СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ / SYSTEM ANALYSIS, MANAGEMENT AND PROCESSING OF INFORMATION

IDENTIFICATION OF THE GENDER DEPENDENCE OF THE FINAL DIAGNOSIS ON THE EXAMPLE OF A SANATORIUM

Research article

Terlyaeva T.K.[1], Drach V.E.[2,]*, Torkunova J.V.[3]
[2] ORCID : 0009-0009-6280-3160;
[3] ORCID : 0000-0001-7642-6663;
[1, 2] Sochi State University, Sochi, Russian Federation
[3] Kazan State Power Engineering University, Kazan, Russian Federation

* Corresponding author (rpd-admin[at]inbox.ru)

Abstract

The purpose of the presented article is to analyze the database of patients at the "Victoria" sanatorium, located in Kislovodsk (Russian Federation), to determine the relationship between the final diagnosis and the patient's gender. This article discusses the methodology for analyzing healthcare data using the Google Colab environment, the Python programming language and other tools for effectively processing information about patients and their diagnoses. For the analysis, we used data on the final diagnosis of the patient made by the attending physician of the sanatorium, and his gender. Based on the analysis, the most common disease is M42.1 – Osteochondrosis of the spine in adults, the majority of whose owners are men. It is recommended to develop special programs and services aimed at preventing and treating spinal osteochondrosis, as well as carrying out activities aimed at men.

Keywords: gender, sanatorium, data storage, medical care, analysis, python, Colab.

ВЫЯВЛЕНИЕ ГЕНДЕРНОЙ ЗАВИСИМОСТИ ОКОНЧАТЕЛЬНОГО ДИАГНОЗА НА ПРИМЕРЕ САНАТОРИЯ

Научная статья

Терляева Т.К.[1], Драч В.Е.[2,]*, Торкунова Ю.В.[3]
[2] ORCID : 0009-0009-6280-3160;
[3] ORCID : 0000-0001-7642-6663;
[1, 2] Сочинский государственный университет, Сочи, Российская Федерация
[3] Казанский государственный энергетический университет, Казань, Российская Федерация

* Корреспондирующий автор (rpd-admin[at]inbox.ru)

Аннотация

Целью представленной статьи является анализ базы данных пациентов санатория «Виктория», расположенного в Кисловодске (Российская Федерация), для определения взаимосвязи между окончательным диагнозом и полом пациента. В этой статье рассматривается методология анализа медицинских данных с использованием среды Google Colab, языка программирования Python и других инструментов для эффективной обработки информации о пациентах и их диагнозах. Для анализа мы использовали данные об окончательном диагнозе пациента, поставленном лечащим врачом санатория, и его половой принадлежности. Исходя из проведенного анализа, наиболее распространенным заболеванием является М42.1 – остеохондроз позвоночника у взрослых, большинство обладателей которого – мужчины. Рекомендуется разработать специальные программы и услуги, направленные на профилактику и лечение остеохондроза позвоночника, а также проведение мероприятий, ориентированных на мужчин.

Ключевые слова: пол, санаторий, хранение данных, медицинское обслуживание, анализ, python, Colab.

Introduction

The purpose of the presented article is to analyze the database of patients of the sanatorium of "Victoria", located in Kislovodsk (Russia), to determine the relationship between the final diagnosis and the patient's gender.

Medicine plays a key role in society, as people face various diseases and diagnoses that affect their lives and well-being. Every day, millions of people around the world face common diseases such as colds, flu, allergies, diabetes, cardiovascular diseases and many others. These diseases can have various causes and manifestations, and require competent medical intervention for diagnosis and treatment. Medicine, in turn, provides not only treatment, but also prevention, counseling and education to help people maintain and improve their health. It is an integral part of our lives, providing care and support during illness and helping to lead an active and healthy lifestyle.

I wonder if there is a relationship between the diagnosis of the disease and the sex of a person? Knowing the answer to this question can help develop effective disease prevention measures and improve people's quality of life.

Investigating possible links between disease diagnoses and gender can help identify groups of people who are susceptible to certain diseases. For example, if it were found that a certain disease is more common in men, it would allow us to focus on preventive measures and lifestyle that can help men reduce the risk of developing this disease. Similarly, if a link were found between a certain disease and the female sex, specific prevention and treatment strategies for women could be developed.

Establishing a link between the diagnosis of the disease and the gender of a person can also contribute to the development of a more personalized approach to medical care, taking into account the characteristics of each group. This may include adapting screening programs, conducting educational campaigns and providing gender-specific recommendations to prevent the occurrence of diseases and improve overall health.

Data collection and analysis

An important step in preventing diseases and improving people's quality of life is to determine the relationship between the diagnosis of the disease and the sex of a person. This can help the medical community (and society as a whole) develop more effective prevention, treatment and care strategies aimed at reducing morbidity and improving public health.

As already mentioned, health analysis plays an important role in optimizing medical care and improving the quality of life of patients. The grouping approach makes it possible to identify common diseases in different gender groups and make informed health decisions [5]. This article discusses the methodology for analyzing health data using the Google Colab environment and other tools for the effective processing of information about patients and their diagnoses for the period 2023.

Data will be taken for analysis:

- the final diagnosis made by the attending physician of the sanatorium;
- gender.

The final diagnosis is made by the patient's attending physician after all the treatment at the last (final) appointment. The diagnosis code is deciphered according to ICD-10 (International Classification of Diseases of the 10th revision), at the moment there are about 15000 names [6].

The first step will be to download and install the standard Numpy and Pandas libraries and import the database in XLSX format (Fig. 1). It includes quite a lot of data about the treatment of a guest of the sanatorium: the procedures assigned to him, their price list, the type of treatment according to the voucher, primary and final diagnoses, etc [7].

```
In [5]:  import numpy as np
         import pandas as pd
         data = pd.read_excel('C:/Users/Tanya/Desktop/title/UV_2023.xlsx')
         data
```

Out[5]:

	ID	DESCRIPTION	PRICE	CURE_ID	CURE	DGN_FIRST_MKB	DGN_LAST
0	4690.0	RSL Sculpting on the Beautylizer machine	2090.0	47.0	The perfect line	E66.0	
1	4690.0	RSL Sculpting on the Beautylizer machine	2090.0	47.0	The perfect line	E66.0	
2	4690.0	RSL Sculpting on the Beautylizer machine	2400.0	47.0	The perfect line	E66.0	
3	4690.0	RSL Sculpting on the Beautylizer machine	2400.0	47.0	The perfect line	E66.0	
4	4690.0	RSL Sculpting on the Beautylizer machine	2750.0	47.0	The perfect line	E66.0	
...	...	...	...	...	...	...	
548739	NaN	NaN	NaN	46.0	Standard (wellness)	Z00.0	
548740	NaN	NaN	NaN	46.0	Standard (wellness)	Z00.0	
548741	NaN	NaN	NaN	46.0	Standard (wellness)	Z00.0	
548742	NaN	NaN	NaN	46.0	Standard (wellness)	Z00.0	
548743	NaN	NaN	NaN	46.0	Standard (wellness)	Z00.0	

548744 rows × 12 columns

Figure 1 - Loading libraries and database
DOI: https://doi.org/10.60797/IRJ.2024.145.110.1

NumPy is a Python library that is used for mathematical calculations, starting with basic functions and ending with linear algebra. The full name of the library is Numeric Python extensions, or "Numeric Python Extensions" [4].

Pandas is a Python programming library designed to work with data. It works on the basis of the NumPy library and provides special data structures for working with numeric tables and time series. The pandas library provides operations for data management and analysis [1].

Thanks to the drop() function, the database gets rid of unnecessary columns for the current analysis (Fig. 2). As a result, the columns "DGN_LAST_MKB" remain — the final diagnosis of the patient and "SEX" — the gender of the patient, where 0 = male and 1 = female.

Figure 2 - Using the drop() function
DOI: https://doi.org/10.60797/IRJ.2024.145.110.2

First, you need to clear the database. This will help eliminate duplicate data that may occur as a result of errors or repeated entries. It also allows you to maintain data integrity and prevent distortion of the analysis results [2].

In addition, database cleanup may include error correction and data standardization. This ensures the uniformity and correctness of the data, which is important for accurate analysis and reliable results.

In general, database cleanup is an important step before analysis, which helps to ensure data quality, improve performance and reliability of analysis results.

Data cleanup involves removing duplicate records and missing (empty) values. This can be implemented using the dropna() function (Fig. 3).

Figure 3 - Using the dropna() function
DOI: https://doi.org/10.60797/IRJ.2024.145.110.3

After the cleanup, the number of records decreased by 46,996. These were patients, for whatever reason, without a final diagnosis.

Next, you need to group the data by the final diagnosis and count the number of men and women using the groupby() and value_counts() functions (Fig. 4).

```
data3=data2.groupby('DGN_LAST_MKB')['SEX'].value_counts().reset_index(name='count')
data3
```

	DGN_LAST_MKB	SEX	count
0	0	0	15
1	125	0	3
2	.J98.	0	8
3	42.9	0	11
4	63.2	0	20
...	...	...	...
915	Z95	1	7
916	Z96.6	1	13
917	Z99	0	1
918	ZA00	1	4
919	ZOO	1	8

920 rows × 3 columns

Figure 4 - Using the groupby() and value_counts() functions
DOI: https://doi.org/10.60797/IRJ.2024.145.110.4

For convenient data display, it was decided to provide the result in the form of an interactive bar chart. The graphical representation of the data will allow you to visually illustrate the distribution of patients by final diagnosis and gender, which will facilitate the perception of information and the identification of possible patterns [10].

Results and their discussion

Thanks to the plotly.express library, a graph was created (Fig. 5), based on which the following conclusions can be drawn:

1. The most common disease is M42.1 – Osteochondrosis of the spine in adults, the majority of whose owners are men, about 123 thousand people, which exceeds the number of women with a similar diagnosis by almost 2 times (Fig. 6).

To prevent this disease, it can be recommended to develop special programs and services aimed at prevention and treatment of osteochondrosis of the spine, as well as holding events aimed at men.

These packages can include specialized programs of physical rehabilitation, massage, fitness, and expert advice on recommended exercises for the prevention and treatment of diseases of the musculoskeletal system. You can also offer meals specially designed to meet the needs of men and their health, and conduct educational seminars on healthy lifestyles and disease prevention.

2. Analysis of the distribution of final diagnoses by gender of patients revealed that there are differences in the incidence of diseases between men and women. This is useful information for the development of individual medical programs and prevention strategies [8], [9]. According to the graph, men are more likely to be confirmed with the disease, especially those associated with diagnoses of M42 (osteochondrosis of the spine) and I11.9 (hypertension, unspecified). This indicates the need for specialized programs for the prevention and treatment of these diseases in men, as well as the importance of educational activities aimed at maintaining men's health and preventing these diseases.
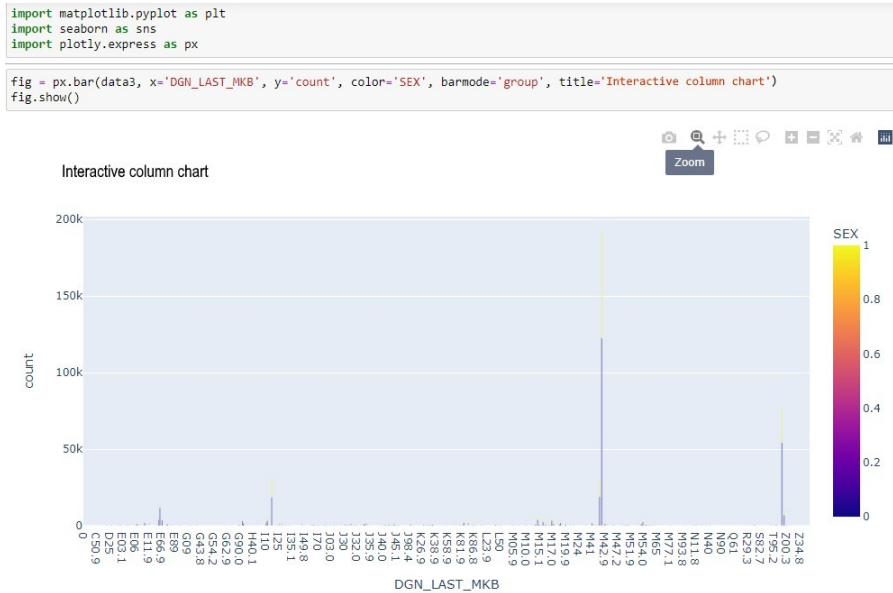
```
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px

fig = px.bar(data3, x='DGN_LAST_MKB', y='count', color='SEX', barmode='group', title='Interactive column chart')
fig.show()
```


Figure 5 - Column diagram of the distribution of patients by final diagnosis and gender
DOI: https://doi.org/10.60797/IRJ.2024.145.110.5
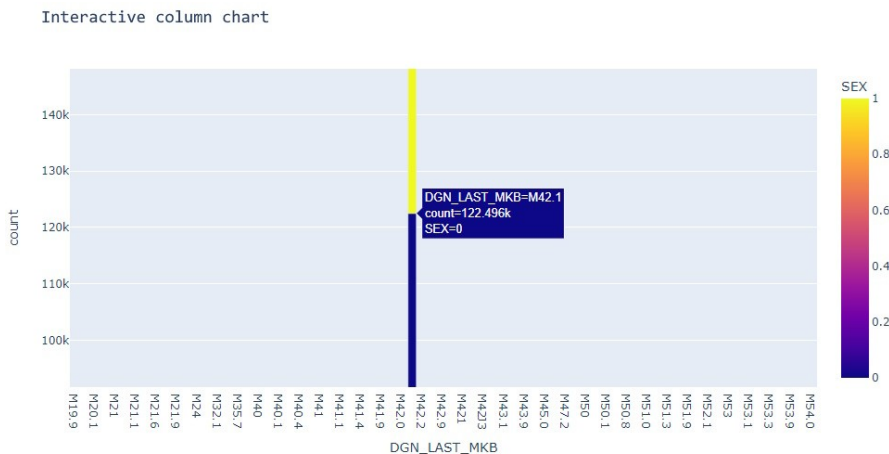
Figure 6 - Number of men diagnosed with M42.1
DOI: https://doi.org/10.60797/IRJ.2024.145.110.6

Among 501,748 patients, only 78,006 people are diagnosed with Z00 – General examination and examination of persons without complaints or established diagnosis (Fig. 7). This diagnosis belongs rather to the category of preventive examinations and consultations, rather than to specific diseases.
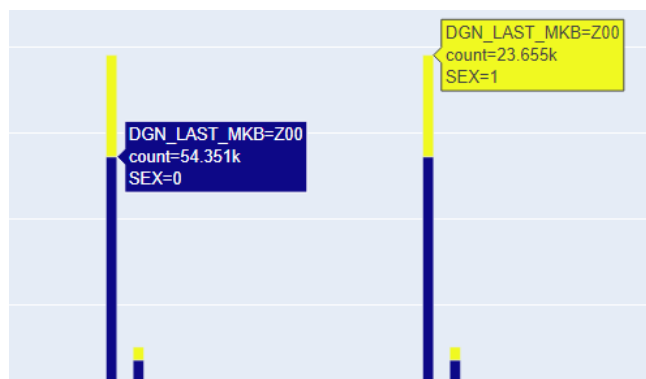

Figure 7 - Number of men and women diagnosed with Z00
DOI: https://doi.org/10.60797/IRJ.2024.145.110.7

Analyzing the number of cases of certain diseases among men and women can help identify possible differences in predisposition to certain diseases depending on gender. This can be important information for a personalized treatment approach.

Also, judging by the graph, it can be concluded that there is a significant prevalence of the diagnosis of E66 (obesity) among the male population (Fig. 8). It can be assumed that this is due to alcohol consumption. Alcohol contains a high number of calories and can contribute to the accumulation of body fat. In addition, alcohol can affect metabolism and eating behavior, which can lead to an increase in appetite and consumption of more food [11].

The diagnosis of E66 is usually associated with overweight and obesity, which may indicate the importance of taking measures to control weight and maintain a healthy lifestyle among men to prevent diseases associated with this diagnosis.
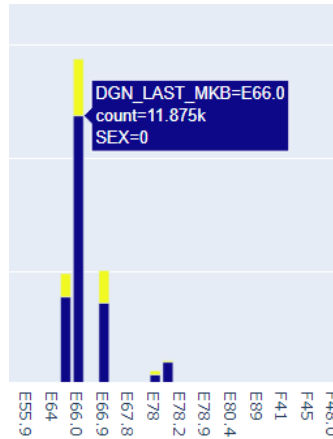

Figure 8 - Number of men diagnosed with E66
DOI: https://doi.org/10.60797/IRJ.2024.145.110.8

Genetic features may play a role in predisposition to certain diseases. For example, genetic factors can influence the tendency to various diseases [3], including cancer, diabetes mellitus, cardiovascular diseases, etc.

Conclusion

The relationship between gender and diagnosis is a complex research issue in the field of medicine. Differences in morbidity and response to treatment may be gender-specific and require additional study and analysis of the causes of this difference. Some diseases may occur more often in a certain gender, which underscores the importance of taking gender into account when conducting medical research.

The analysis of the relationship between the patient's gender and his final diagnosis is important for improving medical practice and developing personalized treatment approaches. Further research in this area may lead to the development of more effective treatment strategies that take into account the gender characteristics of patients.

To solve the problem, various Python functions were used, which made it possible to efficiently process data without resorting to the use of excessive computing power, and create a bar chart, thanks to which they were able to visualize the result.

Список литературы на английском языке / References in English

1. Jain N. Pandas AI: A step towards GEN AI / N. Jain // Interantional journal of scientific research in engineering and management. — 2023. — Vol. 07. — №07.

2. Drach V.E. Analysis of popular non-relational database management systems / V.E. Drach, V.Yu. Ilyichev, A.V. Rodionov // System Administrator. — 2022. — №3(232). — P. 84-88.

3. Duke V.A. Information technologies in biomedical research / V.A. Duke, V.L. Emanuel. — St. Petersburg: Peter, 2003.

4. Grigoriev E. Exploratory data analysis using Python / E. Grigoriev, A. Klimov. — E-Scio, 2020.

5. Torkunova J. Impact analysis of digital transformation of the energy sector economy on hospitality industry / J. Torkunova, M. Khabrieva // Lecture Notes in Civil Engineering. — 2022. — Vol. 190. — P. 105-112.

6. Kobeissi H. National institutes of health grant funding for cerebrovascular diseases / H. Kobeissi, M.S. Jabal, Sh. Ghozy [et al.] // Journal of NeurobInterventional Surgery. — 2024. — Vol. 16. — №2. — P. 209-212.

7. Kuznetsov N.A. Investigation of the possibility of using machine learning models based on the Scikit-learn library to optimize the additive manufacturing process / N.A. Kuznetsov // Innovations in information technologies, mechanical engineering and motor transport (IITMA-2020): collection of materials of the IV International Scientific and practical Conference with online participation. — 2020. — P. 302-304.

8. Nikolenko S. Deep learning. Immersion in the world of neural networks / S. Nikolenko, A. Kadurin, E. Arkhangelskaya. — St. Petersburg, 2020. — 480 p.

9. Shatalova V.V. Big data: how Big data technologies are changing our lives / V.V. Shatalova, D.V. Likhachevsky, T.V. Kazak // Big Data and Advanced Analytics. — 2021. — №7-1. — P. 188-192.

10. Tikhanychev O.V. On methods of coordination in decision support systems / O.V. Tikhanychev // Applied Computer Science Scientific articles. — 2018. — №3(75). — P. 8.

11. He A. Clusters of body fat and nutritional parameters are strongly associated with diabetic kidney disease in adults with type 2 diabetes / A. He, C. Shi, X. Wu [et al.] // Diabetes Therapy. — 2024. — Vol. 15. — №1. — P. 201-214.