HSE / Applied Statistics course

Homework 3: Statistical analysis of "Animal crossing" dataset from Kaggle

Student: Katerina Oleynikova

March 2, 2077

Abstract

This report contains information about the project work of the course "Applied Statics". The main aim of the project is to analyse a relevant dataset in terms of statistics. The dataset that was chosen is dedicated to the "Animal Crossing" analysis. According to the input rules, the data were retrieved from Kaggle (https://www.kaggle.com). The link to this dataset is highlighted in References.

1 Introduction

Animal Crossing is a video game on Nintendo which allows to create an island and grow a little village. Game is started with a random set of 3 animals-villagers, and we can find other animals-villagers to invite to our town by visiting mysterious islands. This project describes distribution of all the possible villager species, gender, personality, and even horoscope sign. Such statistical analysis is useful for Animal Crossing gamers, because it assist them in knowing better their own villagers.

Animals-villagers cohort is one of the most important aspects of the game: for instance, it is possible to have conversations with them, gamers are also able to visit their homes, even to give them gifts, etc. That is why, analysis of animals-villagers personalities/interests is needed to be considered (for example, to choose the best gifts for them).

2 Materials and Methods

Statistical analysis is the collection and interpretation of data in order to uncover the patterns and trends. It is a component of data analytics. Statistical analysis can be used in situations like gathering research interpretations, statistical modeling or designing surveys and studies.

Statistical analysis of the chosen dataset includes several aspects of the course "Applied Statistics" (by S.Spirin) and can be divided into the main sections:
1. Representation of inputs data (pie charts, bar graphs and scatter plot for categorical data are included for graphical demonstration);
2. Solution of Birthday Problem in the context of Animal Crossing villagers analysis (with the use of histogram visualization of simulated data;
3. Rarest cat Raymond analysis (hypothesis testing, test statistic calculation, P-value determination).

All the steps were implemented via Python language (using libraries: pandas, pandasql, plotly, numpy, matplotlib, math, itertools, random, scipy.stats).

3 Analysis and Results

3.1 Data Representation

Def. A pie chart, sometimes called a circle chart, is a way of summarizing a set of nominal data or displaying the different values of a given variable (e.g. percentage distribution). This type of chart is a circle divided into a series of segments. Each segment represents a particular category.

Pie chart shows percentage distribution of animal villagers according to their species (see Figure 1). It is shown the biggest part belongs to cat (5.88%) and rabbit (5.12%) species.

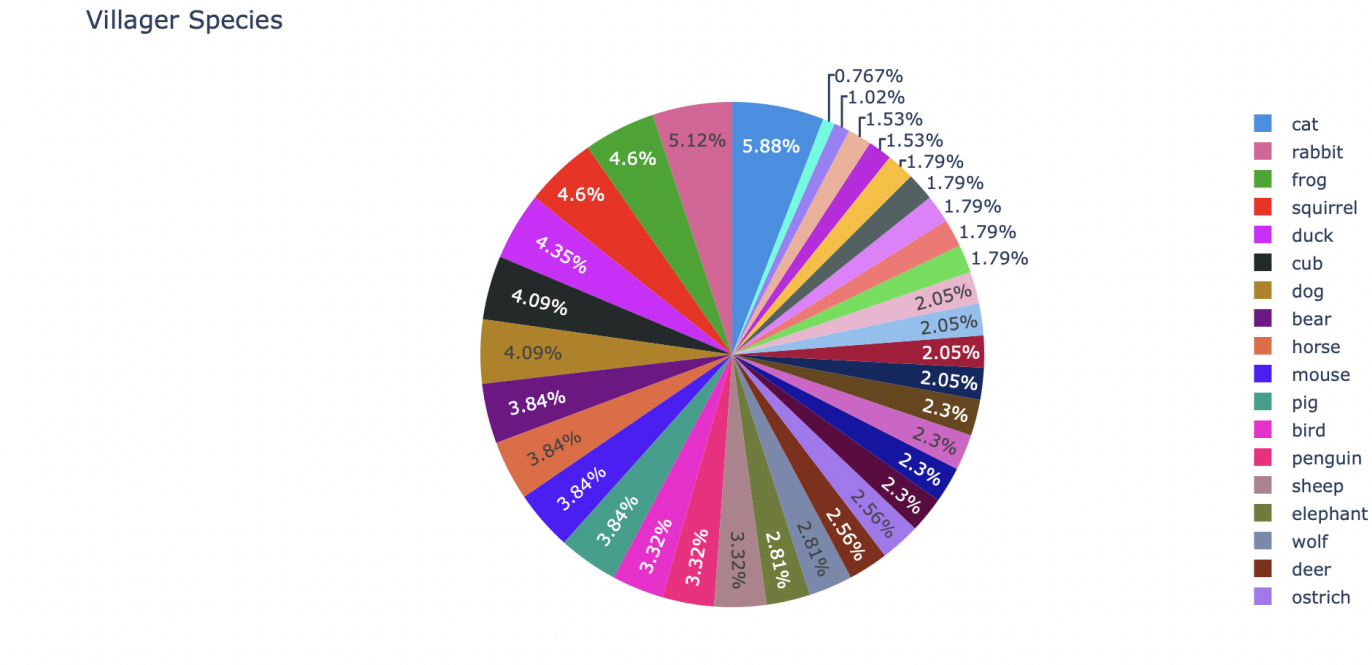

Figure 1: Pie chart of Animal Crossing species distribution.

Another way to represent these data is in a bar chart.

Def. A bar chart or bar graph is a chart or graph that represents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.

The horizontal bar chart that is represented below (see Figure 2) shows the distribution of animal species, too. The smallest part of all the data belongs to octopus (less than 5 types of species).

The pie chart and the bar graph according to gender analysis are shown bellow (see Figures 3-4). It is seen that the distribution of gender among Animal Crossing villagers is nearly equal (52.2% - male, 47.8% - female). It is interesting to note according to the bar graph that the 3 species have only one possible gender type (cow - female, bull - male, lion - male).

Another data representation is illustrated on Figures 5-6 in terms of personality types distribution (the biggest part of all the possible animals-villagers is characterised by lazy type of personality

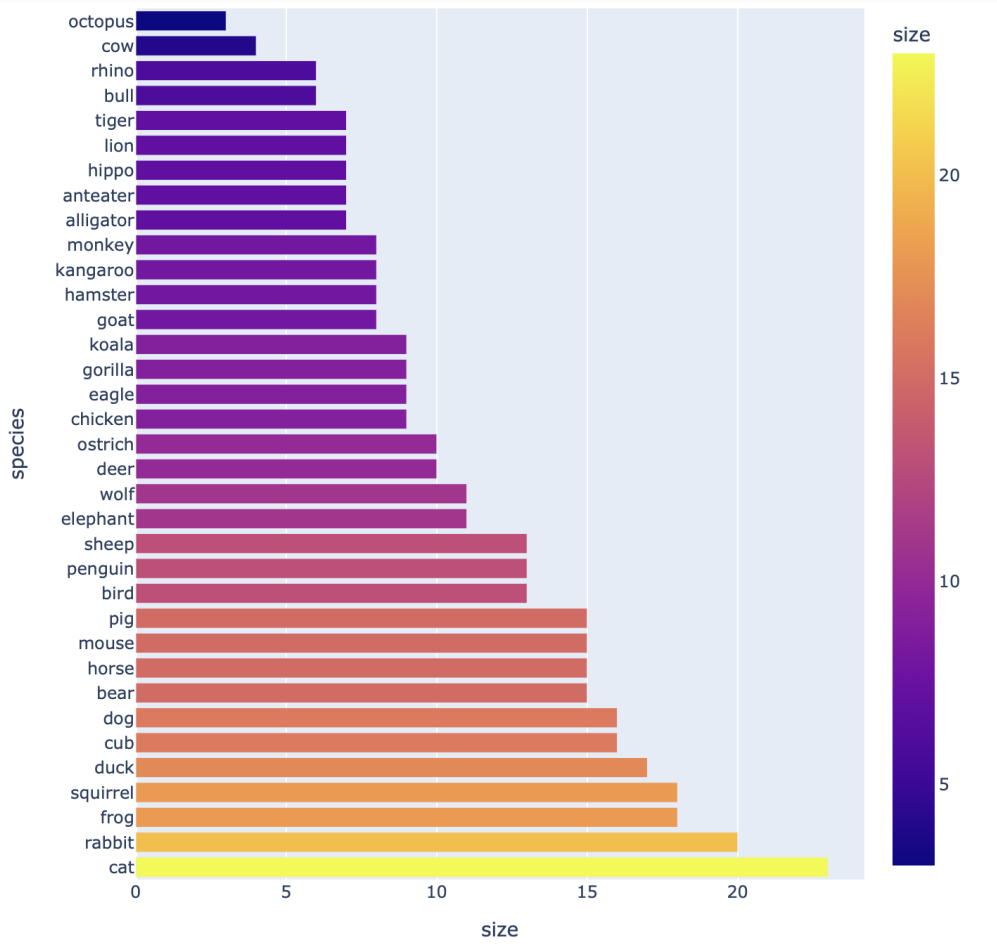

Figure 2: Bar chart distribution of Animal Crossing species.

(15.3%) or normal one (15.1%), while the minority of them has special the 'uchi' type (uchi is a female villager personality type in the Animal Crossing game; they are very caring and not too vain towards the villagers)).

Additionaly, new column (Horoscope sign) to existing dataframe was added according to date of birth of each animal-villager (as a result, the pie chart and the scatter plot for categorical data both on x-/y-axis were made (see Figure 7-8)).

Def. A categorical scatter plot (scatter chart, scatter graph) uses dots to represent values for two different categorical variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

It is shown on Figure 7 (bar graph) that the top 5 zodiac signs among animal-villagers are Leo, Libra, Gemini, Virgo, and Cancer (distributed from 9.72 to 8.7%). The rarest signs of zodiac are Capricorn and Pisces (7.42% and 7.67%, respectively). However, it is needed to highlight that all signs of horoscope are distributed nearly the same (7.42-9.72%).

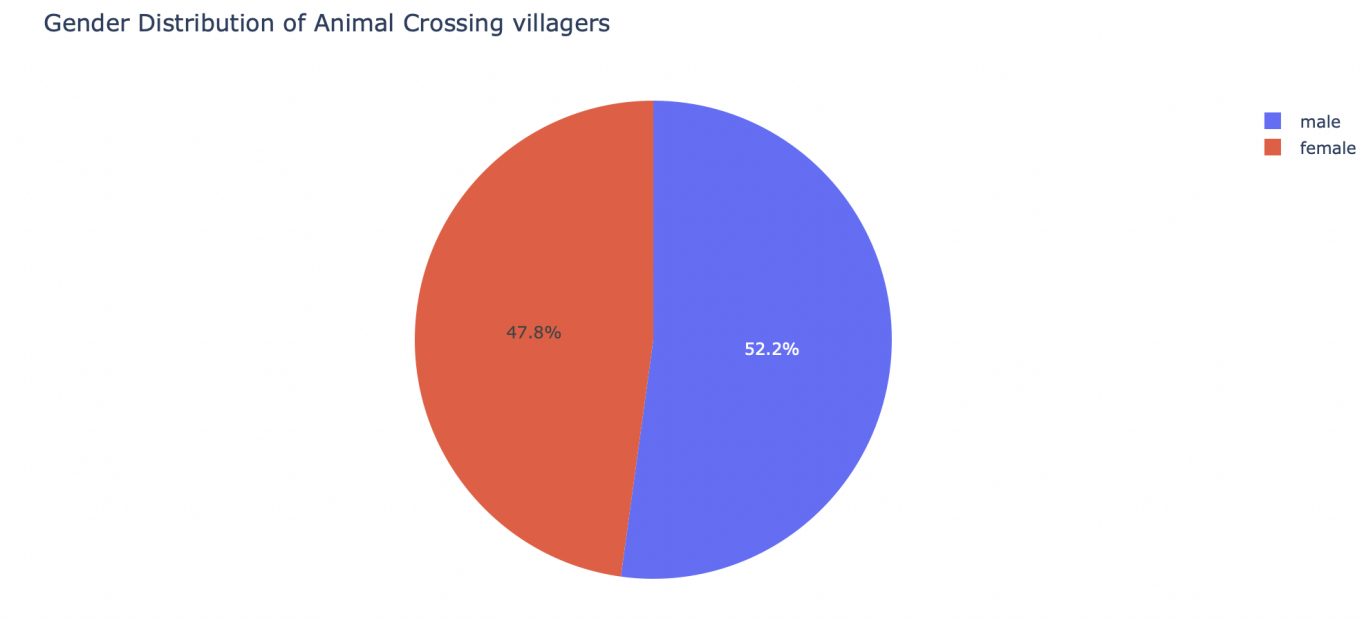

Figure 3: Pie chart gender distribution (among Animal Crossing villagers).

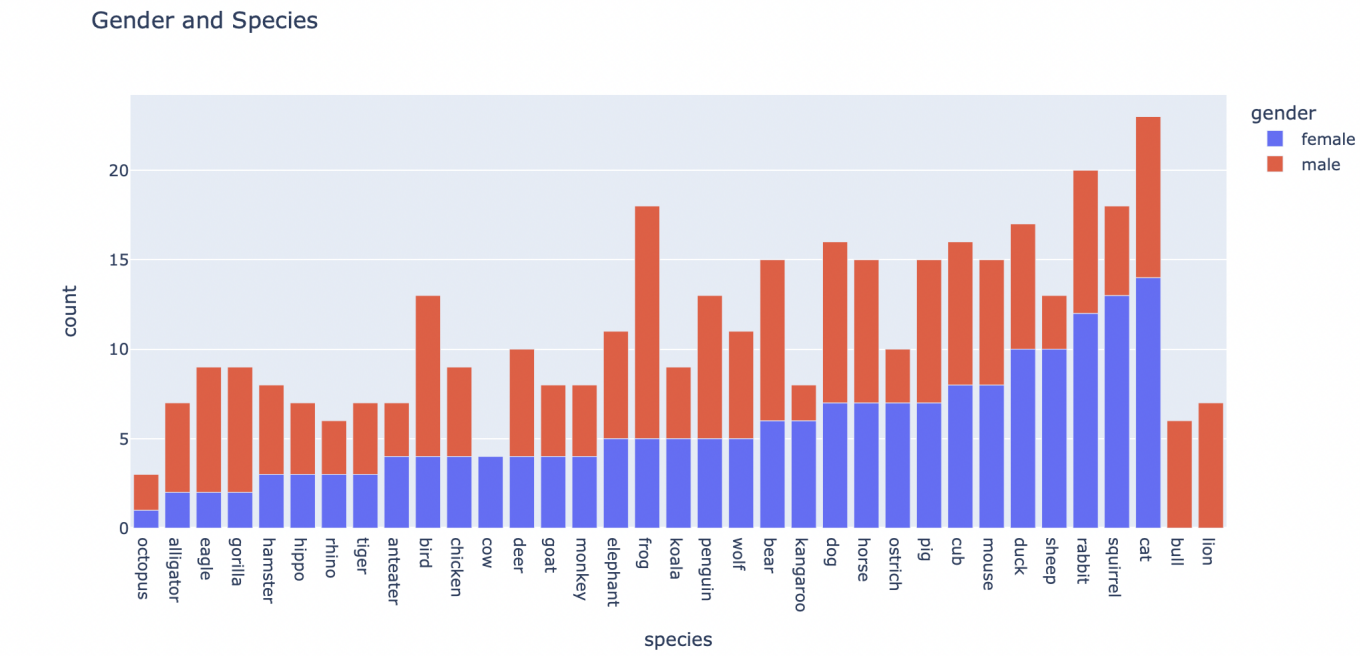

Figure 4: Bar graph of species and gender distribution in Animal Crossing.

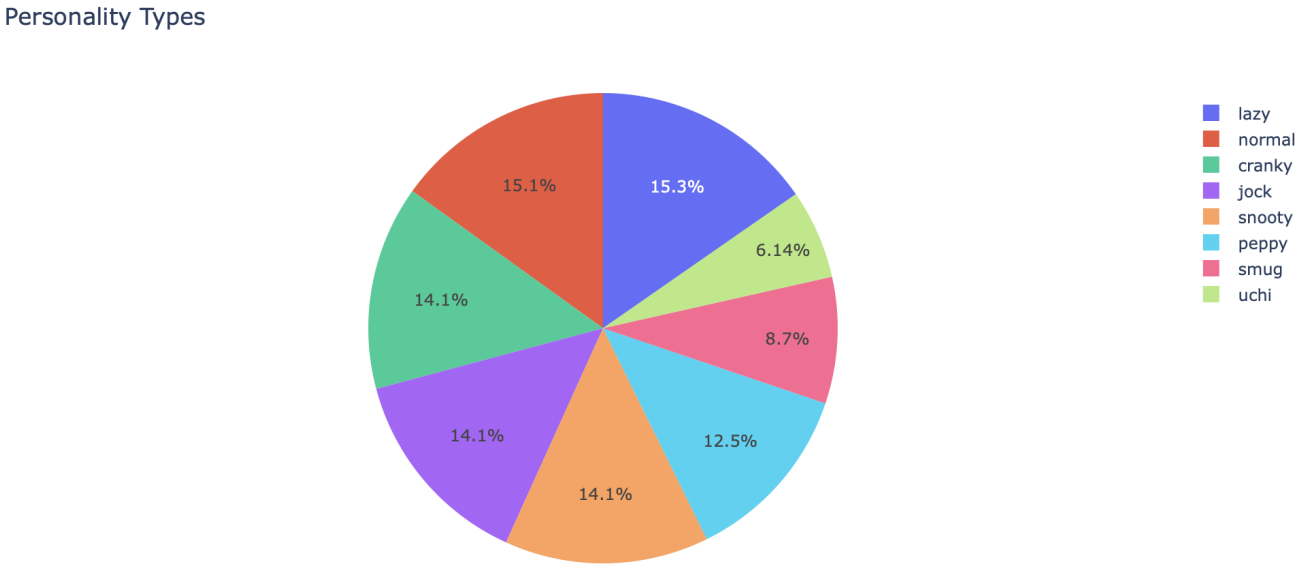

Figure 5: Pie chart personality distribution (among Animal Crossing villagers).

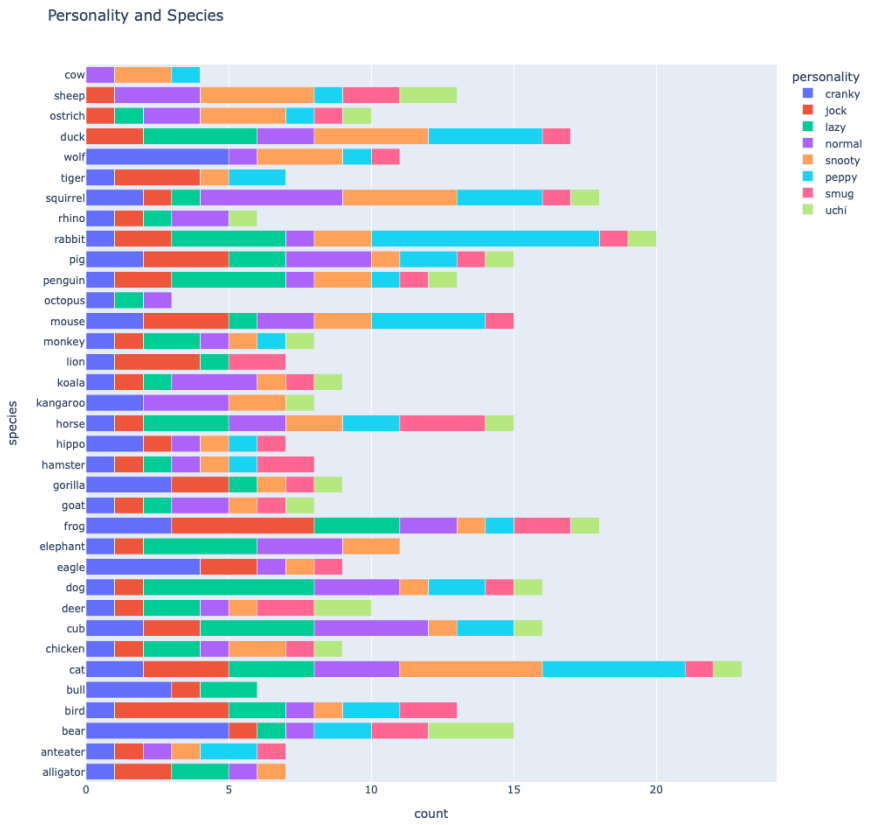

Figure 6: Bar chart personality distribution (among Animal Crossing villagers).

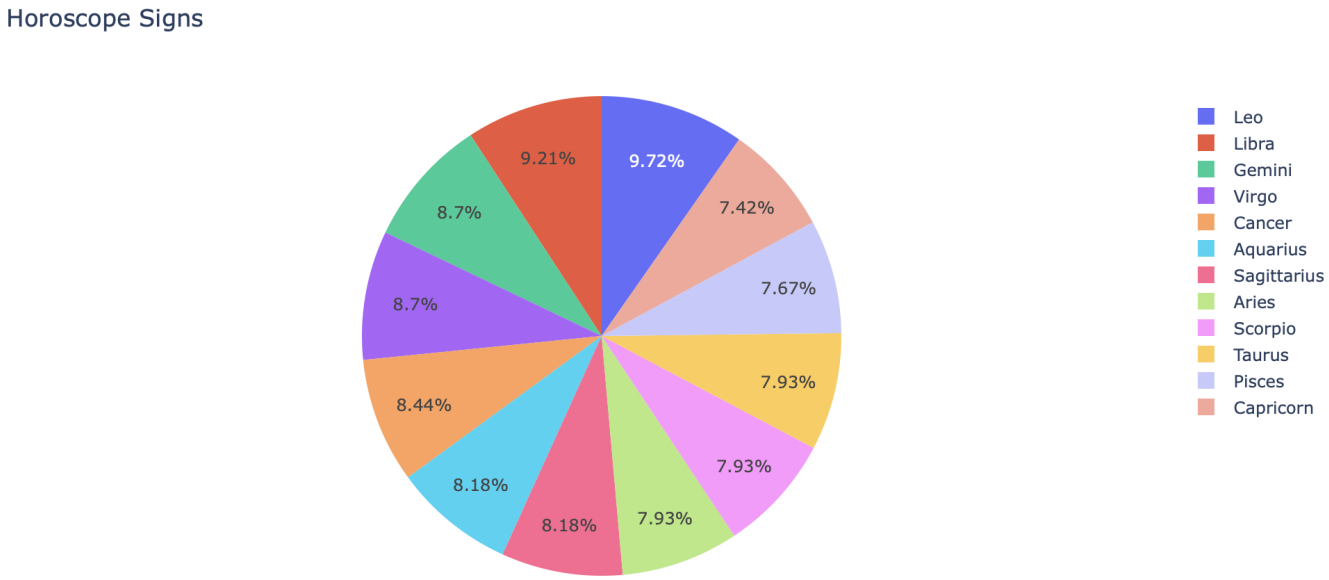

Figure 7: Bar graph horoscope sign distribution (among Animal Crossing villagers).

Figure 8: Categorical scatter plot representing the relation between horoscope sign, personality and species.

3.2 Birthday Problem

Def. In probability theory, the birthday problem asks for the probability that, in a set of n randomly chosen people, at least two will share a birthday. The birthday paradox is that, counter-intuitively, the probability of a shared birthday exceeds 50% in a group of only 23 people.

The birthday paradox is a veridical paradox: it appears wrong, but is in fact true. While it may seem surprising that only 23 individuals are required to reach a 50% probability of a shared birthday, this result is made more intuitive by considering that the comparisons of birthdays will be made between every possible pair of individuals. With 23 individuals, there are (23 × 22) / 2 = 253 pairs to consider, which is well over half the number of days in a year (182.5 or 183).

The trick to calculating it is to start by calculating the complement, - i.e. the probability that no one in the room shares the same birthday. Then, this probability is subtracted from 1:

$$1 - \frac{365!}{365^n(365-n)!}$$

Then, we can move to the Animal Crossing dataset, displaying information about the 391 possible villagers that might move to an island on the game. We have identified from the previous analysis step (3.1), this dataset includes each character's birthday, as a result, there is no problem to do such analysis. As there are 365 possible birthdays (or 366 on a leap year), and 391 villagers, solving the Birthday Problem is possible for the dataset.

According to analysis (implemented in Python) we have identified that 60 villagers share their birthdays (30 dates were found that appear more than once (to be exact, each of one appears two times in the dataset)).

In other words, there are 30 non-unique dates of birth, and 60 animals-villagers share a birthday with another animal-villager, meaning that no more than two villagers share a birthday.

Now, we have a reason to suspect that the birthday date allocation is not random in Animal Crossing (possibly, the game makers will have tried to avoid having multiple villagers on the same island with the same birthday; the fact that we never see more than two villagers with the same birthday definitely may corroborate this).

It does lead to the next questions: if the birthday dates of the villagers were set randomly and independently, how likely would we be to get just 60 villagers with a shared birthday? How many of the 391 villagers would we, on average, expect to see share a birthday?

3.2.1 Histogram visualisation of simulated data

To answer the questions that are marked above we can simulated data and see the result. Simulation of data has been implemented in Python. After that, we plotted the histogram (see Figure 9) showing the number of villagers sharing a birthday across all the generated samples.

The simulated data shows us that on average, we would expect roughly 257 villagers to share a birthday with another villager. Across 50,000 samples, we did not get a number as small as 60 even once; therefore, we may conclude that the real game date of birth allocation is very unlikely by chance.

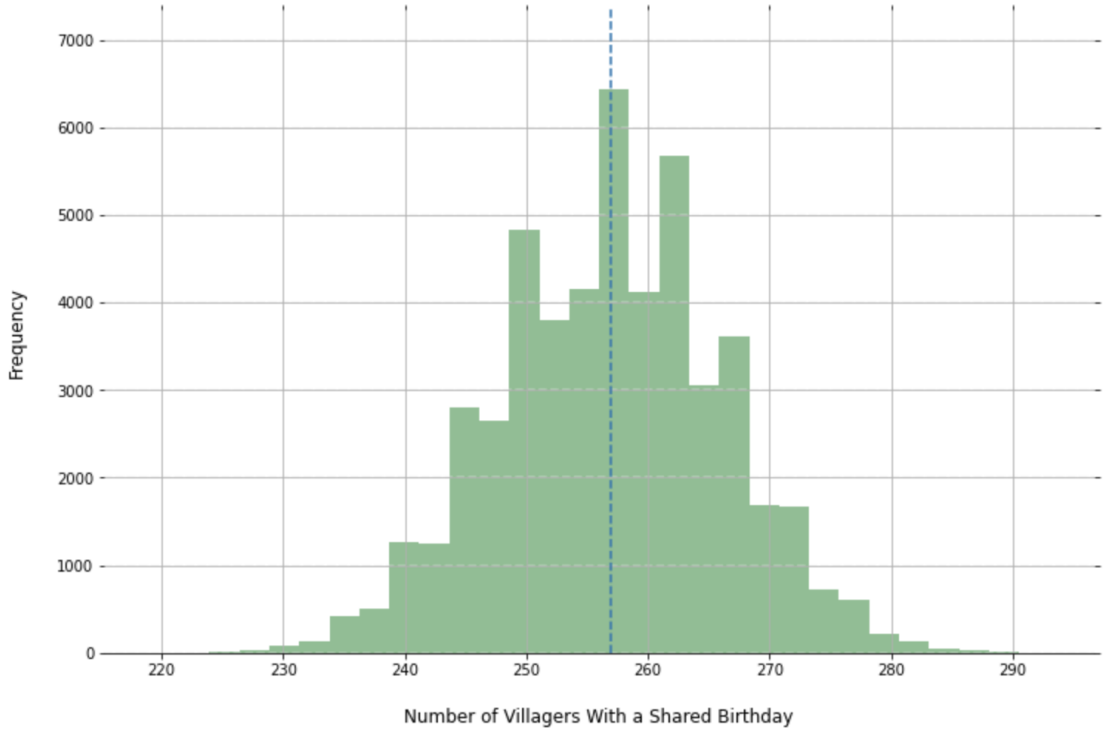

Figure 9: Histogram distribution of number of animals-villagers sharing a birthday.

3.3 Rarest Cat Analysis

Animal Crossing gamers tend to adore the animal-villager Raymond (it is considered as the rarest cat/villager, and almost everyone wants to have it on own island), but he seems very rare to appear.

There are 23 cats in the game at all (according to data analysis of the dataset). We want to determine how likely it is to find Raymond on a mystery island. For easier performance of analysis we would like to assume that *the villager on a mystery island is a cat*, then to determine are the cat-villagers equally likely or not (in other words, are cats equally likely out of all other cats). If they are, then if the villager is a cat, the probability of obtaining Raymond would be $1/23 = 0.0435$. Suppose that p is the true probability of obtaining Raymond knowing the villager is a cat.

To claim that Raymond is equally likely to other cats we need to state hypotheses.

3.3.1 Hypothesis Testing

Def. Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a parameter or a probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H_0. An alternative hypothesis (denoted H_A), which is the opposite of what is stated in the null hypothesis, is then defined. The hypothesis-testing procedure involves using sample data to determine whether or not H_0 can be rejected. If H_0 is rejected, the statistical conclusion is that the alternative hypothesis H_A is true.

A p value is used in hypothesis testing to help us support or reject the null hypothesis. The p value is the evidence against a null hypothesis: the smaller the p-value, the stronger the evidence

that we should reject the null hypothesis.

According to that, we may state the following hypotheses:
H_0: $p_0 = 0.0435$,
H_A: $p_0 < 0.0435$.

Then, we need to simulate data of Raymond appearance among n cat villagers that were presented on mystery islands. Assume that n = 1,000; out of 1,000 cat villagers that appeared, 57 of them were Raymond (the code is implemented in Python):
x = 57,
n = 1000,
$\hat{p} = x/n = 57/1000 = 0.057$ - estimated probability.

To conclude, what hypothesis is relevant, we need to calculate the test statistic and determine the P-value (using a significance level $\alpha = 0.05$).

Def. A test statistic is used in a hypothesis test when we should decide to support or reject the null hypothesis. The test statistic takes data from an experiment or survey and compares results to the results we would expect from the null hypothesis.

According to out dataset, a Z-score was chosen for test statistics:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.057 - 0.0435}{\sqrt{\frac{0.0435 \cdot (1-0.0435)}{1000}}} = 2.0967.$$

Def. Simply put, a z-score (also called a standard score) gives an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.

After that, we may calculate p (by use of scipy.stats Python package) doing the right-tailed test:

$$p = p(Z > 2.0967) = 0.018,$$

$$\alpha = 0.05 > p = 0.018.$$

Thus, we may reject the null hypothesis and confirm that there is sufficient evidence to suggest that **Raymond is less likely to appear than others when a cat appears**.

4 Conclusion

In the present article, several of the statistical analysis options allow us to consider the Animal Crossing dataset in the scientific terms.

The goal of this article was to illustrate how the ideas of statistical analysis and exploratory data analysis can be combined to investigate the data.

We only provided a few analysis options, and many more options could have been included. In addition, our choice of analysis options was highly personal and subjective. Nonetheless, it can be useful both for further analysis of the Animal Crossing game and for gamers who would like to know better about the favourite game.

References

Mostipak, J. (2020). Animal Crossing Reviews dataset. Retrieved from: https://www.kaggle.com/datasets/jessemostipak/animal-crossing.