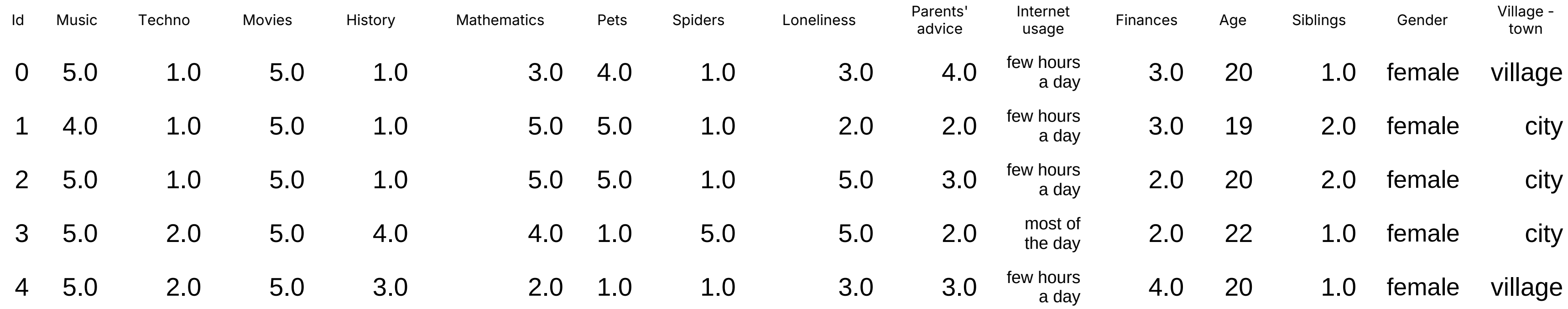
Project 1

Part 1 – Starting Data

The dataset is titled “young-people-survey-responses.csv”, which has 1010 rows and 15 columns (after dropping the ID column in the dataset).

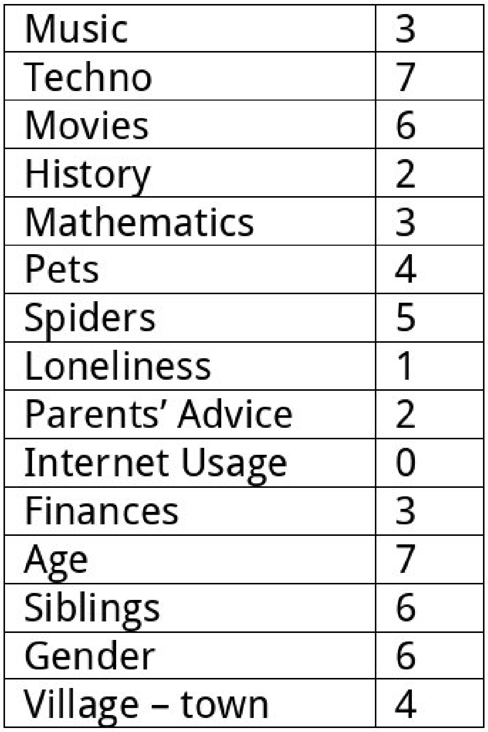


Id	Music	Techno	Movies	History	Mathematics	Pets	Spiders	Loneliness	Parents' advice	Internet usage	Finances	Age	Siblings	Gender	Village - town
0	5.0	1.0	5.0	1.0	3.0	4.0	1.0	3.0	4.0	few hours a day	3.0	20	1.0	female	village
1	4.0	1.0	5.0	1.0	5.0	5.0	1.0	2.0	2.0	few hours a day	3.0	19	2.0	female	city
2	5.0	1.0	5.0	1.0	5.0	5.0	1.0	5.0	3.0	few hours a day	2.0	20	2.0	female	city
3	5.0	2.0	5.0	4.0	4.0	1.0	5.0	5.0	2.0	most of the day	2.0	22	1.0	female	city
4	5.0	2.0	5.0	3.0	2.0	1.0	1.0	3.0	3.0	few hours a day	4.0	20	1.0	female	village

Figure 1

As is clear from the data above, most are numeric, and Internet Usage, Gender, and Village – town are string values. Internet Usage has 4 different strings (no time at all, less than an hour a day, few hours a day, most of the day), and the other two have only two different values.

This is a summary of all the missing values in each column. Each counts only one NaN value, though they don’t necessarily mean it’s the only missing value on that row.



Music	3
Techno	7
Movies	6
History	2
Mathematics	3
Pets	4
Spiders	5
Loneliness	1
Parents’ Advice	2
Internet Usage	0
Finances	3
Age	7
Siblings	6
Gender	6
Village – town	4

Figure 2

Our target variable is the Loneliness column. It has a roughly normal distribution.

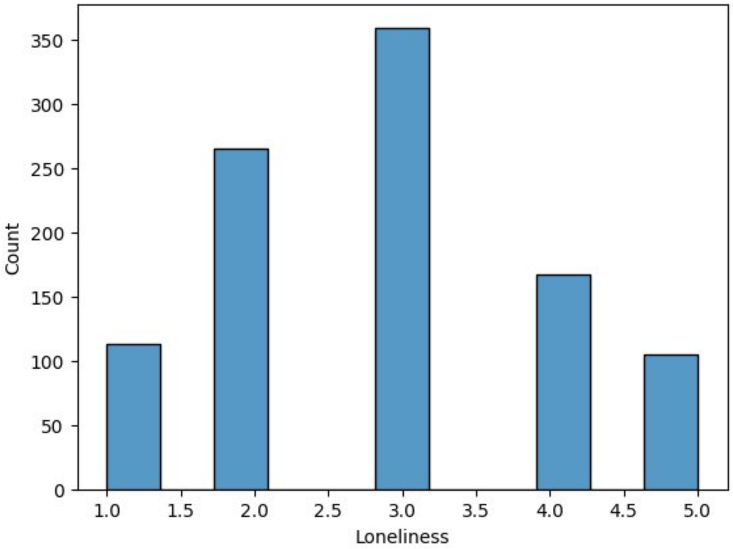


Figure 3

Part 2 - Exploration

Below is a correlation map with the ranges tightly restricted (from 0 to 0.25). None of the values strongly correlate with Loneliness (located in the middle), but some, such as Internet usage, Music, and Spider, interestingly, are slightly positively correlated.

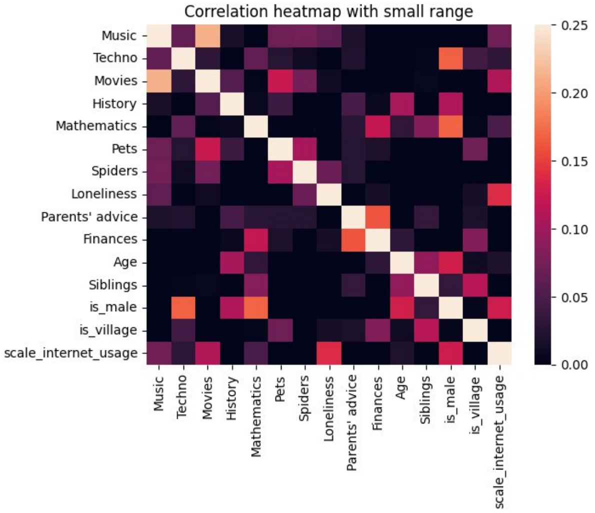


Figure 4

The range was restricted to see the differences in the map, since with no restriction the range is 1 - ~-0.3, which makes the differences much harder to see.

As none of them have basically any correlation, single-variate correlation will make a very poor model.

Below is a new map after the preprocessing step to remove columns with minimal correlation to Loneliness.

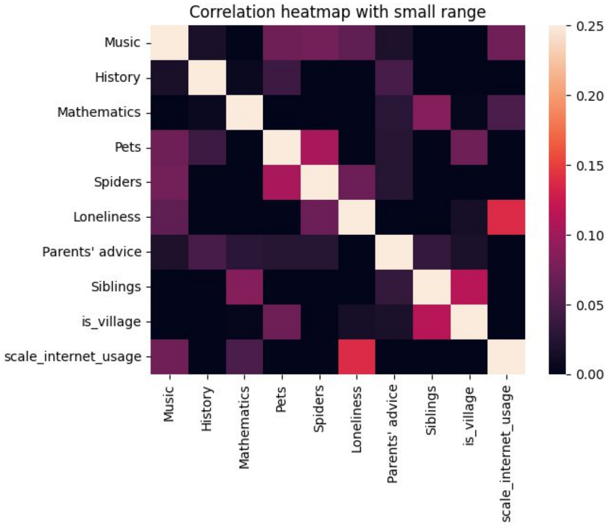


Figure 5

This threshold was determined to be around 1.5%, which was around half of our current columns and seemed like a logical breaking point in the data.

Part 3 – Preprocessing

There are a total of 10 rows that have NaN values in Gender and/or Village – town. These rows were removed since it's only 1 percent of the data and it would be more difficult to impute. The string-based columns are transferred to numeric and are summarized below.

Variable name	Variable type	Encoding method	Justification	Resulting column(s)
Gender	Nominal	One-hot	Not ordered, only 2 values	is_male
Village - town	Nominal	One-hot	Not ordered, only 2 values	is_village
Internet Usage	Ordinal	Ordinal	Ordered values	scale_internet_usage

For the remaining missing values, pairplots were made before and after imputing the mean to fill NaN values, and saw no meaningful difference (as would be expected by changing less than at most 0.6% of data). Only 5% of rows had any missing values. Imputing and saving the rows would be better than dropping all rows in the other numeric values.

All columns were removed which had an absolute correlation of less than 1.5%. Because their contribution was so small, I don't expect their effect on the final prediction to be large enough.

Variable Name	Reason for Removal
is_male	Absolute value of correlation to Loneliness was < 0.15% (extremely small).
Techno	
Movies	
Finances	
Age	

Part 4 – Cross Validation

Dataset	Samples	Features	% of Total
Training	800	10	80%
Testing	200	10	20%

80% is standard and it didn't break the distribution, so I left it. I tested a few different randomness values and found this one gave the most similar distribution.

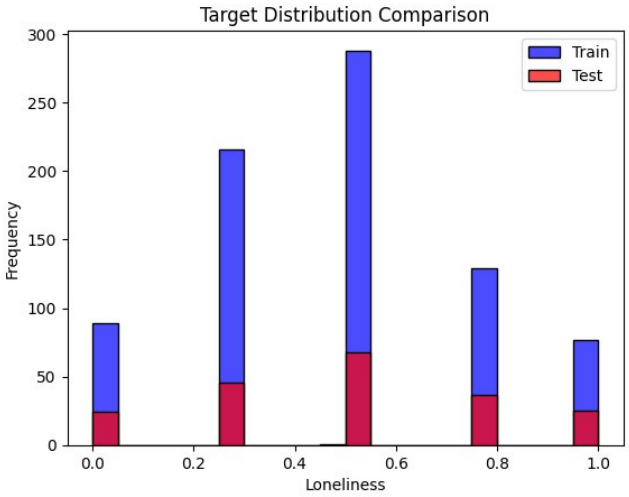


Figure 6

As you can see, the distribution of the Loneliness between the training and testing data is nearly equivalent.

The mean and STD before and after is between 0.03.

Part 5 – Scaling

No features needed to be scaled. The largest-ranging column was Age, which was removed in step 3 for not being correlated enough with Loneliness.

The rest of the columns have ranges from 3 to 10, which is within an order of magnitude and seems to be a reasonable margin.

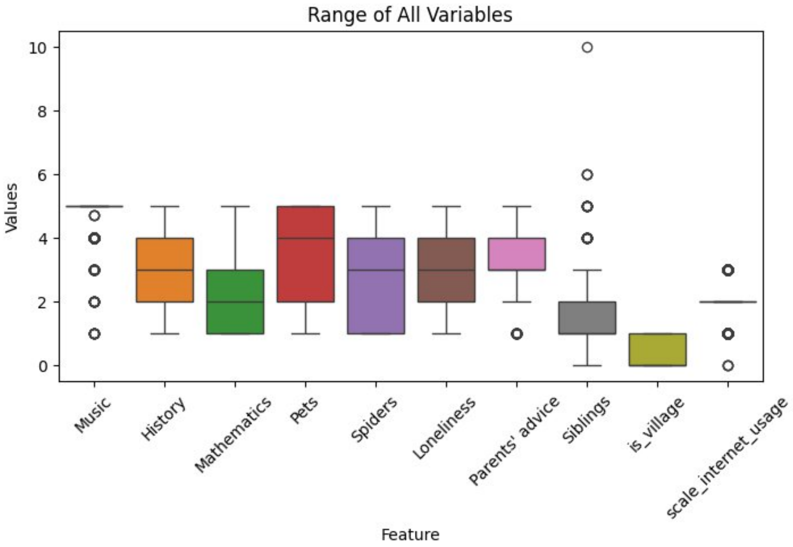


Figure 7

Part 7

Preprocessing step	Initial	Final	Change
Samples	1010	1000	-10
Features	16	10	-6
Missing Values	45	0	-45

This information is not well correlated, as can be seen from Figure 5. The closest is internet usage and loneliness at around 13%, which is not at all highly correlated (expecting > 70%).

The Age, Movies, and Techno features were the ones that were missing the most data, though many of the features were missing slight bits of information which did not meaningfully reflect in the model. In total, 5% of the rows were missing some amount of information.

The dataset has a good amount of information left, but from the analysis at the start, the data never seemed to be well correlated in any direction. I expect any model that was trained from this data will be little more than a random number generator on the expected Loneliness score.

Using a Categorical Naive Bayes solver and the test data in Figure 6, it fit for this confusion matrix:

1	6	16	0	1
3	11	30	2	0
1	13	52	2	0
0	8	28	0	1
0	5	18	1	1

Figure 8

This gives an average precision and F1 score of 0.25, and a recall of 0.33, so it seems to just be randomly guessing which is correct.