A picture containing candle, design, halloween

Description automatically generated

Manipal International University, Malaysia SCHOOL OF ENGINEERING AND COMPUTING

Mini-Project

SCB3374 – DATA SCIENCE TOOLS

Due Date: 23rd JUNE 2023

Instruction: answer ALL the questions. Total marks: 100 marks

**NAME : SOUHAIL MOUSSAIF**

**STUDENT ID :** 1108211003

# PROGRAMME : BCS

ECB4313 – DATA ANALYTICS

|  |  |  |
| --- | --- | --- |
| **Couse Learning Outcome** | **Question** | **Marks** |
| CLO3: Employ cutting edge tools and technologies to gain insights into the data (C3, PLO2) |  |  |
| Total Marks | | **/100** |

SCB3374 – DATA SCIENCE TOOLS

|  |  |  |
| --- | --- | --- |
| **Couse Learning Outcome** | **Question** | **Marks** |
| CLO3: Employ cutting-edge tools and technologies to gain insights into the data (C3, PLO6) |  |  |
| Total Marks | | **/100** |

# Question

In this project, you will use any dataset of your interest. The objective of this project is to use data analysis, data visualization techniques, and machine learning to answer some analytical questions. You need to use jupyter notebook using packages such as Pandas, NumPy and Scikitlearn.

Guide:

1. Use the MJSAT template provided for the report.
2. Individual project only.
3. The report should include an introduction (literature), data analysis, discussions, and conclusion.
4. The report should be at least four pages and at most six pages long.
5. Use IEEE references format. All references must be cited in the paragraph and appear in the references list. References should be at least 15.
6. Use Google Scholar or any scientific database for the literature.
7. The report should be submitted in Ms. Word format.
8. A softcopy of python or excel files must be submitted through google drive.
9. Plagiarized work will result in ZERO marks.

# THE END

*Manipal International University*

Data Science Tool

Mini Project

**Analysis of Stack overflow Developer Survey Data: Providing Insights on Equality**

Author 1Souhaul Moussaif 1108211003

*1Manipal International University, Nilai, Negeri Sembilan, Malaysia*

**Keywords**

**Abstract**

*Analysis*

*Diversity*

*GDP*

*Stack overflow*

This report focuses on the analysis of the Stack Overflow Developer Survey data. The dataset includes information on developers' responses categorized by variables. The objective of this analysis is to extract valuable insights that can inform decision-making processes. The analysis will involve the use of graphs, charts, and statistical calculations to visualize and interpret the survey data effectively, providing valuable insights into the developer community and the broader tech industry.

1. **Introduction**

In today's rapidly evolving technological landscape, understanding the dynamics of equality within the developer community is crucial. The data-set presents an opportunity to delve into the 2022 Stack Overflow Developer Survey results and extract valuable insights on the topic of equality. Developers from various backgrounds and expertise levels were invited to participate, through Stack Overflow web-site to answer a wide range of questions covering different aspects of their professional lives, challenges, experiences, and opinions.

By examining the survey results, we can explore potential disparities and inequalities that exist in the tech industry.

* Representation and Diversity: Investigate the representation of different demographic groups within the developer community and examine any discrepancies in terms of gender, ethnicity, and age.
* Pay Disparities: Exploring if there are any significant differences in pay based on factors such as gender, ethnicity, or geographical location.

the columns that we gona cover and analysis are

1. CompFreq: Frequency of compensation received by respondents (e.g., annually, monthly, weekly, etc.)
2. Country: Country where the respondent is located
3. Employment: Employment status of the respondent (e.g., employed full-time, self-employed, etc.)
4. EdLevel: Highest level of formal education attained by the respondent
5. Gender: Gender of the respondent
6. Age: Age of the respondent
7. Ethnicity: Ethnicity or cultural background of the respondent
8. YearsCodePro: Number of years the respondent has been coding professionally
9. CompTotal: Total compensation or salary received by the respondent
10. Currency: Currency in which the compensation is reported
11. Sexuality: Sexual orientation of the respondent
12. GDP\_per\_capita: Gross Domestic Product (GDP) per capita of the respondent's country
13. Income\_Group: Categorized income level of the respondent
14. **Methodology**

The methodology encompasses the steps taken to collect the data, clean and prepare it for analysis, perform exploratory data analysis, conduct equality analysis, and interpret the findings.

The survey was administered between May 11, 2022, and June 1, 2022, through Stack Overflow.

the data set is carefully examined to remove duplicate entries, handle missing values, removing any unrelated columns and address any inconsistencies or errors.

Exploratory Data Analysis is performed to gain an initial understanding of the data set and explore patterns, trends, and distributions. Descriptive statistics, data visualizations, and summary metrics are used to summarize and visualize key variables of interest.

The results of the analysis are interpreted to draw meaningful conclusions regarding equality in the tech industry.

1. Importing and Formatting Data

In this part we import pandas and numpy library and set them as pd and np respectively. We also import matplot and seaborn as plt and sns respectively for visualization.

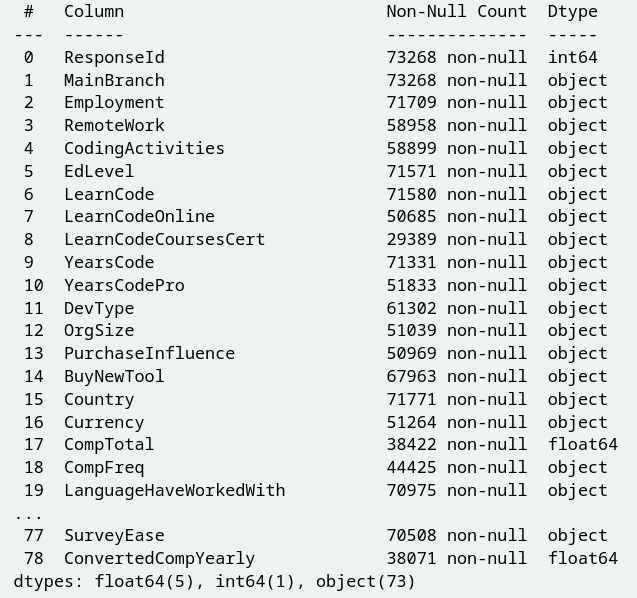
* 1. **Data Information**

After importing the data set, the next step is to understand the types of data we are working with. This initial exploration allows us to gain insights into the structure and content of the data.

Inspecting the Data Set:

Begin by loading the data set and Displaying a summary of the data set to gain a high-level understanding of its structure and contents.

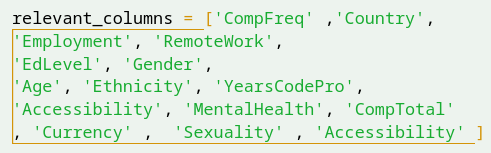
Retrieve basic information such as the number of rows and columns, and any initial insights from a sample of the data.



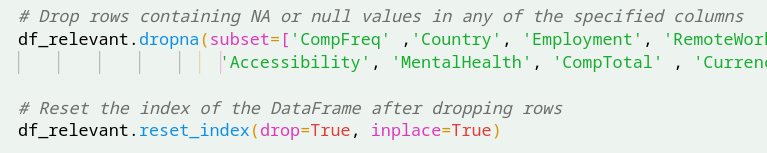
* 1. **Data Cleaning**

Data cleaning to narrow down the data by selecting important columns and removing empty or null rows. Additionally, standardizing certain columns, such as currency conversion and adjusting frequency values, contributes to data consistency and facilitates meaningful comparisons.

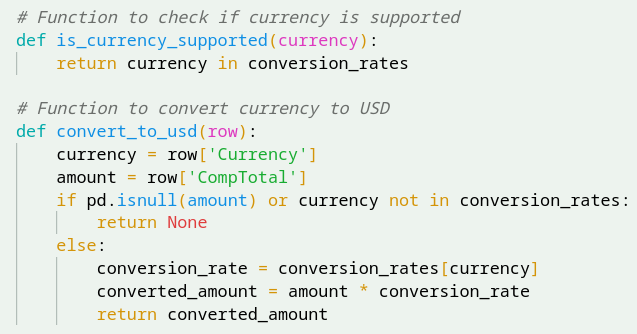
Review the list of columns and identify the ones that are most relevant to the study on equality.



Remove rows with significant missing data, ensuring that the remaining data is representative and reliable.



If the data includes columns representing different currencies, consider converting the values to a common currency, such as the US dollar. Utilize exchange rates or currency conversion factors to standardize the currency values across the dataset.

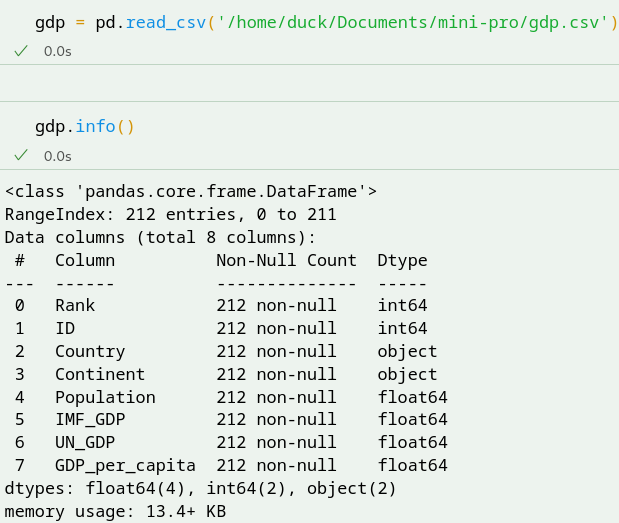


If the data contains a column representing compensation frequency, such as monthly or annually, consider converting it to a standardized frequency, such as yearly.



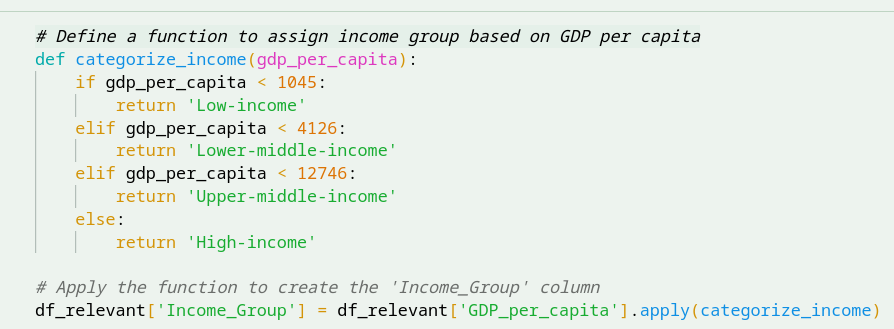
* 1. **Adding a Related Dataset**

Incorporating a related dataset can enhance the depth of information and facilitate answering the research questions on equality. One such dataset that can be imported and joined with the main data set is the GDP per capita dataset. This dataset provides valuable economic information that can be linked to the main dataset based on the common variable of country.



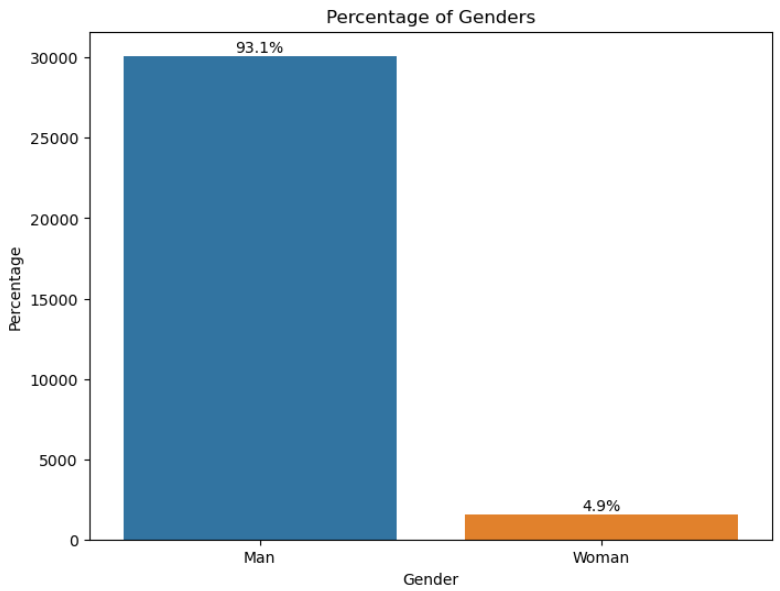
Normalizing country names in the GDP per capita dataset with those in the main dataset. Normalize the country names to ensure consistency and accurate matching during the joining process.

Based on the World Bank's Country Classification by Income report, a new column was created in the dataset to categorize countries into four groups based on their GDP per capita. The income groups include low-income countries (GDP per capita < $1,045), lower-middle-income countries (GDP per capita $1,046-$4,125), upper-middle-income countries (GDP per capita $4,126-$12,745), and high-income countries (GDP per capita > $12,746). This classification enables further analysis of the relationship between income levels and variables related to equality within the tech industry.

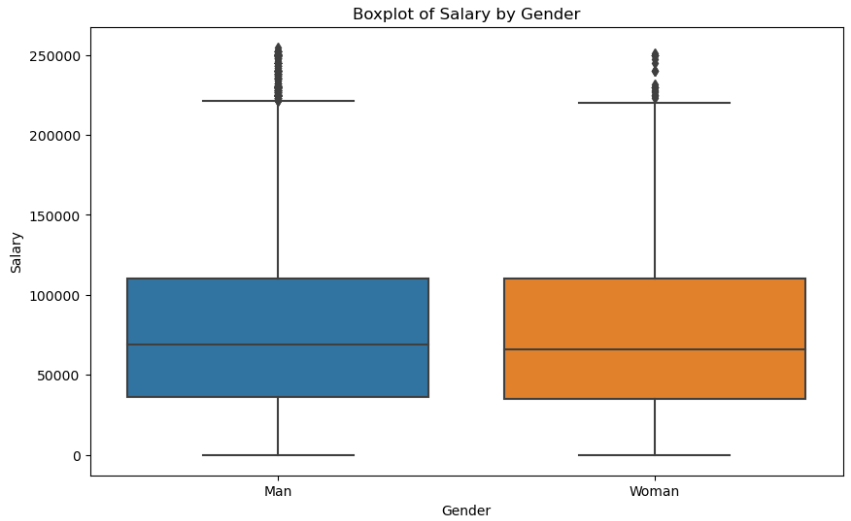


1. Descriptive Statistics
   1. **Gender equality**

After cleaning the dataset, which consisted of over 30,000 survey responses, it was observed that men held the majority in the tech industry, accounting for approximately 93% of the respondents. On the other hand, women represented a smaller portion, comprising around 4.7% of the surveyed population.



Despite men leading in the tech industry, an analysis using boxplots revealed that the median salaries between genders were relatively similar. This suggests that, on average, there is no significant difference in the salaries of men and women in the industry. However, it is important to conduct further statistical tests to determine if there are any statistically significant variations or disparities in salaries between genders.



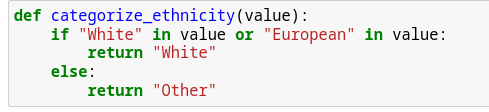
To further support our speculations, additional graphs were examined, revealing that a significant number of women in the tech industry are located in high-income countries. These countries often have stronger legal frameworks and workplace policies in place to promote gender equality and mitigate discrimination.



* 1. **Ethnicity**

In conducting a basic data analysis, we discovered a correlation between the probability of discrimination based on gender in the tech industry and the potential role of ethnicity in inequality. This finding raises the question of whether ethnicity plays a significant role in shaping equality within the industry.

By examining the data, we can explore the distribution of different ethnicities among tech professionals and investigate potential disparities or biases that may exist.

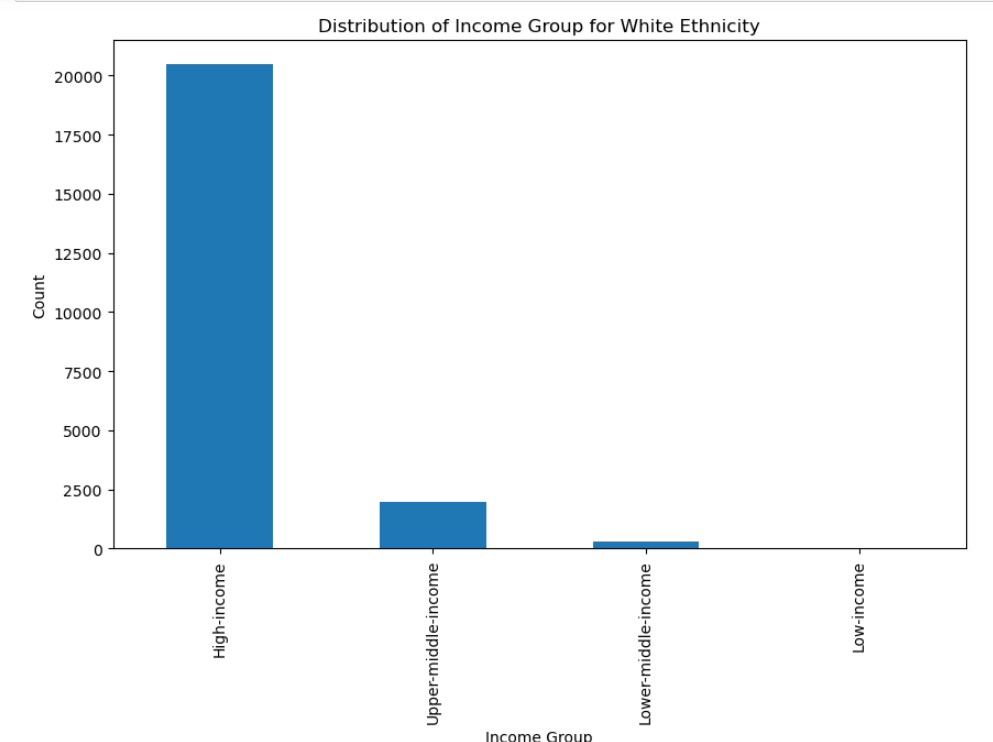


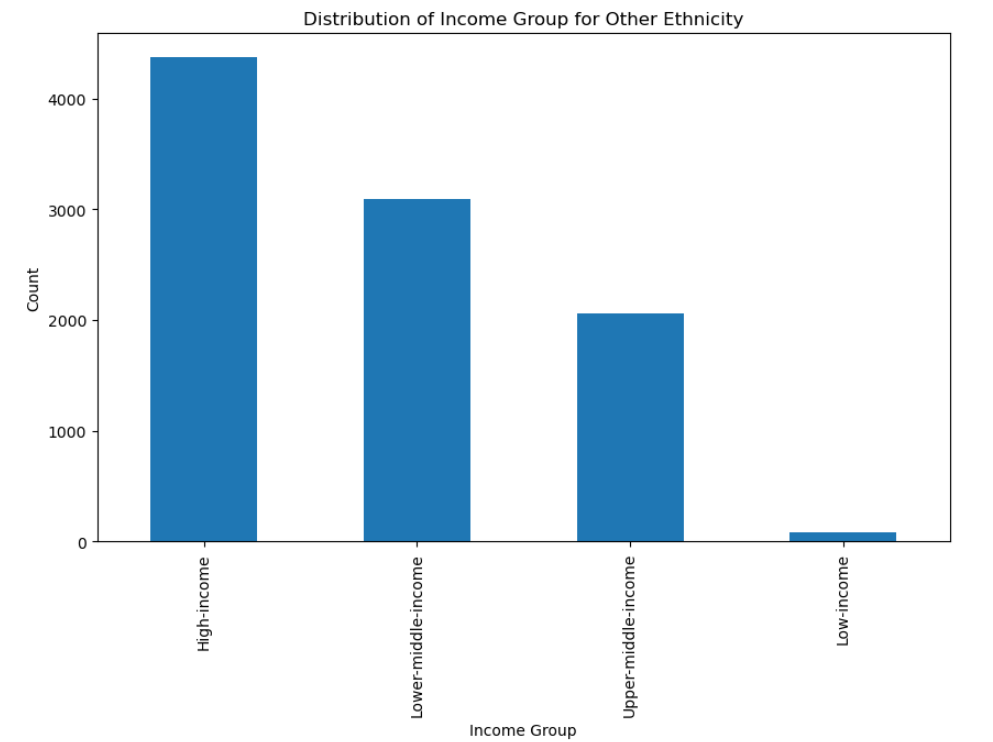
Based on the boxplot analysis, it appears that the median salary for individuals of white ethnicity is higher compared to other ethnicities. This finding suggests the existence of a potential disparity in salaries based on ethnicity within the tech industry.However, it is important to acknowledge that this methodology may have limitations and could potentially introduce biases in the long run. By solely focusing on the comparison between white ethnicity and other ethnicities, we may overlook the nuances and variations within each ethnic group.



Upon further observation and leveraging the GDP data of each user's country based on the World Bank's Country Classification by Income report, it is evident that individuals of white ethnicity predominantly originate from high-income countries. Additionally, there is a smaller representation of white ethnicity in lower-middle-income and upper-middle-income countries, with less than 10,000 users. In contrast, other ethnicities are more evenly distributed across the three income groups, with fewer than 500 users originating from low-income countries.

This observation suggests a potential correlation between ethnicity, income levels, and geographic distribution within the tech industry. The overrepresentation of white ethnicity in high-income countries may reflect economic factor that influence access to education, opportunities, and resources in the tech sector



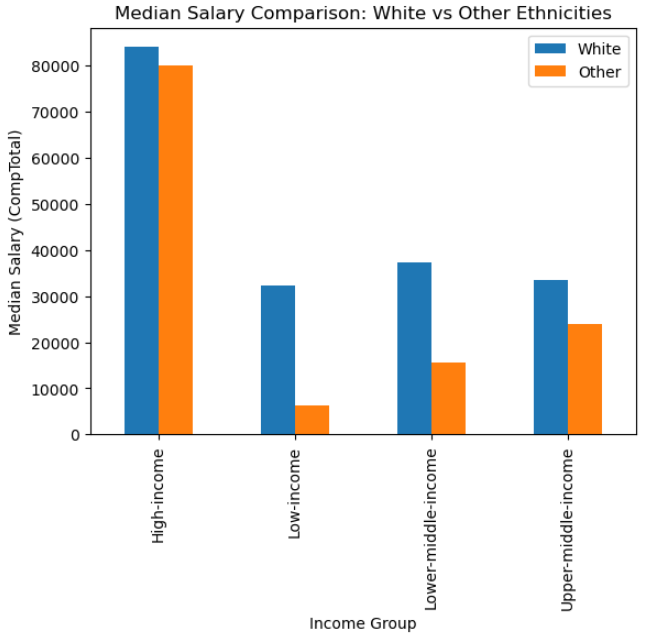


An interesting finding emerges: non-white tech developers have a similar median salary as their white counterparts in high-income countries. This observation challenges the assumption that ethnicity alone is the primary determinant of salary disparities within the tech industry.

The similarity in median salaries between non-white and white tech developers in high-income countries suggests that factors beyond ethnicity, such as education, experience, skills, and job roles, may play a more influential role in determining salaries.

In low-income and middle-income countries, the data reveals a salary discrimination against white individuals within the tech industry. However, it is important to consider the context in these countries. The presence of fewer white and European individuals in these regions may lead to a higher demand for their skills and expertise, which can influence their value in the job market and potentially result in higher salaries.

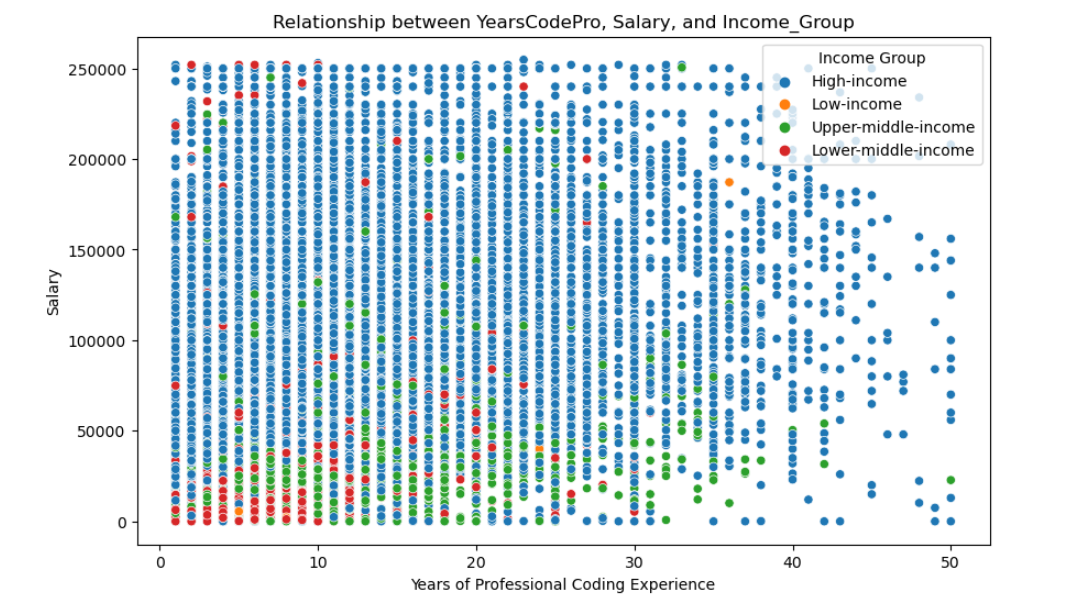
It is worth noting that salary disparities based on ethnicity can also be influenced by various **socio-economic** factors, cultural norms, historical context, and systemic biases present in different countries.



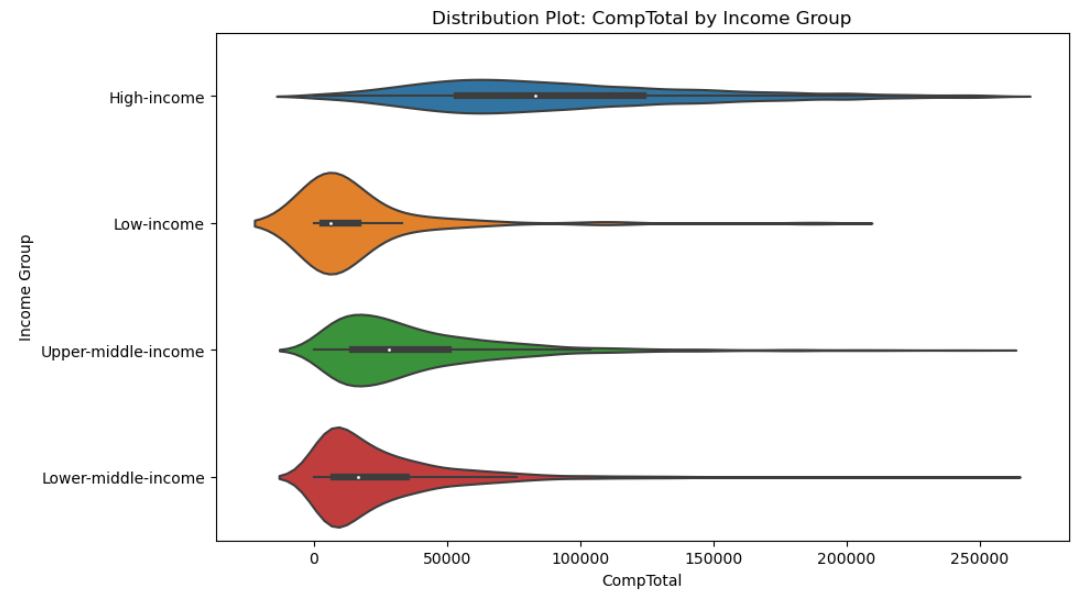
* 1. **GDP**

Through the analysis of both sections, it becomes apparent that GDP can indeed be a significant influencer in the context of our study.

In the scatterplot graph, a clear pattern emerges where the GDP income groups tend to cluster and group together in the bottom left region, irrespective of the years of experience. This indicates a strong correlation between GDP income groups and salary levels. The graph supports the observation that higher GDP income groups, coupled with more years of experience, tend to correlate with higher salaries. This finding highlights the significance of economic factors in shaping salary distributions

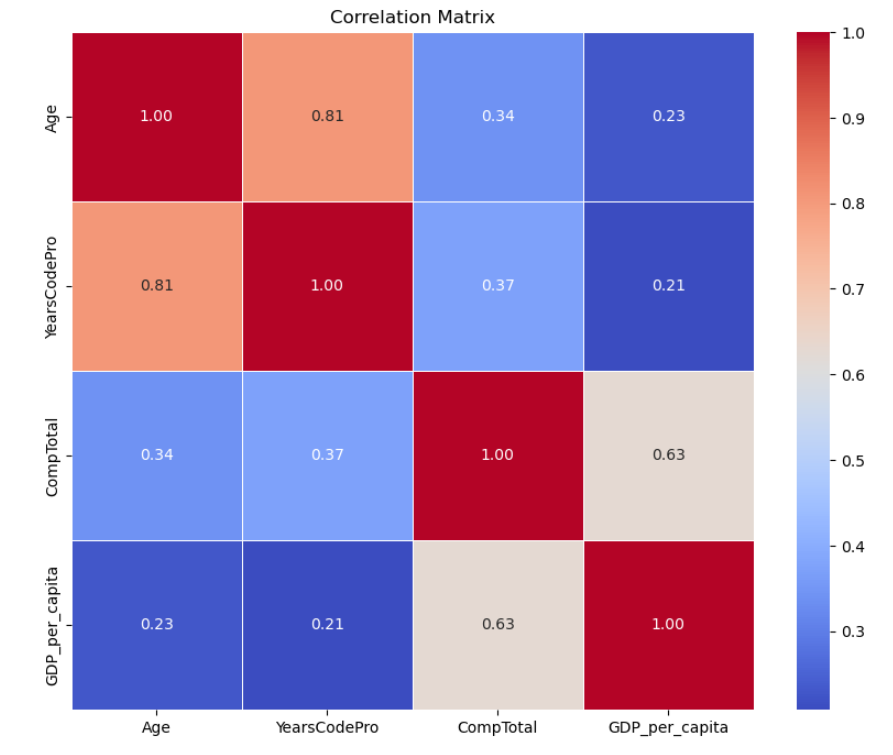


Another graph that supports this theory is a bar graph comparing the mean salaries across the four income groups: low-income countries, lower-middle-income countries, upper-middle-income countries, and high-income countries. The graph reveals a clear trend where the mean salary in high-income countries is significantly higher, approximately double or more, compared to the mean salaries in the other three income groups.



However, it is important to acknowledge that other factors such as skillsets, job roles, and market demands also contribute to salary differentials. Therefore, a comprehensive understanding of the interplay between GDP, experience, and other relevant factors is essential to gain deeper insights into the determinants of salary levels within the tech industry.

* 1. **Heatmap**



Heatmap is used to show the correlation between every variable. In the above diagram the closer to 1 or lighter the colour the correlation with one another is high.

1. **Discussion**

The data cleaning process was crucial in preparing the dataset for analysis, including selecting relevant columns and standardizing certain variables such as currency and frequency.

The analysis revealed that men dominated the tech industry, aligning with expectations, while women represented a smaller percentage. However, it is important to note that this finding may be subject to bias due to the survey's recruitment channels, which primarily targeted highly engaged users on Stack Overflow. Furthermore, the analysis indicated that the median salaries were relatively similar between men and women, suggesting a potential equality in pay within the industry.

To gain deeper insights, additional variables were considered. The examination of ethnicity in relation to GDP income groups showcased interesting patterns. It was observed that white tech professionals were predominantly based in high-income countries, while other ethnicities were distributed across various income groups. Additionally, a graph highlighted that non-white tech developers in high-income countries exhibited similar median salaries as their white counterparts. These findings indicated that economic factors, represented by GDP and income levels, played a significant role in shaping salary disparities within the tech industry.

It is important to acknowledge the limitations of the study, such as potential biases in data collection and the exclusion of other relevant factors that could impact salary disparities.

1. **Conclusion**

The analysis showed that men dominated the tech industry, but median salaries indicated potential pay equality between genders. Ethnicity revealed a concentration of white tech professionals in high-income countries, while other ethnicities were distributed across income groups. GDP income groups correlated with salary levels. Future research should consider additional variables and address study limitations for a comprehensive understanding of tech industry equality.

**References**

[1] J. Babb, N. Terry, and K. Dana, ‘The Impact Of Platform On Global Video Game Sales’, International Business & Economics Research Journal (IBER), vol. 12, p. 1273, 09 2013.

[2] A. Aziz, S. Ismail, M. F. Othman, and A. Mustapha, ‘Empirical Analysis on Sales of Video Games: A Data Mining Approach’, Journal of Physics: Conference Series, vol. 1049, no. 1, p. 012086, Jul. 2018.

[3] J. S. Babb and W. Texas, ‘COMPARING VIDEO GAME SALES BY GAMING PLATFORM’, 2013.

[4] A. Marchand and T. Hennig-Thurau, ‘Value Creation in the Video Game Industry: Industry Economics, Consumer Benefits, and Research Opportunities’, Journal of Interactive Marketing, vol. 27, no. 3, pp. 141–157, 2013.

[5] D. Choi and J. Kim, ‘Why People Continue to Play Online Games: In Search of Critical Design Factors to Increase Customer Loyalty to Online Contents’, CyberPsychology & Behavior, vol. 7, no. 1, pp. 11–24, 2004.

[6] D. Chan, ‘Convergence, Connectivity, and the Case of Japanese Mobile Gaming’, Games and Culture, vol. 3, no. 1, pp. 13–25, 2008.

[7] L. Cadin, F. Guérin, and R. DeFillippi, ‘HRM Practices in the Video Game Industry:: Industry or Country Contingent?’, *European Management Journal*, vol. 24, no. 4, pp. 288–298, 2006.

[8] J. Barnett and M. Coulson, ‘Virtually Real: A Psychological Perspective on Massively Multiplayer Online Games’, Review of General Psychology, vol. 14, no. 2, pp. 167–179, 2010.

[9] By Andrew Beattie, ‘How the Video Game Industry Is Changing’, <https://www.investopedia.com/articles/investing/053115/how-video-game-industry-changing.asp>

[10] H. Liu, ‘Dynamics of Pricing in the Video Game Console Market: Skimming or Penetration?’, Journal of Marketing Research - J MARKET RES-CHICAGO, vol. 47, pp. 428–443, 06 2010.

[11] R. Baer, ‘Videogames: In the Beginning’, 06 2023.

[12] V. Landsman and S. Stremersch, ‘Multihoming in Two-Sided Markets: An Empirical Inquiry in the Video Game Console Industry’, Journal of Marketing, vol. 75, pp. 39–54, 11 2011.

[13]M. S. El-Nasr, A. Canossa, T.-H. D. Nguyen, and A. Drachen, Game Data Science. Oxford University Press, 10 2021.

[14] X. Cai, J. Cebollada, and M. Cortiñas, ‘A grounded theory approach to understanding in-game goods purchase’, PLOS ONE, vol. 17, no. 1, pp. 1–27, 01 2022.

[15] V. Shankar and B. Bayus, ‘Network Effects and Competition: An Analysis of the Home Video Game Industry’, Strategic Management Journal, vol. 24, 01 2002.