A DATA SCIENCE EXPEDITION INTO ACNH'S DYNAMIC WORLD

1. Group Members:

1. Kevin Timothy Muller - 201779539
2. Vishanth Suresh - 201669090
3. Aneeta Cherian - 201801715
4. Saaketh Kopuru - 201799272

2. Introduction:

The following report emphasizes on cleaning the data and the detailed exploration of the dataset acquired from the science data bank. The report contains an AI model that can predict a player's environmental perception depending on the player's social demographic profile. Acquired from an external website, the dataset unfolds a narrative of environmental worldviews and behaviors among 640 individual players immersed in the realm of Animal Crossing: New Horizons (ACNH).
This piece of information is divided into six categories: socio-demographic profiles, COVID-19 concerns, environmental perceptions, game-playing habits, in-game behaviors, and the emotional contours of the gaming experience.
The dataset stands as a testament to the combination of entertainment and societal impact. It is aimed to facilitate a deeper understanding of the intricate interplay between virtual behaviors and environmental perspectives.

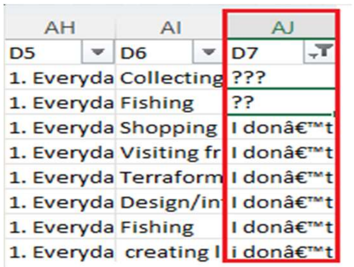
3. Data quality :

After a thorough investigation of the dataset there were multiple instances of data quality issues observed.

- Firstly, the initial column, which lacked a proper name, has been renamed as "Gamer_ID." This column, holding unique identifiers for each record, has been reorganized in ascending order based on Gamer_ID, giving it a more coherent dataset structure. Additionally, in the A4 column, replacing "nan" with "none" is proposed for consistency and clarity.

- Significant gaps in data completeness are observed in several columns. Notably, the D1, D2, and D3 columns exhibit empty cells for specific Gamer_IDs. Addressing these missing values is crucial to ensuring the dataset's integrity and reliability.

- Reportedly in the D4 and D7 columns, for certain Gamer_IDs in D4, unrealistic values are reported, such as 45 hours of gameplay per day. Similarly, in the D7 column, select Gamer_IDs provide responses that are non ascii values.(see figure 1.1)



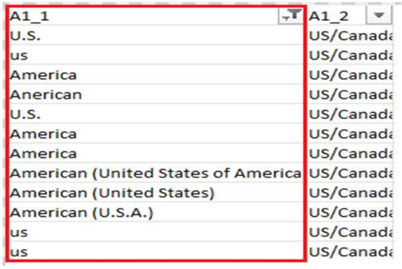
AH	AI	AJ
D5	D6	D7
1. Everyda	Collecting	???
1. Everyda	Fishing	??
1. Everyda	Shopping	I donâ€™t
1. Everyda	Visiting fr	I donâ€™t
1. Everyda	Terraform	I donâ€™t
1. Everyda	Design/in	I donâ€™t
1. Everyda	Fishing	I donâ€™t
1. Everyda	creating l	i donâ€™t

Figure 1.1

- A nuanced issue arises in the age and education-related columns. For Gamer_IDs 27, 36, 50, and 593, the reported age contradicts the indicated postgraduate educational qualification, prompting a validation of such discrepancies.

- Anomalies are detected in the A1 column, where Gamer_ID 195 provides an irrelevant answer. Furthermore, a typographical error is noted in the spelling of "Canadian" for Gamer_ID 251, initiating a correction for data consistency.

- User entries for the geographical nationality they belong to are not properly grouped or categorized. For example, people who belong to the nationality ‘American’ use various instances describing the country they originate from.(see figure 1.2)



A1_1	A1_2
U.S.	US/Canad
us	US/Canad
America	US/Canad
Anerican	US/Canad
U.S.	US/Canad
America	US/Canad
America	US/Canad
American (United States of America	US/Canad
American (United States)	US/Canad
American (U.S.A.)	US/Canad
us	US/Canad
us	US/Canad

Figure 1.2

- Moving to the D1 column, where empty cells are noted for specific Gamer_IDs, a suggestion is made to fill these gaps with values representing more than 3 years.

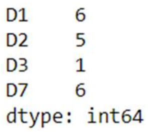
- In the D2 column, the proposal involves creating a new category labeled 'other' and subsequently splitting each response to assign binary values.

- Similarly, in the D3 column, where empty cells are observed, the recommendation is to refill them with the label "everyday," which will provide a standardized representation of the frequency of a particular behavior.

- In the D4 column, where certain Gamer_IDs report unrealistic playing times, a suggestion is made to convert these entries to a categorical format, such as 3-4 hours.

- The D5 column warrants attention to streamline responses by combining "no" and "not yet played," while extreme values are suggested to be transformed into the label "everyday" for consistency.

- The dataset in total has 18 null values out of which the columns “D1, D2, D3, D7” are the ones that consist of null values.(see figure 1.3)



```
D1    6
D2    5
D3    1
D7    6
dtype: int64
```

Figure 1.3

- Below figure displays white spaces which indicate the missing values on random when displayed in a matrix format. (see figure 1.4)

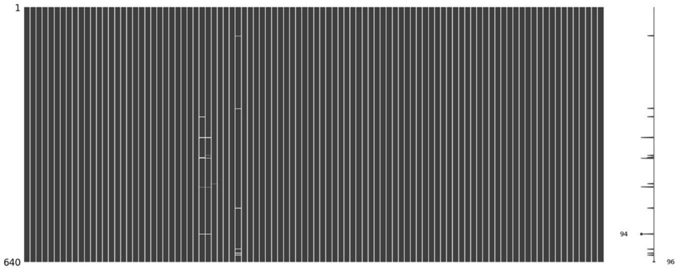


Figure 1.4

4. Detailed analysis:

a) Exploratory Data Analysis:

Age Distribution of Players

The analysis focuses on understanding the age distribution of players in the dataset which involves categorizing player ages into specific ranges and presenting the frequency distribution within each range. The age distribution is segmented into nine ranges: [10, 15), (15, 20], (20, 25], (25, 30], (30, 35], (35, 40], (40, 45], (45, 50], and (50, 55]. Each range represents a group of ages, and the frequency column indicates how many players fall within each respective group. (see figure 1.5)

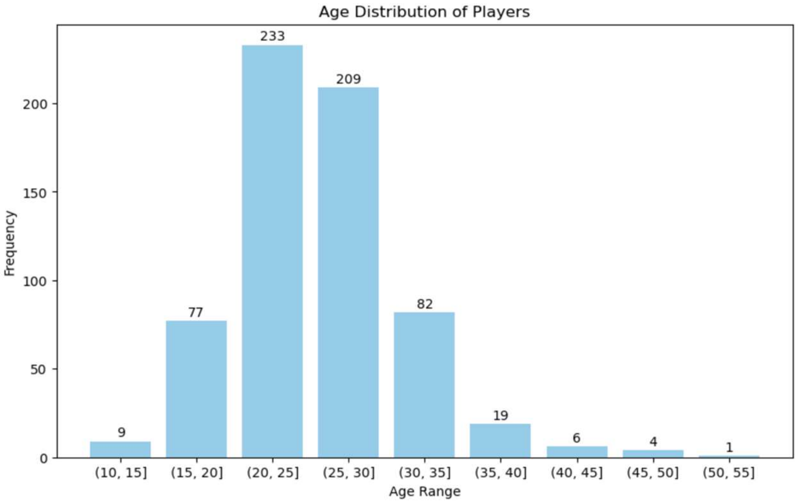


Figure 1.5

Key Findings:

- The majority of players fall within the age range of 20 to 25, with a frequency of 233, indicating a concentration of players in their early to mid-20s.
- The second-largest group consists of players aged 25 to 30, with a frequency of 209.
- The age groups beyond age 30 see a decline in frequency, with (30, 35] having 82 players, (35, 40] with 19 players, (40, 45] with 6 players, (45, 50] with 4 players, and (50, 55] with 1 player. The decline in frequencies for older age groups may be an indication of a decline in player engagement among older players.
- There is a limited representation of players beyond the age of 50, with only one individual recorded in the (50, 55] range.

The relationship between the biological sex as defined by the dataset and the players' environmental perception.

The focus is on the understanding of how the biological sex of players can have an impact on their attitudes towards environmental issues and utilizes a set of survey statements related to environmental perception. (see figure 1.6)

Key Findings:

- Female players, on average, show a higher level of agreement with the idea of getting closer to the population limit. (C1:We are approaching the limit of the number of people the earth can support)
- Male players in comparisons show greater acceptance of humans modifying the natural environment to suit themselves. [Female Players Mean**:** 2.78 and Male Players Mean**:** 3.44] (C2).
- Female players show a high level of environmental awareness, strongly believing in the equal rights of plants and animals. (C7: Plants and animals have as much right as humans to exist)
- Female players are slightly more concerned about the possibility of a major ecological disaster if current trends continue, having a mean value of 4.43 compared to Male player with a mean of 4.17 (C15)
- Female players, on average, disagree with the idea that humans should have dominion over nature (C12), while male players show a higher level of agreement with the statement.

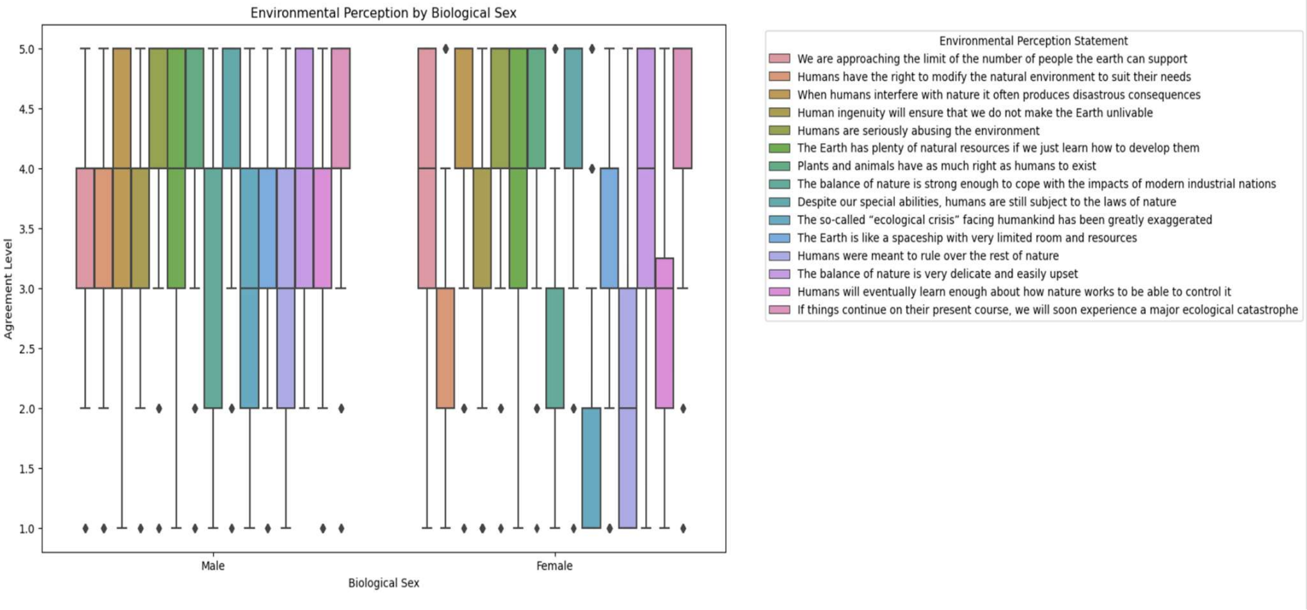


Figure 1.6

- Male players show a relatively higher level of agreement compared to female players with the statement, suggesting that humans will eventually learn enough about how nature works to be able to control it. Female players are more skeptical on the statement (C14)

Comparison of Cutting Down Trees Frequency by Gender

This analysis explores the comparison in the frequency of cutting down trees in the virtual world of Animal Crossing: New Horizons between female and male players. And the frequencies are categorized into different levels (1, 2, 3, 4) (see figure 1.7)

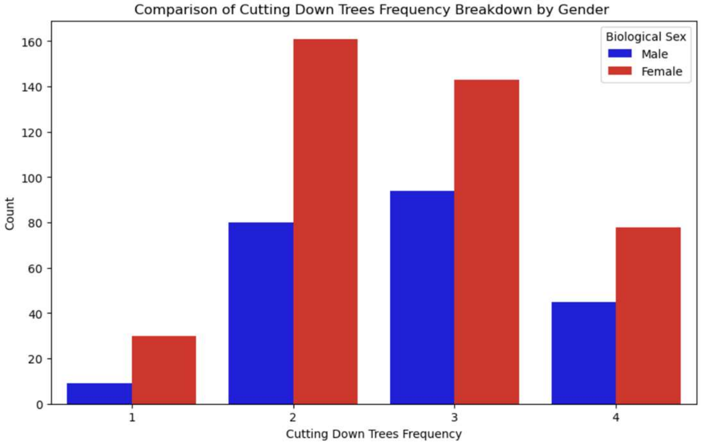


Figure 1.7

Key Findings:

- Across all frequency levels, female players consistently show a higher engagement of cutting down trees. Male players, on the other hand, generally show a lower frequency of involvement in each of the categories. This suggests that female players are more involved in the in-game behavior of cutting down trees compared to male players.
- The majority of players, regardless of gender, cut trees more often, with levels 2 and 3 being the most common.
- The most significant difference in response between male and female players are from the frequency level 2, with female players having a significantly higher count.

b) Identify the most important socio-demographic variables to indicate the environmental perception of the players.

The process in identifying the most important socio-demographic variables involves calculating a correlation matrix, employed a correlation analysis, specifically using Spearman's rank correlation coefficient. This matrix is then visualised using a heatmap. Spearman's rank correlation was chosen due to the presence of ordinal or categorical data. This non-parametric method assesses monotonic relationships, making it suitable for variables that may not have a linear association(see figure 1.8)

From the correlation coefficients it is possible to identify which socio-demographic variables may potentially be an indicator of environmental perception. Spearman's correlation assesses monotonous relationships, which are more general than linear ones. This flexibility is crucial when delering with socio-demographic factors that may influence environmental percerce in different ways.It's robust to outliers and doesn't assume a linear relationship. This makes it appropriate for assessing associations between categorical and numerical variables, as well as between categorical variables.

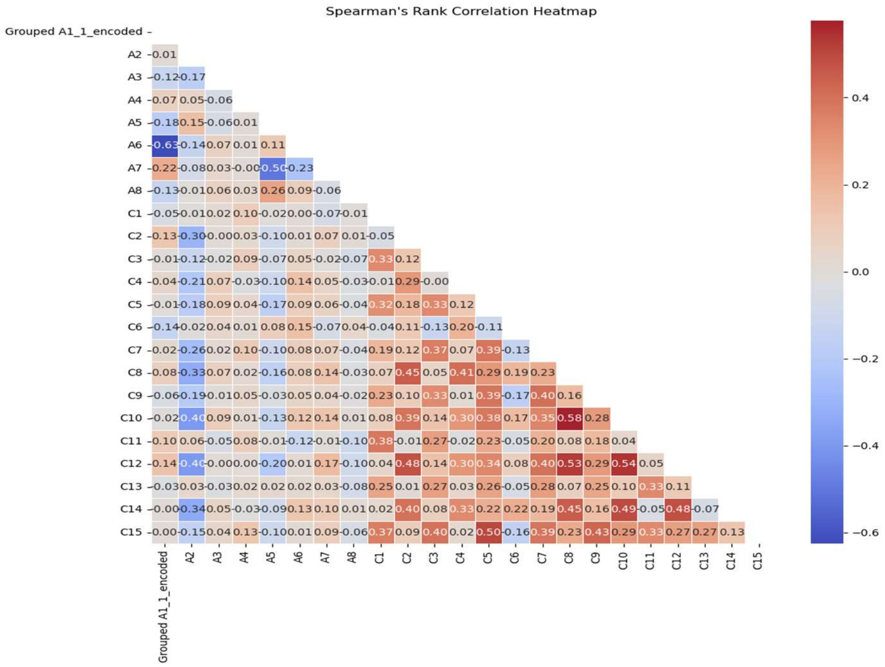


Figure 1.8

- There exists a weak negative correlation between A2 (Biological Sex) and Environmental Perception, indicated by the coefficient of 0.300178 with a p-value of 8.576602e-15.
- A4 to the question do you have a pet or a garden at home,there are positive correlations with several variables, including C1, C3, C4, C5, C6, C7, C9, C11 and C13.
- A5 (Age) vs Environmental Perception:A weak negative correlation with C5 (Humans are seriously abusing the environment).
- A8 (Employment Status) vs. Environmental Perception:Weak positive correlation with C2 (Humans have the right to modify the natural environment): 0.008281 (p-value: 8.343872e-01)Weak negative correlation with C6 (The Earth has plente of natural resources if we just learn how to develop them): -0.033234 (p-value: 4.012729e-01)

c) Modeling:

Predicting the players' environmental perception (C1–C15) using sociodemographic data (A1_1–A8) as input variables is the main goal of the modeling. For modeling purposes, it will be best to derive a single weighted score from the approximately fifteen dependent variables. To ensure that all the variables have the same measuring scale, the values of C2, C4, C6, C8, C10, C12, and C14 that are measured in reverse order are deducted from 6. The mean value of the fifteen dependent variables is then used to produce the weighted score, which is rounded to an integer number. The table below provides the setting of the weighted score.(see table 1.1)

Weighted Score	**Environmental Perception**
1	Very poor perception
2	Poor Perception
3	Good Perception
4	Very Good Perception

Table 1.1

The weighted score is now taken to be a categorical dependent variable, and the sociodemographic variables are normalized using the minmax scaling function after being label encoded to numerical values. The dataset is split into a train set (70%), a test set (15%) and then an evaluation set (15%).

Justification for the model :

The RandomForest Classifier considers many decision trees which each contain a specific instance of the data. The random forest technique goes through each of the instances, taking the one with the majority of votes as the selected prediction.

This categorization technique is highly adaptable, scalable, and inherently less susceptible to bias and outliers. As a result, it is both the most accurate and incredibly efficient in adapting to changes in the future. Due to these reasons, we have chosen to employ the RandomForest Classifier for this dataset.

Training metrics :

The training set consists of 70 percent of our complete dataset which is a total of 448 elements out of the total 640. The RandomForest Classifier is given a number of estimators of 1400, a maximum depth of tree of 500 and no limit on the maximum number of features in a particular branch or leaf of the tree.

With these specifications, the model takes less than a minute to compile. Increasing the estimators or the depth above these values yields no difference to the overall accuracy of the model and hence we finalized these metric values.

Testing performance :

The testing set is 15 percent of our total dataset which amounts to 96 elements. With these test values, the model is observed to classify the socio demographic variables into 4 sets of environmental perception from levels 1 to 4. 1 being poor (no regard for the environment) perception and 4 being excellent (a top tier conservationist and friend of the environment) perception of the environment.

The model gives an overall accuracy of 74 percent. The f1 score of 2's being 0 percent, 3's being 84 percent and that of 4's being 47 percent. The f1 score of 1's are unknown as there are no 1's in our test dataset.

The confusion matrix given below (see figure 1.9) gives a clearer visualization of the above paragraphs :

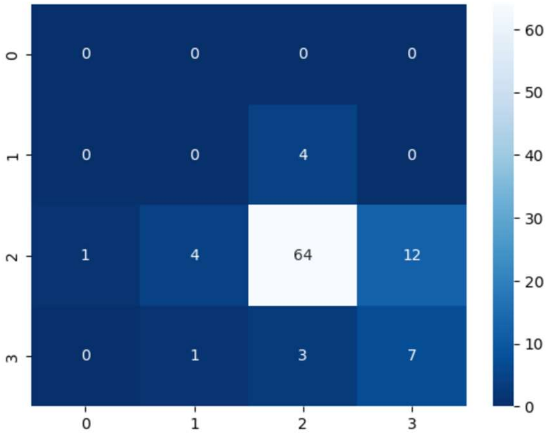
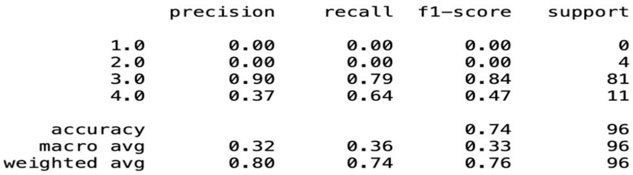


Figure 1.9

In the confusion matrix, 0 represents the real score 1, 1 represents the score 2, 2 represents the score 3, and 3 represents the score 4. This is because of the separate label encoding done by the confusion matrix.

Overall classification report :

The performance metrics is calculated and the classification report is provided below (see figure 1.10):

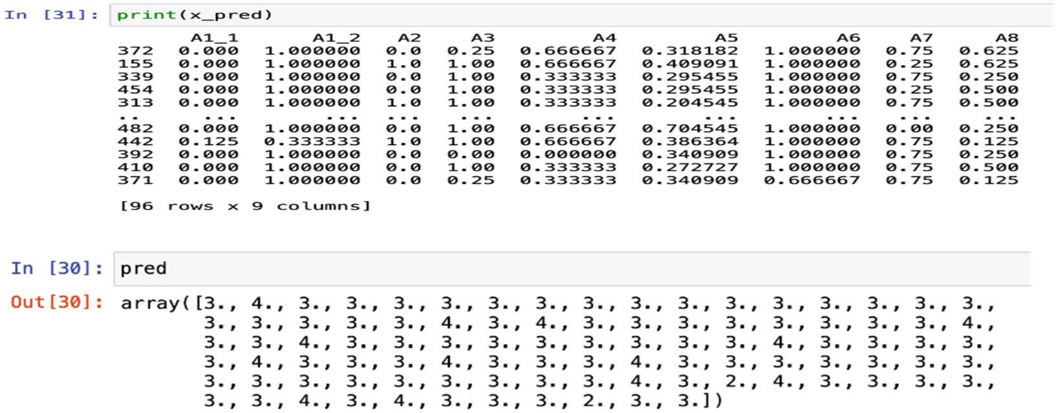


```
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00         0
         2.0       0.00      0.00      0.00         4
         3.0       0.90      0.79      0.84        81
         4.0       0.37      0.64      0.47        11

    accuracy                           0.74        96
   macro avg       0.32      0.36      0.33        96
weighted avg       0.80      0.74      0.76        96
```

Figure 1.10

Evaluation:

The model is now used to predict the output for the evaluation set. The input variables and the predicted output are given below (see figure 1.11) :



```
In [31]: print(x_pred)

           A1_1      A1_2   A2    A3        A4        A5        A6    A7     A8
     372  0.000  1.000000  0.0  0.25  0.666667  0.318182  1.000000  0.75  0.625
     155  0.000  1.000000  1.0  1.00  0.666667  0.409091  1.000000  0.25  0.625
     339  0.000  1.000000  0.0  1.00  0.333333  0.295455  1.000000  0.75  0.250
     454  0.000  1.000000  0.0  1.00  0.333333  0.295455  1.000000  0.25  0.500
     313  0.000  1.000000  1.0  1.00  0.333333  0.204545  1.000000  0.75  0.500
     ..     ...       ...  ...   ...       ...       ...       ...   ...    ...
     482  0.000  1.000000  0.0  1.00  0.666667  0.704545  1.000000  0.00  0.250
     442  0.125  0.333333  1.0  1.00  0.666667  0.386364  1.000000  0.75  0.125
     392  0.000  1.000000  0.0  0.00  0.000000  0.340909  1.000000  0.75  0.250
     410  0.000  1.000000  0.0  1.00  0.333333  0.272727  1.000000  0.75  0.500
     371  0.000  1.000000  0.0  0.25  0.333333  0.340909  0.666667  0.75  0.125

     [96 rows x 9 columns]

In [30]: pred

Out[30]: array([3., 4., 3., 3., 3., 3., 3., 3., 3., 3., 3., 3., 3., 3., 3., 3., 3.,
                3., 3., 3., 3., 3., 4., 3., 4., 3., 3., 3., 3., 3., 3., 3., 3., 4.,
                3., 3., 4., 3., 3., 3., 3., 3., 3., 3., 3., 3., 4., 3., 3., 3., 3.,
                3., 4., 3., 3., 3., 4., 3., 3., 3., 4., 3., 3., 3., 3., 3., 3., 3.,
                3., 3., 3., 3., 3., 3., 3., 3., 3., 4., 3., 2., 4., 3., 3., 3., 3.,
                3., 3., 4., 3., 4., 3., 3., 3., 2., 3., 3.])
```

Figure 1.11

5. Conclusions :

All things considered, the dataset has been cleaned, exploratory analysis has been completed, and the model to forecast environmental perception has been constructed effectively. The fact that about 19% of gamers don't even want to damage the environment when playing video games is really noteworthy. Furthermore, since their behavior in video games may differ from their real-life behavior, the remaining players' actions don't always indicate that they wish to destroy the ecosystem. Nonetheless, the biosphere will always be threatened by the dominant nature of humans.