Honors Peer-graded Assignment

Project Title: IT Salary survey for Europe

Objective: The purpose of this project is to learn the competitive value of a skill set for IT specialists depending on: years of experience, position, languages, etc. with a stronger focus on Germany

Table of contents:

Brief description of the data set and a summary of its attributes

An anonymous salary survey has been conducted annually since 2015 among European IT specialists with a stronger focus on Germany. 1238 respondents volunteered to participate in the survey. The data has been made publicly available by the authors.The dataset contains rich information about the salary patterns among the IT professionals in the EU region and offers some great insights.The purpose of this survey is to learn a competitive value of a skillset for IT specialists depending on: years of experience, position, languages, etc.

Dataset from Kaggle for year 2020

Summary of attributes: Most fields are self explanatory.

- `'Timestamp', 'Age', 'Gender', 'City'` - the city employee live in, `'Position '`,
- `'Total years of experience', 'Years of experience in Germany',`
- `'Seniority level'` - level of responsibility varies to each employer
- `'Your main technology / programming language'` - main skillset or toolset
- `'Other technologies/programming languages y`ou use often', - add-on skills
- `'Yearly brutto salary (without bonus and stocks) in EUR',` - the sum of salary before the deduction of tax and insurance(s)
  `'Yearly bonus + stocks in EUR',` - annual bonus for employee
- `'Annual brutto salary (without bonus and stocks) one year ago` - the sum of salary before the deduction of tax and insurance(s)
- `'Annual bonus+stocks one year ago. Only answer if staying in same country',`

- `'Number of vacation days'` is number of vacation days for each employee
- `'Employment status',` - status of employment like full-time , part-time, self-employed
- `'Contract duration',` = length of contract, if any like unlimited, temporary
- `'Main language at work',` - used language for work like english
- `'Company size',` - range of employees define company size like 51-100, 100- 1000, etc.
- `'Company type', –` types of services like product based company, services etc.
- `'Have you lost your job due to the coronavirus outbreak?',`

Additional qualitative data:

- `'Have you been forced to have a shorter working week?` If yes, how many hours per week',
- `'Have you received additional monetary support from your employer due to Work From Home?` If yes, how much in 2020 in EUR

Actions taken for data cleaning and feature engineering

Here are the steps taken to clean the data:

1. Explored at the current state of the `salary_data` and change the Column names for better readability
2. Dropped unwanted columns that have more count of non-null values like short_workweekhrs, wfm_benefit
3. Reformatted column values or fields for further analysis using regular expressions
4. Transformed data types for each column to appropriate ones like year field from object to datetime, total_experince from object to numeric
5. Handled missing values using using fillna() method like imputing mean values for nulls in Age of employees
6. Handling outliers - explored numeric fields with data visualization methods. Removed outliers and analyzed the impact on the dataset.
7. Exploratory Data Analysis using visualization:
    a. Used `describe()` function to reveal statistical information about the numeric attributes like the count, mean, min, max of the Salary attribute.
    b. Used `value_counts()` function to reveal some information about our categorical (object) attributes like position, age
    c. Used the corr() function to list the top features based on the Pearson correlation coefficient for numeric attributes (measures how closely two sequences of numbers are correlated). Built pair plots, scatter plots, hexbin plots, and correlation matrices
    d. Graphical representations to understand relation between target (Annual salary) and independent variables
        - *Average years of Experience by Seniority Level*¶
        - *Number of employees by Total Experience groups*
        - *Histogram of Age distribution*
        - *Salary distribution over two years*

- *Bonus and Stocks Distribution over two years*
- *Gender distribution of annual Salary versus Years of Experience*
- *Employee concentration across cities*
- *Salary distribution of employees across work locations*
- *Average salary distribution by Company type*
- *Salary distribution based on total experience by gender*

8. Established a correlation between the response variable (Annual Salary) and other predictor variables,as some of them might not have any major impact in determining the salary of the employee and will not be used in the analysis.
    a. Used the corr() function to list the top features based on the pearson correlation coefficient,
    b. Built a pairplot() and heatmaps() to visually inspect the correlation between some of the features and the target variable and possible ways to spot the outliers that might be present in the data.
9. Performed log transformation using the distplot() function to make the 'annual salary ' attribute normally distributed. The assumption of the normal distribution must be met in order to perform any type of regression analysis.The log method transformed the 'annual_salary_without_tax' distribution into a more symmetrical bell curve and the skewness level now is -1.11, well within the range.
10. Performed feature engineering methods to come up with a good set of features to train on.
    - Feature transformation:
        - reduce categorical values to minimum categories to perform one-hot coding
        - one hot encoding categorical column values (new col for each col value)
    - Feature selection
        - select only those attributes which best explain the relationship of the independent variables with respect to the target variable, 'Annual Salary' using heatmap() and correlation coefficients
    - Feature Scaling: Feature scaling reduces distortions caused by variables with different scales.Used the StandardScaler() function
    - Feature Extraction using PCA:
        - Dimensionality reduction is part of the feature extraction process that combines the existing features to produce more useful ones.
        - Applied the fit_transform() function to reduce the dimensionality of the dataset down to 5 dimensions.
        - Calculated 'Explained variance ratio for 5 PCA dimensions

Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

Before data cleaning:

- The Column names need to be cleaned as they are mix of upper case and lower case letters and some columns have long name making it hard to read
- Most of the columns has null values like Gender, Main language at work, etc
- The Column values need to be cleaned as they are alphanumeric characters that doesn't make sense

- change datatypes of columns to operate on them like changing salary columns from object to float64 or Int64
- Replacing columns values into appropriate form
- Extracting Year information from timestamp columns
- Most of the columns have null values more than 10
- We can drop columns that have more count of non-null values like short_workweekhrs, wfm_benefit

Insights:

- 54 cities have records that occur only once so it's not worth excluding them as we further reduce out data
- 152 distinct positions occur less than 10 times in data so it's not worth excluding them as we will further reduce our data size. Data size is important for training ML models
- linguistic diversity, we have a large population who knows English, German and very few knows Russian and multiple languages.
- not large population was impacted by COVID and we only have around 68 impacted from 1247 population
- Age of employees have a total 18 missing values/ nulls and now has been imputed with the mean value based on this information - Range of age: 49.0, Mean of age: 32.54, Median of age: 32.0, Mode of age: 30.0.
- Last year bonus/stocks, annual bonus stocks, last year salary, Other Tech/skills, main tech, vacation has the highest number of missing data
- There are 279 employees with 80K+ annual salary and 50% of employees took 34 days vacation while employees with salary below 80K took 30 days average vacation time. This implies better work life balance for high paying roles. however , there are some outlier like 350 days vacation time and annual salary above 250K
- There are values above 300 for total_experience_years that deviate from the rest of the population and do not seem to follow the trend. That needs to be removed as it’s unbelievable to have more than 80 years of experience
- Annual Salary data - the minimum value is greater than 0. Also,the difference between the minimum value and the 25th percentile is smaller than the 75th percentile and the maximum value. This means that our data might not be normally distributed but it's left skewed
- Age of employees -large working population falls under 40.the minimum start age of employment is 20 years. While 25% of the working population is under 29 years and 50% of the working group is under 32 years. Hence, working population age is heavily skewed on younger age groups. Their average retirement age is 60 and hence we can see there is no limited data for employees above 60.
- Based on Univariate Analysis, we found more than 800 employee have <10 years of experience and 350 employees have 1-20 years of experience
- Roles like lead, Senior, Head has highest experience above 35 years
- Average Annual salary of employees increase by 4K over two year period from 2020 to 2021 in europe while annual bonus/stocks dropped by 1K
- UK employment market is dominated by male with 1000+ male employees , 186 women employee and only 2 diverse

- Berlin and Munich are popular working cities with around 690 employees work based out of berlin location, over 250 employees work from Munich and very few employees prefer to work from dusseldorf location.
- Highest Average salary are paid in Munich, berlin, stuttgart and frankfurt hence making, these four cities as top working location for employees
- More than 60K+ employees work in corporation, media, fintech, product based, E-commerce, start-up, consulting, and banking companies
- Employees with engineering roles like software engineer,Data engineers, frontend developer have high paying jobs over two year time period
- From Pearson's Correlation Coefficients and pair plots and correlation heatmap, we can draw some conclusions about the features that are most strongly correlated to the 'Annual Salary without tax'. They are: 'lastyear_salary_without_tax','total_experience_years', 'lastyear_bonus_stocks','annual_bonus_stocks', 'germany_experince_years', 'age'
- We have more employees that are less/more likely to be paid a higher salary based on distribution of total employee experience and distribution of yearly salaries to employees.
- Hex bins show us the densities, more employees with less experience months plus medium yearly salary. also , few employees with less experience in months have higher yearly Salary

Formulating at least 3 hypothesis about this data

- The annual salary of females is different from that of males.
- The average salary of female employees are lesser than or equal to male employees in product based companies.
- The annual salary of employees who know other tech/skills is higher than only one with no other skills/tech
- Product based companies pay higher salary than other company types

Conducting a formal significance test for one of the hypotheses

***Hypothesis 1*:** the annual salary of female is different from that of male employees

Null Hypothesis (H0)
There is no difference between the annual Salary of male and annual salaries of females.

Alternate Hypothesis (Ha)
There is a difference between the annual Salary of male and annual salaries of females.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_A : \mu_1 - \mu_2 = 0$$

- The equal sign in the null hypothesis indicates that it is a 2-tailed test.
- For our project, We will choose a 5% significance level. Therefore, our α=0.05. Since we have a 2-tailed test, we have to divide alpha by 2, which gives us α=0.025. So, if the calculated p-value is less than alpha, we will reject the null hypothesis
- Used the t-test statistics to evaluate our results.
- Based on the plot , the distribution of 'salary' values for females and males using seaborn's `distplot()` function is not same

- Average of annual salary for male is much greater than female employee (female = 58999 and male = 73038)

Conclusion:

since p_value 1.4121759392001073e-14 is less than alpha 0.05 . We rejected the null hypothesis that there is no difference between the annual salary of females and males. Hence,there is a difference between the annual salary of females and male employees.

Suggestions for next steps in analyzing this data ;summarizes the quality of this data set and a request for additional data if needed

Use more heavy data cleaning techniques to understand the other categorical columns in depth like skills, language used for work. Be cautious to select a large data set so that we don't fear reducing the dataset size during data cleaning. Start the project or analysis with a certain goal like whether it is a classification problem or prediction. This gives a direction for data preprocessing.

References:

http://localhost:8888/notebooks/Desktop/Course/IBM%20Machine%20Learning%20Certification/Salary%20EDA%20for%20IT%20European%20Specialists..ipynb#