Black Fur Colored Cats and Their Relationship with Shelter Outcome: Project Report

Baily M. Jepsen

November 22, 2023

Table of Contents

A. Project Overview

A.1 Research Question or Organizational Need

This data analysis project is aimed at understanding if there is a relationship between the fur color of a cat and their potential outcome when moving through a shelter. The project focuses on cats with black fur due to their historically negative associations, thus leading to the modern belief that black cats are a less desirable option for adoption. When an animal is deemed undesirable, they spend more time in the care of shelters and fosters, resulting in more costs for the caregiver.

A.2 Context and Background

The history of the church and feeling threatened by other branches of Christianity is a long one, but the history of black cats and their association with the Devil/Satan can be traced back to the 12th century and the condemnation of the Waldenses, Cathars, and Templars (as cited in Vocelle, 2013). In the 13th century, Pope Gregory IX issued an official church document titled *Vox in Rama,* declaring cats to be incarnations of the Devil/Satan and to "use any association with cats as a tool to discredit those who threatened power and control of the papacy by labeling them witches and/or heretics" (as cited in Vocelle, 2013). Christianity is currently still a practiced religion, and therefore this negative association with black cats may still be present to this day. This history explains why black cats have been seen as less desirable compared to others.

This prejudice is modernly labeled as "Black Cat Bias", which encompasses "religiosity, superstitious beliefs, and prejudicial racial attitudes" as well as "difficulties… in reading the emotions of black cats" (Jones & Hart, 2020). The American Society for the Prevention of Cruelty to Animals (ASPCA) has recognized this phenomenon and responded with the creation of events such as Black Cat Appreciation Day. The ASPCA offers a promotion on this day,

which is intended to gain the interest of the public in efforts to increase black cat adoptions ("Celebrate Black Cat," 2017).

A.3 and A3A Summary of Published Works and Their Relation to the Project

Review of Work 1

What is Agile Methodology? (A Beginner's Guide) was written by Sarah Laoyan, a content, Search Engine Optimization, and marketing specialist. The article entails "a high-level overview of Agile project management" methodology (Laoyan, 2022). In it, Laoyan explains the "four values and twelve principles of Agile" and that it is an iterative approach and how that can make it more effective than linear methodologies (2022). The article concludes with a list and explanations of other variations of Agile.

The Agile methodology was the most appropriate for my approach to this project. This is largely due to the iterative nature of the methodology, as the hypothesis for the research question was adjusted multiple times during wrangling and analyzing the data. This article assisted in that decision, due to its breakdown of the values and principles of Agile. It also includes a graphic of the six sprints (phases) of Agile (plan, design, develop, test, deploy, and review), which became the backbone of this project.

Review of Work 2

"The Chi-Square Test of Independence" was written by Mary L. McHugh of the Department of Nursing at National University. The purpose of this article is to demonstrate how the Chi-Square Test of Independence "is a valuable analysis tool that provides considerable information about the nature of research data" and how it "enables researchers to test hypotheses about variables measured at the nominal level" (McHugh, 2013). The article also covers Cramér's V, which is the strength test that is typically used alongside Chi-Square.

This article was used as the basis for the statistical tests utilized in this project, as the data that is analyzed to answer the research question is nominal and categorical. The statistical tests utilized in this project are the Chi-Square Test of Independence and Cramér's V.

Review of Work 3

How to Run the Chi-Square Test in Python was written by George Pipis, who is the senior director and a data scientist at Persado. The article provides "a practical example of how we can run a Chi-Square Test in Python" (Pipis, 2020). The example tests "if there is a statistically significant difference in Genders (M, F) population between Smokers and Non-Smokers" using a contingency table, a heatmap, and a Chi-Square test (Pipis, 2020).

This article was used as the basis for the tools used in this project. The coding language Python and a Jupyter notebook are tools that allow for the collection, wrangling, and analysis of data in a manner that produces an answer for the research question.

A.4 Summary of Data Analytics Solution

To understand if there is a relationship between cat fur color and outcomes at a shelter, the Agile methodology was utilized. This methodology allowed for an iterative approach, ensuring that all steps of the project were relevant to the research question. Inferential statistics requires the use of statistical tests to draw conclusions from data and therefore answer the research question. The statistical models utilized to carry out the tests for this project include the Chi-Square Test of Independence and Cramér's V, since the data is categorical and nominal. These two models provide information on the presence of a relationship and the strength of the relationship. Collection, wrangling, and analysis of the data was performed in a Jupyter notebook using the coding language Python. The use of Python provided an efficient analysis, while a

Jupyter notebook provided a platform for code, notes, and explanations that combined in the same document.

A.5 Benefits and Support of Decision-Making Process

The proposed solution for understanding if there is a relationship between cat fur color and outcomes at a shelter is beneficial, as it can determine if a relationship is present and the strength of the relationship if there is one. Since this project focuses on cats with black fur color, shelters can use this information to make decisions on their approach to black cats in their care. These decisions could include the cost of adoption, advertisement, promotions, length of time in care before transfer, and education for the public. The ASPCA is an example of this, as it pioneered National Black Cat Appreciation Day (August 17th) and offers promotions for the adoption of black cats on this day ("Celebrate Black Cat," 2017).

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

The goal of this project is to determine if there is a relationship between fur color and outcome, with a focus on black cats versus all other fur colors.

- Objective 1: Data Gathering
    - Deliverable 1.1: Research and obtain a relevant dataset from a source of reliable credibility.
    - Deliverable 1.2: Load the dataset into a Jupyter notebook using Python.
- Objective 2: Data Wrangling
    - Deliverable 2.1: Assess the general properties of the dataset.
    - Deliverable 2.2: Observe and explain any duplicate/missing data.

    - Deliverable 2.3: Clean the dataset of missing and duplicate values, formatting that is not clear or concise, extraneous rows and columns, conversion of values pertaining to analysis, and creation of necessary additional columns.
- Objective 3: Data Analysis
    - Deliverable 3.1: Visual representation of the number of cats in each outcome (Adoption, Euthanasia, and Transfer) in the form of a bar graph.
    - Deliverable 3.2: Visual representation of the number of black and non-black cats for each outcome in the form of a contingency table.
    - Deliverable 3.3: Visual representation of the number of black and non-black cats for each outcome in the form of a heat map.
    - Deliverable 3.4: Performance of a Chi-Square Test of Independence.
    - Deliverable 3.5: Performance of a Cramér's V Test.
- Objective 4: Project Report
    - Deliverable 4.1: A project report.
- Objective 5: Conclusion Report
    - Deliverable 5.1: A conclusion report.
    - Deliverable 5.2: Informational charts and visualizations that illustrate the results of the tests performed on the data and insights.

B.2 Scope of Project

B.2.A Included in Project Scope

This project entails an investigation into the presence of a relationship between the fur color of a cat and their outcome in a shelter, with a focus on black cats versus all other fur colors. It utilizes data originating from Bloomington Animal Shelter Animal Care and Control that was

the result of a management software migration in 2017 (John Snow Labs, 2022). The data gathering, wrangling, and analysis occurs in a Jupyter notebook using the coding language Python. The notebook contains text explaining the process, visualizations, and code. The analysis utilizes inferential statistics, in which tests are conducted on nominal categorical data to answer the hypothesis. This report and a conclusion report are the result of the analysis process, in which the analytical process is explained in varying degrees of detail.

B.2.B Not included in Project Scope

While the migration occurred in 2017, the dataset includes dates extending into 2019 and many missing dates, so the timeline during which the data was collected is unknown. The dataset contains columns that are outside the scope of the project. These include dates, shelter information, additional animal characteristics, and additional information about deceased animals. The dataset also contains animals and outcomes that are outside the scope of the project. These include species that are not cats and outcomes such as “Escaped”, “Foster”, “Released to Wild”, and “Stolen” (John Snow Labs, 2022). During the cleaning process, it was noted that some animals appear twice in the dataset with two different outcomes. These animals had been adopted after transfer. For the sake of analysis and ensuring an animal only appears once in the dataset, the “Transfer” outcome for these animals was deemed outside the scope of the project, because they were ultimately adopted.

B.3 Standard Methodology

While this project is not a collaborative effort, the Agile methodology allows for an iterative approach to the project. This results in “the ability to shift strategies quickly, without disrupting the flow of a project” (Laoyan, 2022). The process that is followed during this project utilizes the six sprints of Agile:

1. Plan
    a. Brainstorming and deciding on a research question.
    b. Creating a hypothesis that would provide an answer for the research question.
    c. Researching an appropriate dataset that fit the research question (Deliverable 1.1).
2. Design
    a. Collecting the dataset (Deliverable 1.2).
    b. Exploring the dataset to understand what needs to be done with it (Deliverable 2.1).
    c. Wrangling the dataset (Deliverable 2.1 - 2.3).
    d. Creating a visualization of the data that is not used for testing (Deliverable 3.1 & 3.3).
3. Develop
    a. Researching inferential statistics models to be used in the testing of the data after understanding the characteristics of the dataset.
    b. Creating a visualization of the data that is used for testing (Deliverable 3.2).
4. Test
    a. Performing a Chi-Square Test of Independence on the dataset (Deliverable 3.4).
    b. Performing a Cramér's V Test on the dataset (Deliverable 3.5).
    c. Analyzing and interpreting the results.
5. Deploy

   a. Ensuring that the solution implemented answers the research question. If it does not, this results in returning to one of the previous phases.

6. Review
   a. Explaining the process in the form of a project report and a conclusion report (Deliverable 4.1 & 5.1).
   b. Creating informational visualizations and charts illustrating the results and insights gained (Deliverable 5.2).

B.4 Timeline and Milestones

Milestone or Deliverable	Projected Start Date	Anticipated End Date	Duration (days or hours)
Data Gathering	11/6/2023	11/7/2023	1 day
Data Wrangling, Cleaning, & Analysis	11/9/2023	11/16/2023	7 days
Project Report	11/17/2023	11/20/2023	3 days
Conclusion Report	11/21/2023	11/22/2023	1 day

B.5 Resources and Costs

1. Hardware
   a. Computer: $500 (an average, models vary)
2. Software
   a. Jupyter Notebook: No cost
   b. Microsoft Word: No cost
   c. Panopto: No cost

   d. Tableau Basic: No cost
      i. Tableau Pro: $14.99 (monthly subscription)

3. Data
   a. John Snow Labs data (for non-commercial use): No cost
      i. John Snow Labs data (for commercial use): $2,989 (yearly subscription)

4. Work Hours
   a. 112 work hours: $4796.96 (112 hours at $42.83 per hour ("Data Analyst," 2023))

B.6 Criteria for Success

The specific metrics used for evaluating the success of the project includes a p-value resulting from the Chi-Square Test of Independence and a value resulting from the Cramér's V test. A p-value indicates whether a relationship is present, while the Cramér's V value indicates the strength of the relationship. Obtaining a p-value is the main criteria, as it is used to answer the hypothesis that addresses the research question.

C. Design of Data Analytics Solution

C.1 Hypothesis

H_0 = Fur color (black vs non-black) is not related to the outcome of a cat.

H_A = Fur color (black vs non-black) is related to the outcome of a cat.

C.2 and C.2.A Analytical Method

Two statistical tests were chosen to be performed for analysis of the data. These include The Chi-Square Test of Independence and Cramér's V. These tests were chosen due to the limited options available for nominal categorical data. The Chi-Square Test of Independence produces a p-value that indicates whether a relationship is present, while Cramér's V produces a value that indicates the strength of a relationship.

The Chi-Square Test of Independence is adequate for analysis due to the non-parametric nominal categorical nature of the data and its "flexibility in handling data from… multiple group studies" (McHugh, 2013). Flexibility is necessary due to the 2x3 structure of the data used for analysis. The p-value produced from this test can be used to answer the hypothesis. Cramér's V was chosen due to it being "the most common strength test used to test the data when a significant Chi-square result has been obtained" (McHugh, 2013). The Cramér's V test is necessary in understanding the strength of the relationship between fur color and outcome.

C.3 Tools and Environments

Collection, wrangling, and analysis of the data entails a Jupyter notebook and the coding language Python. Multiple third-party packages are utilized throughout this process. These include Matplotlib for formatting of statistical data visualizations, NumPy for scientific computing, Pandas for data manipulation, Seaborn for statistical data visualization, and SciPy for the statistical and strength tests. The environment used for producing visualizations is Tableau, which is a program that is used for creating visualizations directly from a dataset.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

The Chi-Square Test of Independence

- H_0 = Fur color (black vs non-black) is not related to the outcome of a cat.
- This test is performed on a contingency table generated from the dataset. The table contains two Boolean rows pertaining to whether a cat has black fur, and three columns indicating the outcome of a cat at the shelter ("Adoption", "Euthanasia", and "Transfer").
- A p-value is the metric generated from the test.

- The *alpha* value used to determine statistical significance is the critical Chi-Square value. This involves a threshold and a degree of freedom. The threshold chosen for this test was the common threshold of 5%. These two values can then be used to locate the critical Chi-Square value in a Chi-Square table (Glen, n.d.).
- If the p-value is greater than the critical Chi-Square value, statistical significance is found to be present in the data. The analysis would fail to reject the alternative hypothesis. If the p-value is less than the critical Chi-square value, statistical significance would not be present. The analysis would fail to reject the null hypothesis.

C.5 Practical Significance

The practical significance of the analysis performed is assessed based on the value derived from Cramér's V. Though a statistical significance may be found, the strength of that relationship determines how important that statistical significance is practically. If the strength of the relationship is strong (the value calculated is closer to 1), there is a case for practical significance. Otherwise, if the strength of the relationship is weak (the value calculated is closer to 0), there is no case for practical significance.

C.6 Visual Communication

The two graphical representations that will be used for visual communication include a scatter plot and a bar graph. The scatter plot is a visual representation of the number of occurrences of each outcome (Adoption, Euthanasia, and Transfer) for each fur color of cats in the dataset. This visualization aids in providing an overview of the dataset before it is condensed into whether a cat has black fur. The bar graph is a visual representation of the number of cats in each outcome and is separated into two sections indicating whether a cat has black fur. This

visualization aids in providing an overview of the data used for analysis. Both the scatter plot and bar graph are created using Tableau.

D. Description of Dataset

D.1 Source of Data

The data used for this project is from Bloomington Animal Shelter Animal Care and Control. The dataset resulted from a management software migration in 2017 (John Snow Labs, 2022).

D.2 Appropriateness of Dataset

The dataset is appropriate for the stated goals of this project because it contains data on 3,603 unique cats of various fur colors and their outcome after their time spent in the shelter. Fur color and outcome both apply to the research question, and the dataset is large enough for the performance of statistical tests.

D.3 Data Collection Methods

The data was collected by downloading the .csv file from John Snow Labs (2022). It was then loaded into a Jupyter notebook using Python and the Pandas package.

D.4 Observations on Quality and Completeness of Data

Quality Observations

- The column titles did not have clear and concise formatting.
- An outcome of “Euthanasia” was reported in a column separate from the rest of the outcomes.
- The column “basecolor” (later becoming “fur_color”) contained repeating values (“Brown and Black” and “Black and Brown”), inconsistent formatting (“Grey

Black and White" versus "Grey, Black and White"), and repeating values due to inconsistent capitalization ("grey and Black" versus "Grey and Black").

Completeness Observations

- The dataset contained several columns that included missing data. These included "intakereason", "identichipnumber", "breedname", "returndate", and "deceaseddate". These columns were not needed for analysis and were dropped from the dataset.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

The four pillars of data governance are data quality, stewardship, protection and compliance, and management. Since data quality was covered earlier in this report, this section will focus on the other three pillars. In terms of data stewardship, multiple copies of the dataset were found on various websites. The dataset used for analysis was retrieved from the legal owner of the data, John Snow Labs. Retrieving the dataset from the original source ensured no loss of information and a consistent and reliable source. Compliance required by John Snow Labs involved attribution of the dataset to John Snow Labs and non-commercial use under a "Research Data License" (2022). There were no requirements for protection or management of the data after retrieval.

There were no requirements for privacy or security concerning the data. The dataset does not contain any personally identifiable information for humans or animals, so managing privacy or security of the data was not required by John Snow Labs.

The dataset is regulated under a Copyleft license for free content. This "is a general license agreement granted by a copyright owner permitting anyone to freely use copyrighted

property but under specific terms" (Friedman, 2023). The specific terms required by John Snow Labs are non-commercial use unless a yearly subscription is purchased and attribution to John Snow Labs (2022).

For ethical reasons, the conclusions drawn from this analysis may have the potential to adversely impact black cats. More tests on various datasets are recommended before making concrete decisions on the handling of black cats in shelters.

References

Celebrate black cat appreciation day on August 17! (2017, August 15). ASPCA. https://www.aspca.org/news/its-black-cat-appreciation-day

Data analyst salary in Washington state. (2023, November 15). Indeed. https://www.indeed.com/career/data-analyst/salaries/WA

Friedman, K. (2023, September 8) Copyleft. Encyclopedia Britannica. https://www.britannica.com/topic/copyleft

Glen, S. (n.d.). *Chi squared table (Right tail)*. Statistics How To. https://www.statisticshowto.com/tables/chi-squared-table-right-tail/

John Snow Labs. (2022, November 29). *Bloomington animal care and control adopted animals.* John Snow Labs. https://www.johnsnowlabs.com/marketplace/bloomington-animal-care-and-control-adopted-animals/

Jones, H. D., & Hart, C. L. (2020). Black Cat Bias: Prevalence and Predictors. *Psychological reports*, *123*(4), 1198–1206. https://doi.org/10.1177/0033294119844982

Laoyan, S. (2022, October 15). *What is Agile methodology? (A beginner's guide)*. Asana. https://asana.com/resources/agile-methodology

Pipis, G. (2020, October 24). *How to run the Chi-Square test in Python*. Medium. https://medium.com/swlh/how-to-run-chi-square-test-in-python-4e9f5d10249d

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/bm.2013.018

Vocelle, L. (2013, February 8). *History of the cat in the Middle Ages (Part 2)*. THE GREAT CAT. https://www.thegreatcat.org/history-of-the-cat-in-the-middle-ages-part-2/