The first thing we noted was in our dataset there were 10 rows and 16 columns.

Unnamed: 0	Music	Techno	Movies	History	...	Finances	Age	Siblings	Gender	Village - town
0	5.0	1.0	5.0	1.0	...	3.0	20.0	1.0	female	village
1	4.0	1.0	5.0	1.0	...	3.0	19.0	2.0	female	city
2	5.0	1.0	5.0	1.0	...	2.0	20.0	2.0	female	city
3	5.0	2.0	5.0	4.0	...	2.0	22.0	1.0	female	city
4	5.0	2.0	5.0	3.0	...	4.0	20.0	1.0	female	village
5	5.0	1.0	5.0	5.0	...	2.0	20.0	1.0	male	city
6	5.0	5.0	4.0	3.0	...	4.0	20.0	1.0	female	village
7	5.0	3.0	5.0	5.0	...	3.0	19.0	1.0	male	city
8	5.0	1.0	5.0	3.0	...	2.0	18.0	1.0	female	city
9	5.0	1.0	5.0	3.0	...	4.0	19.0	3.0	female	city

Here are the data types in our dataset

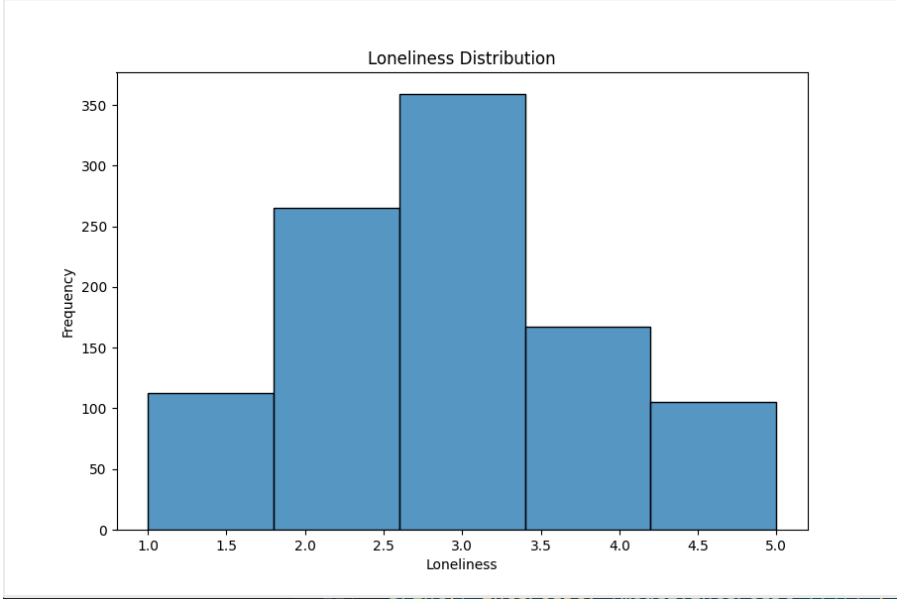
Data columns (total 16 columns):

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	1010 non-null	int64
1	Music	1007 non-null	float64
2	Techno	1003 non-null	float64
3	Movies	1004 non-null	float64
4	History	1008 non-null	float64
5	Mathematics	1007 non-null	float64
6	Pets	1006 non-null	float64
7	Spiders	1005 non-null	float64
8	Loneliness	1009 non-null	float64
9	Parents' advice	1008 non-null	float64
10	Internet usage	1010 non-null	object
11	Finances	1007 non-null	float64
12	Age	1003 non-null	float64
13	Siblings	1004 non-null	float64
14	Gender	1004 non-null	object
15	Village - town	1006 non-null	object

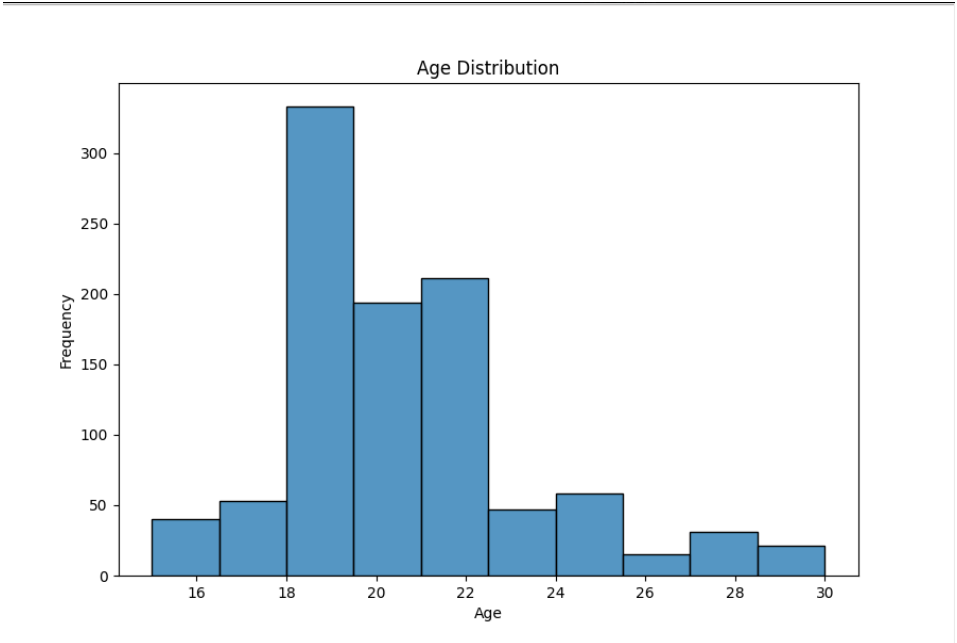
The variable we are trying to predict using this dataset is loneliness.
We began by getting the missing values per column as seen in this image.

	missing_count	missing_percent
Unnamed: 0	0	0.00
Music	3	0.30
History	2	0.20
Pets	4	0.40
Spiders	5	0.50
Loneliness	1	0.10
Internet usage	0	0.00
Age	7	0.69
Gender	6	0.59
Village - town	4	0.40

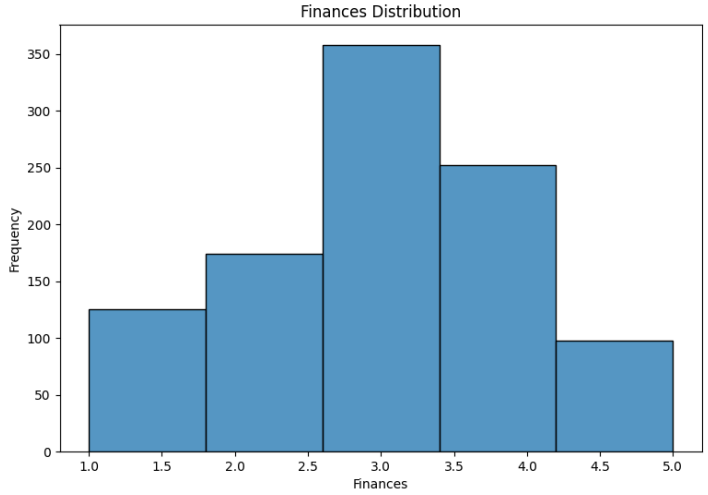
Distribution of Loneliness



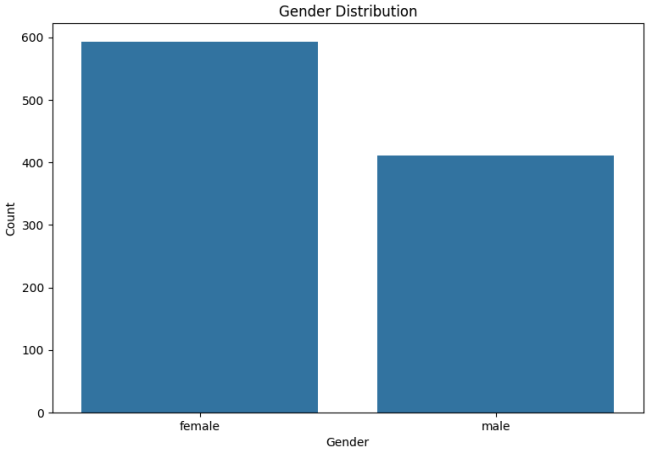
While exploring our dataset we looked at the distribution of many variables to see if any preprocessing was needed.



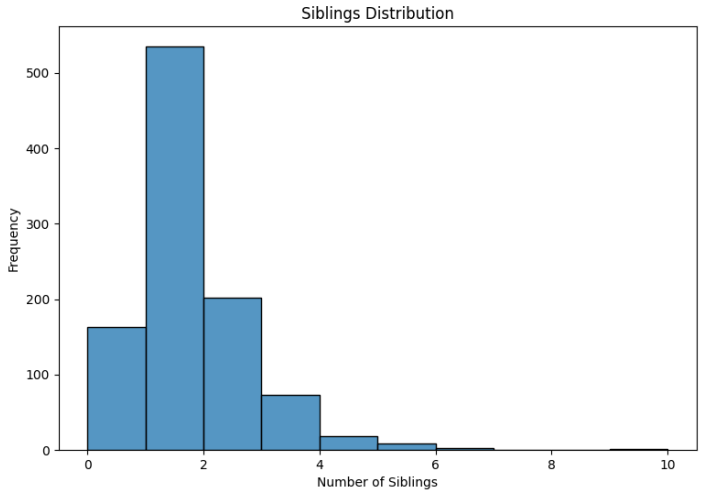
In our Age Distribution we found that the variable is right skewed.



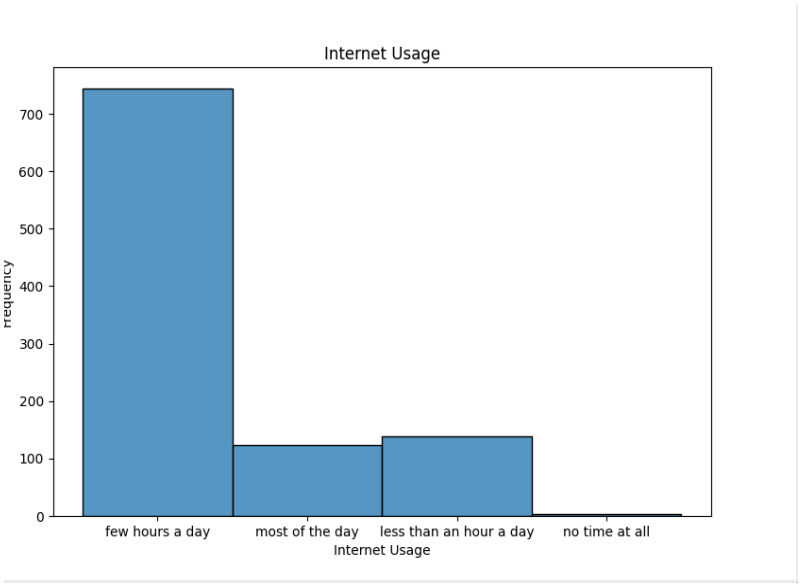
In our Finances Distribution we found that the variable was almost normally distributed.



In our Gender Distribution we found that there are more females in this dataset than males.

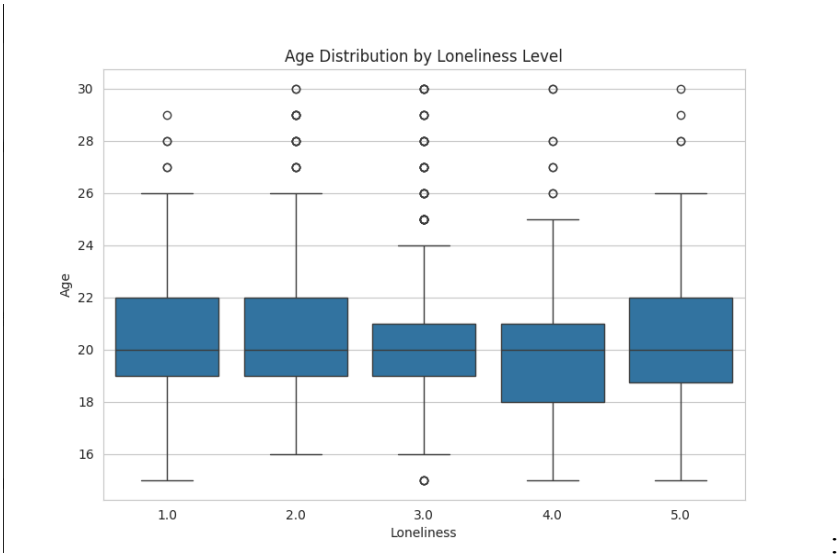


In our Siblings Distribution we found that the variable is right skewed. We also found that there was an outlier.

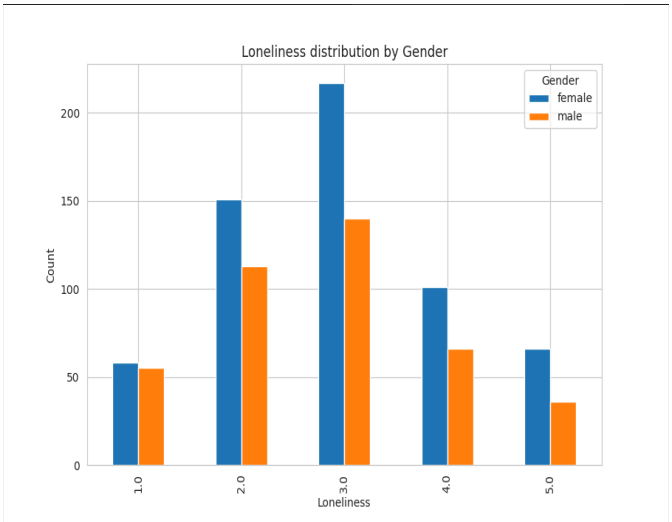


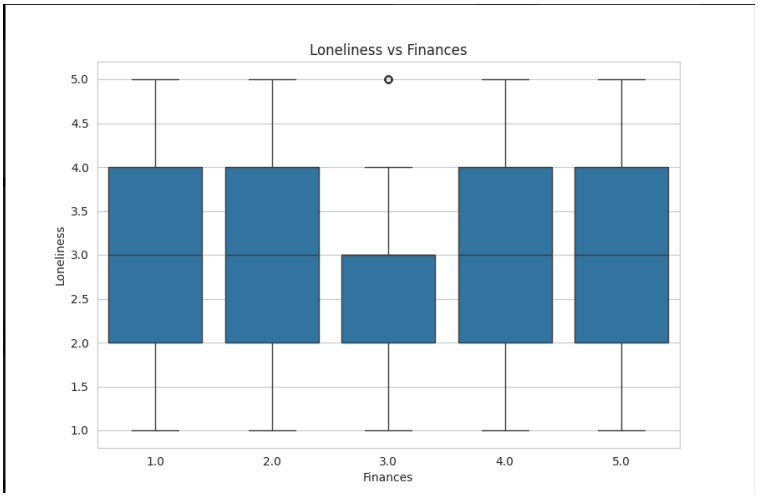
The internet usage of young adults showed that most are using the internet for a few hours a day. Interestingly there were only a couple of respondents who said no time at all.

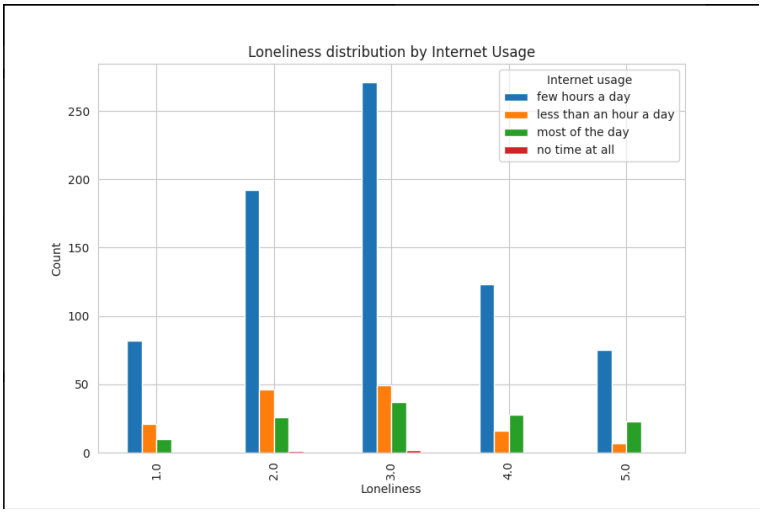
Relationships to Target Variable

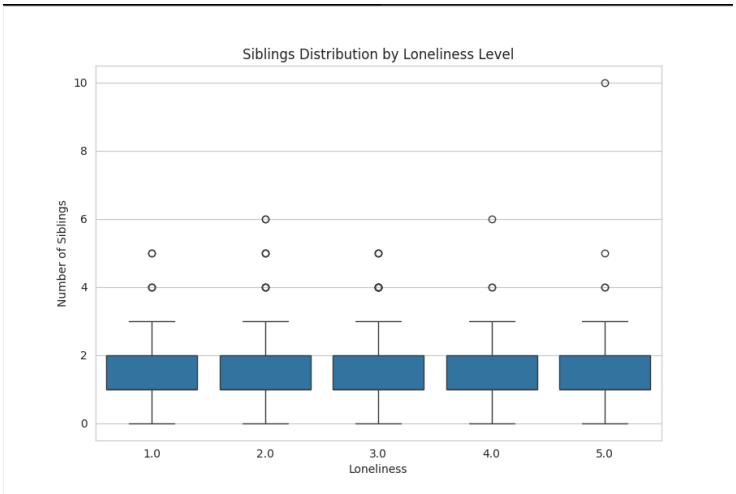


:


Loneliness distribution by Gender
Gender
female
male
Count
Loneliness
0
50
100
150
200
1.0
2.0
3.0
4.0
5.0

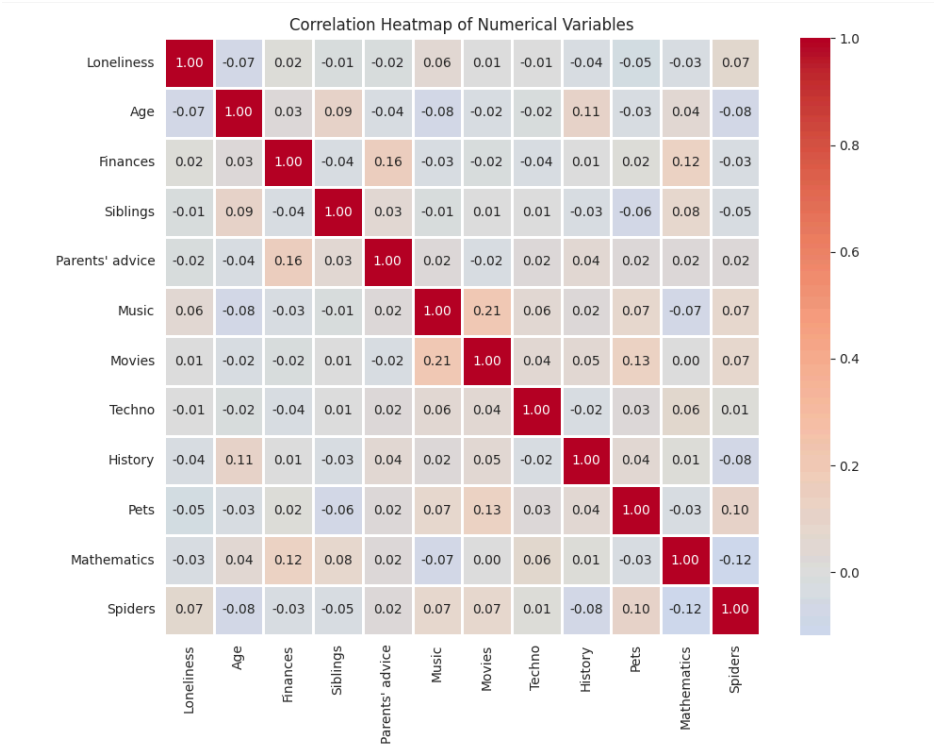

Loneliness vs Finances
Loneliness
Finances
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Loneliness distribution by Internet Usage
Internet usage
few hours a day
less than an hour a day
most of the day
no time at all
Count
Loneliness
0
50
100
150
200
250
1.0
2.0
3.0
4.0
5.0



Looking at the relationships of our variables to our target we found it interesting that those who have a middle of the road in finances are less lonely. We found that those who use the internet more are ever so slightly more lonely. We found that people who are less lonely tend to be slightly older.

For our preprocessing we decided to drop the Techno column, drop any missing loneliness rows and drop any missing finances rows. We filled the gender column by having a function to decide whether it would be male or female using a 50/50 chance. We filled the age column using the calculated average age. We used ordinal encoding on Internet Usage as it needed special handling. Then finally filled the missing siblings values using the calculated average.



There weren't any highly correlated variables. The most correlated variable was spiders although this still wasn't a very strong correlation it was what we used primarily.

Removing variables:

Variable Name	Reason for Removal
Techno	has no correlation with target
Id 586	missing loneliness value
Movies	no strong correlation with target
Siblings	no strong correlation with target
Finances	no strong correlation with target
Parents Advice	no strong correlation with target
Mathematics	no strong correlation with target

After removing these columns from the dataset there were only 8 left of the 15 that were originally part of the dataset.

Imputation Strategy Table

Variable Name	Variable Type	% Missing	Imputation Method	Justification
Loneliness	Numerical	0.1%	Dropped	Takes small amount
Age	Numerical	0.69%	Median	Skew in age
Gender	Categorical	0.59%	Random	To try and show stronger correlation
Music	Numerical	0.3%	Dropped	Takes small amount
History	Numerical	0.2%	Dropped	Takes small amount
Pets	Numerical	0.4%	Dropped	Takes small amount
Spider	Numerical	0.5%	Dropped	Takes small amount

Encode Strategy Table

Variable Name	Variable Type	Encoding Method	Justification	Resulting Columns
Gender	Nominal	One-Hot	No Inherent Order	3 Columns
Internet Usage	Ordinal	Ordinal	Has Inherent Order	2 Columns

Scaling Decision Table

Feature Name	Data Type	Scaling Applied?	Justification
Music	Discrete	No	Already in range we were working with.
Gender	Binary	No	Already in range we were working with.
History	Discrete	No	Already in range we were working with.
Pets	Discrete	No	Already in range we were working with.

Internet Usage	Discrete	No	Already in range we were working with.
Age	Continuous	No	No difference in results

Processing Summary Table

Preprocessing Step	Initial	Final	Change
Samples	1,010	991	-19
Features	16	9	-7
Missing Values	32	0	All Handled

There wasn't anything super insightful from this dataset to predict loneliness. While there were very minor correlations nothing stood out that could be used to predict loneliness. The most challenging part of the preprocessing was finding the correct solution for our missing data. Age had the most missing data which we solved by filling it using the calculated average. We got new features from our hot encoding which were: gender_male, gender_female and internet_usage_ord. We don't believe we were able to get an accurate representation of what the data was trying to show as well as not providing any strong predictors. Through our EDA process we decided to remove the finances column and keep the gender column.