Goodreads – book popularity analysis

Motivation and problem description

Goodreads is a social network for book fans with over 90 million users. The Goodreads database consists of an exhaustive amount of books, authors and user comments. Among other things, users can search a large catalog of books and rate them. With the help of ratings and the created profile network, users can easily find new, interesting publications.

Dataset description and guidelines for creating the project

The given dataset contains books from the Goodreads Best Books Ever list (the largest list on the network). Various information can be read for each book, among others: author, average rating, publication date and genre.

When creating the project, follow these research questions:

- Are there differences in book ratings based on genre?
- Are books with less pages cheaper?
- Can you determine the book's popularity based on available variables?
- Are there differences in the popularity of books with regard to their age?
- Based on the available variables, can you determine whether the book is award-winning?

Moreover, think about your own research questions and hypotheses that you want to analyze using the available data.

Note: If you have any additional questions concerning this specific project, please contact tessa.bauman@fer.hr.

Real estate market analysis

Motivation and problem description

Real estate buyers often want to satisfy various needs (e.g., the number of rooms because of numerous children or the need for a larger garage because of numerous cars), but they do not want to pay too much for such "luxuries". Moreover, real estate prices are often inflated for various reasons, while it is in the interest of banks to objectively assess the value of real estate for mortgage lending. Therefore, it is necessary to collect real estate data.

Dataset description and guidelines for creating the project

The data collected is information about sold real estate in the city of Ames (USA). Each property is described with 79 features, and the total number of data collected is 1460. Real estate are described with features such as square footage, proximity to public transportation, number of bedrooms, basement size, orientation, and so on.

When creating the project, follow these research questions:

- Does the shape of the lot determine the number of stories of the house?
- Does the size of the basement depend on the neighborhood in the city?
- Does the number of bedrooms determine the price per square foot of a property?
- Can available features predict real estate price?

Moreover, think about your own research questions and hypotheses that you want to analyze using the available data.

Note: If you have any additional questions concerning this specific project, please contact tomislav.kovacevic@fer.hr.

Anthropometry

Motivation and problem description

Anthropometry is the study of how to measure humans. General anthropometry includes the complete process of data collection, documentation, summarization, and analysis. In a narrower sense, anthropometry can be defined as the science of body measurement, where lengths, breadths, heights, and circumferences are used to numerically describe body segments and the overall body shape. Body measurement is essential in quantifying the variations in and between populations of different countries, ethnicities, cultures, and ages, and it strongly impacts various industries, such as: medicine, fashion, fitness, and entertainment.

Dataset description and guidelines for creating the project

The dataset is comprised of various anthropometric measurements of 133 subjects gathered in two phases. The first phase gathered anthropometric measurements from subjects located in the United States, whilst the second phase gathered the same measurements from subjects located in Spain. Every measurement is estimated multiple times either by expert antrhopometers or by (semi)-automatic 3D scanners and their corresponding software measuring solutions. Togeather with the mentiond antrhopometric measurements, every subject has its appropriate 3D scan, obtained with 8 different 3D scanners.

When creating the project, follow these research questions:

- Do specific 3D scanners measure a given anthropometric measurement differently?
- Do some antrhopometric measurements have statisticly significant differences between those gathered in the United States and those gathered in Spain?
- Can we regress a body measurement given in the dataset by using other body measurements, also given in the dataset?
- How would you measure an anthropometric measurement of your interest given a 3D scan of the person? Could you regress such a measurement from the already given measurement data?

Moreover, think about your own research questions and hypotheses that you want to analyze using the available data.

Napomena: If you have any additional questions concerning this specific project, please contact david.bojanic@fer.hr.

Premier League player stats

Motivation and problem description

We are aware that statistics have become a part of our everyday life and that quality statistical data analysis is needed in everywhere we go. One area where statistics has been applied for a very long time is sports. Over time, sports statistical analysis advanced a lot and today we have huge amounts of data for various sports that need to be properly analyzed. Quality analysis of these data has become an extremely important factor in the development of sports and individuals, and no one can afford not to do analysis of collected data. Most clubs nowadays have a statistics team to analyze the performance of their players, and this is especially the case in the English Premier League.

Dataset description and guidelines for creating the project

The dataset used as part of this project contains various statistics for all Premier League players for the 2021-2022 season. Each row of the file refers to a player, and in the columns we can find various statistics related to the players, such as the age of the player, the number of goals, the number of minutes played, the number of received cards etc.

When creating the project, follow these research questions:

- Is there a difference in the number of minutes played by young players (under 25) among Premier League teams?
- Who gets more yellow cards on average, forwards or midfielders?
- Can you determine the performance of an individual player based on the given parameters?
- Do ”domestic” players (i.e. players of English nationality) or foreign players contribute more to the overall success of their team?

Moreover, think about your own research questions and hypotheses that you want to analyze using the available data.

Napomena: If you have any additional questions concerning this specific project, please contact krunoslav.jurcic@fer.hr.

Household budget survey analysis

Motivation and problem description

The household budget survey provides annual information on the nature and destination of consumption expenditures, as well as on various characteristics related to household living conditions. The data are collected by the Spanish National Statistic Institute (*Instituto Nacional de Estadística* or INE) to assess the standard of living in the country.

Dataset description and guidelines for creating the project

The dataset contains 24000 responses from the households surveyed between 2009 and 2019. The survey questions are categorized into four different groups: geographical data, general information on household members, income data, and expenditure data. Each category is defined by a specific set of features, e.g. general information on household members category includes the number of household members, their age distribution, members' job sectors, etc.

When creating the project, follow these research questions:

- Does the water heating method depend on the number of household members?
- Are high-income household members more likely to be real-estate owners?
- Is food delivery more frequent in large households?
- Can we predict household income using other available features?

Moreover, think about your own research questions and hypotheses that you want to analyze using the available data.

Note: If you have any additional questions concerning this specific project, please contact andro.mercep@fer.hr.