**Milestone 4**

**SOEN 6611**

**Dr. Rodrigo Morales Alvarado**

**Group Number:**

**1**

**Empirical Study Type-1**

**1. List the statistical tests that you will perform and explain why those tests are suitable for your data. E.g., “We will perform logistic regression because the dependent variable is categorical.”**

Variable is any characteristics or number that can be measured and, a major division about type of variables are dependent and independent.

The dependent metrics are the ones which we can evaluate them based on their relationship with other metrics like fault-proneness or change-proneness of a program. On the other hand, the independent metrics are the ones which we can evaluate freely and with no limitation like size, complexity of program, coupling, etc.

In this project, our collections are based on Tools Computations. As we can describe our variables to numeric (including continuous and discrete) and categorical (ordinal and nominal), we can divide our own metrics as follow:

The number of bugs, Weighted Method Per Class (WMC), Depth Inheritance Tree (DIT), Number of Children (NOC), Lack of Cohesion in Metric(LCOM), Line of Comments(LCM), Line of Code(LOC), Number of physical line(NL) are in the discrete numeric division.

Comment percentage (ratio comment to code) and quality of comments are in the continuous division.

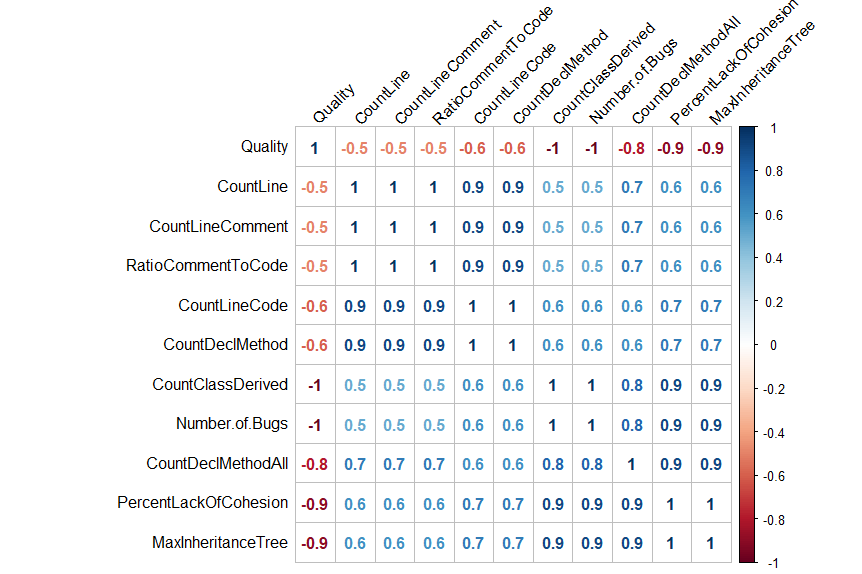
To test our data, regression analysis is used as a method of figuring out which of the independent variables are related to the dependent ones, and finding the form of the relationship between them. As we have discrete numeric variables and continuous variables in our data, we follow the linear regression analysis to predict the relationship between two variables by fitting a linear equation of the data range we have. Also the main coefficient function that we used was Spearman function.

**2. For every pair of independent variables with the dependent variable, compute the Pearson/Spearman coefficient, and discuss the strength of the correlation. If the correlation is strong perform linear regression and include the resulting plot.**

Our dependent variables are number of bugs and quality of comments, hence we performed Spearman function coefficient in this process, we checked the correlation coefficient between our independent variables and these two dependent variables separately by function “cor”. The coefficient matrix for our hypothesis is based on quality, quantity of comments and number of bugs prove the following results:

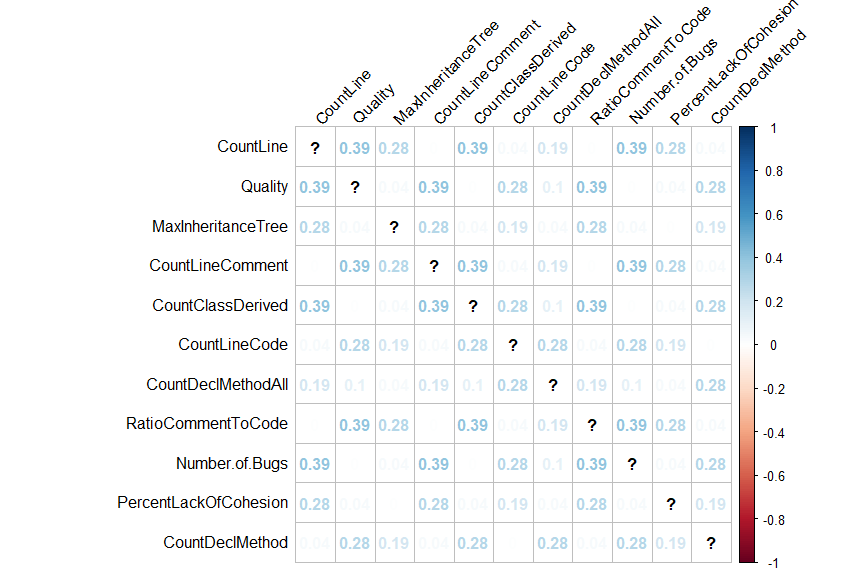
* Quality has a highly negative correlation with “number of bugs” and “count class derived”. It also has a weak negative correlation (-0.5) with “ratio comment to code”.
* Ratio comment to code has a highly positive correlation with “count line” and “count line code”. It also has a weak negative correlation (-0.5) with “quality” and a weak positive correlation (0.5) with “number of bugs”.
* Number of bugs has a highly negative correlation with “quality” and highly positive correlation with “count class derived”. It also has a weak positive correlation (0.5) with “ratio comment to code”.

The correlations between other variables also show interesting information which we skip them here for further discussion due to not having a direct impact to our main goal.



All correlation values with Spearman

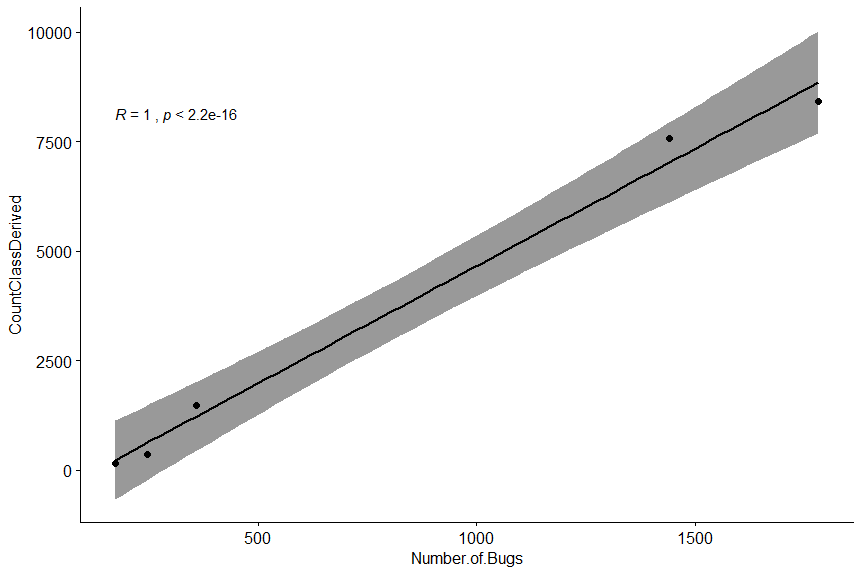
The p-value results are as following:



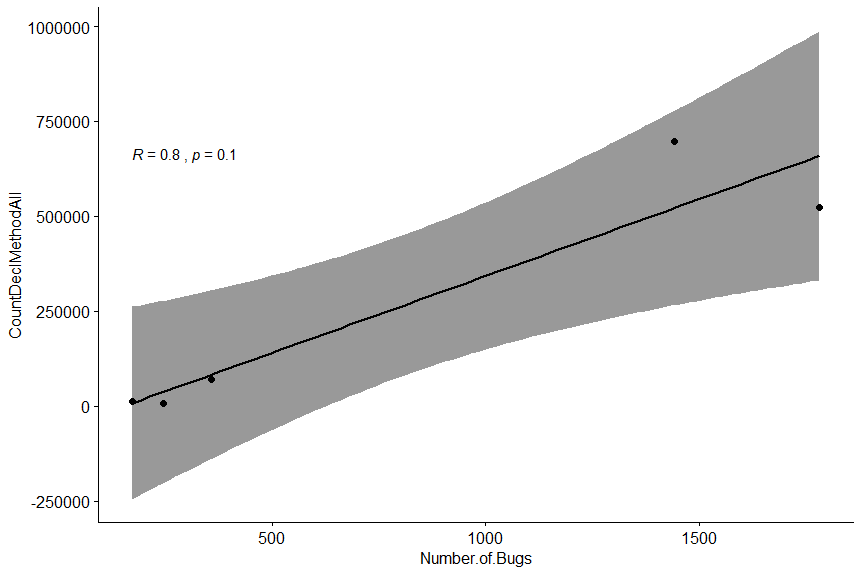
P-value results

We implemented correlation coefficient among all of our independent variables and dependent ones and performed a linear regression between them as well. The following represents the strong relation between independent variables with both comment quality and bugs.

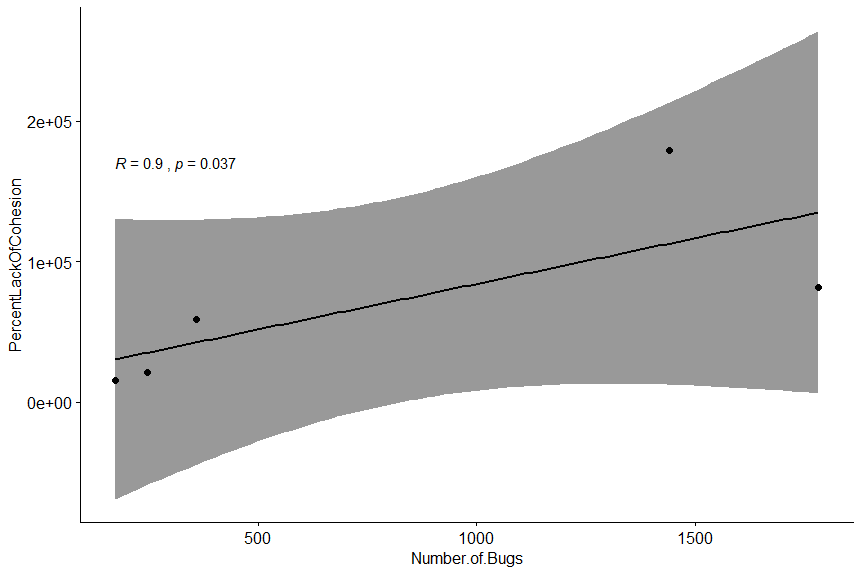
Linear regression of correlation of number of bugs and some independent variable with significant results is provided first including:



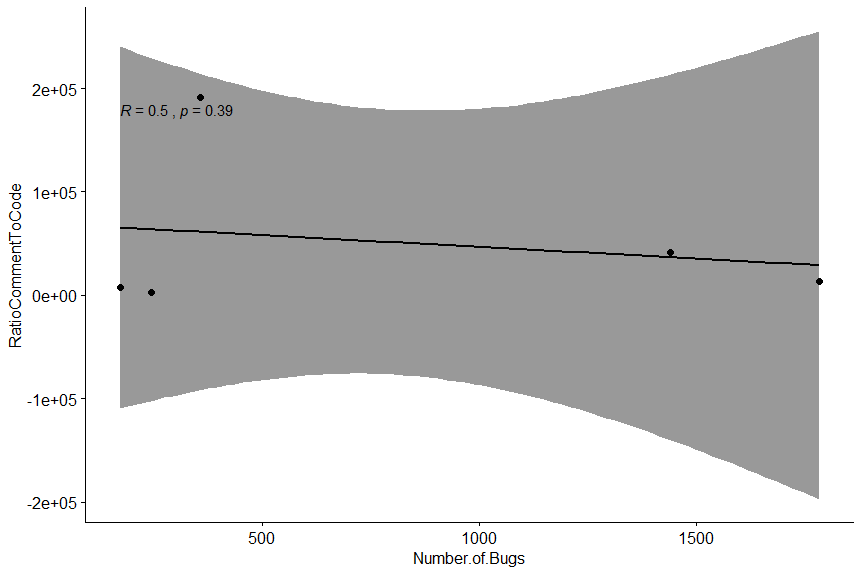
Linear Regression of Number of Bugs and CountClassDrived

****

Linear Regression of Number of Bugs and CountDecMethodAll

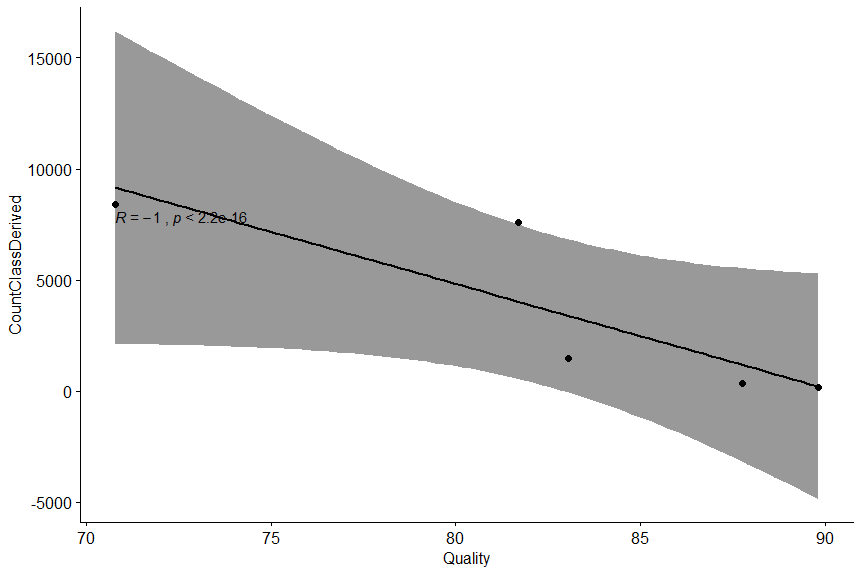
****

Linear Regression of Number of Bugs and PercentLackOfCohesion

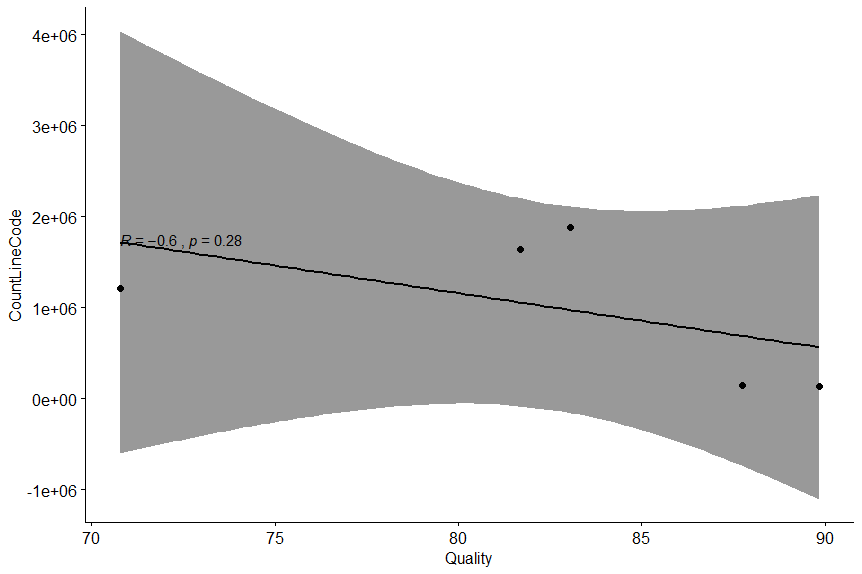


Linear Regression of Number of Bugs and RatioCommentToCode

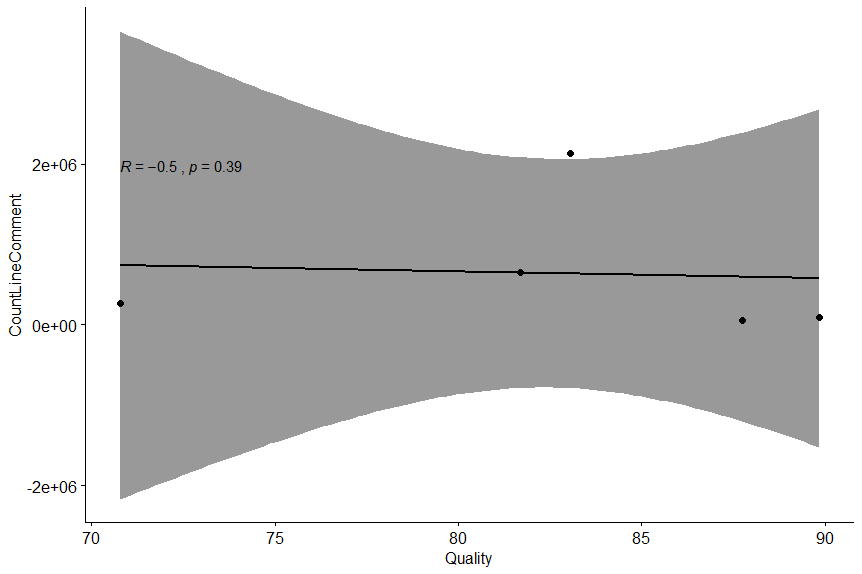
Here is the linear regression plot of some of our independent variables and quality of comments describing noticeable correlation between them:



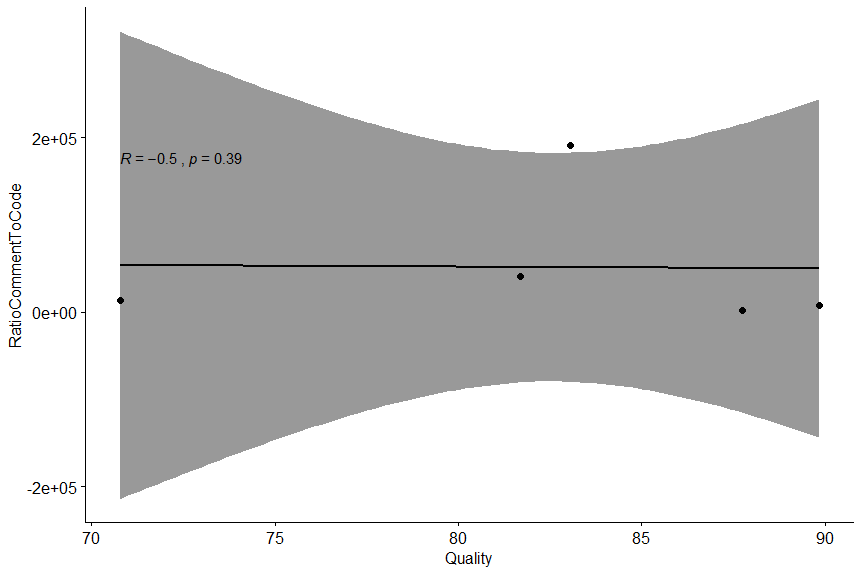
Linear Regression of Comment Quality and CountClassDrived



Linear Regression of Comment Quality and CountLineCode



Linear Regression of Comment Quality and CountLineComment



Linear Regression of Comment Quality and RatioCommentToCode

**3. Regression with all independent variables in the model:**

**Logistic regression is useful when you are predicting a binary outcome from a set of continuous predictor variables. Hence, perform logistic regression with all independent variables in the model, and discuss that statistical significance of each independent variable.**

For this purpose, we split the data into two chunks: training set (including the first four projects) and a testing set (which is Geotools). The training set will be used to fit our model which we will be testing over the testing set. The following tables show the predicted value that we got using our model with all our independent variables one by one, by comparing the actual value of the test set and the predicted value, we can analyze the statistical significance of each independent variable, the closest predicted values are highlighted in green in each table.

Model to predict “**Comment Quality**” with all independent variables (actual value: **83.051**)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **CountLineComment** | **CountLineCode** | **RatioCommentToCode** | **PercentLackOfCohesion** | **CountDeclMethodAll** | **CountDeclMethod** | **MaxInheritanceTree** | **CountLine** | **CountClassDerived** | **Number of Bugs** |
| **predict()** | 62.930 | 74.110 | 61.822 | 83.247 | 86.64 | 68.376 | 85.071 | 68.854 | 86.848 | 87.733 |

Model to predict “**Number of Bugs**” with all independent variables (actual value: **359**)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **CountLineComment** | **CountLineCode** | **RatioCommentToCode** | **PercentLackOfCohesion** | **CountDeclMethodAll** | **CountDeclMethod** | **MaxInheritanceTree** | **CountLine** | **CountClassDerived** | **Quality** |
| **predict()** | 3491 | 2546 | 1110 | 689 | 339 | 710 | 503 | 6501 | 288 | 684 |

**4. Based on the results of the statistical tests, decide on whether you accept or reject your hypotheses.**

**Hypothesis 1:**

Null Hypothesis 1: There is not a specific correlation between number of bugs and quality of comments in the selected projects.

Alternate Hypothesis 1:There is a correlation between number of bugs and quality of comments in the selected projects.

Based on the gathered csv files from our metrics calculated by scitools understand, eclipse/ IntelliJ findbugs plugin and MonkeyLearn, we have interesting results of the relation between number of bugs and quality of comments. The results for these five selected projects show that when the comments have a higher quality, the number of bugs are less. For example RXJava has almost 1782 bugs (the most number of bugs in our selection) while its quality is the lowest amount (70.77 percent). On the other hand we have Paho, which has fewer amount of bugs (it is calculated 157 by findbugs plugin) and it has the highest quality of comments between our projects (89.831).

**Hypothesis 2:**

Null Hypothesis 2: There is not a specific correlation between comment percentage and quality of comments in the selected projects.

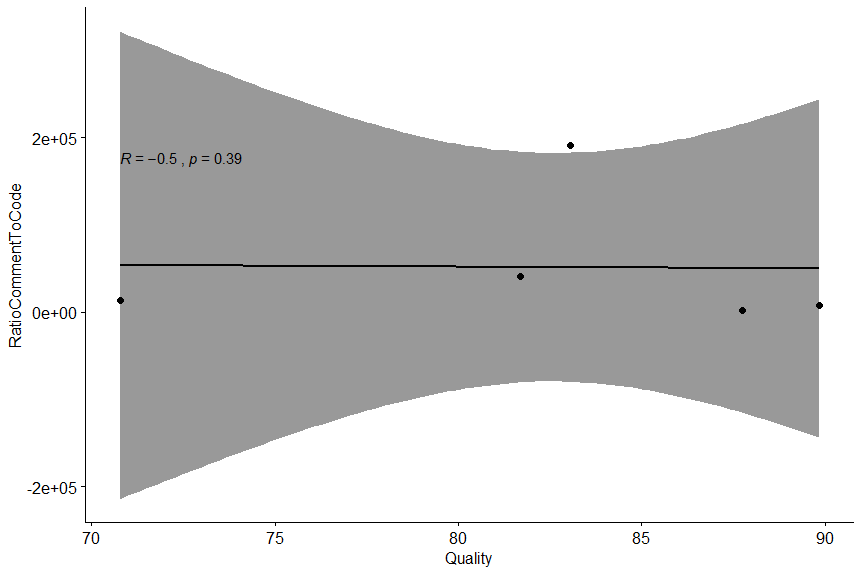
Alternate Hypothesis 2:There is a correlation between comment percentage and quality of comments in the selected projects.

According to the derived results out of our estimation, Null Hypothesis is accepted meaning there is no correlation between percentage of comments and their quality and the Alternative Hypothesis is rejected. As it is shown in the table below comment percentage does not necessarily affects the quality of comments and in various projects its value is varied with a different result. As a result, in the range of selected projects, there has not been an evidence of relationship between quality of comments and the percentage of comments.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **CountLineComment** | **CountLineCode** | **RatioCommentToCode** | **PercentLackOfCohesion** | **CountDeclMethodAll** | **CountDeclMethod** | **MaxInheritanceTree** | **CountLine** | **CountClassDerived** | **Number of Bugs** | **Quality** |
| **RXJava** | 263393 | 1209156 | 13305.92 | 81890 | 522961 | 94212 | 17172 | 1718276 | 8414 | 1782 | 70.77 |
| **Paho** | 94957 | 131542 | 8053.32 | 15632 | 12087 | 9840 | 503 | 251233 | 157 | 173 | 89.831 |
| **Jung** | 49731 | 150895 | 2406 | 20995 | 6109 | 10959 | 1672 | 226574 | 351 | 246 | 87.756 |
| **Vaadin** | 654407 | 1644298 | 41032.15 | 178893 | 697980 | 123381 | 32761 | 2628875 | 7580 | 1441 | 81.691 |
| **Geotools** | 2131286 | 1878530 | 191419.09 | 58900 | 70480 | 202573 | 4722 | 4362907 | 1479 | 359 | 83.051 |

Correlation Values between every individual independent variables and Comment Quality and Number of Bugs

This linear regression plot represents the correlation coefficient values between comment quality and percentage of comments and indicates a 0 correlation which means a weak linear relationship between them and rejecting the idea of the relation between comment quantity and comment quality .



Correlation with Comment Quality and Comment Percentage