**COMP5122M DATA SCIENCE (33862) - COURSEWORK 1**

**REPORT ON A MULTINATIONAL DATASET OF GAME PLAYERS’ BEHAVIORS IN A VIRTUAL WORLD AND ENVIRONMENTAL PERCEPTIONS:**

**GROUP MEMBERS:**

1. Saadhana Ganesa Narasimhan [201703255]
2. Yaser Selvam [201751039]
3. Yuvraj Singh [201799843]
4. Shrichand Bhuria [201800797]

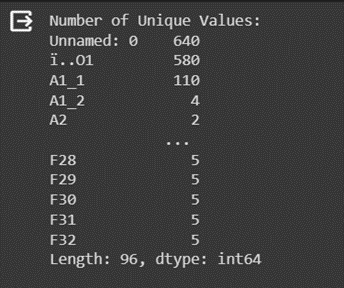
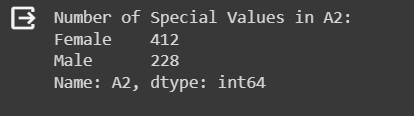
# **INTRODUCTION:**

We aim to analyze the human behavior in the virtual world and in the environment derived from their behaviors in the game. Our main objective is to analyze based on factors like utilitarian, mutualistic, pluralistic, and distanced. The dataset can be downloaded from ([https://www.scidb.cn/en/detail?dataSetId=cb5d36cce29f4e5695a586c9b85d04b6#p2](https://www.scidb.cn/en/detail?dataSetId=cb5d36cce29f4e5695a586c9b85d04b6%23p2))

The dataset contains 96 columns and 640 entries. The aim of this group case study is to do an Exploratory data analysis (EDA) for the age distribution of the players and the relationship between the biological sex which is defined by the dataset and the players environmental perceptions, also a comparison of male and female's players in-game behavior frequency. We also aim to identify and develop a classification model that the most important socio-demographic variables to indicate the environmental perception of the players.

**DATA QUALITY:**

* The primary investigation was to clean up the data information. Upon checking both missing values and null values, there were no such values present in the dataset.
* The shape of the player’s dataset was 96 columns and 640 rows. And most of the column’s data type considered between Integer64 and Object20.
* Upon checking the null and missing values, our second investigation was to check the special and unique values and was able to conclude that there were no special values (e.g., N/A or UNKOWN) but found some exceptional value counts that occurred in a specific value and unique values present in the dataset as well.

   *Figure 1: Unique Values*  *Figure 2: Special Values*

* We decided to check and identify for duplicate values, there were no such values in the dataset.
* As our third investigation, we checked for valid values by defining the MALE and FEMALE as categorical column are valid or not.
* On top of it, we did our investigation on outliers, valid and invalid range, inconsistent and consistent format, and values for the specific column value.
* Additionally, we considered investigating summary statistics, granularity and coverage issues, and extreme values.

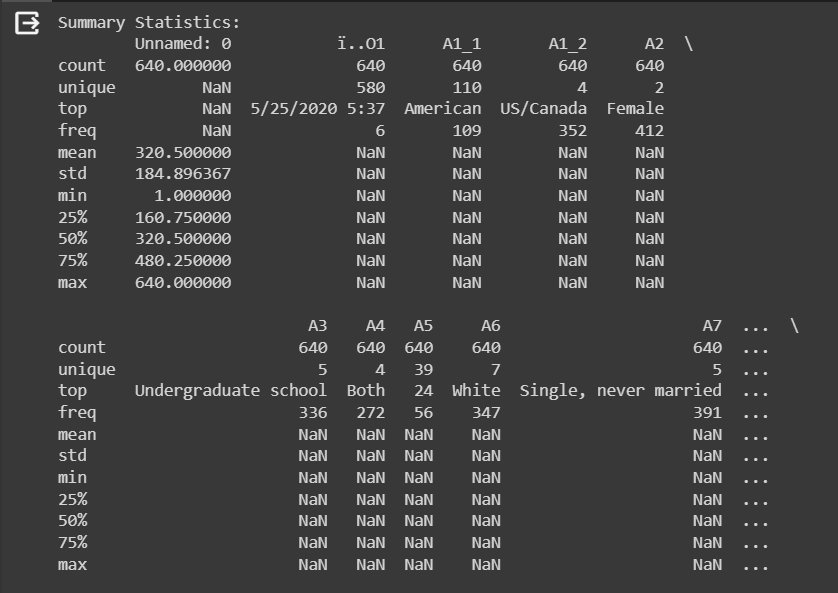


Figure 3: Summary Statistics

**DATA CHARACTERISATION:**

File Details:

|  |  |  |  |
| --- | --- | --- | --- |
| File Name | File Size | Number of Columns | Number of Rows |
| data\_640\_validated | 283 KB | 96 | 640 |

**DETAILED ANALSIS:**

1. **Do exploratory data analysis to show:**
   1. **Age distribution of the players**

**Result for the Age Distribution of the Players:**

The analysis of the age distribution among players reveals intriguing patterns. As illustrated in Figure 1, a predominant portion of players falls within the adult age range, specifically between 21 and 30 years old, as indicated by the elevated frequency within this age group. This observation implies that Animal Crossing: New Horizons (ACNH) holds substantial appeal among young adults.

On the other hand, there is a noticeable decline in the number of players below the age of 15 and those above the age of 40. This observation may indicate that the game is less popular among incredibly young players and older individuals. The physical meaning derived from this dataset implies that the game's design, activities, or themes might be more aligned with the preferences and interests of the young adult demographic, potentially influencing the age distribution of its player base. Further investigation into game features and marketing strategies could provide additional insights into this age-related trend.

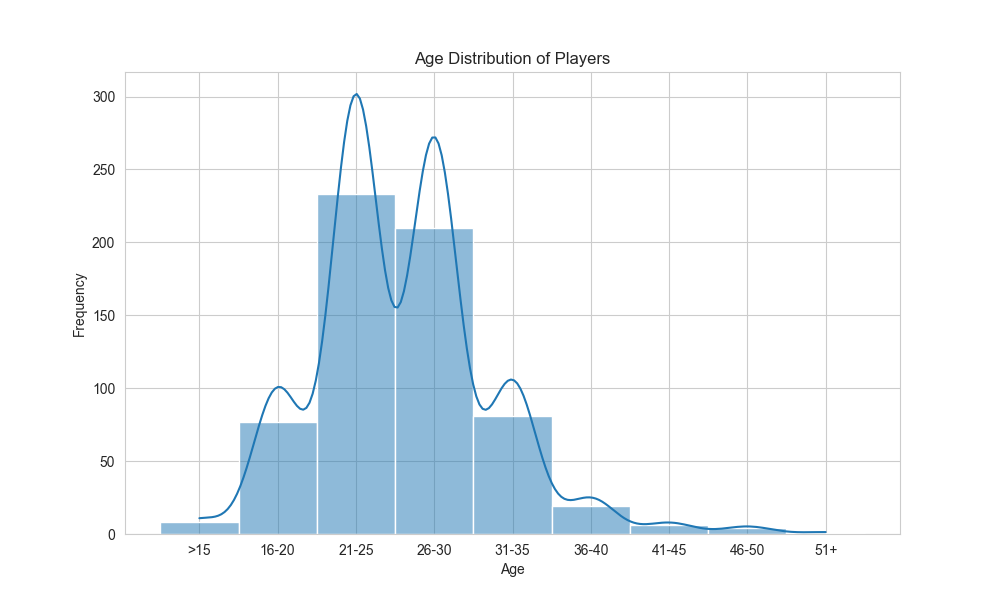


Figure 4: Age Distribution of Players

* 1. **The relationship between the biological sex as defined by the dataset and the players’ environmental perception:**

Analysis of the Relationship between Biological Sex and Environmental Perception:

**Notable Trends:**

Several trends emerged from the analysis, shedding light on the relationship between biological sex and environmental perception:

1. Average Ratings:
   * Female players, on average, tended to rate most environments slightly higher compared to male players.
   * The most significant differences were observed in environments C10, C12, and C14, where males rated their experience lower by approximately 1 point on average.

**Commonalities and Differences:**

1. Commonly High-Rated Environments:
   * Both males and females consistently rated certain environments very highly. Notable examples include C15, C7, and C5, suggesting universal positive perceptions.
2. Variability in Ratings:
   * Environments such as C10, C12, and C8 exhibited more variability in ratings between individuals, indicating diverse personal experiences.
3. No Major Disparities:
   * While there were some differences in average ratings, no major disparities emerged between the environmental perceptions of male and female players.

**Variability in Ratings:**

1. Greater Variability in Specific Environments:
   * Environments C10, C12, and C8 displayed larger ranges and standard deviations for both sexes, indicating greater diversity in how players experienced these environments.

In conclusion, while there were subtle differences in the average ratings of certain environments based on biological sex, there were also significant commonalities. The positive universal perceptions of specific environments suggest that the game's design or content resonates positively across genders. The larger variability in ratings for specific environments emphasizes the individualized and diverse nature of players' experiences. Overall, biological factors alone may not be strong determinants of one's environmental belief within the game, highlighting the importance of individual preferences and individual experiences.

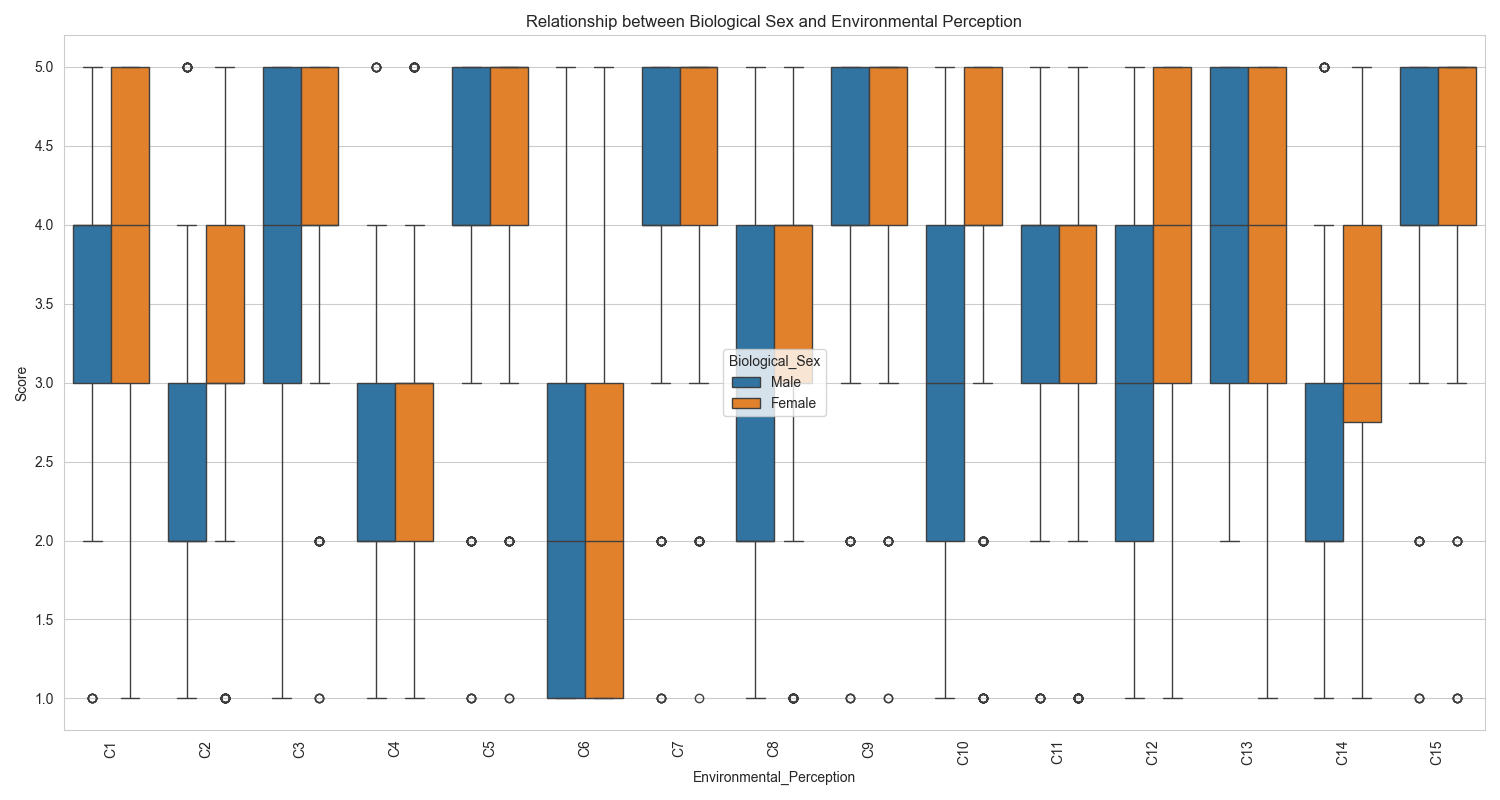


Figure 5: Relationship between Biological Sex & Environmental Perception

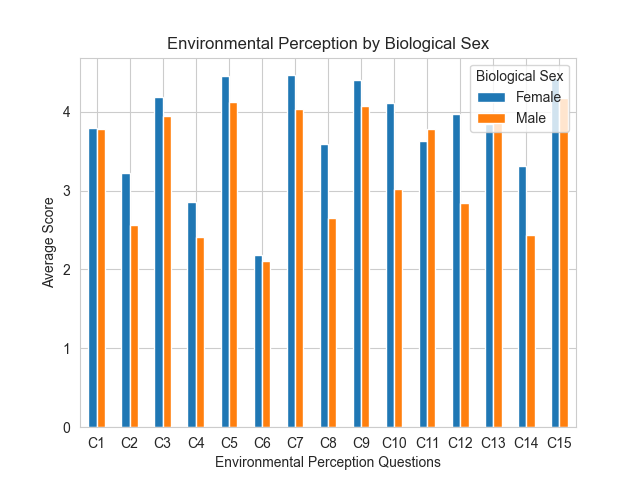


Figure 6: Environmental Perception by Biological Sex

* 1. **A comparison of the frequency of the male and female players’ in-game behavior: cutting down the tree.**

The plot writes down that females tend to engage more often in the activity of cutting down trees compared to males. Specifically, the counts for females in the "Seldom" and "Sometimes" categories are higher than their male counterparts.

Interestingly, a common trend is seen among both male and female players—they cut down trees "Seldom" or "Sometimes." This suggests a shared behavior of infrequent tree-cutting among players, irrespective of gender. The dataset implies that players, regardless of their biological sex, show a preference for less frequent engagement in the activity of cutting down trees.

This shared pattern may be indicative of a cautious approach to altering the in-game environment, emphasizing a balanced and sustainable approach to tree-cutting activities. The data reflects a general tendency among players to avoid excessive or unnecessary tree removal, aligning with an environmentally conscious approach within the virtual world of the game.

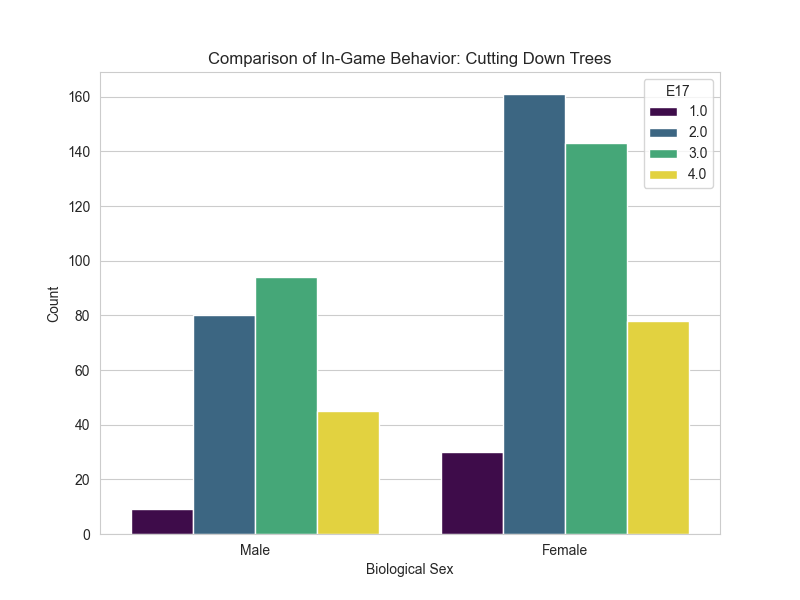


Figure 7: Comparison of In-game Behavior: Cutting Down Trees

1. **Identify the most important socio-demographic variables to indicate the environmental** **perception of the players.**

In the analysis using a Random Forest model to predict environmental belief variables based on socio-demographic variables, the feature importance plot highlighted 'Age (A5)' as the most crucial socio-demographic variable. The plot illustrated that 'Age' consistently had the highest importance scores across multiple environmental perception variables (C1 to C15). This suggests that 'Age' plays a significant role in predicting various aspects of players' environmental perceptions in the context of the given dataset.

**Interpretation:**

* The feature important scores say the relative contribution of each socio-demographic variable to the prediction of environmental perception.
* 'Age (A5)' appeared as the most influential variable, implying that a player's age significantly influences their environmental perceptions in the game.

**Methodology**:

* Used a Random Forest model as it can handle non-linear relationships and interactions among variables effectively.
* Employed a Multioutput Classifier to predict multiple environmental perception variables simultaneously.
* Feature importance was extracted from the trained model to identify the key socio-demographic variables.

**Physical Meaning:**

* The dominance of 'Age' in importance suggests that players' age is a critical determinant of how they perceive and interact with the in-game environment.
* Older players might approach the game differently, potentially valuing certain in-game aspects or activities more than younger players.

**Motivation for Method:**

* Random Forest is robust and capable of handling complex relationships, supplying insights into variable importance.
* Multioutput Classifier allowed simultaneous prediction of multiple environmental perception variables, offering a holistic view.

This analysis provides valuable insights into the socio-demographic factors influencing players' environmental perceptions in the game, with 'Age' identified as the most significant variable.

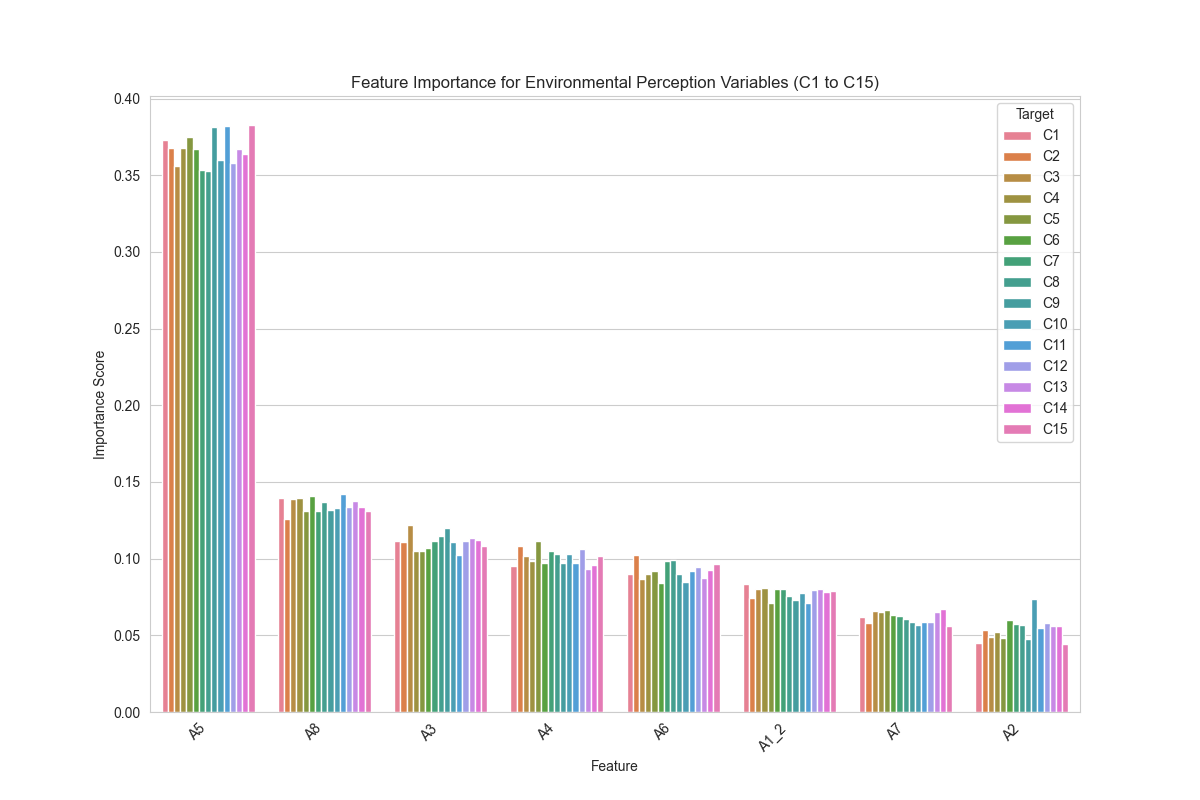


Figure 8: Feature Importance for Environmental Perception Variables

1. **Develop a classification model that can be used to predict a player’s environmental perception based on socio-demographic variables only.**

[To do this, you will need to split the dataset into training (%70), test (%15), and evaluation sets (%15). Evaluate your classification model’s performance using proper metrics. Please describe the method you used for classification, and your motivation to choose the method.]

Classification Model Development for Environmental Perception Prediction

**Data Preparation:**

1. Load the Dataset:
   1. Load the dataset and select socio-demographic variables (prefixed with 'A') as features.
   2. Select environmental perception variables (prefixed with 'C') as the target
2. Clean the Data:
   1. Handle missing values, encode categorical variables, etc.
3. Split the Dataset:
   1. Split the dataset into training (70%), test (15%), and evaluation (15%) sets.

**Model Training:**

1. Use RandomForestClassifier as the base estimator:
   1. Wrap it with MultiOutputClassifier to handle multiple target variables.
   2. Train the model on the training set.

**Model Evaluation:**

1. Predict on the test set:
2. Calculate and report the accuracy and other relevant metrics.
3. Use the evaluation set:
4. Further assess the model's performance.

**Motivation:**

1. RandomForest is chosen for its robustness and ability to handle non-linear relationships.
2. It performs well with a mix of categorical and numerical data.
3. MultiOutputClassifier is used to extend RandomForest for multi-target classification.

**Dataset Split Information:**

The dataset has been split into training, test, and evaluation sets with the following number of samples:

1. Training set: 448 samples
2. Test set: 96 samples
3. Evaluation set: 96 samples

**Task Execution Plan:**

To continue with the task using RandomForestClassifier with MultiOutputClassifier:

1. Convert the average of the environmental perception variables into discrete class labels:
2. Define a threshold to categorize the average.
3. Train the RandomForestClassifier wrapped in a MultiOutputClassifier:
4. Use the training set for model training.
5. Evaluate the model's performance
6. Assess the model on the test and evaluation sets.
7. Use accuracy as the metric.

**Accuracy on the test set:**

Accuracy: 0.71875

The `RandomForestClassifier` with `MultiOutputClassifier`, after converting the average to discrete class labels using a median threshold, achieved an accuracy of approximately 71.88% on the test set. This shows a significant improvement in the model's ability to predict the class labels for the environmental perception based on socio-demographic variables.

**CONCLUSION:**

In conclusion, our analysis of human behavior in the virtual world, particularly within the context of Animal Crossing: New Horizons (ACNH), has unveiled several key insights. The age distribution among players shows a substantial appeal of the game among young adults, with a majority falling between 21 and 30 years old. The examination of the relationship between biological sex and environmental perception revealed nuanced trends, suggesting that while there are subtle differences in average ratings, commonalities prevail across genders.

The frequency of in-game behavior, such as cutting down trees, highlighted a shared tendency among players, irrespective of gender, to engage less often in this activity, reflecting a cautious and environmentally conscious approach. Furthermore, our Random Forest model highlighted 'Age' as the most influential socio-demographic variable in predicting environmental perceptions.

This underscores the significant role age plays in shaping players' virtual experiences. The developed classification model achieved an accuracy of approximately 71.88% in predicting environmental perception based on socio-demographic variables, displaying the effectiveness of the chosen method. Overall, our comprehensive analysis supplies valuable insights into the interplay of socio-demographic factors and virtual environmental perceptions in ACNH.