SALARY SURVEY

2021

Shahin Sulaiman m

2021

**Project Overview:**

This project aims to analyze a salary survey dataset to explore career trajectories across various global industries. By examining key factors such as age, job title, industry, years of experience, education, and compensation, the study seeks to uncover trends and patterns that influence salary growth and career progression.

One of the primary objectives is to understand how salaries vary across industries and how experience impacts earnings over time. For instance, some industries may offer higher starting salaries, while others provide steady salary growth with experience. The analysis will also consider the role of education in shaping career outcomes, determining whether advanced degrees lead to significantly higher salaries.

Additionally, the project will explore compensation structures, identifying disparities based on job roles, industries, or geographic regions. By identifying outliers and trends, the analysis can provide valuable insights into salary expectations, helping professionals make informed career decisions.

The cleaned and processed dataset will serve as the foundation for generating visualizations and statistical models, offering a comprehensive view of salary trends across different industries. Ultimately, this study aims to provide actionable insights for job seekers, employers, and policymakers in understanding career growth and salary dynamics.

**Dataset Overview**:

1. Age Range: The age group of the individual.

2. Industry: The sector in which the individual works.

3. Job Title: The official position held by the individual.

4. Clarification of Job Title: Additional details about the job title.

5. Annual Salary: The individual’s annual salary or earnings.

6. Additional Monetary Compensation: Extra earnings beyond the base salary (e.g., bonuses, commissions).

7. Currency: The currency in which salary and compensation are reported.

8. Other Currency: A custom currency used when it does not match the standard options.

9. Income Clarification: Additional context regarding the income structure.

10. Country: The country where the individual works.

11. State: The state or province of employment.

12. City: The city where the individual works.

13. Years of Professional Experience Overall: The total number of years the individual has worked professionally.

14. Years of Professional Experience in Field: The number of years the individual has worked in their specific field.

15. Highest Level of Education Completed: The highest degree or educational level attained.

16. Gender: The gender identity of the individual.

**Data cleaning and Preprocessing:**

Data cleaning and preprocessing are crucial steps in preparing data for analysis, ensuring accuracy and reliability. The final cleaned dataset refers to the version of the data after applying various cleaning techniques, making it ready for further processing, such as visualization, statistical analysis, or machine learning modeling.

Handling missing values is an essential first step. Missing numerical values, such as those in the "Annual Salary" column, can be filled using the mean or median to maintain data consistency. Alternatively, if a large portion of a column has missing values, removing such rows might be necessary to prevent misleading analysis. Similarly, for categorical data like "Industry" or "Job Title," missing values can be replaced with "Unknown" instead of removing them, preventing unnecessary data loss.

Standardizing data types ensures that each column has an appropriate format. For example, salary values should be numeric to allow calculations, while categorical attributes like industry names should remain as text. Any inconsistencies in data types can cause errors in further processing.

Handling inconsistent values is also vital. Variations in text entries, such as "researcher" and "Researcher," should be standardized to avoid duplicate categories. Ensuring uniform capitalization for categorical values improves data integrity and avoids unnecessary fragmentation in analysis.

Detecting outliers in numerical columns, such as extreme salary values, helps maintain data quality. Outliers can be identified using statistical methods like the interquartile range (IQR) or z-scores. Depending on their significance, they can either be removed or adjusted to prevent skewed results.

Once all cleaning steps are completed, the dataset is saved in CSV format. CSV is widely used for storing structured data, ensuring compatibility with various analytical tools. The cleaned dataset serves as a reliable foundation for generating insights and making data-driven decisions.

**Incorporating the Cleaned Data into MySQL**

After cleaning and preprocessing the salary survey dataset, the next step is to store it in a structured database for further analysis and querying. MySQL, a popular relational database management system, provides a robust platform to store, manage, and analyze data efficiently. This process involves three key steps: creating a database, designing a table, and importing the cleaned dataset.

**Create a Database in MySQL**

To begin, open **MySQL Workbench**, a graphical user interface for MySQL, and create a new database to store the salary survey data. The database acts as a container for tables and other objects. You can create it using the following SQL command:

CREATE DATABASE SalarySurvey2021;

After executing this command, select the newly created database to ensure that all subsequent operations occur within it:

USE SalarySurvey2021;

**Create a Table**

Next, define a table to store the cleaned data. The table structure should align with the dataset, ensuring that each column has an appropriate data type

CREATE TABLE SalarySurvey2021 (

AgeRange VARCHAR(50),

Industry VARCHAR(255),

JobTitle VARCHAR(255),

AnnualSalary INT,

AdditionalMonetaryCompensation INT,

Currency VARCHAR(10),

Country VARCHAR(100),

State VARCHAR(100),

City VARCHAR(100),

YearsOfProfessionalExperienceOverall VARCHAR(50),

YearsOfProfessionalExperienceInField VARCHAR(50),

HighestLevelOfEducationCompleted VARCHAR(100),

Gender VARCHAR(50)

);

QUERIES:

-- a. Average Salary by Industry and Gender

SELECT

Industry,

Gender,

AVG(AnnualSalary) AS Average\_Salary

FROM salarysurvey2021.salarysurvey2021

GROUP BY Industry, Gender

ORDER BY Industry, Average\_Salary DESC;

-- b. Total Salary Compensation by Job Title

SELECT

JobTitle,

SUM(AnnualSalary + AdditionalMonetaryCompensation) AS Total\_Compensation

FROM salarysurvey2021.salarysurvey2021

GROUP BY JobTitle

ORDER BY Total\_Compensation DESC;

-- c. Salary Distribution by Education Level

SELECT

HighestLevelofEducationCompleted AS Education\_Level,

AVG(AnnualSalary) AS Average\_Salary,

MIN(AnnualSalary) AS Minimum\_Salary,

MAX(AnnualSalary) AS Maximum\_Salary

FROM salarysurvey2021.salarysurvey2021

GROUP BY HighestLevelofEducationCompleted

ORDER BY Average\_Salary DESC;

-- d. Number of Employees by Industry and Years of Experience

SELECT

Industry,

YearsofProfessionalExperienceOverall AS Experience\_Level,

COUNT(\*) AS Employee\_Count

FROM salarysurvey2021.salarysurvey2021

GROUP BY Industry, YearsofProfessionalExperienceOverall

ORDER BY Industry, Experience\_Level;

-- e. Median Salary by Age Range and Gender

WITH RankedSalaries AS (

SELECT

AgeRange,

Gender,

AnnualSalary,

ROW\_NUMBER() OVER (PARTITION BY AgeRange, Gender ORDER BY AnnualSalary) AS RowAsc,

ROW\_NUMBER() OVER (PARTITION BY AgeRange, Gender ORDER BY AnnualSalary DESC) AS RowDesc

FROM SalarySurvey2021

),

MedianSalaries AS (

SELECT

AgeRange,

Gender,

AVG(AnnualSalary) AS MedianSalary

FROM RankedSalaries

WHERE RowAsc = RowDesc

OR RowAsc + 1 = RowDesc

OR RowAsc = RowDesc + 1

GROUP BY AgeRange, Gender

)

SELECT

AgeRange,

Gender,

ROUND(MedianSalary, 2) AS MedianSalary

FROM MedianSalaries

ORDER BY AgeRange, Gender;

-- f. Job Titles with the Highest Salary in Each Country

WITH RankedJobs AS (

SELECT

Country,

JobTitle,

AnnualSalary,

RANK() OVER (PARTITION BY Country ORDER BY AnnualSalary DESC) AS rnk

FROM salarysurvey2021.salarysurvey2021

)

SELECT Country, JobTitle, AnnualSalary

FROM RankedJobs

WHERE rnk = 1

ORDER BY Country;

-- g. Average Salary by City and Industry

SELECT

City,

Industry,

AVG(AnnualSalary) AS Average\_Salary

FROM salarysurvey2021.salarysurvey2021

GROUP BY City, Industry

ORDER BY City, Industry;

-- h. Percentage of Employees with Additional Monetary Compensation by Gender

SELECT

Gender,

COUNT(CASE WHEN AdditionalMonetaryCompensation > 0 THEN 1 END) \* 100.0 / COUNT(\*) AS Percentage\_With\_Compensation

FROM salarysurvey2021

GROUP BY Gender

ORDER BY Percentage\_With\_Compensation DESC;

-- i. Total Compensation by Job Title and Years of Experience

SELECT

JobTitle,

YearsofProfessionalExperienceOverall AS Experience\_Level,

SUM(AnnualSalary + AdditionalMonetaryCompensation) AS Total\_Compensation

FROM salarysurvey2021.salarysurvey2021

GROUP BY JobTitle, YearsofProfessionalExperienceOverall

ORDER BY JobTitle, Experience\_Level;

-- j. Average Salary by Industry, Gender, and Education Level

SELECT

Industry,

Gender,

HighestLevelofEducationCompleted AS Education\_Level,

AVG(AnnualSalary) AS Average\_Salary

FROM salarysurvey2021.salarysurvey2021

GROUP BY Industry, Gender, HighestLevelofEducationCompleted

ORDER BY Industry, Gender, Average\_Salary DESC;

**Importing Data into Tableau and Creating Insights**

After processing and storing the salary survey data in MySQL, the next step is to analyze and visualize it using **Tableau**, a powerful data visualization tool. This process involves importing the data, building an interactive dashboard, and summarizing key insights to understand salary trends, gender disparities, and the impact of education on earnings.

**1. Import Data into Tableau**

First, export the cleaned dataset or query results from MySQL as **CSV files** and import them into Tableau. This can be done through the **“Connect”** panel by selecting **“Text File”** and loading the CSV file. Alternatively, Tableau can connect directly to the MySQL database for real-time data updates.

**2. Create an Interactive Dashboard Story**

Once the data is loaded, Tableau’s **Dashboard** feature allows for the creation of an interactive visualization. Key components may include:

* **Salary Distributions**: Histograms or box plots showing salary variations across industries and experience levels.
* **Gender Comparisons**: Bar charts or side-by-side comparisons illustrating salary differences between genders.
* **Education & Salary Impact**: A scatter plot or trend analysis showing how education levels influence salary growth.

Using Tableau’s **filters, parameters, and interactive elements**, users can explore salary trends by different categories like **age, location, and job title**.

**3. Write Insights**

After building the dashboard, a detailed **summary of key findings** should be documented. Insights may include:

* **Salary Trends**: Higher salaries in tech and finance compared to retail or education sectors.
* **Gender Disparities**: Possible wage gaps between male and female employees across different industries.
* **Education Impact**: Advanced degrees generally leading to higher salaries, but varying by industry.

By integrating these insights, the Tableau dashboard provides a **clear, data-driven story** that helps professionals, policymakers, and researchers understand career and salary trajectories effectively.

**KEY INSIGHTS:**

**Education Level:**

* + A significant number of respondents hold a Master's degree or higher, indicating a highly educated workforce.
  + Higher education levels often correlate with higher salaries and more senior positions.

**Age and Experience:**

* + The majority of respondents fall within the 25-34 age range, indicating a younger workforce.
  + Many professionals have 5-7 years of overall experience, suggesting a mid-level career stage.

**Geographical Distribution:**

* + The data includes professionals from various countries, including the United States, United Kingdom, Canada, and others.
  + Within the U.S., states like Massachusetts, California, and New York have a high concentration of respondents.

**CONCLUSION**:

* Salaries vary significantly across industries, with Tech, Engineering, and Senior Management roles offering the highest compensation. This highlights the lucrative nature of these fields.
* Higher education levels and more years of experience generally correlate with higher salaries and more senior positions. This underscores the importance of continuous learning and career development.
* While there is a strong presence of women in various roles, certain industries still show a gender disparity, particularly in senior technical positions. Efforts to promote gender diversity and inclusion in these fields are essential.