A blue circle with white text and wings in a circle

Description automatically generated

**Excel/XlSTAT**

**Mahsa Dorri**

**Hamid Esmaeeli**

**June 2024**

We have a dataset that shows some features of employees. The first step is to understand the features and perform initial preprocessing steps.

Now let's continue with the documentation.

**Step-by-Step Documentation:**

**Understanding the Initial Dataset:**

We have a dataset that shows some features of employees. The first step is to understand the features and perform initial preprocessing steps.

**Formatting Dates and Calculating Age:**

Ensure that the bdate (birthdate) column is correctly formatted as a date. Compute the age of each employee by finding the difference between today's date and the bdate.

A screenshot of a computer

Description automatically generated

**Transforming Education Years to Education Category:**

Convert the number of education years to education using conditional formulas.

A screenshot of a computer

Description automatically generated

**Identifying and Removing Noise or Null Values:**

Use the filter option to check for null values in the dataset. This allows us to see if there is any noise data in a new category for each column.

Remove rows with any null values to ensure the dataset is clean and ready for further analysis.

**Creating Frequency Tables and Charts:**

Create a frequency table to show the count and proportion of individuals in each education category. Use the frequency data to create bar charts for education levels and relative frequencies. Create a Pareto chart combining the frequency data and cumulative frequency percentage.

A screenshot of a graph

Description automatically generated

**Creating a Pivot Tables** :

**Based on Minority Status and Gender:**

We created a pivot table to analyze the distribution of minority status across different genders, resulting in a bar chart that shows 170 male minority employees and 31 female minority employees.

A screenshot of a graph

Description automatically generated

**Creating a Pivot Table Based on Job Categories**:

We created a pivot table to analyze the distribution of employees across different job categories, resulting in a pie chart that shows Jobcat-1 with 316 employees, Jobcat-2 with 23 employees, and Jobcat-3 with 71 employees.

A screenshot of a graph

Description automatically generated

**Creating a Pivot Table Based on Education Categories:**

We created a pivot table to analyze the distribution of employees across different education categories, resulting in a pie chart that shows PhD (1), Masters (21), School (46), Bachelors (72), College (103), and High School (167).

A pie chart with numbers and a graph

Description automatically generated

**Histogram:**

For understanding each feature and how to handle it, we use bar charts available in Excel. Additionally, XLSTAT provides more detailed information and options for analysis, showing the distribution of the data.

* The age data is not normally distributed. The statistical details include: Max (95), Min (53), Skewness (0.93343652), QRT1 (58), QRT2/Median (62), QRT3 (75), IQR (17), Deviation (131.762112), and StdDev (11.4787679).

A graph with blue squares

Description automatically generated

* The histogram of salary shows it is not normally distributed, with the following statistics: Max (86,250), Min (16,200), Skewness (1.66214127), QRT1 (24,300), QRT2/Median (28,800), QRT3 (36,000), IQR (11,700), Deviation (201,184,288.7), and StdDev (14,183.94475).

A screenshot of a computer

Description automatically generated

* The histogram shows that the starting salary (Bsalary) is not normally distributed. Detailed statistics include Max (45,000), Min (9,750), Skewness (1.962550765), QRT1 (12,750), QRT2/Median (15,000), QRT3 (17,250), IQR (4,500), Deviation (40,545,368.28), and StdDev (6,367.524502).

A graph with numbers and a bar

Description automatically generated with medium confidence

* The histogram of previous experience (PrevExp) shows it is not normally distributed, with the following statistics: Max (432), Min (0), Skewness (1.45759502), QRT1 (19), QRT2 (52), Median (52), QRT3 (81), IQR (62), Deviation (201184288.7), and StdDev (14183.94475).

A graph with a bar and a bar graph

Description automatically generated with medium confidence

* In this section, we also look at the boxplot, which helps us make decisions based on the IQR, QRT1, and QRT3. In the future, we might use the 99th percentile for more precise decision-making.
* A point to consider is that QRT2 (the median) and the mean are equal.

**Normality Test Results:**

The normality tests performed show that for all features, the p-value is less than 0.05, confirming that the data is not normally distributed. This result aligns with the observations made from the histograms.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A white grid with black text

Description automatically generatedA table with numbers and letters

Description automatically generated

* A key point to note is that normalizing and standardizing the data does not change the distribution, as can be observed in the charts.

A table of numbers with numbers

Description automatically generated

A group of graphs showing different colored and numbers

Description automatically generated with medium confidence

**Evaluation of Chi-Square and Gamma Distributions for Data Fit Using XLSTAT:**

We check the Chi-square and Gamma distributions using XLSTAT and find that the data does not follow these distributions. The reason for rejecting these distributions is that the p-value is less than 0.05, leading to the rejection of the null hypothesis (H0) that the data follows these distributions.

**Identification and Removal of Outliers Using Quantiles:**

The next step is to identify outliers, which can be done using the quantiles option. I used the 0.99 and 0.01 quantiles to remove extreme data points.

A screenshot of a computer screen

Description automatically generated

A graph on a white sheet

Description automatically generated

A screenshot of a spreadsheet

Description automatically generated

**Correlation Matrix**

In the numeric variables, only salary and job category have a significant correlation, as seen in the correlation analysis. For the other variables, even if the p-value is positive, there is no meaningful relationship, indicated by a low Spearman correlation.

A screenshot of a spreadsheet

Description automatically generated A screenshot of a spreadsheet

Description automatically generated A screenshot of a report

Description automatically generated

We used biserial correlation to analyze the relationship between salary and minority status, finding a correlation coefficient (r) of -0.150 and a p-value of <0.0001. Since the p-value is lower than 0.05, we reject the null hypothesis, indicating a significant association between salary and minority status.

A screenshot of a calculator

Description automatically generated

We used biserial correlation to analyze the relationship between salary and gender, finding a correlation coefficient (r) of -0.445 and a p-value of <0.0001. Since the p-value is lower than 0.05, we reject the null hypothesis, indicating a significant association between salary and gender.

A screenshot of a data sheet

Description automatically generated

**Final Step: Modeling**

In the final step, we performed regression analysis to build a model. The regression of the variable norm-salary was carried out using the variables norm-salbegin and jobcat.

