**Data Analysis on "Ask A Manager Salary Survey 2021"**

**Introduction:**

This report presents the findings from the analysis of the "Ask A Manager Salary Survey 2021" dataset, which contains information about various aspects of employment, including age group, industry, job title, salary, education level, gender, and race. The dataset was obtained from a real-world, messy source (<https://oscarbaruffa.com/messy/>), providing an opportunity to practice data cleaning and analysis on a real-life dataset.

**Data Cleaning and Preprocessing:**

Before conducting the analysis, the dataset underwent several cleaning and preprocessing steps to ensure its quality and usability. The following steps were taken:

1. The column names were modified using Excel for better readability and consistency. The original column names were:
   * "Timestamp" -> "Timestamp"
   * "How old are you?" -> "Age Group"
   * "What industry do you work in?" -> "Industry"
   * "Job title" -> "Job title"
   * "If your job title needs additional context, please clarify here:" -> "job context"
   * "What is your annual salary? (You'll indicate the currency in a later question. If you are part-time or hourly, please enter an annualized equivalent -- what you would earn if you worked the job 40 hours a week, 52 weeks a year.)" -> "annual salary"
   * "How much additional monetary compensation do you get, if any (for example, bonuses or overtime in an average year)? Please only include monetary compensation here, not the value of benefits." -> "Bonus"
   * "Please indicate the currency" -> "currency"
   * "If "Other," please indicate the currency here:" -> (dropped)
   * "If your income needs additional context, please provide it here:" -> (dropped)
   * "What country do you work in?" -> "country"
   * "If you're in the U.S., what state do you work in?" -> "U.S State"
   * "What city do you work in?" -> "City name"
   * "How many years of professional work experience do you have overall?" -> "total professional work experience"
   * "How many years of professional work experience do you have in your field?" -> "professional work experience in field"
   * "What is your highest level of education completed?" -> "highest level of education"
   * "What is your gender?" -> "gender"
   * "What is your race? (Choose all that apply.)" -> "What is your race? (Choose all that apply.)"
2. The "Timestamp" column was converted to a datetime format.
3. Columns with a significant amount of missing values ("Bonus", "U.S State", and "job context") were dropped.
4. Missing values in categorical columns were filled with the value "Unknown".
5. The "annual salary" column was converted to a numeric data type, and commas were removed.
6. Categorical columns were converted to the "category" data type for efficient storage and operations.

**Key Findings:**

1. Top Working Age Groups:
   * The top three working age groups in the dataset were 25-34 (12,646 entries), 35-44 (9,895 entries), and 45-54 (3,188 entries).
2. Most Common Industries and Job Titles:
   * The top three most common industries were Computing or Tech (4,692 entries), Education (Higher Education) (2,463 entries), and Nonprofits (2,418 entries).
   * The top three most common job titles were Software Engineer (286 entries), Project Manager (230 entries), and Director (198 entries).
3. Highest Salaries in USD:
   * The top three highest salaries in USD were $102,000,000 (index 3605), $10,000,000 (index 28021), and $5,000,044 (index 26466).
4. Industry with Highest Average Annual Salary:
   * The industry with the highest average annual salary was "ESL Teacher".
5. Salary Distribution Across Age Groups:
   * The age group with the highest median salary was 45-54 ($88,000), followed by 35-44 ($85,000) and 55-64 ($83,000).
   * The age group with the highest maximum salary was 55-64 ($870,000,000), followed by 25-34 ($180,000,000) and 35-44 ($120,000,000).
6. Impact of Education Level on Annual Salary:
   * The highest average annual salary was observed for individuals with a PhD ($241,019.89), followed by those with a Professional degree ($156,049.85) and a College degree ($170,142.49).
7. Gender Salary Differences:
   * The average annual salary was highest for individuals who identified as "Other or prefer not to answer" ($232,572.52), followed by "Man" ($203,142.74) and "Woman" ($134,141.19).
   * The average annual salary was lowest for individuals who identified as "Prefer not to answer" ($88,000.00) and "Non-binary" ($100,025.49).

**Challenges and Limitations:**

While working with this real-world dataset, several challenges and limitations were encountered:

1. Missing Data:
   * The dataset contained missing values in several columns, which required appropriate handling techniques, such as filling with a designated value or dropping columns with a significant amount of missing data.
2. Data Quality:
   * As the dataset was obtained from a live source and constantly growing, the data quality may vary, and some fixes made during the analysis might not hold for all new entries.
3. Free-form Text Entries:
   * Six of the variables in the dataset were free-form text entries, which required extensive data cleaning and normalization to ensure consistent analysis.
4. Potential Outliers:
   * The dataset contained some extremely high salary values, which could potentially be outliers or data entry errors. Further investigation and domain knowledge might be required to handle such cases appropriately.

**Conclusion:**

The analysis of the "Ask A Manager Salary Survey 2021" dataset provided valuable insights into the salary landscape, highlighting the impact of factors such as age group, industry, job title, education level, and gender on annual salaries. The dataset's real-world nature and inherent messiness presented opportunities to practice data cleaning and preprocessing techniques, which are essential skills for data analysts.