Exploratory Data Analysis for Machine Learning: Honors Peer-graded Assignment

Julián Alberto López Sánchez

Description of Dataset

Like humans, cats have personalities, with behavior differences between individuals. The identification of the personality type is important as with those differences, the environmental needs also change to reach a good quality of life.

The dataset that is going to be worked on is "Reliability and Validity of Seven Feline Behavior and Personality Traits" and its objective was to study the different personality and behavior of over 4300 cats based on the answers of their owners using an online questionnaire and study its validity and reliability, concluding that they are a valid source of behavior data. The feline personality and behavior include seven traits: fearfulness, activity/playfulness, aggression toward humans, sociability toward humans, sociability toward cats, excessive grooming and litterbox issues.

Attributes Summary

There is one hundred fifty-one (151) attributes, for this project there is going to be a focus on twelve, the other one hundred thirty-nine (139) attributes consist in behavioral and personality questions from owners where the answers are in a range from 1 to 5, where one means "disagree" and five means "strongly agree". The twelve features that are the focus of the project are going to be:

- Breedgroup: defines the breed group of the cat, there is twenty-six different groups, those being: "Landrace cat shorthair", "Landrace cat longhair", "House cat", "European", "Maine Coon", "Bengal", "Siberian and Neva Masquerade", "Oriental", "Sphynx and Devon Rex", "Abyssinian", "Ocicat", "British", "Norwegian Forest Cat", "Sacred Birman", "Cornish Rex", "Russian Blue", "American Curl", "Somali", "Siamese and Balinese", "Burmese", "Turkish Angora", "Korat", "Persian and Exotic", "Turkish Van", and "Others".
- Sex: defines the sex of the cat where it could be "Male" or "Female.

- Age Behavior: defines the age when the personality section was answered, its values are continuous and in years.
- Other Cats: defines if there are other cats in the same household, its values could be “Yes” or “No”.
- Problematic Behavior: defines if the owner feels if the cat has any problematic or unwanted behavior, its values could be: “No”, “A little”, “Some”, “A lot”.
- Fearfulness: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate a higher fearfulness.
- Human Aggression: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate a higher aggression towards humans.
- Activity/Playfulness: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate a higher activity/playfulness.
- Cat Sociability: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate a higher sociability towards cats.
- Human Sociability: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate a higher sociability towards humans.
- Litterbox Issues: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate more litterbox issues.
- Excessive Grooming: one of the personality traits, defines a factor score calculated using the answers from the survey, its values are continuous, where the higher score indicate more excessive grooming.

Initial Plan for Data Exploration

The initial plan for the data exploration was to check the different columns and review which were relevant to work with. After that transform features like the “OTHER_CAT” column into a Boolean and “problematic_behavior” into a range using ordinal encoder. Once all the relevant features are into a better format, start to work with the missing data, reviewing the best ways to impute the data of the different types of features, then check for skew data and outliers, finally check for correlation between the different attributes using plots.

Actions Taken for Data Cleaning and Feature Engineering

Started by reviewing the attributes of the dataset and notice the most of them are personality and behavior questions that were used to calculate the last seven attributes, the dataset don’t provide the way that they were calculated, so even if there were some rows with unknown data it the personality features, imputing the data wouldn´t allow me to calculate the new factor scores, so a discard most of the columns and just focus on the factor score attributes and the second to fifth attributes that are background information about the cat.

Once this is done a copy of the dataset is created with only the features that are going to be work on, for each column it was counted number of unique values and it was noticed that in three features there was missing data marked with the String “unknown”, so a line of code was added to replace the word “unknown” of the data frame with a NumPy Nan value, as this would facilitate the process of imputing data, all of this using the Pandas library for Python.

The first attribute to change was “AGE_BEHAVIOR”, there was a small amount of missing data, one hundred twenty-three (123) entries, so it was decided to use the “SimpleImputer” from the “Scikit-Learn” library from Python and replace the missing data with the mean of the column values which was 5.826055.

The next attribute was “OTHER_CATS”, there was nine hundred ninety-seven (997) missing values, being that the feature only had two possible values, the “SimpleImputer” was used again, but this time with the “most_frequent” strategy where all missing values were changed to “Yes”, and then using Pandas and the method “.map” the values of “Yes” and “No” were changed to the Booleans True and False, respectively.

The last feature that needed imputation was “problematic_behavior”, there was again nine hundred ninety-seven (997) entries with missing values, the same rows with the missing values in the “OTHER_CATS” attribute. First the values were changed into a range from 0 to 3 and a NumPy Nan value, this was made possible with the “OrdinalEncoder” from the “Scikit-Learn”, then imputation was needed for the Nan values, as previously said, these rows were the same as the “OTHER_CATS” missing value rows, but because of the way that the latter were changed, it didn´t seems like a good feature to try and find a correlation, so it was decided to use the factor score features and find an attribute that was related with the changes of the problematic behavior of the known values.

The feature with one of better correlation was “fearfulness”, so to impute the missing data it was decided to use the K-Nearest Neighbor Algorithm or KNN, this is a machine learning-based, non-parametric, multi-variant method which uses proximity to classify or predict a grouping of an individual data point. To make this possible the “KNNImputer” of “Scikit-Learn” was used with two neighbors, at the end the result were values with decimals, so they had to be rounded using NumPy round so they could fit in the range made with the ordinal encoder.

Once all the missing values were imputed, I started to inspect skews and outliers, being the nature of the dataset a questionnaire, it was decided to not change the skew, for example: most owners didn’t feel like their cat had any problematic behavior, that caused a positive skew within that feature, changing that to a normal distribution would make that most cats have some problematic behavior and would ignore the impute of the owners. In case of outliers, they were also untouched, most of the features with values that could be considered outliers were the factor scores, and those being calculation based on the behavior attributes means that they could provide valuable data and insight about special cases.

Key Findings and Insights

Firstly, checking the different breed groups we can see that there are some much commons than other, being the Landrace cat Shorthair the most common one, followed by the House cat, and the Turkish Van the most uncommon breed group of the dataset, as seen in Figure 1.

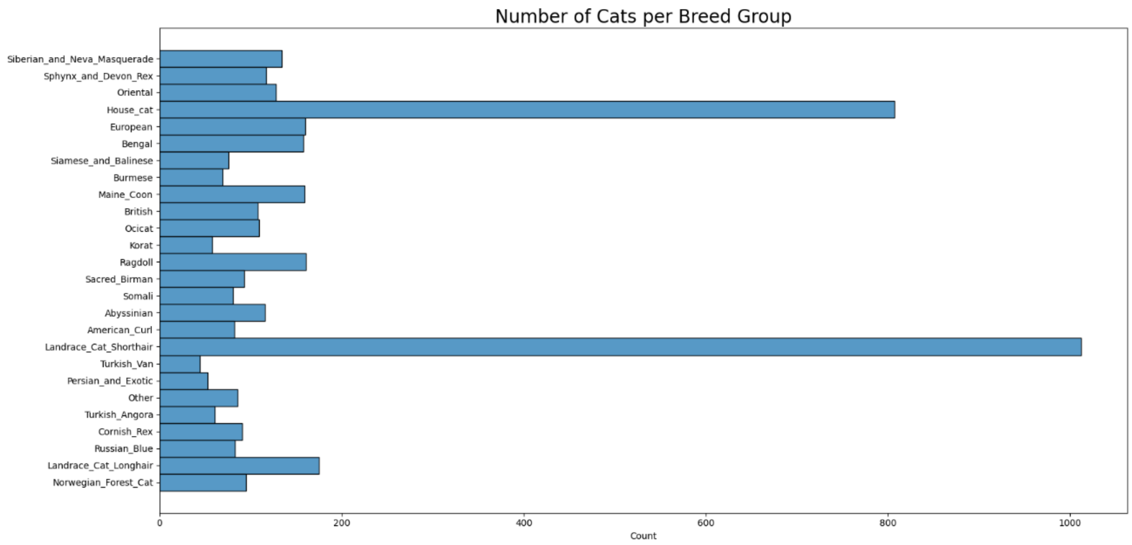


Figure 1:Number of cats per Breed Group.

Regarding the sex of the cats, most of them are males, but not by a considerable amount as can be seen in Figure 2.

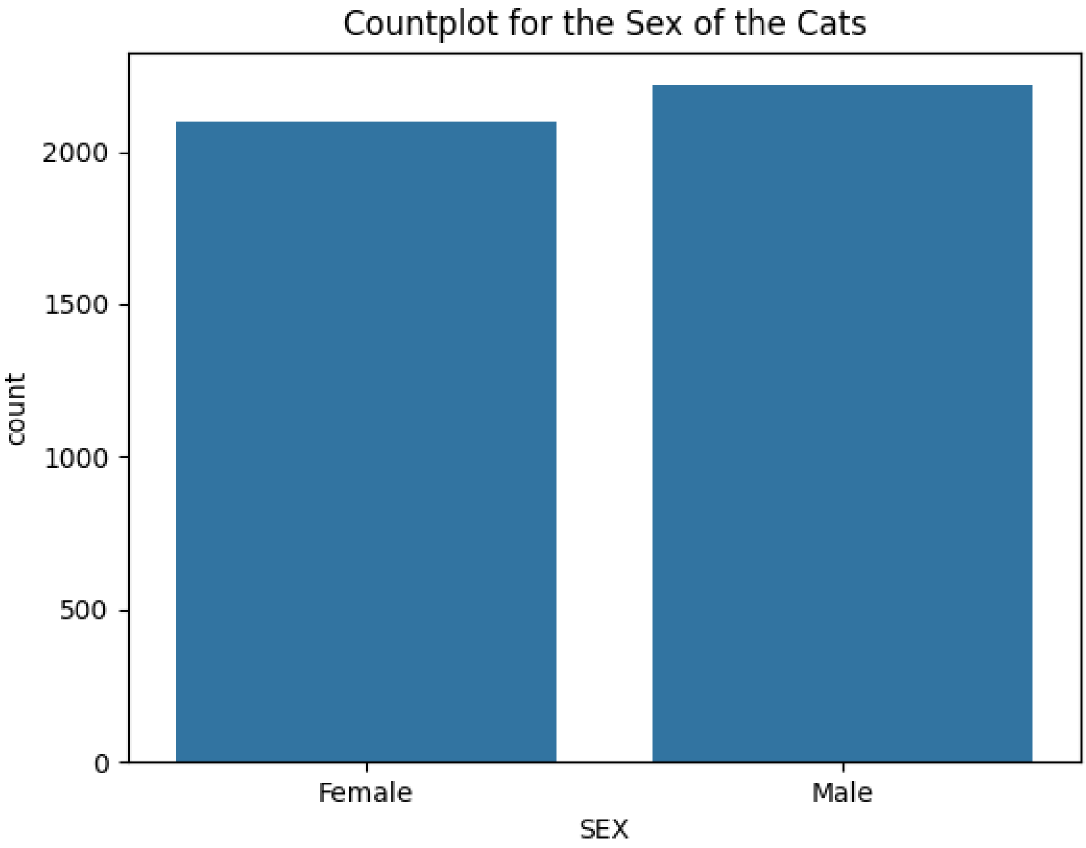


Figure 2: Number of male cats vs. female cats.

Reviewing the age behavior of the cats from the survey, it can be seen in the density plot that the mean is around 5.83, and there is a positive skew because of outliers, removing these outliers doesn't change much about the mean because it is a very small amount compared to the total population. Regarding of the shape of the density plot, even if the mean is 5.83, there is a lot of cats with ages between 0 and 5, this distribution makes it that there is a peak within this range, as seen in Figure 3.

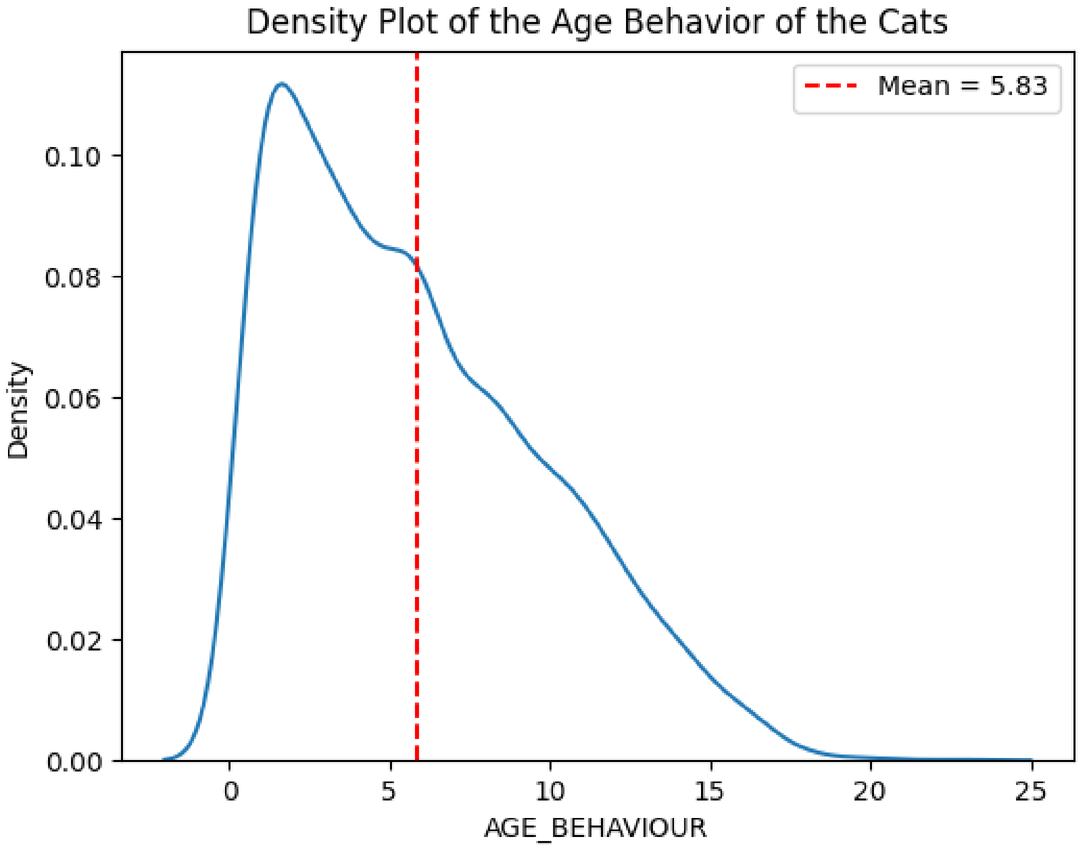


Figure 3: Density plot showing the age behavior of cats.

In the case of the “Other cats” attribute, there is a significant difference. Most owners adopt of buy multiple cats so they can play with each other and have company when the owner is not around, so the difference between the “True” column, with three thousand six hundred thirty-six entries (3636), and the “False” column with six hundred eighty entries (680) is weighty, as seen in Figure 4. Even before imputing the data making 997 Nan entries “True”, there was still a difference of two thousand six hundred and thirty-nine (2639) vs. six hundred eighty (680) entries.

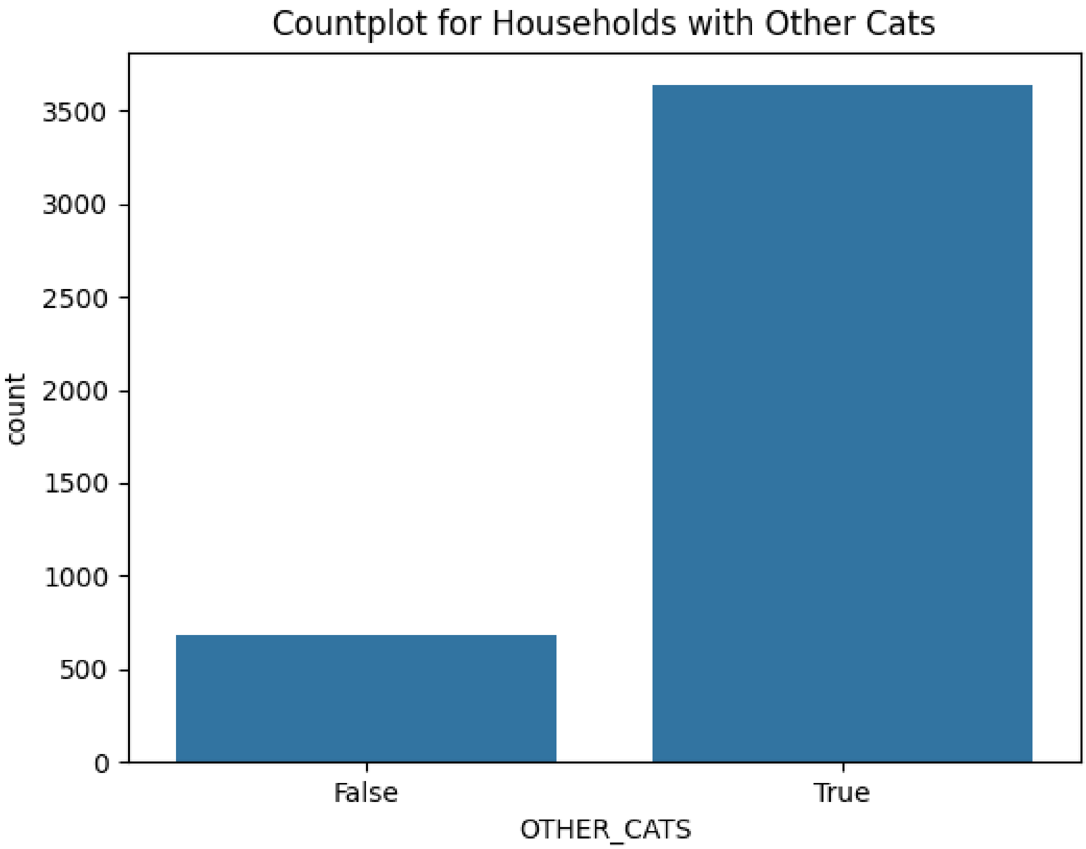


Figure 4: Number of households with more than one cat.

Lastly there is the "problematic behavior" feature where it can be seen in the Figure 5, that most owners said that their cats don't have any problematic behavior, this creates a positive skew, either because most cats really don't have any problematic behaviors or because most of the owners feel like there is no problematic behavior within their cat activities, like mention before, being the nature of the dataset a survey, this really cannot be changed. This distribution is maintained before and after the KNN imputation.

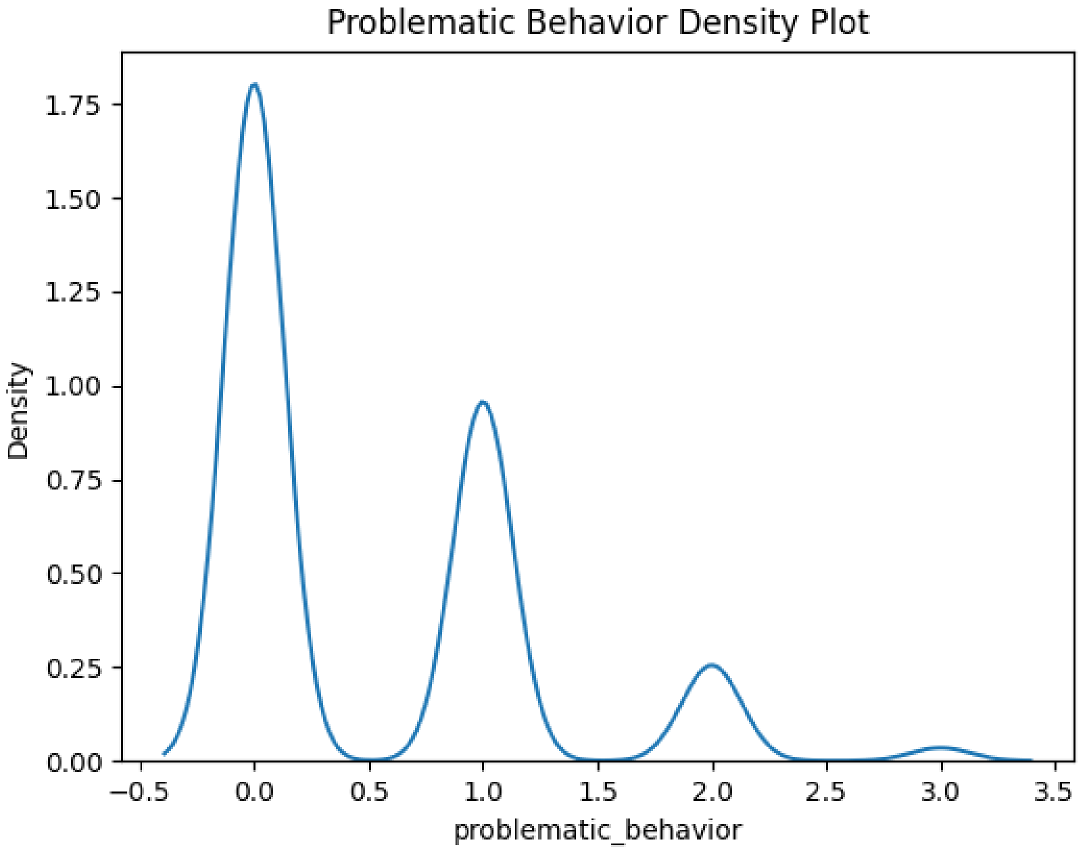


Figure 5: Density plot of the problematic behavior of the cat's population.

There is also the subject of the relationship between the problematic behavior and fearfulness attributes, like it was said these features have a correlation where if the problematic behavior of the cat is greater, the fearfulness score also increases, as seen in Figure 6. In the box plot it can be seen some outliers, but these are from the fearfulness factor that was calculated between all the behavior question, so there is the possibility of special depending in the value of this attributes.

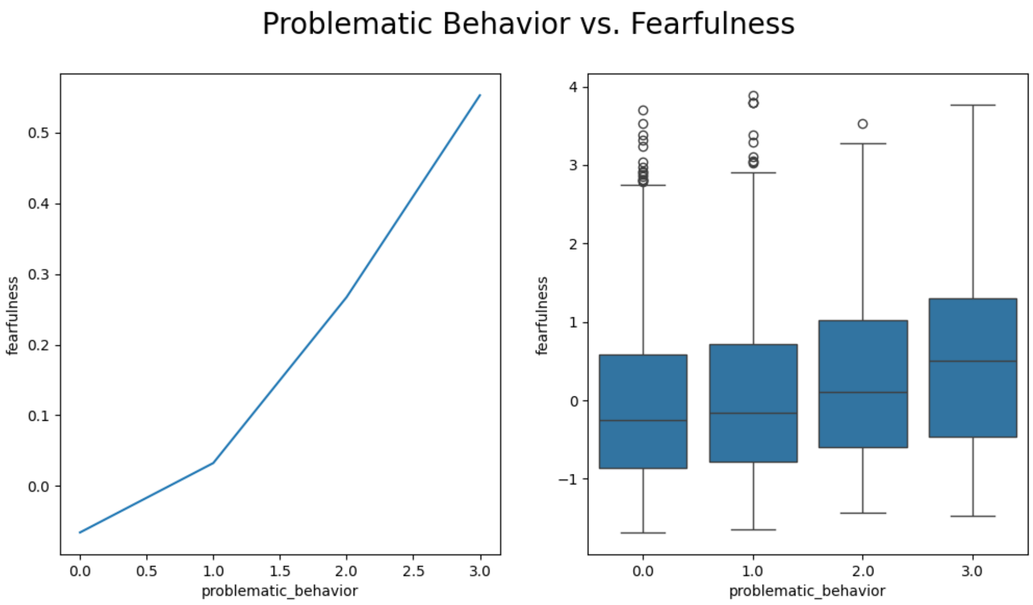


Figure 6: Relationship between Problematic Behavior and Fearfulness.

Hypothesis

The three hypothesis that were formulated are:

1. Female cats have a bigger “litterbox issue” score than male cats.
    a. H0: The sex of the cat doesn’t affect the “litterbox issue” score.
    b. HA: Female cats have a bigger “litterbox issue” score than male cats
2. The age of the cat affects it’s “cat sociability” and “activity/playfulness” score.
    a. H0: The age of the cat doesn´t it's “cat sociability” and “activity/playfulness” score.
    b. HA: The age of the cat affects it's “cat sociability” and “activity/playfulness” score.
3. Having other cats in the household affects the “human sociability” score of the cat.
    a. H0: Having other cat in the household doesn't affect the “human sociability” score.
    b. HA: Having other cat in the household affects the “human sociability” score of the cat.

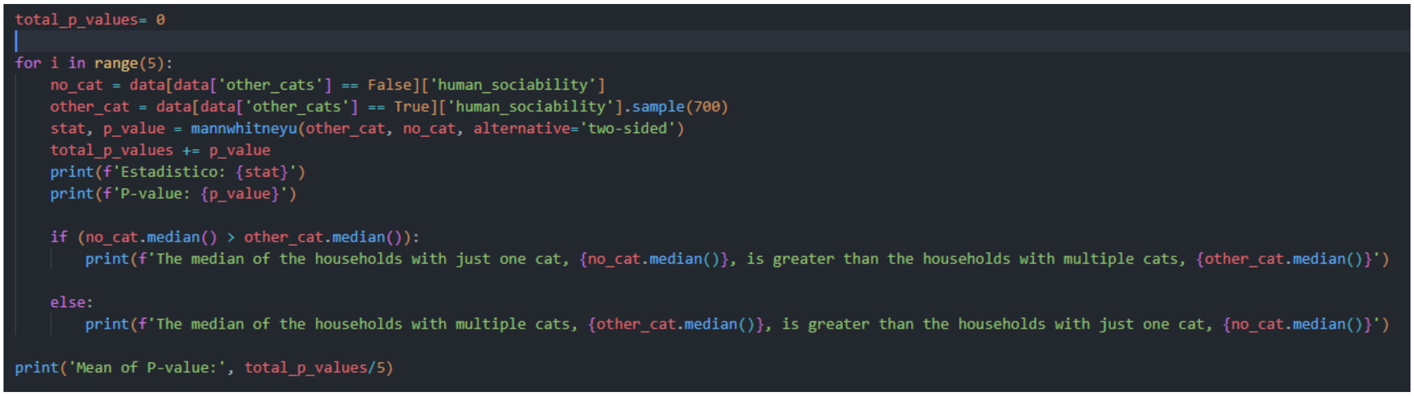
Significance Test

In this case, the third hypothesis is the one that are going to be worked on, we are going to check if having more than one cat in the household improves the sociability towards humans of the cats. The p-value for the significance level is going to be 0.05 and for the test we are going to use the Mann Whitney U test, this allows us to compare whether there is a difference between the two groups, with and without multiple cats in the household, is statistically significant.

First it is necessary to separate the data into two groups, for that we create two new variables: “other_cat” and “no_cat”, these are going to store the values of the “human_sociability” feature when there are other cats and when it is not the case, but because of the difference in size between the groups, where there is more than five times the amount of household with more than one cat that those with just one, we are going to created multiple subsamples with the larger group, to do this we are going to insert the “other_cat” variable with the “.sample” method where we are going to pass the parameter 700 and replace=False, this will ensure that each sample size is seven hundred entries long and will prevent duplicates within each sample.

We are also going to create the variables “total_p_values” to store the sum of the p-values of each test with the different samples. Once that is done, inside the loop we create two variables “stat” and “p_value” and use them to initiate the Mann Whitney U function and pass it the “other_cat” and “no_cat” variables, and set the alternative parameter as “two-sided”, this is because the distributions are not equal.

Lastly, we make an “if” conditional to compare the median of the “other_cat” sample and the median of the “no_cat” group to see which one is greater in each loop, then we print the mean of the “total_p_values” to get a general answer.



```python
total_p_values= 0

for i in range(5):
    no_cat = data[data['other_cats'] == False]['human_sociability']
    other_cat = data[data['other_cats'] == True]['human_sociability'].sample(700)
    stat, p_value = mannwhitneyu(other_cat, no_cat, alternative='two-sided')
    total_p_values += p_value
    print(f'Estadistico: {stat}')
    print(f'P-value: {p_value}')

    if (no_cat.median() > other_cat.median()):
        print(f'The median of the households with just one cat, {no_cat.median()}, is greater than the households with multiple cats, {other_cat.median()}')

    else:
        print(f'The median of the households with multiple cats, {other_cat.median()}, is greater than the households with just one cat, {no_cat.median()}')

print('Mean of P-value:', total_p_values/5)
```

Figure 7: Code made to test the difference in the human sociability score depending if there are or no other cats in the household.

The result of running this test multiple times is a p-value in the range of ~5.83e-08 and ~5.39e-05, the mean value of the total p-value was 1.25e-05. This means that in all the cases the p-value was smaller than the significance level p-value of 0.05, in other words, we can reject the null hypothesis and say that having another cat in the household affects the human sociability score of the cat. When comparing the median of each sample in the loop, all of them gave the same result, the median of the households with just one cat, 0.25, was greater that the household with multiple cats, that means that the human sociability score is better when there is just one cat.

Next Steps for Analyzing this Data

The results from the significance test may suggest that the household where cats have the company of other cats are not that dependent of their owners or humans, that doesn’t have to mean they don’t like their owners, but instead, that their social necessities can be satisfied by a feline companion. To further study this we could check the correlation of the groups with and without other cats, and the cat sociability feature, we could also check based on the age or breed group, the necessities of the cat may change depending of their age or if they need special cares for their health, for example: Sphynx cats doesn’t have hair, so they need special care for their skin like applying oils or even put clothes on them when the temperature drops, this could make a stronger bond with their owner even if there were other cats.

Quality of Dataset

The data wasn’t that bad, there was enough to work with and impute the ones that were missing, the problem was maybe in columns like “OTHER_CATS”, being Boolean data the best option for imputing was using the most frequent one, this means it’s likely that some data for cats that lived alone has been lost, this also adds skewness to that feature, also with the factor scores attributes there was some skewness and outliers, those being calculations means that the information is still valuable so there was no change for those values, but if we had the data of the equation used to calculate each score based on the answers of the behavioral question of the survey, then we could have a better insight of that data and even be able to change it imputing some missing data of the multiple behavioral questions.

Bibliography

Mikkola, S., Salonen, M., Hakanen, E., Sulkama, S., & Lohi, H. (2021). Feline behavior and personality survey data (Version 2). figshare. https://doi.org/10.6084/m9.figshare.14899077.v2