Q1: see code.

Q2: answer question

Chart

Description automatically generated

Q3: answer question

Chart, line chart, histogram

Description automatically generated

Q4: answer question

Q5: answer question

Q6 : answer question

Unlike in other assignments, we’ve decided not to normalize the data, as we will be dealing with regression.

The features we decided to delete:

|  |  |
| --- | --- |
| Reason | Features |
| No extra information, ID feature | ID |
| Highly correlated to data that’s taken | Conversations Per Day |
| Extremely low score on decision tree. | Discipline Score |
| Highly correlated to data that’s taken | Happiness Score |
| Highly correlated to data that’s taken | Nr Cousins |
| Extremely low score on decision tree. | PCR\_10 |
| Extremely low score on decision tree. | PCR\_32 |
| Extremely low score on decision tree. | PCR\_46 |
| Extremely low score on decision tree. | PCR\_83 |
| Extremely low score on decision tree. | Social Activities Per Day |
| Extremely low score on decision tree. | Social Media Per Day |
| Highly correlated to data that’s taken | Sports Per Day |
| Highly correlated to data that’s taken | Steps Per Year |
| Highly correlated to data that’s taken | Studying Per Day |

Q7:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Section | Train MSE | Valid MSE |
|  |  | Cross Validated | |
| Dummy | 2 | 0.011190 | 0.011196 |

Q8: