Mathematics For Computer Science Engineers

UE23MA242A

Teaching Assistants : Archishman VB, Suchir M Velpanur, Neha Bhaskar

Salary Analysis Case Study

Introduction

The two sections below, Background and Case Study provide context for the data science hackathon. This exercise will allow you to test your skills in using the Python programming language to effectively explore the characteristics of a dataset and analyze the features using descriptive statistics such as summary statistics, tables, and graphs. Happy coding!

Background

The given dataset captures a diverse range of information on individuals' professional profiles, including age, job roles, salaries, additional compensation, industry sectors, and educational levels across multiple countries. By analyzing this data, researchers and policymakers can gain valuable insights into global compensation patterns, disparities across demographics, and how variables such as age, education, and experience impact income. Additionally, this dataset enables companies to benchmark salaries, address gender pay gaps, and identify industry-specific compensation trends. Such insights are valuable for both individual career planning and organizational decisions related to recruitment, compensation structures, and employee retention strategies.

Case Study

A researcher, having acquired a certain dataset, would like to view it from the lens of statistical analysis to analyze various factors affecting each of Annual salary, additional compensation w.r.t other factors given in the dataset

Dataset Description

The different variables involved in the dataset include :

1) Timestamp : Date and time of data entry, in the format MM-DD-YYYY HH:MM:SS
2) Age : Age of the respondent at the time of entry
3) DOB : Date of birth of the respondent, in `DD-MM-YYYY` format
4) Industry : The industry sector in which the respondent works (e.g., Education, Computing, Nonprofits).

5) Job Title : The respondent’s specific job role or position (e.g., Librarian, Manager)
6) Additional Context : Any extra information provided by the respondent about their job or work conditions
7) Annual Salary : The respondent's base annual salary
8) Additional Monetary Compensation : Any additional income the respondent earns, like bonuses or freelance income
9) Currency : The currency in which the salary is expressed
10) Country : The country where the respondent is based.
11) City : The specific city of the respondent’s workplace
12) Years of Professional Work Experience Overall : Range of years representing the respondent's total work experience
13) Highest Level of Education Completed : The highest degree or level of education attained by the respondent
14) Gender : The respondent’s gender

Problem Set

Unit 1

1)Classify the features in the Salary dataset into their appropriate data types (ordinal, nominal, interval, or ratio). Provide a rationale for each classification.
2) Identify and describe any data quality issues or inconsistencies within the dataset. What steps would you take to clean and preprocess the data to ensure its accuracy and reliability for further analysis?
(Every feature has some anomaly or the other)
3) A summary statistic provides a numerical summary of a specific feature within the dataset. There are two commonly used categories of summary statistics: those that indicate the central tendency and those that indicate the spread of the data. Identify the most appropriate measure of central tendency for each attribute in the dataset and state its corresponding value. Additionally, calculate the standard deviation and range of values for each numerical column.
4) Plot histogram and box plot for 'Age' and 'Annual salary' variables. From this,
i) Identify the type of distribution each of the variables follow
(Hint : limit scale of visualizations for both histograms and box plots)
ii) Identify the number of outliers for each variable.
iii) Bonus question : After identifying type of distribution followed by both variables, try to plot a new histogram and boxplot for 'Annual Salary' variable after correcting any anomalies noticed in your visualizations

(Hint : Try to rescale histogram and box plot according to distribution noticed from previous visualization)
5) What actions would you take to resolve the presence of outliers? Visualize the changes.
(Hint: Use boxplot and histogram to visualize either using capping)
6) Examine the normal probability plot (Q-Q plot) for the 'Annual salary'' variable in the dataset. Based on the shape and trend of the plot, what conclusions can be drawn? Provide a rationale for your conclusions.
7) Calculate the correlation between ‘Age’ and other numerical variables. Identify variable having highest correlation with ‘Annual salary’
8) Sample 10,000 rows randomly and generate a pairplot that includes the variables 'Age,' and 'Annual salary' while using 'Additional monetary compensation' as the hue in the dataset. What insights can be gained from the pair plot, and how does it help in visualizing the relationships between these variables?

Unit 2

9) Use hypothesis testing to answer the following:

Define a null and alternative hypothesis to investigate whether there is no significant difference in the median age associated with different values of Annual Salary. Use Wilcoxon rank-sum test / Mann-Whitney U test to analyze the relationship between these two variables. Plot a histogram to analyze your hypothesis and its results. Assume significance level as 0.05.

10) Calculate the margin of error to quantify the precision of the analysis done previously and what you can infer from the results.

Unit 3

11) Perform linear regression to predict Age using Annual salary and Additional monetary compensation. Plot the predicted vs actual number age. Also calculate MSE and RMSE. Explain what each of these metrics signify.

12) Given the variables 'Age', 'Annual Salary', and 'Additional Monetary Compensation', which represent demographic and financial attributes, what additional features could be engineered from these variables to improve the prediction of 'Age'? For example, consider how combining or transforming these variables could highlight patterns related to financial stability or career progression over time. Provide 2 such feature aggregations.