Report: Analysis of Teaching Staff Data (2015) using PySpark

M. SAI KIRAN

This report summarizes the analysis performed in the Jupyter Notebook Untitled.ipynb on the U-2015-DCF-Block_1D-Teaching_Staff_Summary-2015.csv dataset. The analysis utilized PySpark for data processing and aggregation, alongside Pandas, Matplotlib, and Seaborn for visualization.

Dataset Description

- **Source File:** U-2015-DCF-Block_1D-Teaching_Staff_Summary-2015.csv
- **Content:** Contains summary statistics for teaching staff across various institutions for the **survey year 2015**. Each row typically represents a specific designation within an institution.
- **Original Columns:**
  - institution_id: Identifier for the institution.
  - name: Name of the institution.
  - survey_year: The year the survey data pertains to (2015 in this dataset).
  - designation: The specific teaching post (e.g., Professor, Assistant Professor).
  - sanctioned_strength: The approved number of positions for that designation.
  - in_position_direct: Number of staff in position through direct recruitment.
  - in_position_cas: Number of staff in position through Career Advancement Scheme (CAS).
  - no_of_phd_teachers: Number of teachers in position holding a PhD.
- **Data Cleaning:**
  - Rows with designation "ALL" (likely representing totals) were filtered out.
  - String "NA" values were interpreted as nulls during loading.
  - Null values in in_position_direct and in_position_cas were filled with 0.
  - Numeric columns were cast to FloatType.
- **Added Columns:**
  - total_in_position: Calculated as in_position_direct + in_position_cas.
  - vacancy: Calculated as sanctioned_strength - total_in_position. *Note: This calculation results in null if sanctioned_strength is null.*

Insights and Findings

Based on the PySpark aggregations and generated visualizations:

1. **Top Institutions by Staff Size (Bar Chart):** There is significant variation in the total number of teaching staff across institutions. The bar chart identifies the 10 institutions with the largest staff complements in 2015, although the specific names are truncated in the output image. Marathwada Agricultural University and Lovely Professional University are visible among the initial rows.
2. **Dominant Designations (Bar Chart & Pie Chart):** The analysis of staff distribution by designation shows that certain roles are much more prevalent than others. **Assistant Professor**

and **Associate Professor** appear to be the most numerous designations, contributing significantly to the overall staff numbers. The pie chart visually confirms this, showing these two categories make up the largest slices among the top 5. □

3. **Staff Count Distribution (Histogram):** The histogram (filtered for entries with 1 to 499 staff) suggests that most institution-designation combinations have relatively small numbers of staff. The distribution is skewed to the right, indicating a large number of entries with fewer staff members and progressively fewer entries with very high staff counts within this range.
4. **Correlations Between Staff Metrics (Heatmap):** The correlation matrix reveals expected and some interesting relationships:
   - Strong **positive correlations (0.9+)** exist between **sanctioned_strength** and **total_in_position**, as well as its components (in_position_direct, in_position_cas), and no_of_phd_teachers. This suggests institutions with higher sanctioned strengths generally have more staff in position and more staff with PhDs.
   - **total_in_position** is strongly positively correlated **(0.9+)** with **no_of_phd_teachers**, indicating that larger staff complements tend to include more PhD holders.
   - **vacancy** shows a very strong positive correlation **(0.9)** with **sanctioned_strength**. This implies that higher sanctioned strength is strongly associated with a higher absolute number of vacancies.
   - Moderate **positive correlations (around 0.4)** exist between **vacancy** and the in_position columns and no_of_phd_teachers. This might suggest that even institutions with many staff in place still face significant vacancies if their sanctioned strength is high.
5. **Data Quality Notes:** The presence of **null values** in sanctioned_strength and no_of_phd_teachers limits some analyses. For instance, vacancy could only be calculated where sanctioned_strength was non-null. The correlation analysis also had to drop rows containing nulls in the selected columns.

Recommendations

1. **Investigate Null Values:** Determine the reason for nulls in sanctioned_strength and no_of_phd_teachers. Are they truly missing, not applicable (e.g., for certain temporary posts), or data entry issues? This understanding is crucial for more accurate vacancy and qualification analysis. Consider targeted imputation strategies if appropriate. □
2. **Deeper Vacancy Analysis:**
   - Calculate **vacancy rate** (vacancy / sanctioned_strength) instead of just absolute numbers for better comparison across institutions of different sizes.
   - Identify institutions or specific **designations with chronically high vacancy rates**.
   - If location or institution type data were available, analyze vacancy patterns geographically or by sector (public/private).
3. **PhD Qualification Insights:**
   - Analyze the **percentage of PhD holders** within each designation (no_of_phd_teachers / total_in_position).
   - Compare PhD percentages across different institutions or types of institutions (if possible).
4. **Longitudinal Study:** If data for multiple years exists, conduct a **time-series analysis** to track trends in sanctioned strength, positions filled, vacancies, and PhD qualifications over time.
5. **Refine Groupings:** For visualizations with many categories (like designation), consider logically grouping similar or less frequent roles (e.g., "Lecturer/Instructor", "Senior Roles") to enhance clarity.