Analysis of Characteristics of Top Dog Breeds

Alyson M Dunkle

Western Governors University

**Table of Contents**

[A. Project Overview 3](#_Toc58328659)

[A1. Research Question or Organizational Need 3](#_Toc58328660)

[A2. Context and Background 3](#_Toc58328661)

[A3. Summary of Published Works 3](#_Toc58328662)

[A3a. Relation of Published Works to Project 3](#_Toc58328663)

[A4. Summary of Data Analytics Solution 3](#_Toc58328664)

[A5. Benefit to Organization and Decision-Making Process 3](#_Toc58328665)

[B. Data Analytics Plan 3](#_Toc58328666)

[B1. Goals, Objectives, and Deliverables 3](#_Toc58328667)

[B2. Scope of Project 3](#_Toc58328668)

[B3. Standard Methodology 4](#_Toc58328669)

[B4. Timeline and Milestones 4](#_Toc58328670)

[B5. Resources and Costs 4](#_Toc58328671)

[B6. Criteria for Success 4](#_Toc58328672)

[C. Design of Data Analytics Solution 4](#_Toc58328673)

[C1. Hypothesis 4](#_Toc58328674)

[C2. Analytical Method 5](#_Toc58328675)

[C2a. Justification of Analytical Method 5](#_Toc58328676)

[C3. Tools and Environments of Solution 5](#_Toc58328677)

[C4. Methods and Metrics to Evaluate Statistical Significance 5](#_Toc58328678)

[C4a. Justification Of Methods and Metrics 5](#_Toc58328679)

[C5. Practical Significance 5](#_Toc58328680)

[C6. Visual Communication 6](#_Toc58328681)

[D. Description of Datasets 6](#_Toc58328682)

[D1. Source of Data 6](#_Toc58328683)

[D2. Appropriateness of Dataset 6](#_Toc58328684)

[D3. Data Collection Methods 6](#_Toc58328685)

[D4. Data Quality 6](#_Toc58328686)

[D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance 6](#_Toc58328687)

[D5a. Precautions 6](#_Toc58328688)

[E. Sources 6](#_Toc58328689)

# Project Overview

# A1. Research Question or Organizational Need

This project will be used to answer the question what the top three dog breeds with the most desirable characteristics for a dog breeder to start breeding. This data analyze will process the data needed to discover the top three dog breeds from seven categories and then the top three dog breeds overall. This information will be given to the breeder to allow them to choose the next dog breed they will breed to sell.

# A2. Context and Background

A dog breeder has decided to expand their business by selecting a new breed of dog to start breeding and selling. They have asked for an analysis to be done to compile the top three dog breeds using characteristic comparison. The breeder wants to know what characteristics are most sought after and which are not. These characteristics will be used to compare the dog breeds in each of the seven categories to obtain the top three dogs. These top three dogs from each category will be compared to obtain the top three dogs overall. The breeder can then use this list to choose which dog breed will be the most profitable for them to begin breeding.

# A3. Summary of Published Works

In this section summarize the content of the 3 articles related to the project.

# A3a. Relation of Published Works to Project

In this sub-section, discuss how each published work relates to the project and its planning.

# A4. Summary of Data Analytics Solution

A dog breeder who would like to buy a new breeding pair of dogs and would like to have a small selection to choose from. This project will take a dataset of dog breeds and clean up the data to allow an easier analysis of the information. The cleaned data will be compared to find the top several dogs of each of the seven categories: herding, hound, non-sporting, sporting, terrier, toy and working dogs. These dogs will be compared using eight different characteristics, four positive and four negative. This will allow for a comparison of the dog breeds to discover the top three dogs in each category. The top three overall dog breeds will be chosen from the list of top three dogs in each category. By using this analysis, the dog breeder will be able to chose from a list to select the most profile dog to begin breeding.

# A5. Benefit to Organization and Decision-Making Process

The benefit of this analysis is the dog breeder being given a list of choices to choose the most profitable dog to begin breeding. The dog breeder will not need to do all the research and analysis on their own, which can be time-consuming and could possibly lead them to making an incorrect choice. By outsourcing the analysis, the dog breeder will get top dog breeds and be able to make an educated decision on which dog to begin breeding.

# Data Analytics Plan

# B1. Goals, Objectives, and Deliverables

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. Scope of Project

The scope of this project is 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. The scope of this project does not include every single dog breed in the world. There are just too many variables to consider when including that many dogs. The dataset used has a large number, enough to be viable for this project.

# B3. Standard Methodology

This project will use the SEMMA methodology. This methodology consists of five steps – 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 choose the data file best\_in\_show from Kaggle which contained enough information to work with in this data analysis.

Explore: In this step, I explore 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 use Python programming language to clean up the data. The data could have many cleaning and tidying issues to get it ready for assessing.

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

Assess: In this step, I will use Microsoft Excel to compile all the relevant characteristics of the top seven dog breeds to discover the top three dogs in each of the categories. Then I will take the list of top dogs and find the top three dogs overall.

# B4. Timeline and Milestones

Present a table showing for each milestone its projected start and end dates, and its projected duration:

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

# B5. Resources and Costs

|  |  |
| --- | --- |
| **Personal, technology, or infrastructure** | **Cost** |
| Jupyter Notebook – Python code | N/A |
| Tableau | N/A |
| Microsoft Excel | N/A |

The resources for this project are Jupyter Notebook, Tableau, and Microsoft Excel. These platforms are all free to the public and will accrue no cost to me or the dog breeder.

# B6. Criteria for Success

The criteria for success I have chosen for this project as the output from the three different applications I will use to discover the answer to my question. If all these criteria are met, the project will be deemed a success.

Once the dataset has been clean in Jupyter Notebook, the data will need to be able to be processed in Tableau. If the data is easily put into graphical form in Tableau, the criterion is a success.

The data must be easy to read in graphical from in Tableau. If this is so, the criterion is a success.

The final criterion is to output the top dog breeds overall for the breeder to choose from. If there is a list of top dog breeds for the breeder to choose from, the criterion is a success.

|  |  |  |
| --- | --- | --- |
| **Criterion/Metric** | **Required Data** | **Cut Score for Success** |
| Data cleaned for analysis? | Output of Jupyter Notebook | Success is if and only if all data is able to be processed |
| Clear output in graphs? | Output of Tableau | Success is if and only if all data is clear to read |
| Top three dogs produced? | Output of Microsoft Excel | Success is if and only if there is an end output of top dog breeds |

# Design of Data Analytics Solution

# C1. Hypothesis

Clearly state the hypothesis/need identified in A1.

What are the top three dog breeds with the most desirable characteristics for a dog breeder to start breeding.

# C2. Analytical Method

Clearly identify the descriptive or predictive method to be used.

Examples of acceptable descriptive methods:

* K-means clustering
* Hierarchical clustering
* Other clustering methods (max distance, min distance, etc.)
* PCA (all variables need to be numeric)
* MCA (all variables need to be factorized)
* FAMD
* Logistic regression with interpretations of estimated coefficients
* Multi-linear regression with interpretations of estimated coefficients

Examples of appropriate non-descriptive methods:

* Logistic regression
* Decision trees
* Random forest
* Neural network
* Multi-linear regression (if they provide a strong justification – like ease of interpretability)

The data analysis method that I will use for this project will be descriptive analysis.

# C2a. Justification of Analytical Method

Discuss why you chose this method. What makes it appropriate for this data and the question it is intended to address.

# C3. Tools and Environments of Solution

Identify the tool to be used for data extraction and why it is selected over other tools.

# C4. Methods and Metrics to Evaluate Statistical Significance

What are the methods and metrics you will use in the tool, and how is statistical significance determined?

# C4a. Justification Of Methods and Metrics

Justify the appropriateness of these metrics and methods as the correct ones to use.

# C5. Practical Significance

Statistical significance indicates the existence of an effect, while practical significance deals with the magnitude of the effect. How practical is this result in the context of decision-making? Present examples of how this result will matter in practical use.

# C6. Visual Communication

What graphical representations will you use to communicate the findings? Why use these over others that may be chosen?

# Description of Datasets

# D1. Source of Data

List the dataset(s) used for analysis. List them in **E. Sources** and cite them here, all in APA format.

# D2. Appropriateness of Dataset

Justify the choice of this dataset, or these datasets, as appropriate for addressing the project goal(s).

# D3. Data Collection Methods

Offer a discussion of how data was collected, including the steps taken to insure accuracy and quality. What are the advantages and disadvantages of collecting data using this methodology? Also discuss any challenges in the collection process and how they were overcome.

# D4. Data Quality

Discuss issues with the quality of the data and the need for interventions. Missing data, outliers, formatting issues, dirty data, and other data quality categories could be discussed.

# D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance

These considerations could relate to the problem/need, data sensitivity and nature, the industry within which the problem/need exists, etc.

# D5a. Precautions

Discuss guidelines related to data preparation, analysis, storage, access, and dissemination.

# Sources

Michael Worboys, Julie-Marie Strange, & Neil Pemberton. (2018). *The Invention of the Modern Dog: Breed and Blood in Victorian Britain*. Johns Hopkins University Press.

Ghirlanda, S., Acerbi, A., Herzog, H., & Serpell, J. A. (2013). Fashion vs. Function in Cultural Evolution: The Case of Dog Breed Popularity. *PLoS ONE*, *8*(9), 1–6. https://doi.org/10.1371/journal.pone.0074770