DOI: 10.37943/21SUAS7119

Aruzhan Shoman
PhD, Scientific Director of the Research Center AgroTech
a.shoman@astanait.edu.kz, orcid.org/0000-0002-7844-8601
Astana IT University, Kazakhstan

Assel Smaiyl
PhD, Assistant Professor, Department of Computer Engineering
assel.smaiyl@astanait.edu.kz, orcid.org/0000-0002-6215-932X
Astana IT University, Kazakhstan

Aliya Kalykulova
Master's student, Research Center "Big Data and Blockchain Technologies"
aliyakalykulova@mail.ru, orcid.org/0009-0006-5641-3797
Astana IT University, Kazakhstan

Aliya Nugumanova
PhD, Director of the Research Center "Big Data and Blockchain Technologies"
a.nugumanova@astanait.edu.kz, orcid.org/0000-0001-5522-4421
Astana IT University, Kazakhstan

Aidar Mukhametkaliyev
Master of Veterinary Sciences, Assistant of the Department of Clinical Veterinary Medicine
mukhametkaliyev.aidar@kaznaru.edu.kz, orcid.org/0009-0009-4261-5330
Kazakh National Agrarian University, Kazakhstan

LEVERAGING BIG DATA FOR DOG HEALTH ANALYSIS: AN EXPLORATORY STUDY USING "TANBA" IN KAZAKHSTAN

Abstract: In the era of artificial intelligence, collecting and analyzing data about dog health through electronic medical cards and passports has become a key factor in improving the quality of life for pets. In this study there was analyzed 93,922 records about dogs contained in Kazakhstan's pet registration information system "Tanba". The research focused on the demographic characteristics of dogs, including breed, age, and region of residence. Explanatory Data Analysis was conducted using descriptive statistics, and Natural Language Processing (NLP) methods were applied to standardize breed names, improving data consistency. Additionally, an ANOVA test was performed to assess the impact of factors such as gender, region, breed, and breed size on dogs' lifespan. Based on the data analysis, there are highlights of key aspects such as the predominance of young dogs (average age 5.52 years), the high proportion of dogs without breed, and the high concentration of stray animals in some regions, which emphasizes the need for increased efforts to control the population and improve living conditions for stray dogs. This study presents an analysis of the dog population for 2024 based on data from the Tanba national registration system. Unlike previous studies that focused on the prevalence of individual diseases or were limited to data from specific regions, this study covers the entire country and provides a general overview of the dog population. The findings indicate a high proportion of mixed-breed and stray dogs in Kazakhstan, as well as significant regional differences in canine lifespan. Breed and regional factors have a statistically significant impact on lifespan, emphasizing the importance of considering these characteristics

Received: 20.11.2024 Accepted: 24.03.2025 Published: 30.03.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

when developing programs to improve animal welfare and veterinary care. In the future, it is planned to improve data processing algorithms and expand the use of additional sources of information, which will allow for more accurate assessment of dog health risks and development of more effective preventive measures.

Keywords: Electronic records of non-productive animals, Veterinary medicine, Explanatory data analysis (EDA), Big data.

Introduction

With economic growth, more people are choosing to adopt pets, and, as noted in [1], owners are no longer limited to simply feeding and walking their pets but aim to provide them with a comfortable and high-quality living environment. This has led to an increasing demand among pet owners for high-quality veterinary and household services (such as grooming and spa treatments), various accessories, enrichment products, and nutritious, tasty foods. With the growth in volume and diversity of services in the pet industry, the need has arisen for effective management of the accompanying information and data generated through interactions between owners and various services, salons, clinics, and more. This creates a demand for efficient data management systems, including electronic health passports for dogs, recommendation platforms, specialized websites, and online forums. These tools allow for the accumulation and analysis of large volumes of data, giving pet owners access to quality services and support in pet care, expert consultations, and the opportunity to exchange experiences with other owners. For service providers and veterinary clinics, they provide essential data on pets, such as health conditions, vaccinations, diets, habits, and more.

On the other hand, traceability systems for dogs, based on mandatory electronic identification and registration in a digital database, are among the most effective ways to enhance the utility of dog population management and disease control programs [2]. In Kazakhstan, the implementation of dog registration information systems is especially significant due to the vast territory and variety of climate zones, which affect the spread of various diseases. Through mandatory electronic registration and identification, relevant authorities can access up-to-date data to respond promptly to epidemiological threats and develop preventive measures. Additionally, these systems facilitate the return of lost or stolen animals to their owners and increase owner responsibility for pet care.

This study examines the Kazakhstani pet registration information system, Tanba [3], from which 194,914 records of microchipped dogs (both domestic and stray) were extracted at the time of inquiry. An explanatory data analysis was conducted to identify the most common dog breeds and age groups, as well as the distribution of animals across regions of Kazakhstan.

In the context of Kazakhstan, previous studies focused on analyzing the prevalence of individual diseases or limited data for specific regions. This study provides a comprehensive analysis of dog demographic data for 2024 based on the national Tanba registration system covering the entire country.

The structure of the work is as follows. A review of related works is presented in Section 2, while the materials and methods of the research analysis are described in Section 3. Section 4 discusses the results obtained, and the article concludes with the main findings.

Related works

Technologies for regulating and monitoring pet health are rapidly evolving due to the introduction of innovative areas such as artificial intelligence (AI), big data and the Internet of Things (IoT). These technologies are used to monitor the health status of pets for timely response when diseases occur, as well as to manage pet populations, including stray pets, and provide diagnostic advice. In most cases, IoT systems are involved in collecting large amounts of data, which is then analyzed and interpreted with the help of AI.

IoT technologies and big data. A variety of IoT devices such as automatic feeders, drinkers, toilets and multifunctional collars are used to assess the health of animals. These devices record various parameters such as heart rate, sleep duration, activity level, food and water intake, and other important metrics. In article [4] presented a diagnostic Health Score system that analyzes dog behavior using data from collars, including scratching, licking, swallowing, and sleep frequency. From this data, the system generates an AI-assisted Health Score that matches veterinarians' diagnoses at 87.5%.

However, these systems often analyze the condition of the animal in different aspects separately from each other. According to the article [5], [6], such technologies face a number of limitations: some devices have limited Bluetooth range, animal cameras do not always recognize all behavioral cues, and automatic feeders and drinkers do not always consider individual nutritional needs. In this regard, the authors [5] propose an improved IoT-based technology that considers all system data in an integrated manner. This approach can significantly improve accuracy in animal condition assessment.

Optimizing nutrition. Proper nutrition of pets is also an important part of their health and requires careful attention. When designing a nutritional system, it is necessary to consider the characteristics of each breed of animal and calculate the required amount of nutrients. Currently, AI is used to calculate feed dosage with metabolic energy [7] or to predict growth and its impact on animal health [8].

Nutritional management of animals with chronic diseases is an important aspect. Proper nutrition plays a key role in slowing the progression of chronic kidney disease (CKD), which is a common disease in aging pets. CKD in dogs and cats leads to impaired kidney function, which can cause metabolic changes such as elevated urea and creatinine levels, markers of deteriorating kidney function [9].

The use of AI to predict feed composition is still a challenge. According to [8], mechanistic models, which are based on clear principles and allow control and adjustment of parameters through experimental data, are more often favored. Also relevant is the issue of cost optimization in food production, where waste can be used to create feed. However, this requires a lot of experimentation to ensure feed efficiency and storage, as well as considering organoleptic characteristics.

In addition, there is interest in different nutrient extraction methods such as protein hydrolysates, which help to improve feed digestibility and increase nutritional value [10]. Sosa-Holwerda et al (2024) noted that the role of AI in nutrition is still under development. The main focus is on nutritional assessment, and to a lesser extent on predicting malnutrition, lifestyle interventions and the study of nutrition-related diseases [11].

Animal population control. Scientific studies demonstrate that stray dogs have three times the risk of viral infections compared to domestic animals [12]. Factors such as overpopulation and pollution cause stress, weakening the immune system and increasing the likelihood of viral and bacterial infections. These data emphasize the importance of systematic monitoring of the stray dog population, timely vaccination and implementation of preventive measures to minimize epidemiological risks.

With age, the efficiency of the immune system of dogs decreases, making them more susceptible to infections, inflammatory processes and cognitive impairment [13], [14]. According to a study [15], the average age of dogs has increased by 0.5-1 year between 2013 and 2019, indicating a trend of increasing longevity. However, the increase in the number of older dogs is associated with an increased risk of infectious diseases, which may pose a threat to epidemiological safety.

Sterilization and vaccination are the most effective measures to regulate the dog population in the long term, while capture and euthanasia methods show low efficacy and are only effective in the short term [16], [17].

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

A mobile platform, HSIApps, has been developed for dog population management in India [16]. It optimizes the registration, sterilization and vaccination of stray dogs, and tracks their population by region using GPS coordinates. However, the effectiveness of this technology depends on the human factor, as data collection and entry are done manually.

A study [17] assessed the number of dog owners, registration rates and factors affecting dog ownership in Italy. It was found that about 30% of owners do not register their pets. The largest number of dog owners live in rural areas and among families with children. Regular visits to the veterinarian are more common among owners who buy commercial food compared to those who prepare their own. Control of the number of dogs contributes to the timely assessment of the epidemiological situation and taking the necessary measures in case of disease outbreaks.

Modern technologies allow monitoring and analyzing data from livestock farms and holdings [18]. Time series analysis and artificial intelligence-based forecasting methods are used to detect disease outbreaks and generate hypotheses, which significantly improves the accuracy of epidemiological control.

Telemedicine and AI chatbots. Pet owners do not always have the opportunity to promptly seek help from a veterinary specialist. Telemedicine and AI chatbots are actively used for pre-diagnosis of animal diseases. After the COVID-19 pandemic, the demand for remote veterinary consultations has increased significantly. However, according to a study [19], 68% of veterinarians believe that remote consultation should only take place after a face-to-face examination of the animal. On the other hand, 79.2% of pet owners are willing to use telemedicine services but are aware of their limitations [20].

Despite the convenience of using AI chatbots to pre-diagnose diseases, there is a risk of making incorrect diagnoses and prescribing the wrong treatment. To minimize such risks, it is necessary to educate animal owners about the possibilities and limitations of AI in veterinary medicine [21].

Findings on Canine Diseases in Kazakhstan. Canine infectious diseases remain a significant concern in Kazakhstan, particularly in regions with specific climatic conditions that may contribute to the spread of infections. Several studies have analyzed the epidemiology of the most common diseases affecting dogs, including rabies, canine distemper, and parvoviral enteritis. These diseases exhibit seasonal patterns and tend to be more prevalent in certain regions, emphasizing the need for continuous monitoring and preventive measures.

One of the most critical zoonotic diseases in Kazakhstan is rabies, which is actively monitored. According to [22], from 2013 to 2021, the highest incidence of rabies was recorded in southern and eastern Kazakhstan, affecting various categories of animals.

A study by [23] titled “Monitoring of Rabies Outbreaks in Kazakhstan” examines cases of human rabies infections from 1997 to 2015. The findings indicate that the highest concentration of outbreaks occurred in the South Kazakhstan region, underlining the persistent risk of rabies transmission in this area.

Another study, [24], focuses on canine distempering, describing laboratory investigations into the disease. The results demonstrate a seasonal distribution, with peak incidence occurring in spring and autumn, while summer has the lowest number of cases. Among the most affected breeds are German Shepherds, Central Asian Shepherd Dogs, Tazy, and Tobet.

Similarly, research [25] on canine parvoviral enteritis (CPV) presents veterinary clinic data from 2015 to 2020. The virus follows a seasonal pattern, with peak infection rates observed in April-May and October. The most susceptible breeds include German Shepherds, Rottweilers, and Collies.

The reviewed studies suggest that seasonal climatic conditions, particularly high humidity and rainfall in spring and autumn, contribute to the increased prevalence of infectious diseas-

DOI: 10.37943/21SUAS7119

es in dogs. Additionally, the South Kazakhstan and East Kazakhstan regions emerge as epidemiological hotspots, likely due to their environmental and climatic characteristics, which may facilitate disease transmission.

To control canine diseases, it is essential to use monitoring technologies such as GIS systems, big data analysis, and machine learning for outbreak prediction. Implementing electronic animal tracking can help monitor populations, identify infection hotspots, and improve prevention and veterinary control.

Methods and Materials

General information on data structure and preparation. To conduct an explanatory data analysis on dogs in Kazakhstan, data was obtained from the Tanba website, an animal registration information system. This source provides open data on registered animals in Kazakhstan, including species, breed, gender, location, date of birth, passport number, and more. Table 1 presents information on the structure of the table retrieved from Tanba.

Table 1. General information on the structure of the table containing data on dogs in Kazakhstan (without duplicates)

Attribute	Not null values	Null values	Number of categories
Chip Number/ID	131 375	63 539	unique
Animal Gender	194 914	0	2
Animal Breed	142 310	52 604	346
Passport Number	194 914	0	unique
Animal Registration Date and Time	194 914	0	data
Animal Birth Date	100 781	94 133	data
Animal Status	194 914	0	14
Ownership Status	194 914	0	5
Animal Registration Type	194 914	0	6
Marking Location Name	191 163	3 751	348
Marking Location	191 163	3 751	14
Animal Marking Date	194 276	638	data
Region/District	194 790	124	204
Vaccination Date	32 555	162 359	data
Sterilization/Castration Date	19 842	175 072	data

In addition to duplicates, the data also contained erroneous birth information. For example, some dogs' ages exceeded 31 years (the age of the oldest dog in the world), or, conversely, the birth date was set in a future time, as if from the future. Therefore, for analysis, only entries where the age ranged from 0 to 25 years were included. Age was calculated based on the birth date by subtracting it from October 7, 2024, the date when the data was scraped from TANBA. Dogs were also filtered by animal status. The analysis included dogs with the following statuses:

- Active
- In transit
- In quarantine
- In transfer
- Awaiting marking
- Awaiting owner confirmation
- Awaiting owner instruction
- Created
- Created by owner

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

Dogs with statuses “Retired,” “Deceased,” “Lost,” “Disposed,” and “Draft” were excluded from the data analysis to focus on the current situation of dogs in Kazakhstan. After filtering, the number of records totaled 93,922, with 100,992 rows removed from the table, accounting for just over half of the original data.

To simplify the analysis and reduce the diversity of categories, standardization and unification processes were applied to the attributes “Breed” and “Region/District”. The process involved the use of natural language processing (NLP) techniques and some manual corrections. Different systems and platforms use various spelling variants for the same breed, which complicated the analysis. For example, in TANBA, the same breed’s name could be written in several variants, which also made the data processing more difficult. Moreover, the diversity of breeds added additional challenges, so measures were taken to standardize breed names. There are different methods for grouping breeds, such as based on their origin, like American (AKC standard, American Kennel Club) or British (The Kennel Club) breeds. However, in Kazakhstan, there are breeds that cannot be clearly classified according to these standards. Therefore, standardization of breeds was carried out based on their general names.

Dog breeds were simplified into more general categories. For instance, “German Shepherd” and “East European Shepherd” were combined under the name “Shepherd.” A dictionary was created for this task, where the key is the general breed’s name, and the elements are the breed names presented in TANBA (an example is shown in Table 2).

Table 2. Example of a dictionary for standardizing breed names from TANBA

General name	Breed name from TANBA
Hound	Polish **hound** (Polish **hound**)', 'Bernese **hound**', 'Bulgarian **hound**'
Terrier	Border **terrier**', 'Australian **terrier**', 'Irish **terrier**', 'Biewer Yorkshire **terrier**'
Spitz	German **spitz**', 'Samoyed (Samoyed **spitz**, Sammy, dog breed)', '**Spitz**'
Laika	West Siberian **Laika**', 'East Siberian **Laika**', 'Yakutian **Laika**'
Bulldog	Alapaha **Bulldog**', 'Campbell **Bulldog**', 'Old English **Bulldog**', 'English **Bulldog**'

Figure 1 shows the algorithm for standardizing breed names. All data were converted to lowercase and using part-of-speech tags provided by the SpaCY model (the model for processing Russian text, used for TANBA data — “ru_core_news_sm”), unigrams corresponding to nouns (NOUN) and proper nouns (PROPN) were selected from each breed. At this stage, the obtained unigrams were manually checked and became the general breed names. If words such as “dog” or “breed” were found in some breed names, which occurred in certain breed names from TANBA, these anomalies were excluded to improve the accuracy of processing.

DOI: 10.37943/21SUAS7119

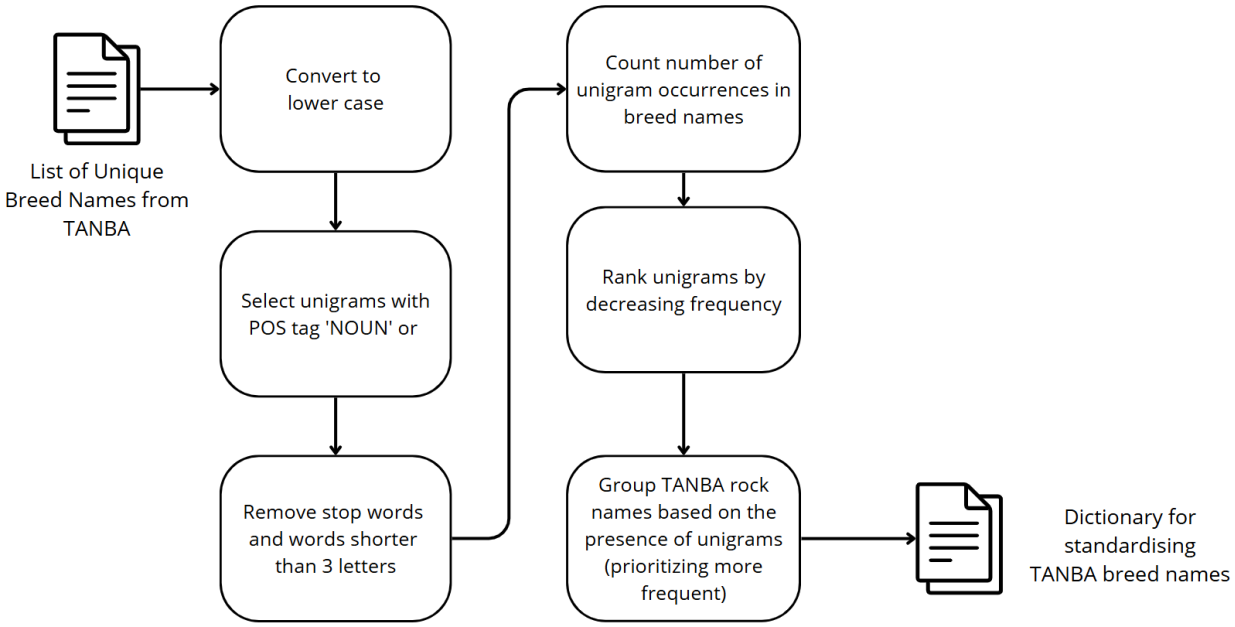

Figure 1. Algorithm for standardising TANBA breed names

Next, a count was made of how many breed names from TANBA contained each unigram. Table 3 shows an example of the 10 most frequent unigrams occurring in TANBA breed names. Breeds were grouped, starting with the most frequently occurring unigrams. If a unigram appeared in a breed name from TANBA, that breed was added to the dictionary, where the key was the general breed's name, and the elements were the specific breeds from TANBA. One unique breed name from TANBA could only be associated with one general name. Thus, the breed names of dogs were standardized.

Table 3. Frequency of unigrams occurrence in the dog breed names

Unigram	Count
shepherd	36
terrier	29
spaniel	20
hound	16
spitz	10
bulldog	10
laika	10
pointer	8
greyhound	8
setter	6

This reduced the number of categories for the "Breed" attribute from 346 to 114 (plus mixed breeds). However, this approach has several limitations. It is not entirely accurate for grouping breeds, as more accurate grouping requires considering additional external and genetic characteristics of the dogs. Furthermore, many breeds retained their unique names. As a result, 34 groups were created, including 266 breed names, while 80 breeds could not be grouped by their names.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

The “Region/District” attribute contained different administrative levels (region, district, city), so they were standardized to larger territorial units: regions and cities of republican significance. Cities of republican significance were treated separately, as their data significantly differs from the region they belong to, which could complicate data analysis. To standardize the region names, data [26] available on stat.gov.kz were used, which provides information on the administrative-territorial units of Kazakhstan. This dataset includes the names of regions, as well as the districts and cities within each region. Cities of republican significance are also listed separately, with information on the districts that fall under each city. As a result, the number of categories was reduced from 204 to 20.

Explanatory data analysis. According to the data, the average age of all dogs registered with TANBA is 5.52 years. The average age of purebred dogs is 3.9 years. A distribution of dogs by age is shown in Fig. 2. The age distribution is left-skewed, indicating a predominance of young animals. Most of the data is concentrated in the 0 to 5-year age range, with a gradual decrease in the number of older animals. Notable peaks around 7 and 10 years suggest the presence of several age cohorts. Animals over 15 years are rare, with numbers dropping significantly, especially between 20 and 25 years, which can be considered age outliers.

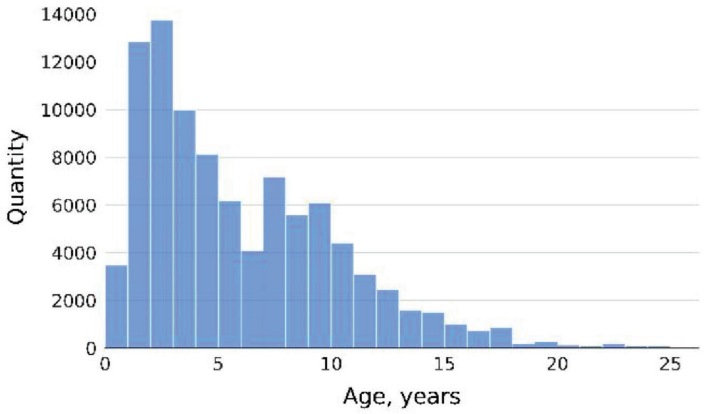

Figure 2. Distribution of dog ages

In addition to age characteristics, the analysis also revealed a significant predominance of males in the sample: they account for 62.6%, while females account for 37.4%. This ratio (about 2/3 in favor of males), reflected in Fig. 3, indicates an imbalance between the sexes, which may indicate accounting patterns or demographic trends in the animal population.

Fig. 4 shows the demographic pyramid of dogs, which illustrates the age and sex distribution. The data were divided into age groups with an interval of 5 years. In total, there were 5 age categories. The largest number of dogs is in the age category from 0 to 5 years. In this group, males account for 32.4% of the total, while females account for 18.9%. The next largest group is 6-10 years old, where the proportion of males is 18.1% and females are 12.9%. In older age categories, the number of dogs decreases significantly: 11-15 years old (9.2% for males and 4.7% for females), 16-20 years old (2.4% for males and 0.9% for females), 21-25 years old (0.4% for males and 0.2% for females). In all age categories, there is a greater number of males compared to females, especially in younger age groups. With age, the number of dogs of both sexes decreases, and the decline occurs faster in females than in males, especially after 10 years.

DOI: 10.37943/21SUAS7119

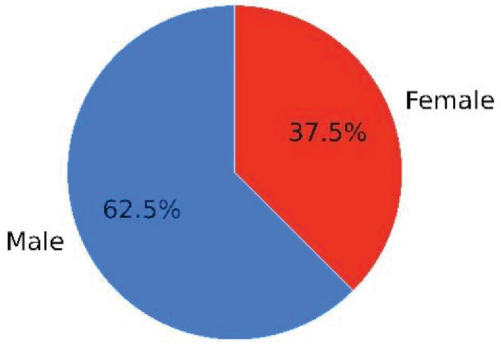

Figure 3. Gender ratio of dogs

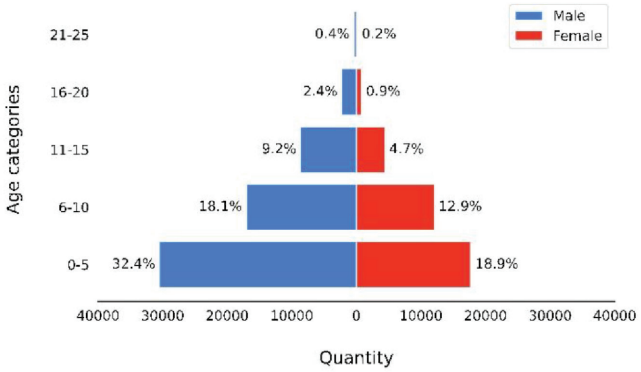

Figure 4. Distribution of dogs by age and gender

Analysis of the distribution of dogs by breed and region. The distribution of the total number of dogs by percentage of mongrel, purebred, unknown and mestizo is shown in Fig 5. The largest group is non-pedigreed dogs (38.9% of the total). The category with an unknown breed ranks second with a share of 38.4%, which is almost identical to the first indicator. The number of pedigree dogs is almost 2 times less than non-pedigreed dogs, which is 17.9%. The smallest share is occupied by "mestizo" with 4.8%, which may indicate the relatively rare presence of this group.

Fig. 6 shows the distribution of all dogs by region of Kazakhstan. The largest number of dogs was registered in the city of Almaty (more than 40,000), which is significantly higher than in other regions. This outlier is explained by the fact that almost all dogs with an unknown breed were registered in the city of Almaty, according to Fig. 7. Turkestan region and the city of Astana are next in number, where about 7,000 to 10,000 dogs are registered. In Aktobe, Kyzylorda, North-Kazakhstan, Almaty, Zhetysu, Abay and East Kazakhstan regions, about 3,000 to 5,000 dogs are registered. In the rest of Kazakhstan, less than 3,000 dogs are registered.

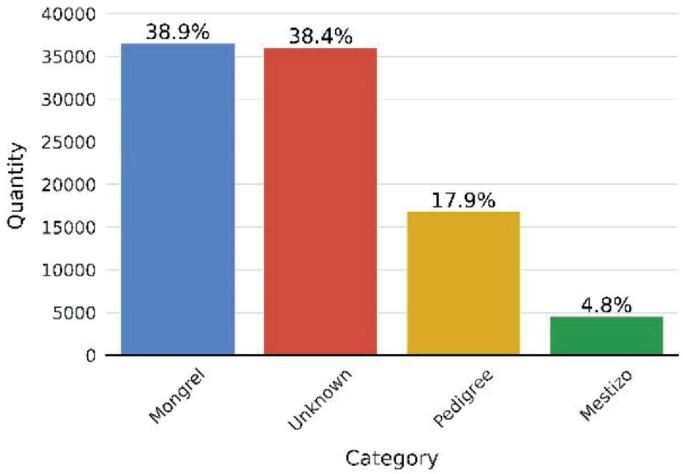

Figure 5. Total number of dogs by category with percentage

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

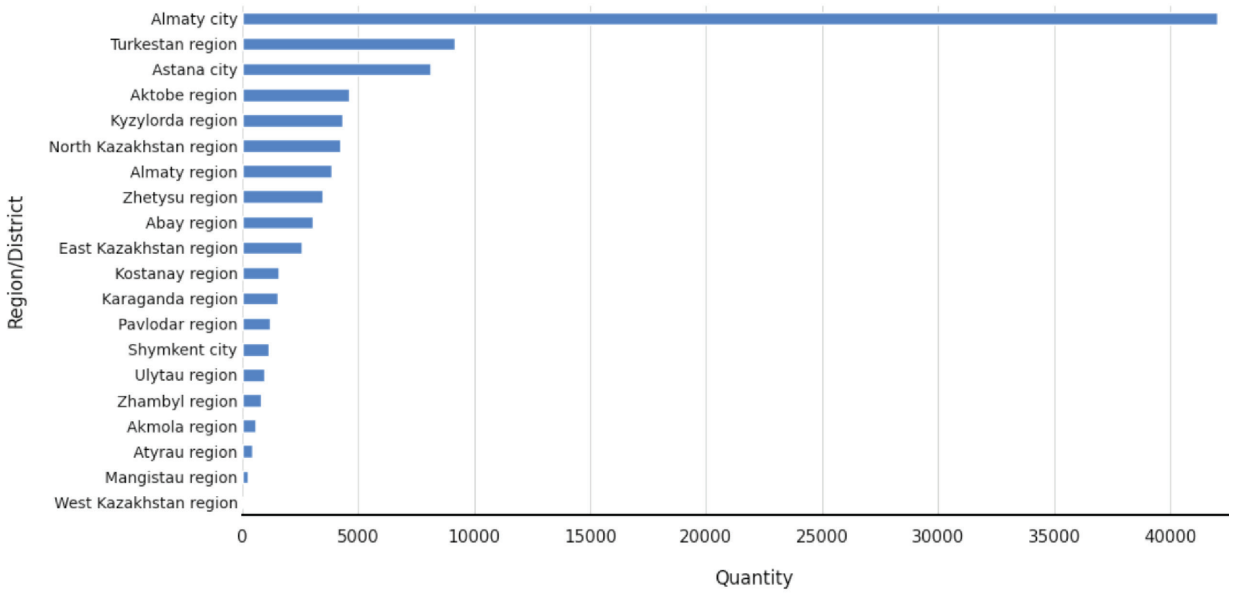

Figure 6. Distribution of dogs by region

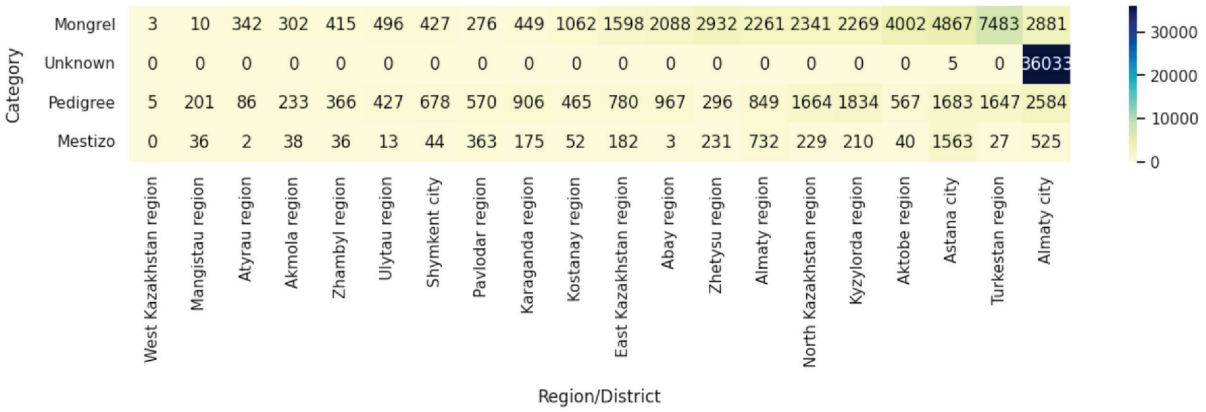

Figure 7. Heat-map of distribution of porous, non-pedigreed, mestizo and unknown dogs by regions of Kazakhstan

Fig. 8 shows the distribution of purebred, mongrel and mixed (mestizo) dogs by region. The largest number of mongrel dogs is registered in the Turkestan, Aktobe regions and the city of Astana. Pedigree dogs are most registered in the city of Almaty, as well as in Kyzylorda, North Kazakhstan, Turkestan regions and the city of Astana. Most of the mestizos were recorded in the city of Astana.

DOI: 10.37943/21SUAS7119

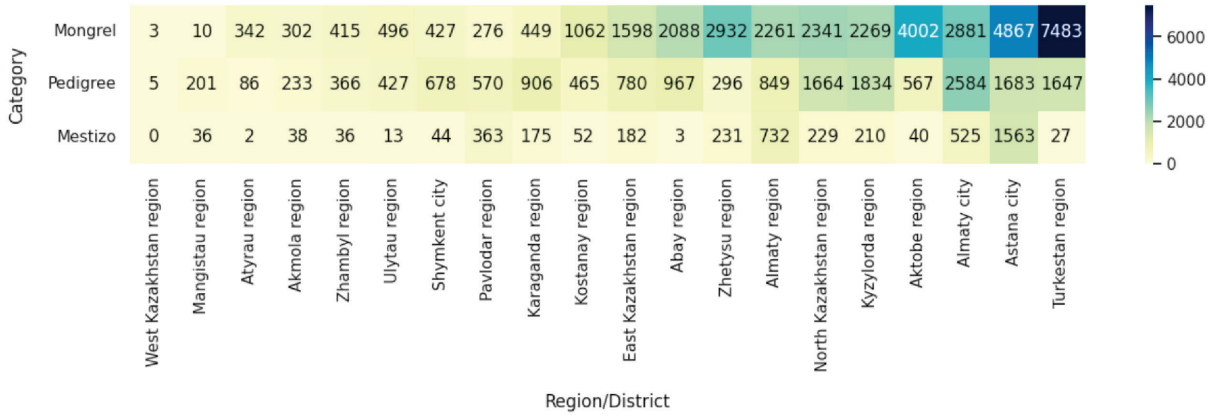

Figure 8. Heat-map of distribution of pedigree, mestizo and mongrel dogs by regions of Kazakhstan

Fig. 9 shows the distribution of dog breeds (15 with the highest frequency of occurrence) by region. The most popular breeds in Kazakhstan are the shepherd, terrier, greyhound, and Tibetan. The largest number of purebred dogs are found in Almaty, Kyzylorda region, Turkestan region, Astana and North Kazakhstan region. Some dog breeds are common in certain regions. The greyhound is most common in Abay and Turkestan regions, and the Tibetan in Kyzylorda and Turkestan regions. The largest number of terriers are registered in Almaty.

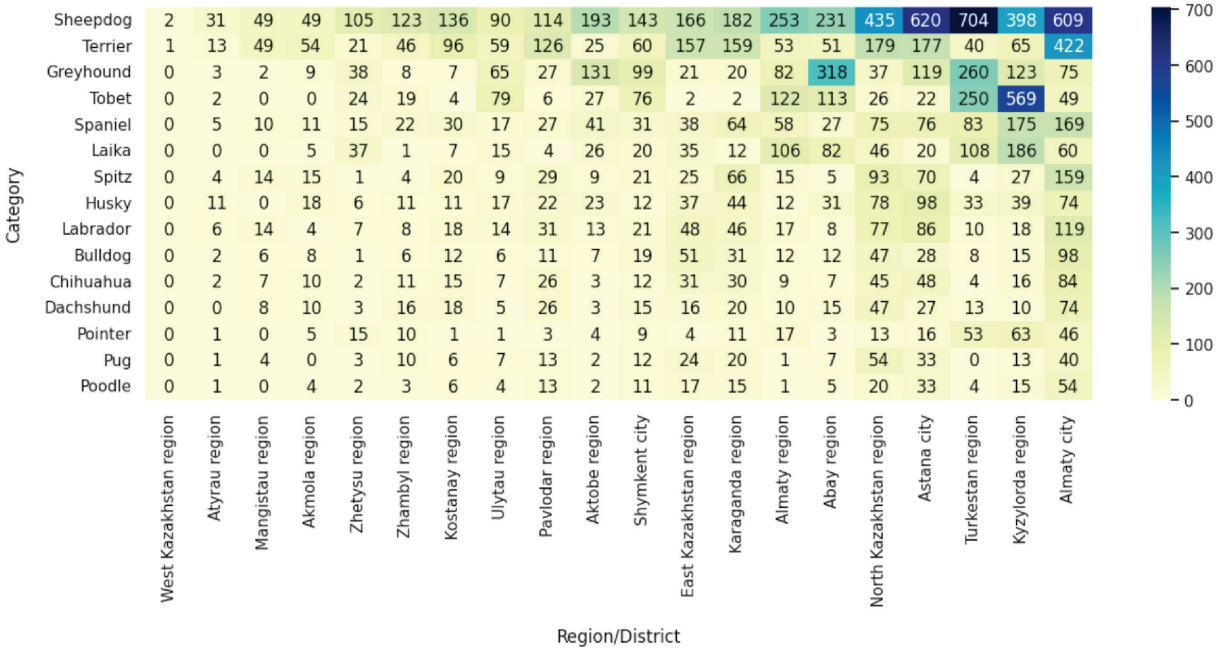

Figure 9. Heat-map of the distribution of dog breeds by regions of Kazakhstan (only purebred)

Figure 10 presents the distribution of dogs by breed size, including only purebred animals. The study categorized breeds into three groups: small, medium, and large. The size of each breed was determined manually based on data from sources [27, 28] by matching general breed names to the corresponding category. According to the graph, large dogs make up the majority, accounting for 58%, while small breeds represent 33%. Medium-sized purebred dogs are the least common, with a share of only 8.6%.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

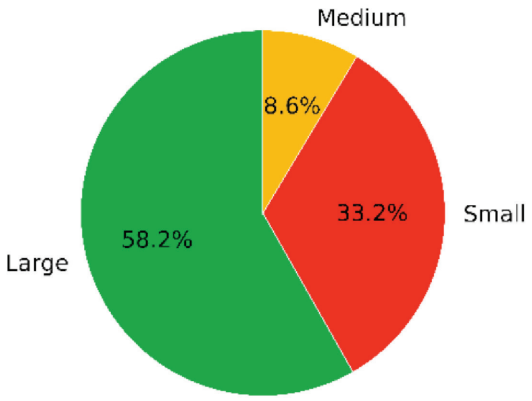

Figure 10. Ratio of purebred dogs by size

Analysis by ownership status. Figure 11 shows the distribution of dogs by their ownership status and region. Most dogs have an owner, especially in regions such as Almaty city (41,730 dogs), Turkestan and Kyzylorda region. The highest number of stray dogs is observed in Astana city (4,764 dogs) and Aktobe region. Responsible persons and volunteers are mostly registered in small numbers, and their activity in the regions is limited. For example, volunteers are mainly observed in Almaty city (104 dogs). In some regions, such as Mangystau and West Kazakhstan regions, there are significantly fewer dogs with an owner or registered as strays, which may be due to lower population density or other regional characteristics.

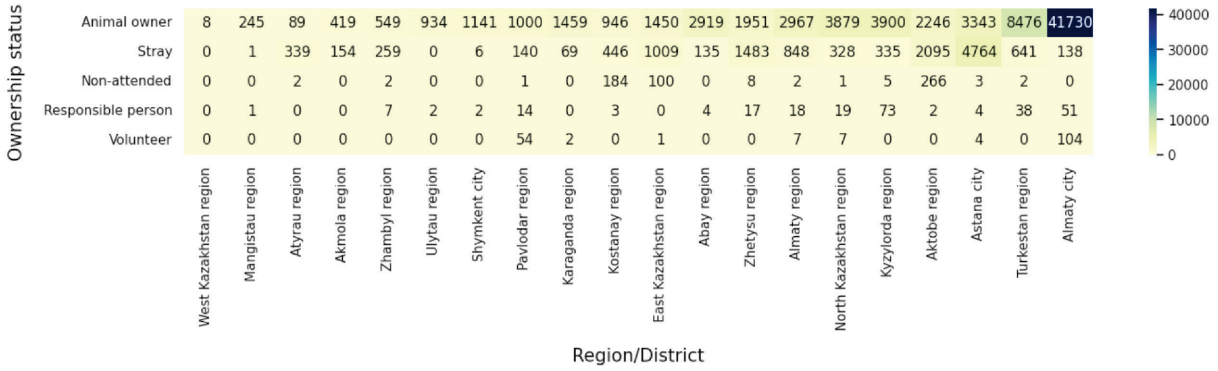

Figure 11. Heatmap by Region and Ownership Status

Figure 12 shows the percentage of ownership statuses. The largest group is dogs with owners, about 85%. Dogs without owners make up 14%. The remaining groups Non-attended (0.6%), Responsible person (0.3%) and Volunteer (0.2%) together make up about 1%.

DOI: 10.37943/21SUAS7119

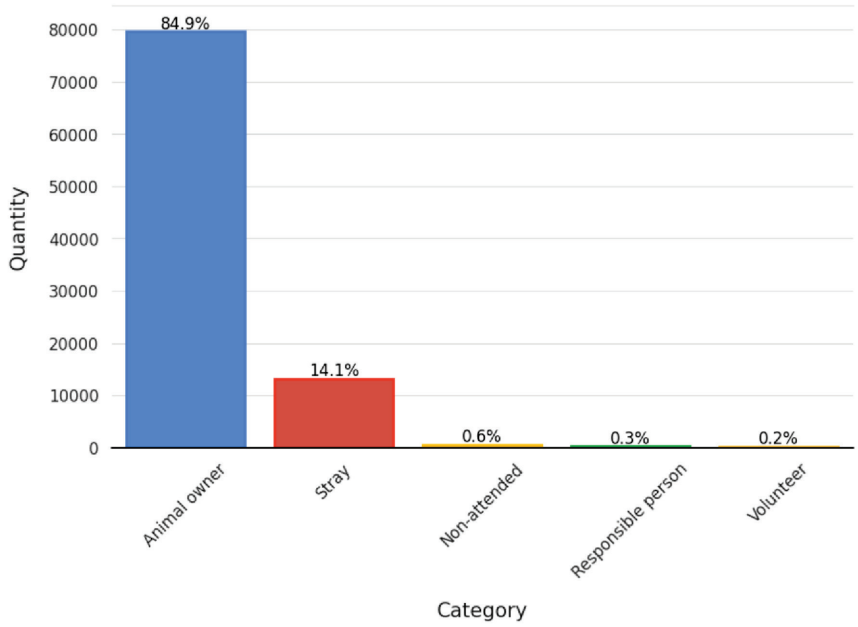

Figure 12. Total number of dogs by ownership status

Analysis of factors influencing dog age. To investigate the influence of categorical factors on the age of purebred dogs, a one way Analysis of Variance (ANOVA) was employed. Withing the framework of this statistical approach, th age of the dog was considered the dependent variable, while the independent variables included region, gender, breed, and breed size, categorized as small, medium, large. The analysis was limited to purebred dogs due to the availability of complete data for all considered categories exclusively within this group. The application of ANOVA enables the identification of factors that have a statistically significant impact on the age variable.

The general linear model underlying this analysis is represented as:

$$Y = \mu + \alpha + \beta + \gamma + \delta + \varepsilon \quad (1)$$

where Y denotes the observed age of a dog, μ is the overall mean age across the population, α reflects the effect of region, β the effect of gender, γ the effect of breed, and δ the effect of breed size. The term ε represents the random error component, which is assumed to follow a normal distribution with a mean of zero.

Each of the categorical effects in the model—such as α, β, γ, and δ – is interpreted as the deviation of a specific group's mean from the overall mean. For instance, in the case of gender, the effect of the male group is computed as the difference between the average age of male dogs and the overall average age: $\alpha = \mu_{male} - \mu$. This formalization allows for the assessment of how each group-level characteristic contributes to variance in dog age.

The statistical significance of these effects is tested through the ANOVA procedure, which evaluates whether the mean age differs systematically between groups. The null hypothesis posits that all group means are equal, implying that the considered factors exert no influence on age and that there are no significant differences between group means. Formally, this hypothesis is stated as H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$.

The ANOVA test is based on the computation of the F-statistic, which serves as a measure of the ratio between the variance observed between groups and the variance within groups. Mathematically, the F-statistic is calculated as the quotient of the mean square between groups ($MS_{between}$) and the mean square within groups (MS_{within}). These mean squares are obtained by

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

dividing the corresponding sum of squares (SS) by their respective degrees of freedom, such that the formula for the F-statistic becomes:

$$\mathrm{F} = \frac{MS_{between}}{MS_{within}} = \frac{\frac{SS_{between}}{df_{between}}}{\frac{SS_{within}}{df_{within}}} = \frac{\frac{SS_{between}}{k-1}}{\frac{SS_{within}}{N-k}} \tag{2}$$

$$SS_{between} = \sum_{i=1}^{k} n_i(\mu_i - \mu)^2 \tag{3}$$

$$SS_{within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\mu_{ij} - \mu_i)^2 \tag{4}$$

where k denotes the number of groups, N represents the total number of observations in the dataset, n_i is the number of observations in the i-th group, μ_i is the meaning of that group, μ is the grand mean, and μ_{ij} denotes the j-th observation in the i-th group.

The F-statistics thus estimate the extent to which variability between groups exceeds that within groups. A higher F-value suggests that the differences between group means are unlikely to be due to random variation alone.

To assess the statistical significance of this result, the corresponding p-value is computed. This value represents the probability that an F-statistic equal to or greater than the observed one would occur under the null hypothesis. Formally, it is expressed as:

$$\text{p-value} = \mathrm{P}\ (\mathrm{F} \geq \mathrm{F}_{\text{observed}} \mid \mathrm{H}_0) \tag{5}$$

where, $\mathrm{F}_{\text{observed}}$ is the observed value of the F-statistic, and the probability term indicates the likelihood of obtaining such a value assuming the null hypothesis is true. If the p-value is smaller than a predetermined significance threshold (commonly set at 0.05), the null hypothesis is rejected, indicating that at least one group mean differs significantly from the others. Conversely, if the p-value exceeds this threshold, the null hypothesis is not rejected, implying insufficient evidence to conclude that the groups differ.

Following the ANOVA, a deviation analysis was conducted to explore how the mean age within each category of the examined factors deviated from the overall mean. The deviation for each group was calculated using the formula:

$$d = \bar{\mathrm{x}} - \mu \tag{6}$$

where $\bar{x}$ represents the mean age within a specific category, and μ denotes the overall mean age across the entire sample. To assess whether these deviations were statistically significant, a One-Sample T-Test was applied to each category independently. The t-statistics used in this test is defined as:

$$\mathrm{t} = \frac{\bar{\mathrm{x}} - \mu}{\mathrm{s}/\sqrt{n}} \tag{7}$$

where s is the standard deviation within the category, and n is the number of observations in that group. A p-value less than 0.05 was considered statistically significant, indicating that the mean age within the category significantly differed from the overall mean.

For the analysis of standard deviations and group-level deviations, the focus was placed on the following categorical variables: breed size, region, and breed. The gender variable was ex-

DOI: 10.37943/21SUAS7119

cluded from this stage of the analysis, as the ANOVA had indicated that it does not have a statistically significant effect on age (as discussed in more detail in the Result analysis section).

The breed size category included three levels: small, medium, and large breeds. The regional variable covered all seventeen administrative regions of Kazakhstan, as well as the three republican cities — Almaty, Astana, and Shymkent. Breed-specific analysis was limited to the ten most frequently registered dog breeds in the TANBA system: Sheepdog, Terrier, Greyhound, Tobet, Spaniel, Laika, Spitz, Husky, Labrador, and Bulldog.

The results of the ANOVA identified which of the investigated factors significantly influenced the age of dogs. Notable differences in age distribution were observed based on region, breed, and breed size. The subsequent One-Sample T-Tests provided a more detailed view by revealing which specific categories within each factor demonstrated statistically significant deviations from the overall average age.

Results

General Characteristics of the Dog Population. Data analysis revealed that the sample included 93,922 registered dogs, with an average age of 5.52 years. The age distribution is left-skewed, indicating a predominance of young dogs. The highest number of dogs is concentrated in the 0 to 5-year age group, followed by small peaks at ages 7 and 10. Animals older than 15 years are rare. The analysis by sex showed that 62.5% of the dogs are male, indicating a significant predominance of males among the registered animals. The male-to-female ratio is approximately 2 to 3. In the age structure of the dog population, there is a trend of a more rapid decrease in the number of females compared to males, especially after age 10.

Distribution by Breed. The largest share of registered dogs belongs to the mixed-breed category (38.9%), followed closely by the unknown breed category (38.4%). Purebred dogs represent only 17.9%, while mestizos account for 4.8%. This may indicate the prevalence of stray and unregistered purebred dogs in the country.

Geographic Distribution. Results show that the highest number of registered dogs is found in Almaty (over 40,000), significantly exceeding the figures in other regions. The second largest area is Turkestan, followed by the city of Astana, with 7,000 to 10,000 dogs registered. In other regions, the number of registered dogs is less than 3,000. There is also a high concentration of dogs with unknown breeds in Almaty, which may be related to inadequate registration and identification of stray animals.

Ownership Status. Analysis of ownership status revealed that 14% of registered dogs are strays, primarily concentrated in the Turkestan region, while 85% have owners, mainly located in Almaty, Turkestan, Kyzylorda, North Kazakhstan, and Astana.

Analysis of factors affecting age. Table 4 presents the results of an analysis of variance (ANOVA) assessing the influence of gender, breed, region, and breed size on dog age. The findings indicate that the factors "breed," "region," and "size" have a statistically significant impact on age, as their p-values are below the conventional significance threshold (typically 0.05). In contrast, the factor "gender" is not statistically significant. Among the examined factors, region and breed have the most substantial effect on age.

Table 4. Analysis of variance (ANOVA) of factors affecting dog age

Factor	sum_sq	df	F	p-value	Significance
C(gender)	2.991	1.0	0.350	0.553	Not significant
C(breed)	10953.727	104.0	12.335	1.2e-193	Significant
C(region)	4874.210	19.0	30.043	6.5e-107	Significant
C(size)	101.443	4.0	2.970	0.018	Significant

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

The following section examines the impact of significant factors—size, region, and breed—on dog age. Table 5 presents the deviations in average age across different regions and their statistical significance.

The results indicate that all categories are statistically significant, with p-values below the conventional threshold of 0.05. Small-breed dogs live approximately one year longer than the overall average for purebred dogs. In contrast, medium and large breeds exhibit slightly lower average ages, suggesting that size plays a notable role in lifespan variations among different breeds.

Table 5. Deviation of average age of purebred dogs by breed size

Size	Deviation	p-value	Significance
Small	01.03	6.5e-97	Significant
Medium	-0.52	6.7e-11	Significant
Large	-0.51	8.7e-73	Significant

Table 6 presents the deviation of the average age of purebred dogs from the overall regional average age across different regions. Almost all regions, except for four, show statistically significant deviations. The oldest dogs are found in Karaganda, East Kazakhstan, and Pavlodar regions, where their average age exceeds the mean age of purebred dogs by about one year. In contrast, the dogs in Turkestan, Kyzylorda, and Zhetysu regions have the lowest average ages, which are approximately 1-1.5 years below the average age of purebred dogs.

Table 6. Deviation of average age of purebred dogs by region

Region	Deviation	p-value	Significance
Karaganda region	1.29	0.00000	Significant
East Kazakhstan region	1.21	0.00000	Significant
Pavlodar region	0.88	0.00000	Significant
Mangistau region	0.76	0.00026	Significant
Kostanay region	0.69	0.00002	Significant
Zhambyl region	0.62	0.00096	Significant
Almaty city	0.59	0.00000	Significant
Akmola region	0.55	0.00814	Significant
Astana city	0.50	0.00000	Significant
North Kazakhstan region	0.39	0.00000	Significant
Atyrau region	0.37	0.21450	Not significant
Shymkent city	0.03	0.80005	Not significant
Ulytau region	0.01	0.91993	Not significant
Aktobe region	-0.41	0.00044	Significant
Abay region	-0.66	0.00000	Significant
West Kazakhstan region	-0.71	0.52209	Not significant
Almaty region	-0.77	0.00000	Significant
Zhetysu region	-0.88	0.00000	Significant
Kyzylorda region	-1.23	0.00000	Significant
Turkestan region	-1.57	0.00000	Significant

DOI: 10.37943/21SUAS7119

Table 7 presents the deviation of the average age of purebred dogs from the average age within each breed. All breeds, except for Bulldogs, show statistically significant deviations from the breed's average age. Terriers and Labradors have an average age that is approximately one year higher than the average age. On the other hand, Greyhounds and Tobets have an average age that is nearly 1.6 years lower than the average age of purebred dogs.

Table 7. Deviation of average age of purebred dogs by breed

Breed	Deviation	p-value	Significance
Terrier	1.00	0.00000	Significant
Labrador	0.76	0.00000	Significant
Spaniel	0.51	0.00001	Significant
Spitz	0.41	0.00113	Significant
Husky	0.30	0.00892	Significant
Bulldog	-0.11	0.47617	Not significant
Sheepdog	-0.29	0.00000	Significant
Laika	-0.87	0.00000	Significant
Greyhound	-1.56	0.00000	Significant
Tobet	-1.58	0.00000	Significant

Discussion

The results of the analysis of dog data in Kazakhstan have identified several key aspects relevant to population management and policy regarding both domestic and stray animals.

First, the predominance of young dogs and the significant proportion of stray animals highlight the need to strengthen population control programs and promote responsible pet ownership. The increasing number of unknown dogs in Almaty indicates the severity of the stray animal issue, which requires a comprehensive approach to registration, identification, and population regulation.

Second, the gender imbalance among dogs may reflect existing preferences in dog selection, influenced by both cultural and functional factors. In Kazakhstan, dogs are traditionally used for hunting and guarding, which may explain the prevalence of large breeds such as sheephdogs, greyhounds, and tobets. Males are generally larger and physically stronger, which could have influenced their selection. Additionally, potential challenges associated with female sterilization and the related stress may have affected their survival rates, particularly among stray animals, for which sterilization programs have been implemented.

A third important aspect is the geographical distribution of dogs. Urbanization, particularly in Almaty, creates unique conditions for population formation, necessitating adaptations in regional animal management strategies. In southern regions such as the Turkestan and Kyzylorda regions, the lower average age of dogs may indicate insufficient conditions for their maintenance or a higher prevalence of diseases reducing their lifespan.

It was also found that the average age of sheepdogs, greyhounds, and tobets is lower compared to other popular breeds. This could be due to their use in hunting and guarding, which increases the risk of injury. Additionally, a significant proportion of these dogs are kept in private homes, often outdoors, which may negatively impact their health and increase disease rates.

Overall, the findings emphasize the need for a comprehensive approach to dog population control, standardization of breed identification systems, and improvement of living conditions for stray animals in Kazakhstan. Further research is recommended to deepen the understand-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

ing of factors affecting population dynamics, with a focus on the health and longevity of dogs.

Regarding the analytical methodology, the data and methods used have certain limitations. First, the TANBA database is a registration system and does not include information on the physiological characteristics of animals (e.g., weight and size) or their medical history, which limits the ability to analyse health and lifespan factors. Second, Kazakhstan lacks a standardized breed classification, making breed-based analysis more challenging. The grouping used in this study is based solely on breed names without considering phenotypic or genetic characteristics. Third, the calculation of dogs' ages was performed by subtracting the birth date from the data collection date (October 7, 2024) and included only dogs with statuses such as "active" or "in transit," which are considered to be alive. This approach may have affected the accuracy of the age distribution assessment.

Future research should focus on refining the impact of demographic, social, and environmental factors on the dog population in Kazakhstan and developing effective strategies for their registration and management.

Conclusion

The study of data from the "Tanba" system has provided valuable insights into the demographic characteristics of the dog population in Kazakhstan. Despite limited data on diseases, the system allows for extensive information on breed distribution, age groups, and regional differences, opening new opportunities for the development of veterinary services. The use of electronic medical records and analytical data can serve as a foundation for creating effective disease prevention and control programs for pets. Ensuring proper nutrition and regular veterinary care should be a priority to improve the health of dogs in the country.

The significant potential of using big data for analyzing dog health in Kazakhstan through the "Tamba" system. Applying big data technologies allows for the identification of patterns and predictions of trends that might remain unnoticed with traditional monitoring methods. One of the key findings is the ability for early disease detection and more accurate health risk assessment, which can contribute to improved veterinary care and preventive measures.

However, this study faces several limitations. One of the main challenges is the quality and consistency of data obtained from various sources. Some data may be incomplete or contain errors, which could affect the accuracy of the analysis. Future research could focus on improving data cleaning and processing algorithms to minimize these risks.

Another important aspect is the ethical and legal framework related to the use of big data in veterinary medicine. Ensuring data privacy and obtaining consent from pet owners are crucial issues that need to be addressed for broader implementation of such systems.

The obtained results can be used to analyze the overall dog population in Kazakhstan and considered in the development of programs aimed at improving dog lifespan and regulating their numbers. To enhance the accuracy of animal health analysis, integrating data from veterinary clinics is recommended. This would help identify the most common diseases in different regions and develop targeted preventive measures.

Additionally, using natural language processing (NLP) methods to analyze textual data from medical records and reports can improve data integration from various sources and automate the detection of important patterns, such as disease trends.

Based on the already identified most common dog breeds in Kazakhstan, an analysis of their predisposition to various diseases can be conducted. This would enable the development of more precise preventive recommendations for pet owners and veterinary specialists, as well as improve the system for early risk detection within dog health monitoring programs.

Prospects for future research include integrating additional data sources, such as information on weather conditions, activity levels, and breed-specific characteristics, which could allow

for the creation of a more comprehensive health risk assessment model for animals. Exploring the applicability of this methodology in other regions and with different animal species also shows promise, potentially expanding data-driven approaches to enhance veterinary care.

In conclusion, this study demonstrates that using big data for monitoring dog health is an effective tool for veterinary medicine, though it requires further refinement and adaptation to maximize benefits and minimize risks.

Acknowledgment

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23488904 Development of scientifically based formulation using unsupervised machine learning in the production of full canned wet feed for unproductive animals).

References

[1] Luo, H., Wen, Y., & Zhang, X. (2021). Research on Intelligent Pet Management Platform System Based on Big Data Environment. *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), 176,* 641–649. https://doi.org/10.1109/aiid51893.2021.9456589

[2] Villa, P. D., Messori, S., Possenti, L., Barnard, S., Cianella, M., & Di Francesco, C. (2012). Pet population management and public health: A web service based tool for the improvement of dog traceability. *Preventive Veterinary Medicine, 109*(3–4), 349–353. https://doi.org/10.1016/j.prevetmed.2012.10.016

[3] *İnformatsionnaya sistema ucheta jivotnyh TANBA [Animal Recording Information System TANBA]*. Retrieved October 3, 2024, from https://tanba.kezekte.kz/ru/

[4] Kim, S., & Kim, S. (2024). Development of a dog health score using an artificial intelligence disease prediction algorithm based on multifaceted data. *Animals, 14(2),* 256. https://doi.org/10.3390/ani14020256

[5] T Chen, Y., & Elshakankiri, M. (2020). Implementation of an IoT based Pet Care System. *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC)*, 256–262. https://doi.org/10.1109/fmec49853.2020.9144910

[6] Tauseef, M., Rathod, E., Nandish, S.M., & Kushal, M.G. (2024). Advancements in Pet Care Technology: A Comprehensive Survey. *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*, 1–6. https://doi.org/10.1109/icdecs59733.2023.10503555

[7] Wang, H., Liu, J., Dong, Z., Song, J., & Zhu, Z. (2023). Artificial intelligence-based metabolic energy prediction model for animal feed proportioning optimization. *Italian Journal of Animal Science, 22(1),* 942–952. https://doi.org/10.1080/1828051x.2023.2236132

[8] Jacobs, M. (2021). 84 the adoption of AI in the core scientific cycle of feed research. *Journal of Animal Science, 99,* 42–43. https://doi.org/10.1093/jas/skab235.074

[9] Parker, V. J. (2021). Nutritional Management for Dogs and Cats with Chronic Kidney Disease. *Veterinary Clinics of North America Small Animal Practice, 51*(3), 685–710. https://doi.org/10.1016/j.cvsm.2021.01.007

[10] Hou, Y., Wu, Z., Dai, Z., Wang, G., & Wu, G. (2017). Protein hydrolysates in animal nutrition: Industrial production, bioactive peptides, and functional significance. *Journal of Animal Science and Biotechnology/Journal of Animal Science and Biotechnology, 8*(1). https://doi.org/10.1186/s40104-017-0153-9

[11] Sosa-Holwerda, A., Park, O., Albracht-Schulte, K., Niraula, S., Thompson, L., & Oldewage-Theron, W. (2024). *The Role of Artificial Intelligence in Nutrition Research: A Scoping review. Nutrients, 16(13)*, 2066. https://doi.org/10.3390/nu16132066

[12] Cardillo, L., Piegari, G., Iovane, V., Viscardi, M., Alfano, F., Cerrone, A., Pagnini, U., Montagnaro, S., Galiero, G., Pisanelli, G., & Fusco, G. (2020). Lifestyle as risk factor for infectious causes of death in young dogs: a retrospective study in Southern Italy (2015–2017). *Veterinary Medicine International, 2020,* 1–10. https://doi.org/10.1155/2020/6207297

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

[13] Wrightson, R., Albertini, M., Pirrone, F., McPeake, K., & Piotti, P. (2023). The Relationship between Signs of Medical Conditions and Cognitive Decline in Senior Dogs. *Animals, 13*(13), 2203. https://doi.org/10.3390/ani13132203
[14] Pereira, M., Valério-Bolas, A., Saraiva-Marques, C., Alexandre-Pires, G., Da Fonseca, I.P., & Santos-Gomes, G. (2019). Development of dog immune system: from in uterus to elderly. *Veterinary Sciences, 6*(4), 83. https://doi.org/10.3390/vetsci6040083
[15] Montoya, M., Morrison, J. A., Arrignon, F., Spofford, N., Charles, H., Hours, M., & Biourge, V. (2023). Life expectancy tables for dogs and cats derived from clinical data. *Frontiers in Veterinary Science, 10*. https://doi.org/10.3389/fvets.2023.1082102
[16] Chaudhari, A., Brill, G., Chakravarti, I., Drees, T., Verma, S., Avinash, N., Jha, A.K., Langain, S., Bhatt, N., Kumar, S., Choudhary, S., Singh, P., Chandra, S., Murali, A., & Polak, K. (2022). Technology for Improving Street Dog Welfare and Capturing Data in Digital Format during Street Dog Sterilisation Programmes. *Animals,* 12(15), 2000. https://doi.org/10.3390/ani12152000
[17] Carvelli, A., Scaramozzino, P., Iacoponi, F., Condoleo, R., & Della Marta, U. (2020). Size, demography, ownership profiles, and identification rate of the owned dog population in central Italy. *PLoS ONE, 15(10)*, e0240551. https://doi.org/10.1371/journal.pone.0240551
[18] VanderWaal, K., Morrison, R.B., Neuhauser, C., Vilalta, C., & Perez, A.M. (2017). Translating Big Data into Smart Data for Veterinary Epidemiology. *Frontiers in Veterinary Science, 4*. https://doi.org/10.3389/fvets.2017.00110
[19] Magalhães-Sant'Ana, M., Peleteiro, M. C., & Stilwell, G. (2020). Opinions of Portuguese Veterinarians on Telemedicine—A Policy Delphi study. *Frontiers in Veterinary Science, 7*. https://doi.org/10.3389/fvets.2020.00549
[20] Akchurin, S.V., Benseghir, H., Bouchemla, F., Akchurina, I.V., Fedotov, S.V., Dyulger, G.P., & Dmitrieva, V.V. (2024). Veterinary telemedicine practicability: Analyzing Russian pet owners' feedback. *Veterinary World,* 1184–1189. https://doi.org/10.14202/vetworld.2024.1184-1189
[21] Jokar, M., Abdous, A., & Rahmanian, V. (2024). AI chatbots in pet health care: Opportunities and challenges for owners. Veterinary Medicine and Science, 10(3). https://doi.org/10.1002/vms3.1464
[22] Kabzhanova, A.M., Muhanbetkaliev, E.E., Esembekova, G.N., Berdikulov, M.A., & Abdrahmanov, S.K. (2022). Prostranstvenno-vremennoj analiz jepizooticheskoj situacii po beshenstvu zhivotnyh v Kazahstane [Spatio-temporal analysis of the epizootic situation of animal rabies in Kazakhstan]. *Herald of science of S Seifullin Kazakh Agro Technical University, 3*(114), 51–58. https://doi.org/10.51452/kazatu.2022.3(114).1118
[23] Ajkimbaev A. M., Tuleuov A.M., Zholshorinov A.Zh., & Bekenov Zh.E. (2015). Monitoring ochagov rabicheskoj infekcii v Kazahstane. Medicina Kyrgyzstana [Monitoring of rabies infection outbreaks in Kazakhstan. Medicine of Kyrgyzstan]. *Medicina Kyrgyzstana,* (3), 26-34. https://cyberleninka.ru/article/n/monitoring-ochagov-rabicheskoy-infektsii-v-kazahstane
[24] Baikadamova, G., Rakhimzhanova, D., Yeszhanova, G., & Seitkamzina, D. (2022). Laboratory studies of Canine Distemper. *3i Intellect Idea Innovation, 4*, 34–41. https://doi.org/10.52269/22266070_2022_4_34
[25] Kaliev, D.S., & Bajkadamova, G.A. (2021). Jepizootologicheskij monitoring veterinarnyh klinik goroda Nur-Sultan po parvovirusnomu jenteritu sobak [Epizootological monitoring of veterinary clinics in Nur-Sultan for canine parvovirus enteritis]. *Sovremennaja Agrarnaja Nauka: Cifrovaja Transformacija*. https://kazatu.edu.kz/assets/i/science/sf17-vet-109.pdf
[26] Administrative-territorial units of the Republic of Kazakhstan. (2024). https://stat.gov.kz/en/industries/social-statistics/demography/publications/207830/
[27] China National Center for Bioinformation. (2019). Dog breed. iDog. Retrieved January 15, 2025, from https://ngdc.cncb.ac.cn/idog/dogph/breed/getBreedByCond.action
[28] The Royal Kennel Club. (n.d.). *Official website*. Retrieved January 15, 2025, from https://www.thekennelclub.org.uk/search/breeds-a-to-z/