Analysis of Characteristics of Top Dog Breeds

Alyson M Dunkle

Western Governors University

**Table of Contents**

[Project Overview 3](#_Toc55905041)

[A. Summary 3](#_Toc55905042)

[Project Plan 3](#_Toc55905043)

[B. Summary 3](#_Toc55905044)

[Methodology 3](#_Toc55905045)

[C. Data Selection 3](#_Toc55905046)

[C1. Data Set Advantages and Limitations 4](#_Toc55905047)

[D. Data Extraction/Preparation Processes – Tools, Techniques, Suitability 4](#_Toc55905048)

[E. Data Analysis Process 4](#_Toc55905049)

[E1. Analysis Methods 4](#_Toc55905050)

[E2. Analysis Tools/Techniques - Advantages and Limitations 4](#_Toc55905051)

[E3. Step-by-Step Explanation for E1 4](#_Toc55905052)

[Results 4](#_Toc55905053)

[F. Project Success 4](#_Toc55905054)

[F1. Statistical Significance 4](#_Toc55905055)

[F2. Practical Significance 4](#_Toc55905056)

[F3. Overall Success and Effectiveness 4](#_Toc55905057)

[G. Key Takeaways 4](#_Toc55905058)

[G1. Conclusions 4](#_Toc55905059)

[G2. Justify Visual Communications Tools 4](#_Toc55905060)

[G3. Findings-based Recommendations 4](#_Toc55905061)

[H. Panopto Video Link 4](#_Toc55905062)

[Appendices 4](#_Toc55905063)

[Sources 4](#_Toc55905064)

# Project Overview

# Project Highlights

A1. Research Question:

The research question this project aimed to answer were what the top three dog breeds with the most desirable characteristics for a dog breeder to start breeding.

A2. Project Scope

The project’s scope was to analyze data about seven different categories of dog breeds. Characteristics were used to compare the different dog breeds from each of the different categories of dog breeds. The top three dogs of each category would be compared to discover the top three dog breeds overall.

A3. Solution Overview – Tools:

The data was found on Kaggle and was a csv file called best\_in\_show. The data was cleaned using Python programming language in a Jupyter Notebook. The cleaned csv file was extracted from the Jupyter Notebook and analyzed in Tableau using eight different characteristics of seven dogs of each breed category. The information gathered from the graphs in Tableau were input into Microsoft Excel to discover what the top seven tops of each category was for each of the chosen characteristics.

A4. Solution Overview – Methodologies:

The project methodology that was used in this project was SEMMA, Sample, Explore, Modify, Model, and Assess. These steps allowed me to extract all meaningful data to come to the proposed conclusion.

# Project Plan

# Project Execution

B1. Project Plan

The goal of this project is to analyze a dataset of top dog breeds to discover the top three dog breeds for which a breeder could choose from to start breeding.

The objectives for this goal are:

* + - Determine eight characteristics, four negative and four positive, that can be used for comparison.
      * The deliverable for this objective is a list of the eight characteristics which can be used.
    - Determine the top three dog breeds of each of the seven dog breed categories.
      * The deliverable for this objective would be a list of the top three dog breeds of each category.
    - Determine the overall top three dog breeds.
      * The deliverable for this objective would be a final list of the overall top three dog breeds.

B2. Project Planning Methodology

The project methodology used was the SEMMA methodology. The steps of SEMMA are Sample, Explore, Modify, Model and Assess. This methodology was chosen because it best fit the process which was needed to complete the analysis of the data that was chosen.

Sample: In this step, I chose the data file best\_in\_show from Kaggle which contained enough information to work with in this data analysis.

Explore: In this step, I explored the data set to see all the different characteristics of the set, of each category of dog breed and the different dog breeds included in the data set.

Modify: In this step, I used Python programming language to clean up the data. The data had incorrect data types are well as columns and rows which were not necessary to the analysis.

Model: In this step, I modeled the cleaned data file in Tableau to see the top seven dog breeds for each of the seven different category of dog breed. I graphed a separate graph set for eight different characteristics, four positive and four negative.

Assess: In this step, I used Microsoft Excel to compile all the relevant characteristics of the top seven dog breeds to discover the top three or four dogs in each of the categories. These final twenty-five breeds were then themselves compared to discover the top four dog breeds.

B3. Project Timeline and milestones

The actual project timeline and milestones followed the proposed timeline in Task 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Projected Start Date** | **Projected End Date** | **Duration (days/hours)** |
| Cleaned data in Jupyter Notebook | 7/5/2022 | 7/8/2022 | 12 hours |
| Created graphs in Tableau | 7/8/2022 | 7/13/2022 | 8 hours |
| Complied data of top dog breeds in Excel | 7/8/2022 | 7/13/2022 | 4 hours |

# Methodology

# Data Collection Process

* Actual data selection vs. planned collection process:

The planned collection process ended up being the actual data selection on the whole. The data was found on Kaggle but only some of the data found in the data set was used in the data analysis process. There were several rows and columns which were not needed in the process.

* Obstacles to data collection

There were no obstacles to the data collection.

* Unplanned data governance handling

There were no unplanned data governance handling.

# C1. Advantages and Limitations of Data Set

The advantage of the chosen data set was all the information was already complied into a single place. This allowed a faster process in cleaning and analyzing the data. Another advantage was the dog breeds already being categorized by different breed types. This allowed for a better analysis of the dog breeds.

The limitation of this data was having only a limited number of breeds in the dataset. This made the final choice of dog breeds to be only the best of the collected data and possibly excluding other dog breeds. Another limitation was there were many null values within which made a complete comparison impossible to complete. This again made the final outcome to only be from what I had to sample and may not have been the entirety of every dog breed.

# Data Extraction and Preparation Processes

The data was already complied into a csv file which was the best format for the data analysis process I used for this task. This csv file was uploaded to a Jupyter Notebook where the data was cleaned in preparation for analysis. The data needed the cleaning process to have a dataset which could easily be read and put into graphical form for analysis.

# Data Analysis Process

# E1. Data Analysis Methods

The data analysis method that was used in this project was descriptive analysis. This method was used because the raw data is manipulated and interpreted to gain insights for the project question being answered.

# E2. Advantages and Limitations of Tools/Techniques

The first tool used was Python programming language within a Jupyter Notebook. The tool was very useful for cleaning the acquired data. The advantage was it allowed me to manipulate the data in many different ways. There are not many limitations with Python programming language, as long as you know the code that is needed to change or clean the data.

The second tool used was Tableau. The tool was used to produce several graphs pertaining to the different selected characteristics of the dog breeds. The advantage of using Tableau is the ease of usage. Tableau is easy to use with the different tools provided and can help the user to create helpful models used to analysis the data. Tableau is only limited by the creativity of the user and the information that is input into it.

The third and final tool which was used was Microsoft Excel. The advantage of Excel is there are many different tools that can be utilized within it. One just has to know the different tricks and tools to do so. The limitation of excel is that the csv file had to be processed and condensed into smaller information to make Excel the platform to finish the data analysis. Large data files are harder to use in Excel and can make analyzing the data much more difficult.

# E3. Application of Analytical Methods

The submission includes a thorough step-by-step explanation of how the analytical methods were applied to the data and how *all* assumptions or requirements were verified.

This project used the following analytical methods and techniques:

* + Obtaining a dataset through the Kaggle database
  + Cleaning the data through Python programming language
  + Using the cleaned data in Tableau to discover the top dog breeds in each of the seven categories for eight characteristics
  + Putting the above information into Microsoft Excel to compare the eight characteristics of the dog breeds to obtain the top three dog breeds

# Results

# Project Success

# F1. Statistical Significance

To determine statistical significance for the chosen characteristics, I have used Pearson’s Correlation Coefficient. I compared the relationship between the Popularity in the US and Trainability as well as Popularity in the US and Lifetime Cost. The Pearson’s correlation coefficient between the Popularity in the US and Trainability was -0.215. This represents a negative correlation with a low degree of correlation. This tells us that the relationship between Popularity in the US and Trainability is such as one increases, the other decreases. The Pearson’s Correlation Coefficient between Popularity in the US and Lifetime cost was -0.503. This also represents a negative correlation but with a high degree. This indicates Popularity in the US and Lifetime Cost have a much stronger correlation.

# F2. Practical Significance

The practical significance of this project is having four different dog breeds for dog breeder to choose from when deciding which type of dog to breed next. The Dandie Dinmont Terrier had four of the four positive characteristics and only one of the four negative characteristics. The top two to four dogs all had three of the four positive and only one of the four negative characteristics. This allows the breeder a range of choices. All four of the top dog breeds are quite different from each other. The have two dogs from the hound category, one from the non-sporting category, and the last from the terrier category.

# F3. Overall Success

I believe this project was a success because in the end it gave me the top dogs from the seven categories. The dog breeds are varied and different enough for the dog breeder to choose from when making their decision. The characteristics were varied and were chosen because they were both negative and positive characteristics.

# Key Takeaways

# G1. Summary of Conclusions

This project was used to discover the top three dog breeds from seven different breed categories in a dataset. The dataset had to be cleaned and analyzed to get to the final discovery of the needed information. The three objectives were to discover eight characteristics for use in the analysis, use those characteristics to discover the top three dogs in each category, and finally, to find the top three dog breeds overall.

The four positive characteristics: trainability, longevity, popularity, and suitability for children, and four negative characteristics: genetic aliments, lifetime cost, average purchase price, and grooming required. These eight characteristics were chosen because they are factored buyers will look into while deciding which dog breed to buy from a breeder.

The top three dog breeds in each category ended up having a couple of top four dog breeds due to some of them having the same number of positive and negative characteristics. The seven breed categories were herding, hound, non-sporting, sporting, terrier, toy and working.

The top twenty-five dog breeds were then compared to find the top four breeds. The top four were chosen because of the top second, third, and fourth dogs all had the same number of positive and negative characteristics.

The top four dog breeds were Dandie Dinmont Terrier, Afghan Hound, Beagle, and Boston Terrier. The gathering of the top four dog breeds makes me conclude this project a success.

# G2. Effective Storytelling

Include logical reasons why the chosen tools and graphical representations for visually communicating the findings support effective storytelling.

The chosen tools and graphical representations for visually communicating the findings were the best methods because it allowed me to analyze the data to find the answer to the research question. The graphical representations of each of the eight characteristics clearly showed the top dogs in each category. This allowed me to combine the data into excel and show the end results.

# G3. Findings-based Recommendations

The research question asks what the top three dog breeds are overall from the chosen dataset. The top four dog breeds were Dandie Dinmont Terrier, Afghan Hound, Beagle, and Boston Terrier. The top dog, the Dandie Dinmont Terrier, was the top dog breed with four positive characteristics and only one negative characteristic. The other three dog breeds all had three positive characteristics and only one negative characteristic. Based on these findings, I recommend breeding the Dandie Dinmont Terrier as the first choice. The Beagle or the Boston Terrier are more well known and would also be very good candidate for a breeding program. Any of the top four dog breeds should be profitable for a breeder to include in their breeding program.

# Panopto Presentation

Provide a link to your Panopto presentation. Include the following in your summary:

• a summary of your research question or organizational need

• a demonstration of the functionality of any code you used for your data analytics solution

• an outline of the findings and implications of your analysis

The summary should be appropriate for a data-analytics audience.

# Appendices

# Evidence of Completion

The following files will serve as evidence of completion for this project:

1. Jupyter Notebook
2. Tableau Public link
3. Excel data file

# Sources

Mooney, P. (2020, December 14). *Best in show (data about dogs)*. Kaggle. Retrieved May 23, 2022, from https://www.kaggle.com/datasets/paultimothymooney/best-in-show-data-about-dogs