**README**

1. **Use or create an original dataset (meaning that you haven't analyzed the dataset for this class before). Be sure to explain where your dataset comes from and where it can be found (or if you made it, how you made it).**

**LINK TO THE DATASET: https://oscarbaruffa.com/messy/**

* The dataset being used is a Salary Survey from AskAManager.org.
* The dataset primarily focuses on salary-related information but also includes additional demographic and professional details.
* The dataset contains 28035 rows and 17 columns.
* Notably, 6 of the variables involve free-form text entry.
* AskAManager.org is a popular advice column and blog run by Alison Green.
* The salary survey is conducted through this platform and collects responses from individuals regarding their salary, job title, industry, location, years of experience, gender, race, and other relevant information.
* The dataset is "live" and constantly growing, as new responses are added over time.

1. **Identify what would make the dataset fit for use.**
2. **Relevance to Analysis Goals:** Firstly, the dataset's relevance to our analysis goals is paramount. Ensuring that the dataset aligns closely with the questions we seek to answer will enhance the efficacy of our analysis.
3. **Currency Standardization:** Secondly, currency standardization is essential for accurate comparison and interpretation of salary data across different regions. Converting all salary values to a common currency (e.g., USD) allows for meaningful comparisons and eliminates potential biases due to currency fluctuations.
4. **Domain-specific Considerations:** Domain-specific considerations, such as industry norms and regional variations, are vital for accurate interpretation. Understanding the nuances of different industries, job roles, and geographical locations enables us to contextualize the data effectively and draw insightful conclusions.

**3. Evaluate if the data are fit for use or if cleaning will need to be done. If cleaning will need to be done, identify the cleaning steps that will be taken.**

Before proceeding with analysis, it's crucial to evaluate the fitness of the dataset for our purposes and determine if any cleaning is necessary. The initial assessment reveals several areas where cleaning may be required.

1. **Renaming Columns:** Column renaming might be needed to ensure clarity and consistency across variables.
2. **Dropping Columns:** Some columns may need to be dropped if they are redundant or contain irrelevant information.
3. **Currency Conversion and Related Manipulation:** Currency conversion and related manipulations are essential to standardize salary data for meaningful comparisons across different regions.
4. **Handling Missing Values:** Handling missing values is crucial in maintaining the integrity of the dataset, potentially involving imputation techniques or the exclusion of incomplete entries.
5. **Label Encoding:** Label encoding may be employed to convert categorical variables into the numerical format for analysis.
6. **Standardization and Similarity Metrics:** standardization and similarity metrics may be applied to ensure uniformity and facilitate comparisons between variables.  
     
   By addressing these cleaning steps, we can enhance the quality and usability of the dataset for our analysis.

**6. Document the cleaning steps that were taken (in a README, or Open Refine log, or some other documentation mechanism).**

1. **Renaming Columns:**

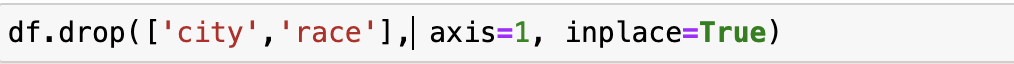
A screenshot of a computer

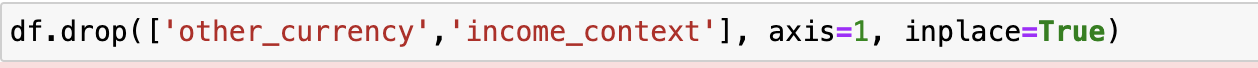
Description automatically generated

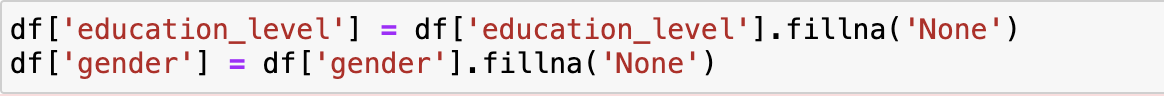
  
  
To enhance clarity and consistency across variables, column renaming was performed. For instance, the column "Job" was renamed to "Job Title" to provide a more descriptive label that accurately represents the information contained within the column. This ensures that users can easily understand the meaning of each variable without ambiguity or confusion.

1. **Dropping Columns:**





  
  
During the data cleaning process, certain columns were identified as redundant or containing irrelevant information for the analysis goals. These columns were subsequently dropped from the dataset to streamline the data and focus on the most relevant variables. By removing unnecessary columns, we ensure that the dataset remains concise and contains only the essential information needed for analysis.

1. **Handling Missing Values:**  
     
   

A screenshot of a computer

Description automatically generated



  
  
Missing values in the dataset were addressed through a systematic approach to maintain data integrity. Techniques such as imputation were employed to fill in missing values where appropriate, ensuring that no valuable information was lost during the analysis. Additionally, incomplete entries containing a significant number of missing values were excluded from the dataset to prevent bias in the analysis results.

1. **Currency Conversion and related manipulation:**



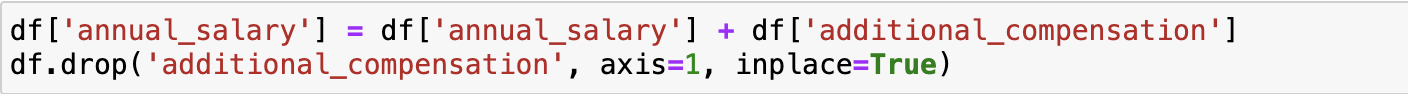
A screen shot of a chart

Description automatically generated A screen shot of a computer code

Description automatically generated

A screen shot of a computer

Description automatically generated

  
  
To facilitate meaningful comparisons across different regions, currency conversion, and related manipulations were performed on salary data. All salary values were standardized to a common currency, such as USD, using current exchange rates. This ensures that salary data is consistent and eliminates potential biases arising from currency fluctuations, enabling accurate analysis and interpretation.

1. **Label Encoding:**

A screenshot of a computer

Description automatically generated

 A screenshot of a computer

Description automatically generated

1. **Standardization and Similarity Metrics:**

A screen shot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

**7. Analyze the data. Were you able to answer your questions? If not, what else might need to be done?**

1. **Which industry pays the most?**

A graph of blue and white bars

Description automatically generated

Computing or Tech emerges as the highest-paying industry with an average salary of around $140,000, while Library appears to be the lowest-paying industry, offering an average salary of approximately $70,000.

1. **How does salary increase given years of experience?**  
     
   A graph of a number of years

   Description automatically generated  
     
   Salary exhibits a linear increase with years of experience up to 10 years, beyond which it appears to plateau, as depicted in the graph.
2. **How do salaries compare for the same role in different locations?**  
     
   A map of the united states

   Description automatically generated  
     
   Salaries for the same role vary across different locations, with California, New York, Washington, Texas, and Illinois emerging as the hotspots, likely due to their status as major markets.
3. **How much do salaries differ by gender and years of experience?**

A graph of different colored lines

Description automatically generated  
  
Based on the graph, salaries increase linearly for all genders as expected. Initially, men and women start with comparable salaries, but as their field experience increases, men tend to earn more than their counterparts.

1. **How do factors like education level correlate with salary?**

A graph of a graph

Description automatically generated with medium confidence  
As indicated by the graph, individuals with professional degrees such as JD or MD typically command higher salaries, followed by those with a PhD. Interestingly, there appears to be only a marginal difference between the salaries of individuals with Bachelor's and Master's degrees. Additionally, it's noteworthy that individuals without degrees also tend to earn comparatively well.

1. **Is there a “sweet spot” in total work experience vs years in the specific field?**  
   A screen shot of a graph

   Description automatically generated  
   The analysis suggests that there exists a "sweet spot" when individuals have 8-10 years of total work experience in the same field they are currently working in. Specifically, having 5-7 years of experience in the field within this total experience range correlates with higher salaries, as indicated by the scatter plot.