**STAT – 650**

**MID-TERM PROJECT**

**GROUP - 6**

**SALARY SURVEY FOR THE EUROPEAN IT SECTOR**

**– A DATA MINING PROJECT PROPOSAL**

**GROUP MEMBERS:**

**Harikrishnan Raghukumar (UIN – 833003243)**

**Sneha Mishra (UIN – 733000826)**

**Seshadri Srinivasan (UIN – 532001767)**

**Swati (UIN - 933006041)**

**Darshan Maharudrappa Vasanth Kumar (UIN – 933005008)**

**Hemashruthi Ganesh (UIN - 532009854)**

**Section 1 - Introduction: Aim and Description of the Project**

* 1. **Aim**
* To aid job seekers in comprehending the prevalent IT-salary trends in Europe, 2020.
* To investigate the impact of various factors on the pay scale in order to get useful conclusions.
  1. **Background and motivation**

The COVID-19 health catastrophe is an unprecedented shock that is reshaping people's lives and means of subsistence all over the world. The IT job market has also been adversely impacted during these challenging times. Considering these uncertainties, finding a stable job has become a major concern for the working population.  Therefore, having a thorough understanding of the numerous trends and factors influencing the salary compensation for various IT-career positions is crucial in making a rational decision for a job seeker. Data mining will be instrumental in achieving this.

We aim to address the following set of questions through our analysis

1.3.1 Univariate analysis:

1. What is the current gender diversity trend in the sample population?
2. Which city in Europe may be regarded as the IT hub?
3. What are the most popular IT job roles in Europe (2020)?
4. What IT-related technologies are in high demand in Europe?
5. How much will I earn as an IT professional in Europe?
6. What is the job experience range of IT professionals?

1.3.2 Bivariate analysis:

1. What are the most lucrative job roles in the European IT sector (2020)?
2. Which technology pays the highest during 2020?
3. What will my salary growth look like as I climb up the ladder?

1.3.3 Multivariate analysis:

1. Does having prior German work experience influence the income?
2. Is there a remarkable gender-based pay disparity in the European IT Industry?
3. Does gender play an integral role as one moves up the corporate ladder?

**Section 2 – Benefits of Data Mining and Dependencies**

**2.1 How data mining can be used**

Data mining aims to discover patterns and relationships in data that help in making informed business decisions. In the scope of this report, it has been implemented to analyze the effects of various variables on the target variable, i.e., yearly gross salary. Crucial insights can be established on the correlation amidst the various attributes to aid in future model-building and training aspects.

**2.2 Dependencies**

There are certain dependencies in implementing data mining techniques to large data sets. They are not always precise, so there is a chance that the derived information is not entirely accurate. This obstacle is especially prevalent when there is a lack of diversity in the dataset. Many data analytics tools are complex and challenging to use. Data professionals need the right training and professional expertise to use the tools efficiently. Large databases are needed for data mining which makes the process cumbersome and inconvenient to manage.

**Section 3 – Analysis Outline**

**3.1 Dataset – Background**

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| Timestamp | STRING | Time at which data was collected |
| Age | INTEGER | Person's Age |
| Gender | CATEGORICAL | Person's Gender |
| City | CATEGORICAL | City of work |
| Position | CATEGORICAL | Person's job role |
| Total\_exp | INTEGER | Person's total work experience |
| German\_exp | INTEGER | Person's work experience in Germany |
| Seniority\_level | CATEGORICAL | Person's work rank |
| Main\_tech | CATEGORICAL | Main programming language used at work |
| Vacation\_days | INTEGER | No: of vacation days offered |
| Employment\_status | CATEGORICAL | Status of employment |
| Contract\_duration | CATEGORICAL | Duration of contract |
| Company\_size | CATEGORICAL | Size of the company |
| Company\_type | CATEGORICAL | Type of the company |
| Yearly Salary | INTEGER | Salary earned |

**3.2 Data Mining Steps**

**3.2.1 Data Collection:** This is a vital component in exploratory data analysis. It describes the procedure for locating and importing data into our system. ‘IT salary survey for EU region 2020’, dataset has been used from Kaggle for the purpose of this report.

**3.2.2 Data Cleaning:** It is the process of eliminating misleading variables and irregularities from your dataset. Such abnormalities may unreasonably distort the data, which will have a negative impact on the outcomes. Some steps that can be taken to clean data are:

* Removing missing values, outliers, and unnecessary rows/columns.
* Re-indexing and reformatting our data.

**3.2.3 Univariate Analysis:** It helps in identifying the quality and distribution of each feature. Some visual techniques used in this report are:

* For categorical attributes: Pie Chart, Bar graph, Word Cloud
* For numerical attributes: QQ Plot, Box plot, Histogram

**3.2.4 Bivariate Analysis:** It inspects the relation between two variables in order to ascertain their empirical relationship. Some visual techniques used in this report are:

* Cat plot
* Scatter plot
* Bar plot
* Hist plot

**3.2.5 Multivariate Analysis:** It takes more than two variables into account. This significantly reduces bias and produces a result that is most accurate. Some visual techniques used in this report are:

* Cat plot
* Scatter plot
* Bar plot
* Hist plot

**3.2.6 Feature Engineering:**This is used to leverage existing data to transform into attributes or features to aid efficiently in model building.

**3.2.7 Data modeling:**To achieve our primary objective, next step in the data mining process will be to build a statistical or Machine learning model to predict the yearly salary. A variety of regression models such as Linear regression, Random Forests, Ridge Regression and SVM can be implemented and evaluated based on criteria such as R-square Error, Mean Square Error and Mean Absolute Error.

**Section 4 – Alignment with the module**

**4.1 Process Model**

The following flowchart demonstrates the process model for this project:

Diagram

Description automatically generated

**4.2 Data Preparation**

The following steps were taken to prepare the data for further analysis:

**Step 1**: **Remove irrelevant data** - the columns that are unrelated to the analysis's goals are eliminated. This includes ‘*Timestamp’*, ‘*Yearly Bonus’*, ‘*Number of vacation days’* etc.

**Step 2**: **Data type correction** - categorical columns identified as ‘Object’ are converted to ‘category’ type. Continuous values identified as ‘Object’ are converted to ‘int’ or ‘float’.

**Step 3: Deduplication -** 16 duplicate rows were found and removed.

**Step 4: Fix errors** - columns with incorrect / misspelled values are converted to the closest / best values. E.g. In column *‘German Experience’*, we can convert ‘4 (in Switzerland), 0 (in Germany)’ to 0.

**Step 5: Handling Outliers** - All the values outside the 1.5\*IQR are considered outliers and removed.

**Step 6: Handling NULL values -** The dataset excludes columns with more than 25% NULL values. We have replaced numerical variables with the column mean, and categorical variables with the mode, if the proportion of NULL values is less than 25.

Chart, scatter chart

Description automatically generated

**4.3 Univariate Analysis**

**4.3.1 Analysis of Yearly Salary**Chart, histogram

Description automatically generated  
Chart, line chart

Description automatically generated

Most of the employees earn within the range of 50000 Euros to 90000 Euros. The annual salary follows a normal distribution.

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Yearly Gross Salary** |
| Mean | 68921.36935 |
| Median | 69000 |
| Standard deviation | 15370.96075 |
| Minimum value | 29000 |
| Maximum value | 110000 |
| Kutosis | 286.079 |
| Skewness | 12.959 |

**4.3.2 Analysis of Age**

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

The employees’ age is a right skewed normal distribution in which around 60% lies between 28 and 35.

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Age** |
| Mean | 32.141851 |
| Median | 32 |
| Standard deviation | 4.945117 |
| Minimum value | 22 |
| Maximum value | 56 |
| Kurtosis | 3.5902 |
| Skewness | 1.2072 |

**4.3.3 Analysis of Gender**

Chart, pie chart

Description automatically generated

Around 87.36% of employees surveyed are male, followed by around 12.54% of females. 

**4.3.4 Analysis of City**

**Chart

Description automatically generated**  
 Berlin has the highest no. of IT jobs in Europe, followed by Munich being the second.

**4.3.5 Analysis of Position**

Chart

Description automatically generated

Software engineer role is the most popular position in the European IT industry, followed by the role of Backend Developer and DevOps respectively. The others category represents all the remaining cities that do not lie within the 8 major cities of employment.

**4.3.6 Analysis of Total Experience**

Chart, bar chart

Description automatically generated Chart, histogram

Description automatically generated

The majority of employees have overall work experience ranging from 5 years to 10 years.

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Total experience** |
| Mean | 8.362634 |
| Median | 8 |
| Standard deviation | 4.567661 |
| Minimum value | 0 |
| Maximum value | 22 |
| Kurtosis | 2.276 |
| Skewness | 1.15 |

**Chart, box and whisker chart

Description automatically generated4.3.7 Analysis of German Experience**

**Chart, bar chart

Description automatically generated**

Most of the employees have German work experience ranging from 1 year to 10 years.

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Total experience in germany** |
| Mean | 3.448994 |
| Median | 3 |
| Standard deviation | 2.967314 |
| Minimum value | 0 |
| Maximum value | 20 |
| Kurtosis | 10.571 |
| Skewness | 2.558 |

**4.3.8 Analysis of Seniority Level**

Chart, pie chart

Description automatically generated

Senior, Middle and Lead roles constitute above 60% of all employees.

**4.3.9 Analysis of Main Technology**

Chart, bar chart

Description automatically generated

Text

Description automatically generated with medium confidence

The most popular programming languages appear to be Python, Java and JavaScript. 38% of total employees use them.

**4.3.10 Analysis of Employment Status**

**Chart, bar chart

Description automatically generated**

Full-time employees are the most common IT employees with 979 individuals.

**4.3.11 Analysis of Contract Duration**

Chart, pie chart

Description automatically generated

More than 95% of the employees are working for an unlimited contract.

**4.3.12 Analysis of Company Size**

Chart, bar chart, histogram

Description automatically generated

Large IT firms are in higher numbers than comparatively smaller firms in size.

**4.3.13 Analysis of Company Type**

Chart, bar chart

Description automatically generated

There are 556 product-based IT companies and 182 startup IT companies in Europe. This shows the presence of newly launched IT companies is close to established product-based IT organizations.

**4.4 Bivariate Analysis**

**4.4.1 Bivariate Correlation Analysis for all Features**

Graphical user interface, application

Description automatically generated with medium confidencePhik’s correlation coefficient has been used for the bivariate analysis in this report. It consistently works with categorical, ordinal, and interval variables, captures non-linear dependency, and falls back to the Pearson correlation coefficient in the event of a bivariate normal input distribution.

From the correlation plot, Yearly\_salary (target) has the highest correlation coefficient with City, Main\_tech, Work\_level and Experience. Among features, Job role is highly correlated with Main\_tech and total\_experience has high correlation coefficient with german\_experience .

**4.4.2 Analysis of Total Experience vs Salary**

**Chart, scatter chart

Description automatically generated**

There is a positive association between years of experience and income. The compensation earned rises in line with experience levels. Additionally, the distribution of years of experience and annual salary is also shown using a histogram. 

**4.4.3 Analysis of Main Technology vs Salary**

**Graphical user interface, chart

Description automatically generated**

Employees with expertise in uncommon programming languages like Go and Scala seem to make more money overall.

**4.4.4 Analysis of Job Role vs Salary**

Chart, bar chart

Description automatically generated

Engineering managers receive the highest compensation, followed by product managers and team leads. Therefore, we can conclude that the job role affects remuneration.

**4.5 Multivariate Analysis**

**4.5.1 Analysis of German Experience and Total Experience vs Salary**

**Chart, bar chart

Description automatically generated**

Having a sizable quantity of work experience in Germany—more than 50% of one's total experience—has a positive impact on the yearly pay received. This may be the case given that German respondents make up a majority of the survey sample, leading us to conclude that German companies value the German experience highly.

**4.5.2 Analysis of Gender and Job role vs Salary**

**Chart, bar chart

Description automatically generated**

There is significant salary parity between male and female respondents for the same job role. It is also noteworthy that there are no women in the top-paid managerial positions, such as engineering manager and product manager.

**4.5.3 Analysis of Gender and Seniority Level vs Salary**Chart, bar chart

Description automatically generated

Men seem to earn more than women at every level of employment, and the gap widens as we ascend the corporate ladder. Additionally, we can see that there are no women working in leadership roles.

**4.6 Feature Engineering**

The following steps were taken to identify and create the right features for future model building.

**Step 1: Drop correlated features** - Correlated features will be removed in order to improve the stability of future models. Age and Total exp, for example, have a Pearson correlation coefficient of 0.7 and should be removed.

**Step 2: Discretization -** Columns like Total\_experience can be discretized to buckets since hiring businesses tend to classify candidates' experience level as a Novice (0 experience), Advanced Beginner (0–2 years experience), etc.

**Step 3: Variable transformation** – The format of all ordinal categorical variables, such as "Seniority level," must be transformed to integer encoding.

**Step 4: Scaling** - To make it easier for future ML models to learn the data, the numerical features will be standardized by subtracting the mean and dividing by the standard deviation.

**Step 5: Creating Features** – The creation of new features from the already existing ones will help with model training. Dummy features must be produced, particularly in the case of nominal categorical variables. For example, the most popular 'Main\_tech' values can be added as a new feature.

Graphical user interface, table

Description automatically generated

Table

Description automatically generated

**Section 5 – Anticipated Results**

Through this project, we aim to predict the yearly salary of an employee in the IT sector. As part of this proposal, we have cleaned, explored and enhanced the data. By implementing machine learning in future, we expect to create a good predictor of yearly salary to help job seekers.