Bitbucket repository:
https://bitbucket.org/erkokaar/datascience-project-gender-statistics-analysis/src/main/
Team members: Timo Kaasik, Sofia Kriuchkova, Erko Käär

Team D10 - Gender statistics analysis

Business understanding

Business goals

- Background

  Gender studies as a scientific field is gaining popularity. At the end of the last century, the new direction of feminism - postmodern feminism - was finally formed, which offered the world a new look on gender in general and the role of women in society in particular. Since Sofia studied gender during her BA studies and for her bachelor thesis, and the rest of the other team members found the topic of gender diversity interesting and relevant, we decided to analyze gender statistics in our project.

- Business goals

  We want to determine the relationship between economic background and gender diversity in the fields of education (specifically STEM) and technology, and to see how technology development contributes (or not) to gender equality. At the end we want to make a prediction about how the gender gap in technology and STEM sciences will change in the next 5 years.

- Business success criteria

  To be successful, we need to ensure the following:
  - we have a complete data set for each of our focus points
  - the machine learning models we have developed are accurate and reliable

- ensure that the findings are understandable and interpretable to a wide audience, including those who are not data scientists or machine learning experts, as well as those who are not experts in gender studies
- the recommendations and predictions from the analysis are of practical value

Current situation

- Inventory of resources

As a data set, we will use gender statistics from Kaggle (https://www.kaggle.com/datasets/joebeachcapital/gender-statistics/). It was gathered from the World Bank on key gender topics for all countries in the last 20 years, which can be found here (https://databank.worldbank.org/source/gender-statistics). We use the python programming language with the pandas and Sklearn libraries in an interactive jupyter notebook environment to process and analyze the data.

- Requirements, assumptions, and constraints

04.12.2023 - final assignment of responsibilities
06.12.2023 - draft code and draft of interpretation of data analysis.
08.12.2023 - final code, second draft of data interpretation and prediction, draft poster.
11.12.2023, 12:00 am - finished project and poster
14.12.2023 - presentation of the project

- Risks and contingencies

Delays are possible if our team members or equipment become incapacitated. So, if someone on our team has problems with their laptop, the rest of the team will either provide them with a laptop for the tasks required by the project, or do a redistribution of work tasks.
In case of illness of one of the participants, a redistribution of work tasks will also take place.

- Terminology

Gender - a term used to exemplify the attributes that a society or culture constitutes as "masculine" (expected from men) or "feminine" (expected from women)
Gender gap - a difference between men and women in terms of social and political attitudes
Machine-learning (ML) - algorithms that learn from data
Postmodern feminism - a type of feminism that emerged in the late 20th century. It is marked by a rejection of traditional feminist ideas and an embrace of postmodern philosophy. Postmodern feminism is critical of essentialism, patriarchy, and binary thinking.
STEM - a common abbreviation for four closely connected areas of study: science, technology, engineering and mathematics.
Panel data - a collection of quantities obtained across multiple individuals, that are assembled over even intervals in time and ordered chronologically.

- Costs and benefits

Since this research is conducted as a project for the university we do not have any financial resources to conduct it, accordingly we are not going to spend financial resources on this project. Our resources will be our time and the intellectual property of people who have done similar research before.
Our benefits: gaining experience in analyzing and processing data.

Data-mining goals

- Data-mining goals

- Data cleaning to eliminate errors, outliers and missing information.
- Convert data into usable formats for the application of modern machine learning models.
- Identify key variables affecting gender diversity, including economic status and educational attainment.
- Analyzing correlations between gender diversity and economic indicators in various fields.
- Applying cluster analysis methods to identify groups of similar patterns of gender diversity.
- Development of gender gap forecasting models in technology and STEM-education for the next 5 years.

- Data-mining success criteria

- Ensure that data meets project standards and requirements
- Ensure that machine learning models are accurate and reliable
- Identify and focus only on important variables affecting gender diversity
- Ensure the quality of visualization of the analysis results to better understand key findings

Data understanding

- Gathering data

The data can be found on many big organizational websites and offers a variety of topics and statistics of different magnitude. Most of it can be exported as .csv files or in similar formats. We are most interested in measurements regarding technology, education, social status/background and economies.
The data is openly available in the World Bank Data Bank as the "Gender Statistics" database. A formatted and easily usable .csv version of it can be found on Kaggle.

- Describing data

The dataset contains an impressive 305,575 rows and 24 columns. The layout of the dataset is a term called panel data (referenced in terminology). Column 1 and 2 ("Series Name", "Series Code) refer to a question and its abbreviated name. Column 3 and 4 ("Country Name", "Country Code") refer to a country's name and its abbreviated name. Columns 5 to 24 refer to specific years. There are 266 unique country names, meaning all the countries are present, with also region totals (i.e Europe or Sub-Saharan Africa) and social classifications (i.e low/middle/high income). There are 1,153 unique attributes ("Series Names") which are actual labels for data. The attributes are from key topics, which can be summarized into 12 categories, which are: assets, economic and social context, education, employment and time use, entrepreneurship, environment, health, leadership, norms and decision-making, population, technology and violence.

- Exploring data

There is a lot of data and some of it is very thorough. It will be interesting to see which useful patterns can be found from. The attributes' coded names (“Series Code”) have a tag for the topic and an abbreviation for the question, meaning they can be grouped easily. Unfortunately, a lot of data is missing, meaning that for some attributes there may be values found for only a few years out of the 19 observed years or the data being entirely missing for this value for a country, but it is important to note that these kinds of statistics are not always gathered annually or that some countries may even lack the ability to perform these kinds of processes. This poses a great problem and requires thorough thinking, as how to use what we have, what to do with what we do not have (as just deleting it might not be too efficient) and how to connect it all together and make it useful. Regarding the different topics, some may turn out to be more meaningful than others, meaning some weighting might need to be introduced. Then there is the question of some attributes having values of 1 or 0 (true or false), whilst others having percent (%) values and some having a flat out integer value related to the country’s population, which is not included.

- Verifying data quality

In theory, this data should be enough to support our goals and be considered usable. We have to be mindful about the data, the intervals of the data gathered and properly group regions, or even discard some of them. It might be best to perform our analyses on separate topics separately and then join them up for a final result. Prioritizing relevant attributes and assessing the applicability of identified categories contribute to the project's success and the potential for meaningful insights and reliable predictions for our exploration of gender disparities over the past two decades.

Project plan

1. Data preparation
   - Estimated time: 20 hours
   - Methods, tools: Pandas dataframe on Jupyter notebook
   - Details: Preparing data before operating with it. Including dropping values, reformatting values.

   Task 1: Download the gender statistics dataset from Kaggle.

Task 2: Explore the dataset to understand its structure and content.

Task 3: Clean the data by addressing missing values, outliers, and errors.

Task 4: Convert the data into usable formats for machine learning models.

Task 5: Identify and extract relevant variables for analysis.

2. Discovering patterns from data

   - Estimated time: 20 hours

   - Methods, tools: pandas, matplotlib

   - Details: discover the most frequent and relevant patterns, based on statistics and group them.

   Task 1: Apply machine learning models to predict gender gap changes over the last 20 years.

   Task 2: Analyze correlations between gender diversity and economic indicators.

   Task 3: Conduct cluster analysis to identify patterns of gender diversity.

   Task 4: Develop forecasting models for gender gaps in technology and STEM fields for the next 5 years.

3. Code Development and Documentation

   - Estimated time: 10 hours

   Task 1: Create a draft code for data analysis and machine learning models.

   Task 2: Document the code for clarity and future reference.

   Task 3: Draft interpretations of the data analysis.

4. Visualization of data

   - Estimated time: 15 hours

   - Methods, tools: matplotlib, pandas, seaborn libraries

   - Details: Visualize relations of the attributes

   Task 1: Finalize the machine learning models and code.

   Task 2: Create visualizations for key findings.

   Task 3: Prepare a draft poster for the project presentation.

5. Project Presentation

   - Estimated time: 5 hours

   Task 1: Prepare a comprehensive presentation summarizing the project.

   Task 2: Rehearse the presentation for clarity and coherence.