MSc in Fundamental Principles of Data Science
Presentation & Visualization

Task 2: Storytelling

Theodoros Lambrou, Georgia Zavou, Ilaria Curzi, Alana Zoloeva

Firstly, we would like to point out that for this delivery, we have used the same dataset as the 1st task, i.e. 'Adults' (found here: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data).

As established in the initial delivery, we assume the primary audience is a government organization or department focused on labor and income-related policies and analysis (i.e. government analysts). Their expertise is in data analysis, statistics, and policy research, specifically in areas related to labor and income, though they may not be experts in chart design. On the other hand, our external company ('Income Equity Hub'), specializes in visually illustrating demographic disparities within the population.

Government Analysts primarily interact with visualizations through desktop computers or internal government systems. These visual tools play a vital role in their routine consultations, aiding in decision-making and policy analysis within their area of expertise. Additionally, we identified a secondary audience comprising policymakers, researchers, and analysts focused on labor and income-related issues.

After completing all validation rounds, we have finalized the following theme for our presentation: Our goal is to explore insights into income disparities, employment trends, and the impact of education, marital status, and other factors on income levels within the population. A key focus will be on analyzing gender inequality.

To shape our message, we propose considering the following questions:

- "Does gender affect income levels, and if so, how?"
- "What does the income distribution look like across different education levels within the population?"
- "How does an individual's marital status impact their income?"
- "Are there significant regional differences in income across the country?"

Additionally, some potential user goals include the following:

1. Understand gender-based income disparities:
    - Analyze how gender affects income levels.
    - Identify gaps in income between men and women.
2. Assess the impact of education on income:
    - Examine income distribution across different education levels.
    - Investigate the correlation between education and income inequality.
3. Assess the impact of occupation and marital status:
    - Identify income trends and disparities linked to various occupations.
    - Analyze how marital status affects individual income.
4. Investigate regional income variations:
    - Examine geographic differences in income distribution.

Pilot Users

In each round of validation and co-creation, we included both a government analyst and a policymaker as pilot users to ensure balanced representation of our target audience. This balance was crucial, as the analyst prioritized data-driven insights, while the policymaker focused on actionable government strategies. This approach enabled us to address the distinct needs of both user groups effectively.

First round

In the first round, the primary goal was to outline the desired outcomes and key questions that will guide the presentation's development. Below are the critical questions raised and the insights gathered during this stage.

The government analyst was first asked about the relevance of the chosen dataset to their typical analytical work and whether there were specific data points or variables they commonly use in their reports. In response, the analyst highlighted that the "Adults" dataset is well-stratified and sufficiently representative, eliminating the need for supplementary data sources. They also identified key variables frequently analyzed, such as gender, education, marital status, occupation, and country of origin. Based on this feedback, it was decided to focus on these critical variables.

The analyst was asked about the types of visualizations or charts they find most effective for presenting analytical insights, aligned with their usual methods of data interpretation. They were also asked to identify any elements of data visualizations that they find challenging to interpret. Their feedback emphasized a preference for simplicity and clarity in visualizations, showing the importance of straightforward designs. This consideration reflects the variety of projects they handle and the potential cognitive demands of their work, ensuring that findings are communicated effectively to the audience.

Next, the policymaker was asked about their objectives, and they outlined their goal of advancing economic equality of opportunity across diverse demographic groups while maintaining voter support for the upcoming election in two years. This feedback prompted a shift in focus towards showcasing income disparities among various demographic segments. Notably, a strategic decision was made to prioritize gender equality, reflecting its relevance as a key public concern in current discussions. Additionally, we opted to structure the presentation logically, beginning with fundamental gender disparities and gradually exploring more complex issues.

When asked about existing policies, the policymaker mentioned ongoing initiatives aimed at addressing income inequality related to country of origin. They recommended conducting a routine review for updates to compare with their current perspective, while avoiding excessive focus on the process itself.

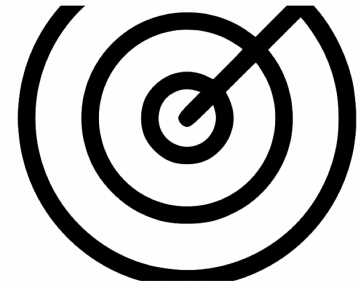
As we move into the next phase, our focus turns to translating these valuable insights into a concrete prototype that reflects the preferences and objectives expressed by both the government analyst and the policymaker. This prototype will form the foundation for the upcoming rounds of validation and co-creation.

Second round

In developing our presentation, we chose a design approach that is based in simplicity and minimalism. Our aim is to provide the audience with clear and impactful content while upholding a professional and structured tone. The following images showcase 2 sample slides from our presentation.

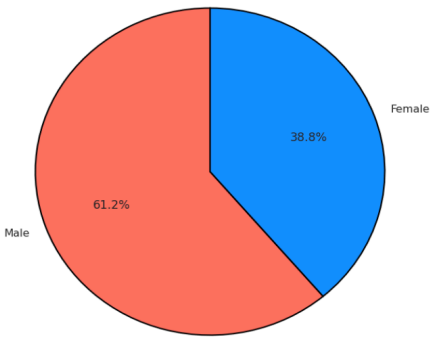
ABOUT US

- We are a team of Data Scientists specializing in statistical analysis, machine learning, and data visualization.
- Our goal is to uncover intricate patterns and trends within the broad scope of socioeconomic data, with a focus on gender

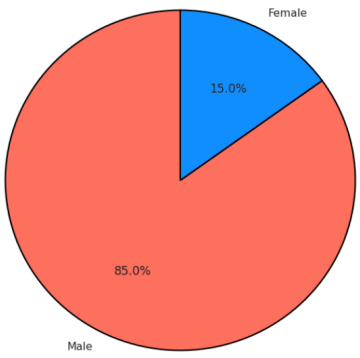


INCOME

Low-Income (<=50K) distribution by Gender



High-Income (>50K) distribution by Gender



We aim to use images with a consistent theme throughout most slides to enhance visual cohesion and create engaging presentations. Alternatively, if this is not feasible, we use a clean, white background to maintain simplicity. Each slide will include a title and the key points relevant to the discussion. This approach is designed to ensure clarity, maintain audience engagement, and effectively convey essential information and details.

After presenting the design to both pilot users, they expressed their approval but advised against overcrowded slides, especially when showcasing analytical insights. So we decided not to use background images on such slides to ensure a clean and focused presentation of the analytical findings.

Subsequently, we began developing the prototype. The slides were not fully populated with all the content; only selected graphs have been included. This approach serves as a guide for the pilot users, offering a preview of the slide design, coherence, and overall structure of the presentation.

The pilot users provided positive feedback on including details about our external company and methodology. This transparency is intended to enhance credibility and build trust in the presentation. Furthermore, explaining the dataset and clarifying how the data is presented helps to eliminate any misunderstandings.

During the brief presentation of the prototype, the plots were thoroughly explained; however, the pilot users recommended revising the titles to improve clarity. They also suggested modifying the color scheme for males and females to avoid potential bias and incorporating consistent legends across all plots. Regarding the map depicting the percentage of high-income individuals by country, they raised concerns about its clarity. As a result, we decided to replace it with a simpler plot, focusing on representing the continent of origin rather than individual countries.

Both pilot users gave positive feedback about the structured approach of the detailed analysis, starting with an overview of general gender disparities and gradually delving into specific factors contributing to these differences. The use of causality-based storytelling enhances understanding of the core concepts while supporting our commitment to clarity and coherence in presenting insights. Their positive feedback validates our strategy of leading the audience through a logical progression, ensuring better comprehension and retention of key ideas, leaving a lasting impact on the audience.

Based on their feedback, we identified an opportunity to enhance audience engagement. Recognizing the value of giving them time to reflect on their initial assumptions about the topic, we decided to include a quiz/questionnaire prior to the detailed analysis. This interactive feature aims to boost engagement while encouraging active participation and critical thinking among the audience. The quiz will be introduced during the third round.

Third round

In the third round, we incorporated all the recommendations from our pilot users. This included revising the color schemes for males and females, improving title clarity, ensuring consistent legends across all visualizations, and replacing the map with a simpler and more intuitive representation, specifically using pie charts.

The color scheme was redesigned to feature two distinct colors with varying shades, depending on the visualization. Notably, these colors are complementary, ensuring clear differentiation between them and against the white background. In keeping with our focus on formality and minimalism, black and white were consistently used throughout the presentation, contributing to a cohesive and polished visual aesthetic. This deliberate adjustment enhanced visual clarity while preserving a professional and consistent tone across the entire presentation.

We introduced the concept of the quiz to the pilot users, featuring the following set of questions:

1. Do you think there is a pay gap among genders?
2. What percentage of each gender earns more than $50K?
3. For married individuals, what percentage of each gender has an income over $50K?
4. Among those with a master's degree, what percentage of each gender earns over $50K?
5. Can you identify the continent of origin with the smallest percentage of individuals earning over $50K?

After each question, participants were shown the correct answer, allowing them to compare their assumptions with the actual data. We also shared the aggregated results of the entire audience, providing insight into the collective viewpoints of their peers. This interactive approach encouraged personal reflection while promoting a shared understanding and heightened awareness among the participants.

The pilot users highlighted a notable improvement in the overall presentation with the addition of the quiz, emphasizing its positive impact and the effectiveness of this interactive feature.

In terms of storytelling, we revised it to align with the following structured sequence:

1. Introduction: Discuss our identity, mission & the methodology we followed.
2. Dataset Explanation: Offer clear insights into the dataset.
3. Quiz: Engage with the audience using a quiz to enhance individual and collective reflection.
4. Gender pay inequalities discussion: Examine gender inequalities and emphasize the demographic factors that play the most significant role in these disparities.
5. Conclusions: Provide a summary of the key findings.

Explanation of the duties of each member of the team

Each team member contributed actively to every aspect of the project, including data analysis, the development of the dashboard, the storytelling presentation, and the reports. Ongoing communication throughout the process facilitated strong collaboration within the team.

Although the roles below identify aspects of the work that the respective team member focused on, everyone helped in other parts of the overall delivery too.

Team members & roles:

- Theodoros Lambrou: Data Analyst
  Responsible for analyzing and interpreting the data. Also in charge of coordinating the team's workflow & communication.

- Georgia Zavou: Presentation Designer
  Responsible for designing an elegant and effective presentation, ensuring consistency in design elements such as color schemes and fonts, and making any necessary changes based on the feedback received from the pilot users.

- Ilaria Curzi: Content Coordinator
  Responsible for creating compelling and informative narratives for the presentation, ensuring the storyline aligned seamlessly with its goals and objectives.

- Alana Zoloeva: Quality Assurance Coordinator
  Responsible for performing quality assurance checks on the presentation, gathering and organizing feedback from pilot users and team members, and overseeing revisions based on the collected input.

Submission

Within our submission, alongside this report we also include the final presentation (`.pptx`) and the python notebook (`.ipynb`) within which the plots were produced.