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

Baily M. Jepsen

November 22, 2023

Table of Contents

A. Project Highlights

This data analysis project was 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 focused 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.

The project entailed 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 utilized 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 occurred in a Jupyter notebook using the coding language Python. The notebook contains text explaining the process, visualizations, and code. The analysis utilized inferential statistics, in which tests were conducted on nominal categorical data to answer the hypothesis. A project report and this conclusion report are the result of the analysis process, in which the analytical process is explained in varying degrees of detail.

While this project was not a collaborative effort, the Agile methodology allowed for an iterative approach to the project. This resulted in "the ability to shift strategies quickly, without disrupting the flow of a project" (Laoyan, 2022). The process that was followed during this project utilized the six sprints of Agile.

Collection, wrangling, and analysis of the data entailed a Jupyter notebook and the coding language Python. Multiple third-party packages were utilized throughout this process. These included 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 created for the sole use of creating visualizations directly from a dataset.

B. Project Execution

The goal of this project was to determine if there is a relationship between fur color and outcome, with a focus on black cats versus all other fur colors. Analysis of the data was successful in determining the presence of a relationship. The Agile methodology allowed for an iterative approach to the project, resulting in a successful execution and no variances from the project plan.

Almost all objectives, milestones, and deliverables were successfully completed within their anticipated dates and durations. The first variation was the data wrangling and analysis milestone, as it took nine days to complete instead of the anticipated seven. After beginning the project report, minor changes were made to the Jupyter notebook to ensure proper alignment. The second variation was the project report deliverable, as minor changes were made to it as visualizations were produced. The third variation was the actual start date for this conclusion report due to ensuring alignment between the two reports. These variations were mostly due to the Agile methodology that allowed for an iterative approach to all steps of the project. This approach was beneficial as it ensured alignment between all milestones and deliverables.

Milestone or Deliverable	Projected Start Date	Anticipated End Date	Actual Start Date	Actual End Date	Duration (days or hours)
Data Gathering	11/6/2023	11/7/2023	11/6/2023	11/7/2023	1 day
Data Wrangling, Cleaning, & Analysis	11/9/2023	11/16/2023	11/9/2023	11/18/2023	9 days
Project Report	11/17/2023	11/20/2023	11/17/2023	11/22/2023	5 days
Conclusion Report	11/21/2023	11/22/2023	11/20/2023	11/22/2023	2 days

C. Data Collection Process

The data was collected by downloading the .csv file from John Snow Labs (2022). The only obstacle encountered in the collection of the data was the presence of multiple copies of the dataset on various websites. This required a small amount of additional research to find the original dataset from John Snow Labs. There were no unplanned data governance issues aside from ensuring the data originated from the rightful owner. The lack of unplanned data governance issues was due to the dataset being publicly available for non-commercial use.

C.1 Advantages and Limitations of Data Set

Advantages

- The dataset contained many unique cats for analysis (n = 3,603).
- The dataset contained characteristics such as fur color and outcome, both of which pertain to the research question.
  - Fur colors include Tortie, Dilute Tortoiseshell, Black, Seal, White and Grey, Black and Brown, Torbie, Various, Grey, Black and White, White and

Orange, White, Buff, Fawn, Orange, Brown, Golden, Tabby, Tabbico, Cream, White and Tan, Lilac, Silver, Black and Grey, Black Tortie, Dilute Calico, Black, White and Brown, Calico, White and Tabby, Blue Point, White and Brown, Lilac Point, Flame Point, Lynx Point, Buff and White, Siver and Black, Tortie and White, Black and Tan, Chocolate Point, Tortie Point, Blue, Chocolate, Brindle and Black, Tan and Brown, Seal Point, Tricolour, Grey, Black and White, Cinnamon, and Smoke.

  - Outcomes include Adoption, Foster, Transfer, Stolen, Escaped, Released to Wild, and the column “puttosleep” (Euthanasia).

Limitations

- The data originates from only one shelter. This is a limitation because it only reports on animals from one geographic location, and therefore is not an accurate representation of the experiences of adoptable animals on a more global scale.
- The data was collected during an underminable span of time. This is a limitation because while the dataset is claimed to originate from a management software migration that occurred in 2017, there are dates included that span into 2019. There are also many entries with missing dates.
- The dataset contained several columns that included missing data. These included “intakereason”, “identichipnumber”, “breedname”, “returndate”, and “deceaseddate”.
- The column “basecolor” (later becoming “fur_color”) contained repeating values (“Brown and Black” and “Black and Brown”), inconsistent formatting (“Grey Black and White” versus “Black, Brown and White”), and repeating values due to inconsistent capitalization (“grey and Black” versus “Grey and Black”).

- The column titles did not have clear and concise formatting.
- An outcome of "Euthanasia" was reported in a column ("puttosleep") separate from the rest of the outcomes.

D. Data Extraction and Preparation

The extraction process was quite simple, and only required the dataset to be loaded into a Jupyter notebook using Python and the Pandas package. This was due to the data being contained in an easily usable .csv file format.

The preparation process took the longest during the collection, wrangling, and analysis portion of the project. Several steps were taken to ensure proper cleaning of the dataset before analysis:

1. Though the dataset contained missing values, those columns that contain the missing values were not pertinent to the analysis. These columns were considered extraneous and were dropped from the dataset.
    a. Extraneous columns dropped: "intakedate", "intakereason", "istransfer", "sheltercode", "identichipnumber", "animalname", "breedname", "animalage", "sexname", "location", "movementdate", "istrial", "returndate", "returnedreason", "deceaseddate", "deceasedreason", "diedoffshelter", and "isdoa".
2. Column labels required formatting for clarity.
    a. "basecolor" became "fur_color"
    b. "speciesname" became "species_name"
    c. "movementtype" became "movement_type"

  d. “puttosleep” became “put_to_sleep”

3. Since the analysis is focusing on cats, any animals that do not have "Cat" in the “speciesname” (“species_name”) column were dropped.
    a. Extraneous species dropped: “Dog”, “House Rabbit”, “Rat”, “Bird”, “Opossum”, “Chicken”, “Wildlife”, “Ferret”, “Tortoise”, “Pig”, “Hamster”, “Guinea Pig”, “Gerbil”, “Lizard”, “Hedgehog”, “Chinchilla”, “Goat”, “Snake”, “Squirrel”, “Sugar Glider”, “Turtle”, “Tarantula”, “Mouse”, “Raccoon”, “Livestock”, and “Fish”.
4. If a cat had been "Reclaimed", that outcome was viewed to be the same as an "Adoption". This required converting "Reclaimed" values to "Adoption" in the “movementtype” (“movement_type”) column.
5. If a cat had the value "TRUE" in the “puttosleep” (“put_to_sleep”) column, the value for “movementtype” (“movement_type”) became "Euthanasia".
6. If a cat had the value "Black" in the “furcolor” (“fur_color”) column, the value for a new Boolean column “is_black” became "True". Otherwise, the value for “is_black” became "False".
7. Rows that contained a “movementtype” (“movement_type”) value that was not needed for analysis was dropped.
8. The “basecolour” (“fur_color”) column contained multiple formatting errors, most likely due to various people inputting and handling the data. These were:
    a. Repeating values
        i. ("Brown and Black" and "Black and Brown").
        ii. ("Tan and Black" and "Black and Tan").

iii. ("White and Black" and "Black and White").

iv. ("grey and Black" and "Black and Grey").

v. ("Grey and Black" and "Black and Grey").

vi. ("Brown, Black and White" and "Black, Brown and White").

vii. ("Brown and White" and "White and Brown").

viii. ("Grey and White" and "White and Grey").

ix. ("Orange and White" and "White and Orange").

x. ("Tabby and White" and "White and Tabby").

xi. ("Tan and White" and "White and Tan").

b. Missing commas

i. ("Grey Black and White" vs "Grey, Black and White").

c. Lowercase letters

i. ("Black and grey" became "Black and Grey").

ii. ("Buff and white" became "Buff and White").

iii. ("Dilute calico" became "Dilute Calico").

iv. ("Dilute tortoiseshell" became " Dilute Tortoiseshell ").

v. ("Lynx point" became "Lynx Point").

9. The dataset did contain a type of duplicate values that needed to be dealt with. For the analysis, the dataset could not contain a cat that had been transferred then adopted. To solve this, a duplicate “id” that had a “movementtype” (“movement_type”) of "Transfer" was dropped. The "Adoption" entry was then preserved in the dataset, resulting in each “id” appearing only once in the dataset.

E. Data Analysis Process

E.1 Data Analysis Methods

Two statistical tests were chosen to be performed for analysis of the data. These included 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 utilized a contingency table and produced a p-value that indicated whether a relationship was present, while Cramér's V produced a value that indicated the strength of a relationship.

The Chi-Square Test of Independence was 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 was necessary due to the 2x3 structure of the data used for analysis. The p-value produced from this test implied statistical significance and was 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 was necessary in understanding the strength and practicality of the relationship between fur color and outcome.

E.2 Advantages and Limitations of Tools and Techniques

Jupyter Notebook

- Advantages
  - Ability to contain code, code results, text annotations, and visualizations in one document.
  - Ability to export the notebook to different formats such as HTML and PDF.

  - Works well with the iterative Agile methodology due to the ability to make a change to one cell of code and rerun it.
  - A variety of packages are available for various uses and are not pre-loaded, so you get to choose what you want to use.
- Limitations
  - No history of previous checkpoints (saves), so there is an inability to revert to an earlier version of the document. If you delete a cell that was not part of the last checkpoint, it is unretrievable.
  - Can run slowly depending on calculations being performed and when handling large/complex datasets.

Matplotlib

- Advantages
  - Ability to create and customize statistical visualizations and graphs.
- Limitations
  - Requires more code than Seaborn, especially when creating complex visualizations or graphs.

NumPy

- Advantages
  - Contains a large library of mathematical functions.
- Limitations
  - Performance is decreased when handling complex datasets.

Pandas

- Advantages

    - Converts a .csv file into an indexed dataframe.
    - Efficient at handling large datasets.
- Limitations
    - Performance is not as efficient as NumPy.
    - This package relies on NumPy.

Seaborn

- Advantages
    - Ability to create and customize complex statistical visualizations and graphs without much effort.
- Limitations
    - This package relies on Matplotlib.
    - Performance is decreased when handling complex datasets.

SciPy

- Advantages
    - Allows for the ability to perform scientific and statistical tests on a dataset.
    - An understanding of the calculations performed is not necessary, as it performs the calculations for you.
- Limitations
    - Use is limited to the Python coding language.
    - If you do not understand the calculations being performed, you may not understand what the calculations produce or the implementation.

Tableau

- Advantages

  - This program is powerful and quick to produce visualizations.
- Limitations
  - There is a small learning curve in understanding how to use the program.
  - Understanding SQL or R is required to produce complex visualizations.

E.3 Application of Analytical Methods

The Chi-Square Test of Independence was performed on a contingency table that was derived from the dataset. The contingency table contained two rows pertaining to whether a cat is black, and three columns for “Adoption”, “Euthanasia”, and “Transfer”. The values contained in this table were frequencies for how many cats were observed in each scenario. The Chi-Square test utilized the “chi2_contingency” function from SciPy and produced a p-value of 0.1044. 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 0.05. The degree of freedom was 2 and was calculated based on the number of rows and columns in the contingency table ((2 rows – 1) * (3 columns – 1) = 2). These two values were then used to locate the critical Chi-Square value in a Chi-Square table, which was found to be 0.0599 (Glen, n.d.).

A Cramér’s V test was then performed to measure the strength of the relationship found. This test was performed to determine practical significance. This test utilized the association function from SciPy, using the “cramer” method parameter. This test produced a value of 0.035, which when combined with a degree of freedom of 1 (2 rows - 1) was used to locate the strength of the relationship using a Cramér’s V table (Zach, 2021).

F. Data Analysis Results

F.1 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.
- H_A = Fur color (black vs non-black) is related to the outcome of a cat.
- A p-value of 0.1044 was the metric generated from the test.
- The *alpha* value used to determine statistical significance is the critical Chi-Square value. With a common threshold of 0.05 and a degree of freedom of 2, these two values were then used to locate the critical Chi-Square value in a Chi-Square table, which was found to be 0.0599 (Glen, n.d.).
- Because the p-value of 10.44% was greater than the critical Chi-Square value of 5.99%, statistical significance was found to be present in the data. The analysis completed failed to reject the alternative hypothesis and found sufficient evidence to support the null hypothesis that black fur color is related to the outcome of a cat.

Cramér's V

- 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.
- A value of 0.035 was the metric generated from the test.
- Since this test measured practical significance, an *alpha* value was not applicable.
- Given a degree of freedom of 1 (2 rows – 1), the value derived from this test indicated a small/weak relationship between black fur color and the outcome for a cat in a shelter (Zach, 2021).

F.2 Practical Significance

The practical significance of the analysis performed was assessed based on the value derived from Cramér's V. Though a statistical significance was found, the strength of that relationship determined how important that statistical significance is practically.

While there was statistical significance found from the analysis performed, there was little practical significance. This means that even though black cats suffer worse outcomes more often than those of other fur colors, the difference is not large enough to warrant any changes in the decision-making process behind black cats versus cats of other fur colors in shelters.

F.3 Overall Success

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

Because the p-value of 10.44% was greater than the critical Chi-Square value of 5.99%, statistical significance was found to be present in the data. Given the Cramér's V value of 0.035 and a degree of freedom of 1, the value derived from this test indicated a small/weak relationship between black fur color and the outcome for a cat in a shelter.

The analysis completed was successful even though there is little practical significance. This analysis was a success because the p-value produced failed to reject the alternative hypothesis and found sufficient statistical significance to support the null hypothesis that black fur color is related to the outcome of a cat.

G. Conclusion

G.1 Summary of Conclusions

The analysis performed for this project was successful in answering the hypothesis for the research question of whether cats with black fur experience worse outcomes in a shelter versus all other fur colors. The dataset, tools, and methodology used were appropriate for the analysis. The dataset provided an appropriate amount of data with characteristics pertaining to the analysis. The tools used were sufficient at carrying out the collection, wrangling, and analysis of the dataset. The Agile methodology allowed for an iterative approach to the process that was beneficial is ensuring the data was properly analyzed and the results were appropriate.

Both statistical and practical significance were properly evaluated. The Chi-Square Test of Independence produced a p-value that was greater than the critical Chi-Square value, indicating statistical significance for the relationship between color of fur (black versus non-black) and outcome at a shelter. The Cramér's V value produced indicated a small/weak relationship between the two, and therefore little practical significance. While the analysis was successful in answering the hypothesis, the results indicated that the relationship was not strong enough to warrant any changes in decision-making for handling of black cats in shelters. Black cats fare slightly worse than those of other fur colors, but not enough to warrant any additional actions in ensuring their adoption.

G.2 Effective Storytelling

The two graphical representations that were 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 (see Appendix A). 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 (see Appendix B). This visualization aids in providing an overview of the specific data used for analysis. Both the scatter plot and bar graph were created using Tableau. Tableau was chosen for the creation of visualizations because it is powerful, fast, and was created purely for the creation of data visualizations.

G.3 Recommended Courses of Action

Based on the findings that there is statistical significance, but not practical significance between the fur color of a cat (black versus non-black) and outcome at a shelter, two recommended courses of action can be made. The first course of action that can be recommended is for the shelter to continue their business as is. The strength of the relationship between a cat with or without black fur and their outcome at the shelter is weak, indicating that this relationship is not practically significant enough to recommend any changes in an approach to increase adoptions of cats who have black fur.

The second course of action that can be recommended is to perform this analysis on other sets of data from the shelter. This recommendation is based on the lack of information concerning when this data was collected, and that there is a relationship present. The shelter may find it advantageous to continue monitoring this relationship, but over a known span of time to understand if there are any fluctuations in the strength of the relationship during various times of the year or between different years. The shelter may find that the strength of the relationship changes during the month of October, a time of the year that black cats have been associated with.

References

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/

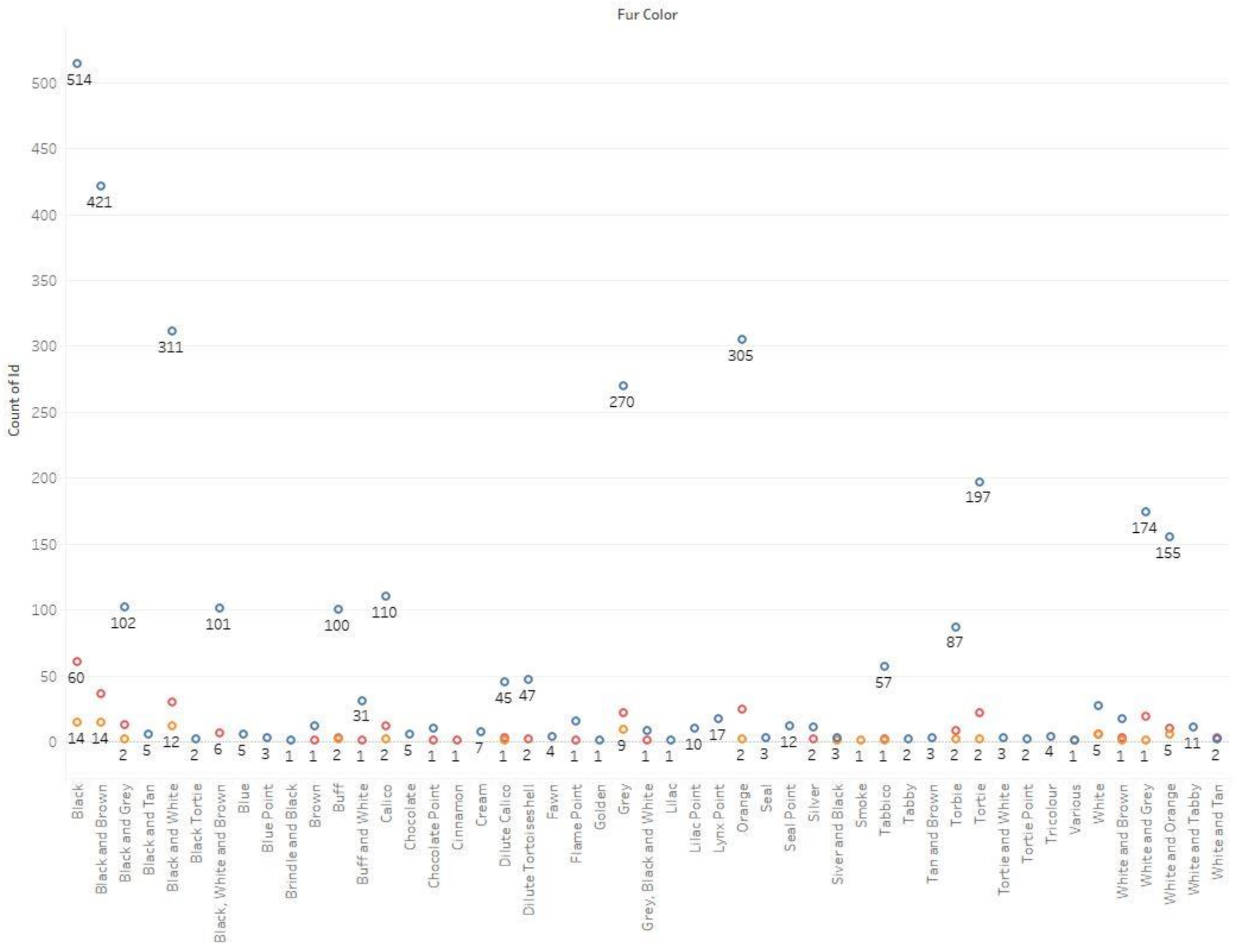
Laoyan, S. (2022, October 15). *What is Agile methodology? (A beginner's guide)*. Asana. https://asana.com/resources/agile-methodology

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

Zach. (2021, September 30). *How to interpret Cramer's V (With examples)*. Statology. https://www.statology.org/interpret-cramers-v/

Appendix A

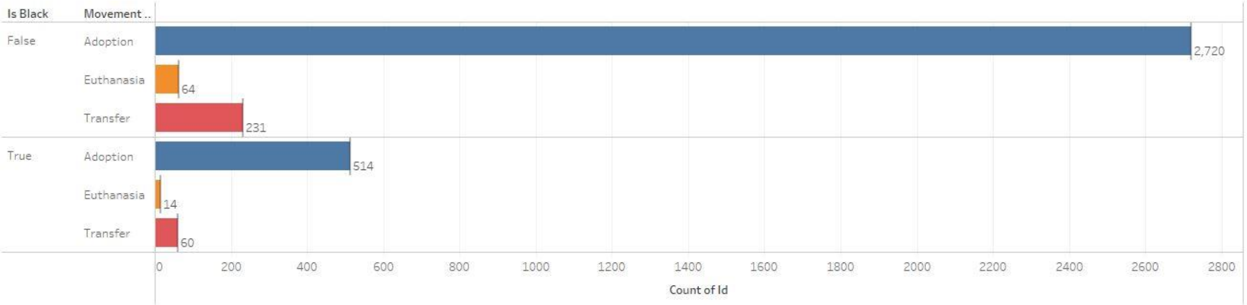
Scatter Plot of the Number of Cats in Each Movement Type by Fur Color



Due to the number of fur colors included in this scatter plot, this image can be hard to read. This image can also be found in the GitHub repository for this project.

Appendix B

Bar Graph of the Number of Cats in Each Movement Type – Separated by Whether Fur Color is Black



Due to the number cats included in this bar graph, this image can be hard to read. This image can also be found in the GitHub repository for this project.