Skill Trends Insight

Anastasiia Sviridova
Porimol Chandro
Łukasz Janisiów

Overview

The report analyses current trends in skills, seniority levels, and salary expectations for Big Data Analyst roles, based on job posting data sourced from Poland (justjoin.it). The goal is to provide actionable insights for helping LearnTech Solutions in designing targeted training programs tailored to market demands.

Tools & Technology

- Programming Languages: **Python**
- Data Processing Libraries: **Pandas, PySpark**
- Visualization Libraries: **Matplotlib**, **Seaborn**
- Cloud Platform: **Kaggle** and **Google Colab Notebooks**
- Version Control Platform: **Github**

Data

- **Data Collection**: The dataset was collected from Kaggle, with the original source being a polish job portal, Just Join IT. The analysis covers job postings from October 2021 to September 2023.
- **Scope of Data**: The dataset includes detailed job descriptions, which were cleaned and preprocessed to ensure relevance and reliability for the analysis.

Methodology

- **Data Filtering:** Job postings selected for analysis included those containing the terms "big" and "data" or "analyst" and "data" in their job titles. This approach targeted roles specifically related to Big Data and Data Analyst positions while avoiding overly restrictive filters, as requiring all three terms significantly reduced the dataset size. Job offers have also been aggregated by ID to avoid considering the same job multiple times.

```
filtered_df = df.filter((df.title.rlike("(?i).*big.*") | df.title.rlike("(?i).*analyst.*")) & df.title.rlike("(?i).*data.*"))
```

- **Date Encoding:** Job posting dates were categorized using a "Quartile-YY" format to enable time-series analysis and identify trends across different quarters.
- **Skill Standardization:** Similar skill names were consolidated for consistency (e.g., "Pyspark" was standardized to "Apache Spark," and "Powerbi" to "Power BI") using conditional transformations. Due to the manual nature of this process, the focus was limited to skills that appeared in more than 100 job postings.

```
skills_df = skills_df.withColumn( "quarter",
    F.when((F.month(F.col("published_at")).between(1, 3)), "Q1")
    .when((F.month(F.col("published_at")).between(4, 6)), "Q2")
    .when((F.month(F.col("published_at")).between(7, 9)), "Q3")
    .when((F.month(F.col("published_at")).between(10, 12)), "Q4")
)
```

```
flattened_skills = flattened_skills.withColumn(
    "skill_name",
    when(col("skill_name") == "Excel", "Ms Excel").otherwise(col("skill_name"))
)
```

- **Skill Categorization:** Skills were divided into six categories (Programming Languages, Big Data, Analytics Tools, Databases, Cloud Platforms, General Skills) to simplify analysis and highlight trends within each group.
- **Analysis:** Multiple visualizations were created to effectively interpret the data. For overall analysis, plots were generated to display average salary trends and the demand for key skills. For skill categories, time-series plots and pie charts illustrated demand fluctuations across different skills, while additional analyses focused on average salaries associated with each skill level.

Future Work

Future work could focus on:
- Analyzing emerging skills and technologies in the Big Data field.
- Expanding the dataset to include more recent and geographically diverse data.
- Developing predictive models to forecast future in-demand skills and salary trends.