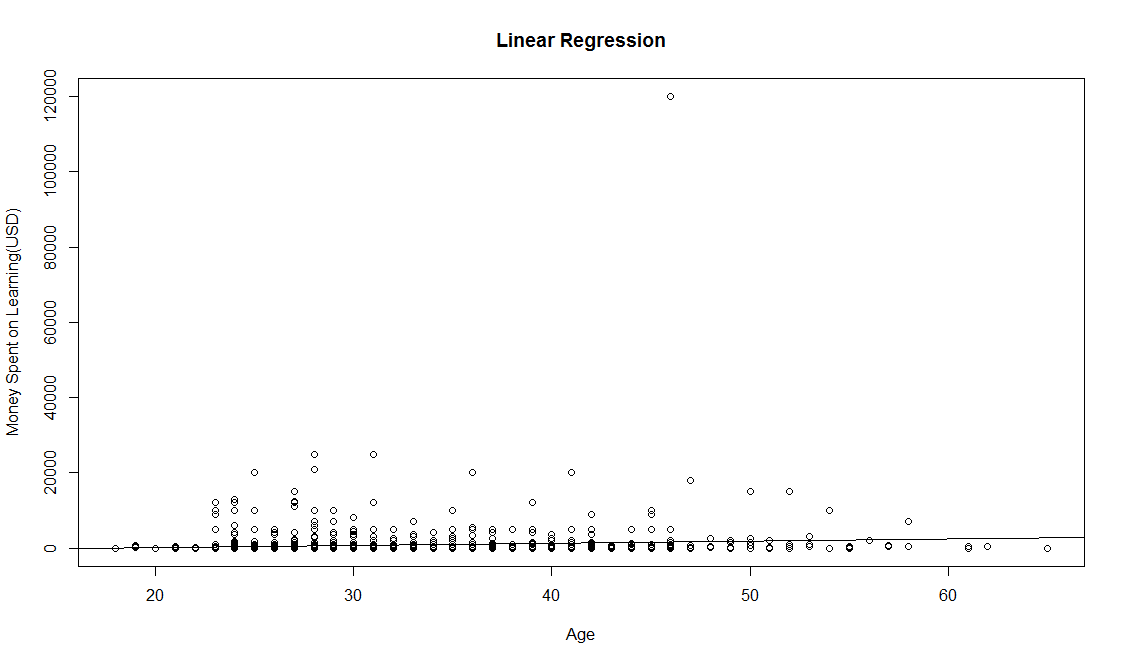
To determine the strongest relationship between Age and Months Programming, we consider the values for Pearson’s correlation coefficient since correlation best describes how well one variable explains the other and vice versa. On comparing values from above table between all predictor variables, the correlation Coefficient for "Age" (0.1013) is found to be higher, positive and closer to 1(Perfect positive relationship). This testifies that the relationship between "Age" and "MoneyForLearning" is stronger than that of “MonthsProgramming" and "MoneyForLearning". The relationship explained demographically with linear regression displays a high correlation between the Age and Money Spent on Learning. 

Figure 1 Plot between Age and the Money spent in Learning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Standard Error | T-value | P-value |
| Intercept | -921.69 | 483.51 | -1.906 | 0.056828 |
| Age | 57.32 | 15.21 | 3.769 | 0.000171 |

* 1. Is there a relationship between employment status and investment?

We intend to find if there exists a relationship between the employment status of people who are learning how to code and the amount of money they spend on learning. Employment status of people is determined by variable called "Employment Field" that comprises of nominal values. The amount of money spent available in ratio scale of measurement is the dependent variable under consideration. Since we need to test relation between one nominal variable with more than two levels, a one-way ANOVA (Analysis of Variance) would be the most appropriate statistical test. Every level was tested for Normality and the ones that highly violated the assumption were eliminated from the final analysis. The filtered data-frame is used to perform Analysis of Variance test between the two variables. Fig. 2 and Fig. 3 show the QQ plot and the histogram for a variable (transportation) that violates normality.

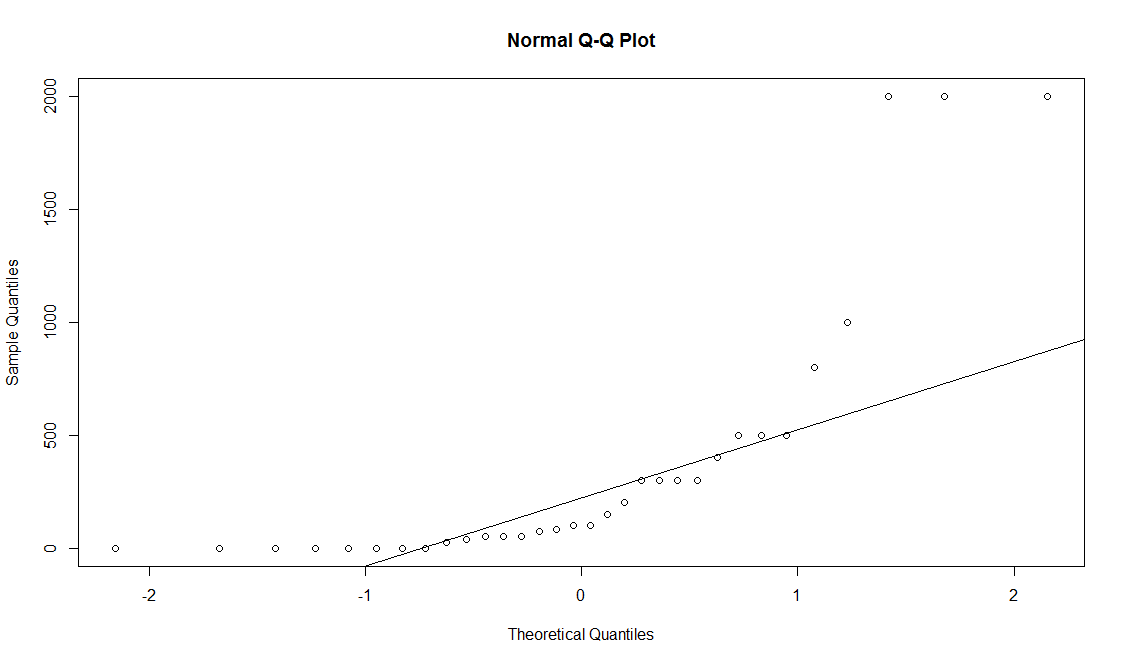


Figure 2 Violation of Normality Test

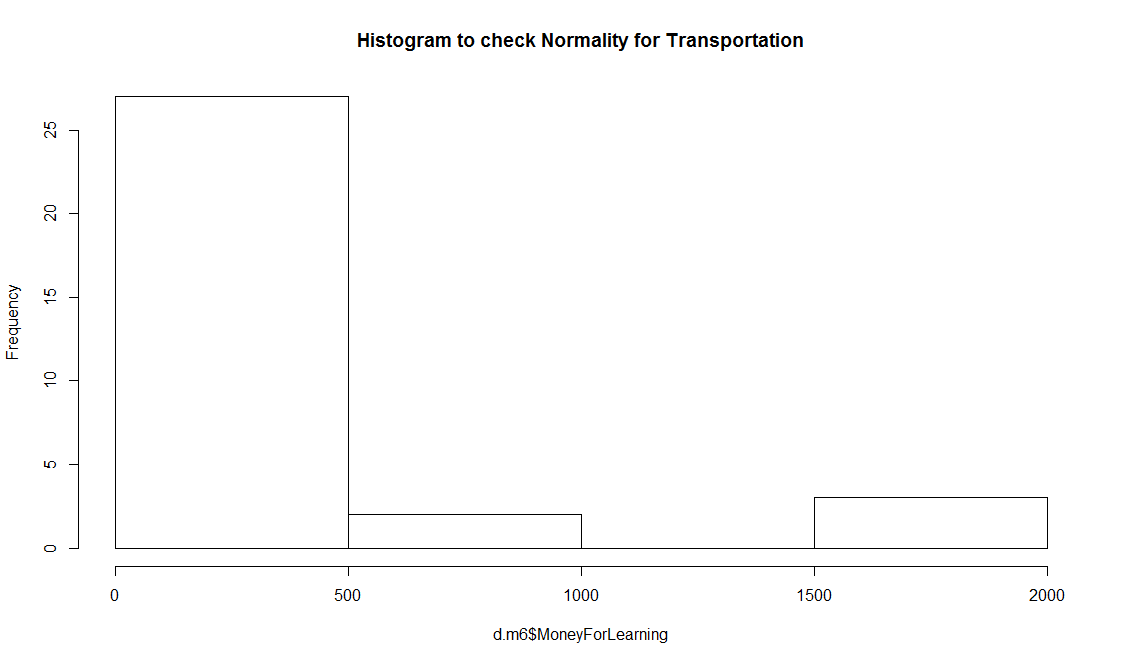


Figure 3 Histogram to check for normality

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Degrees of Freedom | Sum Squares | Mean Squares | F value | Probability(>F) |
| Employment Field | 14 | 2.006e+09 | 143315248 | 9.659 | <2e-16 \*\*\* |
| Residuals | 1356 | 2.012e+10 | 14837878 |  |  |

For this question, the null hypothesis states that the means of all the employments fields are the same and the alternate hypothesis states that at least one of the means of different employment fields is different as compared to the means of other employment fields. Results from the ANOVA test provide a p-value(<2e-16) less than standard significance level (0.05). This enables us to reject the null hypothesis and state that at least one of the means of different employment status is different from the others. A measure of the effect size is provided by R-squared value that is calculated as mentioned below:

|  |
| --- |
| R2 = Sum of squares(between) / Sum of Squares(total) = 143315248 / 153156126 = 0.9061 |

This value for effect size indicates that about 90.61% of the variance is explained by the different employment fields.

Plotted below in Fig. 4 is a bar graph that displays various employment Fields and their frequency. The height of the bars indicate that maximum people are from ‘Software development and IT’ which is intuitive. The employment field ‘Education’ is second highest which was a bit surprising.

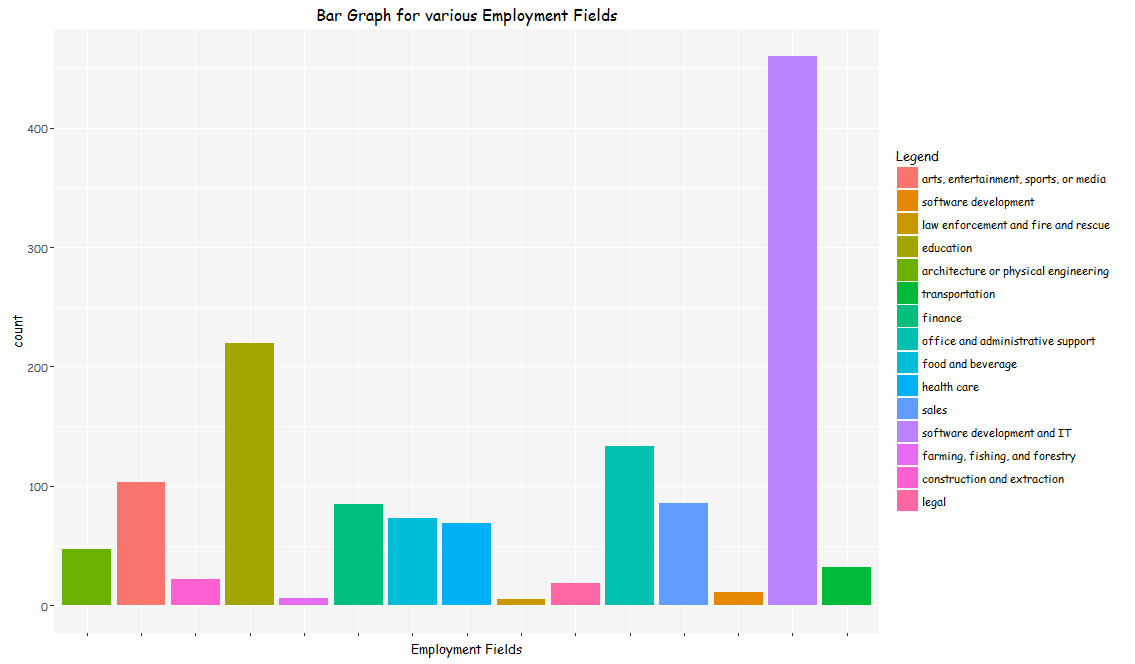


Figure 4 Bar graph of various employment fields

1. Factors that affect the amount of time spent in learning how to code:
   1. Do people spend more time in commuting spend less time learning?

To find the relationship between the amount of time an individual spends learning and the amount of time spent travelling, we carried out a regression test where the dependent variable is time spent learning, given by the variable ‘HoursLearning’ and the independent variable ‘CommuteTime’.

|  |  |  |
| --- | --- | --- |
| Predictor Variables | Outcome Variables | Result: r , p-value |
| Commute Time | HoursLearning | r = 0.02607 | p = 0.3345 |

From the above Pearson’s co-relation coefficient value of r = 0.02607, we can say that there is a weak but positive relationship between the time one takes to commute and the time one spends in learning how to code. The scatter plot in Fig.5 shows that there is a high concentration of observations in the lower range and it keeps on decreasing as we go higher.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Standard Error | T-value | P-value |
| Intercept | 13.47 | 0.407054 | 33.094 | <2e-16 |
| Commute Time | 0.0048 | 0.005070 | 0.966 | 0.334 |

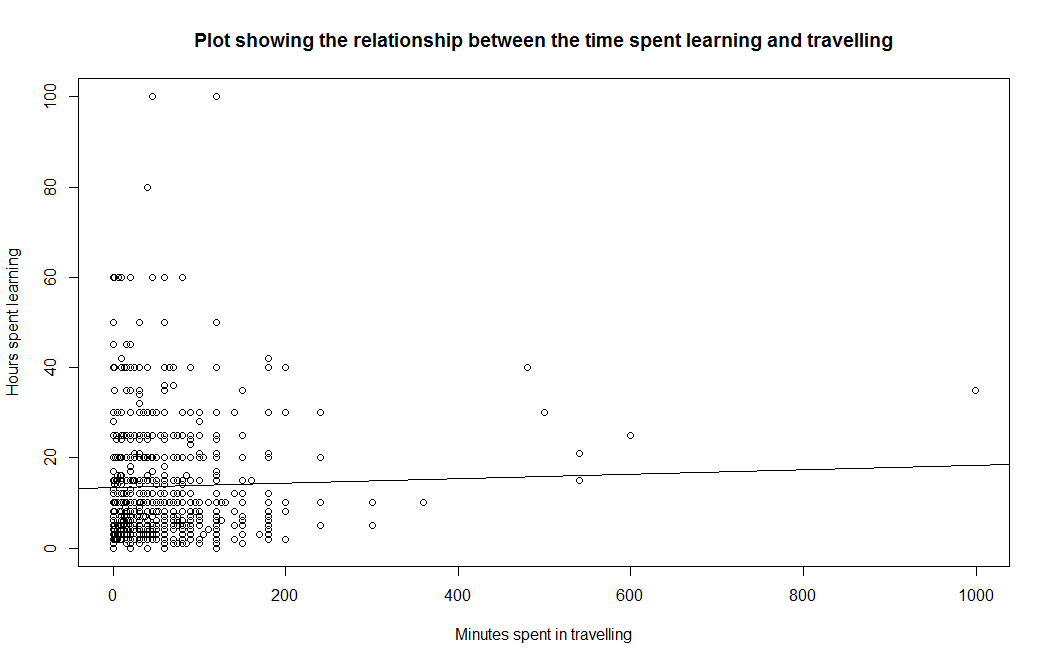


Figure 5 Plot showing the relationship between time spent learning and time spent commuting

* 1. Does a specific major in computer science or information technology or experience in related fields reduce the amount of time spent in learning and practicing?

In our dataset, we have 8 different majors that are related to computer science or information technology. To understand whether individuals with a background in any one of these majors spent less time in learning how to code, we conducted a one-way ANOVA test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Degrees of Freedom | Sum Squares | Mean Squares | F value | Probability(>F) |
| School Major | 7 | 386 | 55.17 | 0.44 | 0.875 |
| Residuals | 118 | 14806 | 125.47 |  |  |

With a p-value of 0.875, we fail to reject the null which states that there is no difference in the mean time spent by individuals from any of these backgrounds. People with backgrounds in computer science or information technology spend almost equal time in learning how to code.

R2 = Sum of squares (between) / Sum of Squares (total) = 55.17/ 180.64 = 0.305

Moreover, R2 value of 0.305 explains that 30% of the variance in the time spent learning to code is explained by the different school majors.

The bar graph is Fig. 6 shows the hours spent by people belonging to different majors in learning how to code. The height of the bar shows that people with ‘Computer Hardware Engineering’ spend the maximum time while people from ‘Computer engineering Technician’ spend the least.

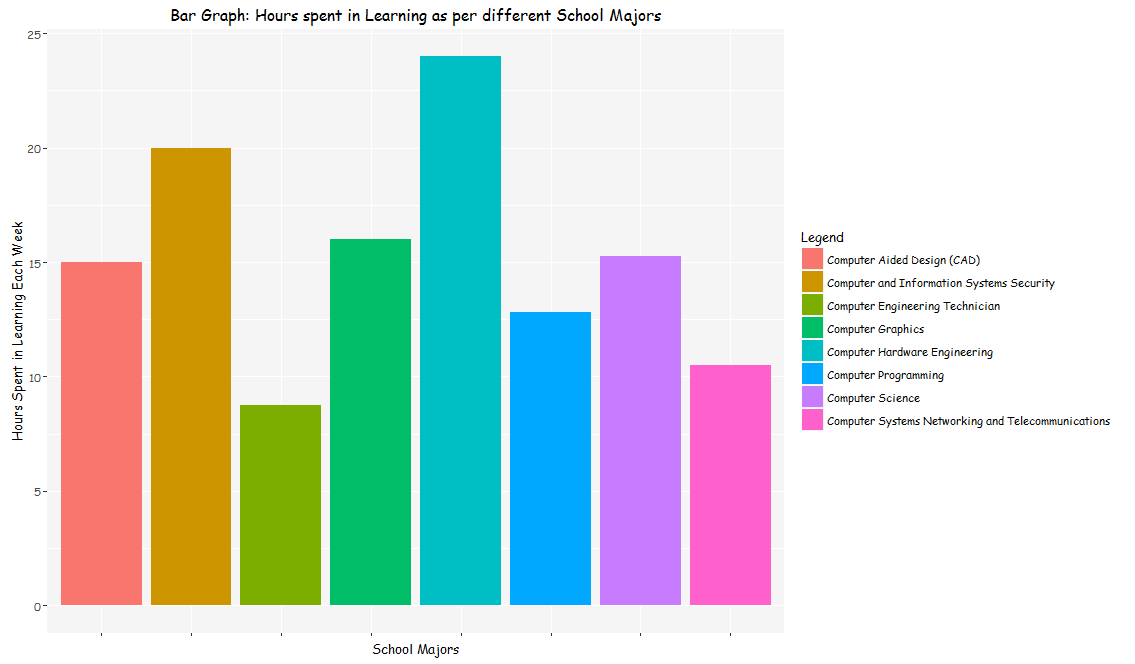


Figure 6 Bar graph of Number of hours spent by people belonging to various school majors

Word Count: 996