**Preliminary Analysis**

**BBangya  
Hansol\_Kim / Youtaek\_Jeon / Youngha\_Lee / Hyoshin\_Soh**

**o How many unique observations to you have?**   
- Math : 395, Language : 649

**o What information/features/characteristics do you have for each observation?**Our data is about student performance in secondary education (high school in Korea). G3 is final score of them and our target variable. The datasets include student demographic, social, school related features) and family environment and it was collected by using school reports and questionnaires. So, all of response are categorical data.

**o What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)? Are there obvious trends in the data (over time, across subgroups, etc.), and are the differences statistically significant?**We organize foundation statistic using Tableau 9.0. We determine eliminate some variables because of logic on analysis. But we numbered variable originally by origin order of datasets. We comment on each variable some of trend.   
  
Origin dataset have 30 explanatory variables. We calculate all of variables (about 30) and 2 subjects. So number of charts is 60. It is so many. So we show some chosen variables to explain our preliminary analysis. Below examples are “Math score of students”.

02. sex - student's sex (binary: 'F' - female or 'M' - male)   
 > We can find difference between students’ sex.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex | Count | Mean | Median | Sd |
| Female | 208 | 9.966 | 10 | 4.622 |
| Male | 187 | 10.914 | 11 | 4.495 |

03. age - student's age (numeric: from 15 to 22)   
> It looks like decreasing with age.

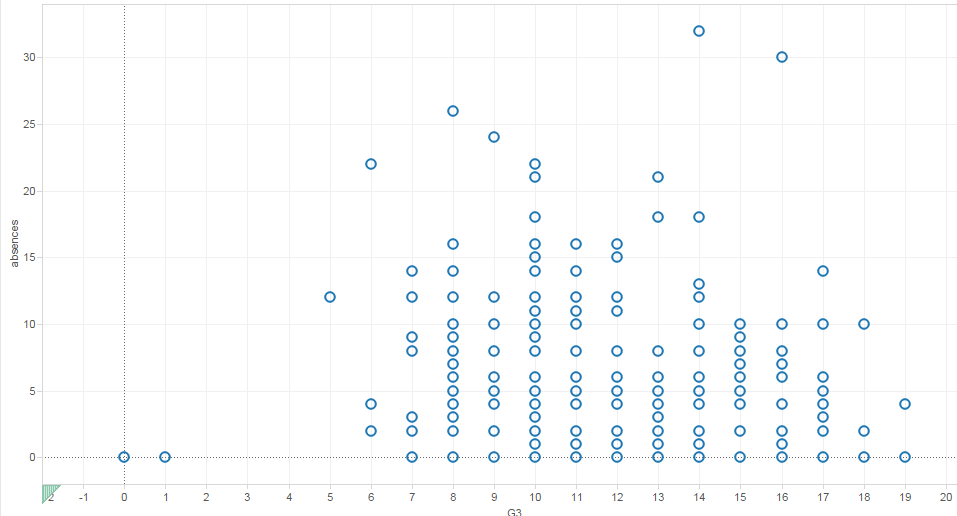
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Count | Mean | Median | Sd |
| 15 | 82 | 11.256 | 11 | 4.597 |
| 16 | 104 | 11.029 | 11 | 4.282 |
| 17 | 98 | 10.276 | 11 | 4.290 |
| 18 | 82 | 9.549 | 10 | 4.994 |
| 19 | 24 | 8.208 | 9 | 4.606 |
| 20 | 3 | 14 | 15 | 4.583 |
| 21 | 1 | 7 | 7 |  |
| 22 | 1 | 8 | 8 |  |

15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)   
 > It looks like decreasing

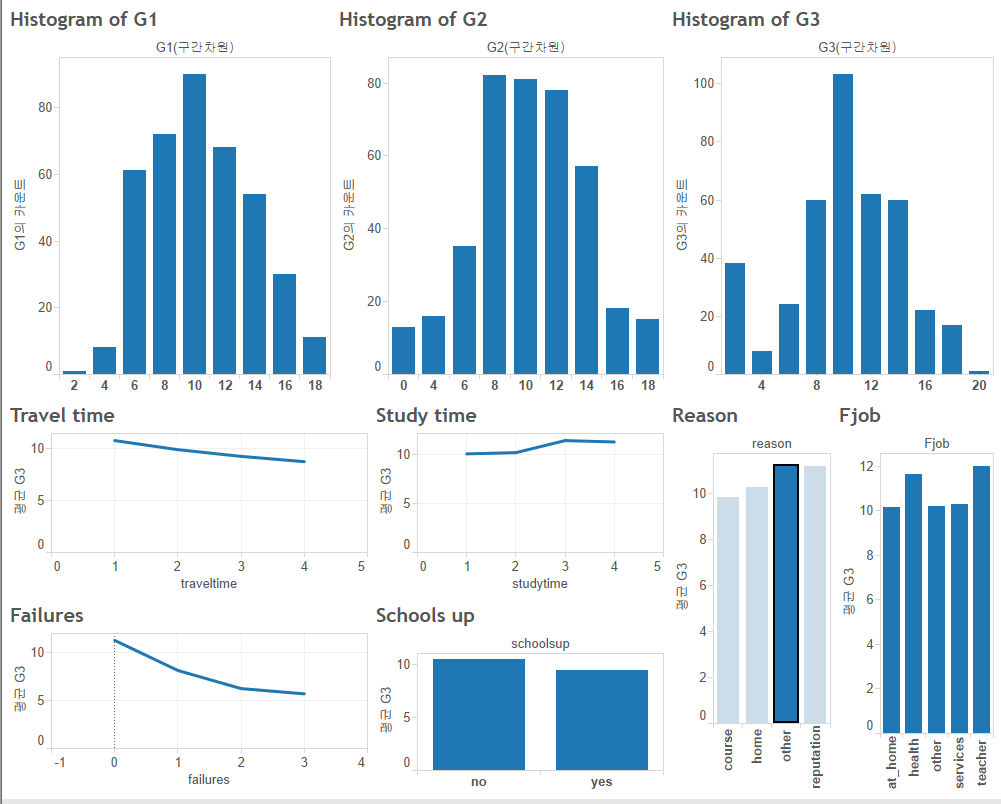
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Failures | count | Mean of G3 | Median of G3 | SD of G3 |
| 0 | 312 | 11.3 | 11.0 | 4.2 |
| 1 | 50 | 8.1 | 9.0 | 4.7 |
| 2 | 17 | 6.2 | 8.0 | 4.8 |
| 3 | 16 | 5.7 | 7.0 | 4.2 |

21. higher - wants to take higher education (binary: yes or no)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Higher | Count of G3 | Avg. G3 | Median G3 | Std. dev. of G3 |
| no | 69.0 | 8.8 | 10.0 | 3.0 |
| yes | 580.0 | 12.3 | 12.0 | 3.1 |

30. absences - number of school absences (numeric: from 0 to 93)   
> It looks like bell shaped.  


* Our target variable is G3 (Final score). So we plotted like as below variable plots.



**o What are the other salient aspects of the data (e.g. geospatial factors, text content, etc.)**From establishing the independence nation system, Portugal’s education system is based on Catholic. Many people prepare technique school and college entrance rate is very low.

**o Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.**We will model based on multiple regression and find best model with variable selection.  
If we can find proper model, we will interpret our model within the framework of Portugal’s culture. And we will compare to Korea’s education system.

**Members’ face ☺**

|  |  |
| --- | --- |
| **Hansol\_Kim** |  |
| **Youtaek\_Jeon** |  |
| **Youngha\_Lee** |  |
| **Hyoshin\_soh** |  |

c