ETC1010-5510 Project Milestone 2 - T11_Wed_skimr

A Report On The LinkedIn-World Bank Digital Data for Development

Hanchen Wang (30704456), Hao Li (32041594), Jiaying Zhang (30930685), Mohammed Faizan(31939872

<img src="header.png">

Contents

TUTOR: Samantha Dawson

LECTURER: Patricia Menéndez

1 Project and Data Description

1.1 The LinkedIn-World Bank Digital Data for Development: Industry Jobs and Skills Trends

"The World Bank Group and LinkedIn have created the Digital Data for Development collaboration to support innovative policy decisions as developing countries grapple with a rapidly changing global economy. With hundreds of millions of members worldwide, LinkedIn has the potential to offer a new, timely, and granular source of data about emerging industries, workers' changing skills composition and how they're engaging with labor markets globally." ~ 1

This collaboration enables government and policy makers to drive better policy implementations, thus creating opportunities to the global work force. The data represents LinkedIn members' data based on four metrics: Industry Employment Shifts, Talent Migration, Industry Skills Needs and Skills Penetration. The records in the data represent over 100 countries having a distribution across six major industry sectors(representing 148 industries): Financial Services, Professional Services, Information & Communication Technology (ICT), the Arts & Creative Industries, Manufacturing, and Mining/Quarrying and possessing skills within the over 50,000 distinct, standardized skills classified by LinkedIn into 249 skill groups, further categorized as: Business Skills, Disruptive Tech Skills, Soft Skills, Specialized Industry Skills and Tech Skills.

For our project, we will be working on the following questions.

Refer Appendix for variable description.

1.2 Questions and Data Cleaning

1) **Which skill category is most common across all Industry Sections and how does it vary between each section? (Hao Li)**

Data description: The data is about what skills are needed most in different industries. The dataset used in this question is “industry-skill-needs”. It showcases the rank of each skill _grp _category depending upon its skill_grp_name being used in an industry. Additionally each industry is categorized among a given industry _section. The file from the source is tidy and does not require further cleaning.

2) **What is the average percentage of net migration for each industry over the past five years and Is the growth rate of immigration within the industry related to the growth rate of the industry? (Jiaying Zhang)**

In the first part of the question, I used the data in migration_industry data and mainly focus on the net immigration in each industry. Therefore, I chose the industry name, industry code, industry group, industry group code and the net immigration rate each year for analysis. First of all, I change the name of the variable and use Pivot_Longer to extract the year. Besides that I calculate the average net migration rate of the industry and add a variable combining industry code and year and then I get the 2_migrationInd_growth_clean1 (26475 * 4). In the second part of the question, I combined migration_industry data with 4_employment_growth data to study whether the number of immigrants in an industry is related to industry development. I chose industry name, industry code and industry net growth per year. As above, I extracted the year and combined it with the industry code as the key to connect both data and then changed year to a numerical variable and then I get the 2_migrationInd_growth_clean2 (380 * 5).

3) **For each common skill_category, which industry has the highest penetration rate and what is the change of the common skill penetration rate over the period of time? (Hanchen Wang)**

Data Description

“2_skill_penetration.csv” is the data showing skill profiles for each global industry and penetration of LinkedIn skill groups at the global industry level from 2015 to 2019, in which penetration means the weight of one skill occupied each industry.

The old data (“2_skill_penetration.csv”) is 30,740 × 7 variables and the cleaned clean data (“2_skill_penetration_clean.csv”) is 4156 × 9.

4) **Find the industry_section that is best to each region/continent. (Faizan and Karan)**

5) **Which industry_name sees the maximum growth over time in each region/continent, depending on the region’s best industry_section ? (Mohammed Faizan)**

The employment growth data represents the overall rate of change of employment between a pair of consecutive years for an industry, across 2015 to 2019. This rate of change is called “growth rate” which is measured by the percentage change in the number of employees for that industry .The sample of LinkedIn members is limited to those that have a company registered on LinkedIn on their profile. For a year, the number of employees working in an industry is the cumulative sum of the shift in the employed industry of the LinkedIn members, that is,the sum of the linked profiles with no shift in industry and the difference of the number of employees entering that industry and the number of employees leaving that industry. For example, if an industry has for the year 2014, 1000 employees and 100 employees enter this industry and 50 employees leave this industry in 2015, then the number of employees in this industry for the year 2015 is 1000 + 100 - 50 and the growth rate in 2015 is 0.05%. The formula for growth rate is,

$$growthrate = (membercount_i - membercount_j/membercount_i) * 100 \tag{1}$$

The growth is described with respect to variables such as, country_code, country_name, wb_region, wb_income, isic_section_index, isic_section_name, industry_id, industry_name, year, and growth_rate. However, the original data is in a wider format with growth rates for different years represented in the same row. As such, in accordance with the tidy data definition, the data was transformed into longer format and hence cleaned. For the question, records are filtered for the appropriate industry section for each region and then analysis is being done for industries within that section.

6) **For each region, which country did the above found industry had had maximum growth? And, what is the income group of that nation? (Karan Garg)**

This data describes the industry wise employment growth across the years, 2015 to 2019. It further showcases the growth rate across variables such as Country, Region/Continent, Industry Section and Industries. The tidy data definition states that, a data set is called tidy, when "each variable is in its own column, each observation is in its own row and each value is in its own cell". In accordance with the tidy data statement,the employment growth dataset is in the wide format and values like Year given in separate columns and the **growth rate** of different years given in the same row. To make the data clean we used functions like **pivot_longer** and **separate** to clean the data and transform it into tidy data.

The dimensions of the cleaned data set are 36675 X 10.

2 References

2.1 Data Source

1) The LinkedIn-World Bank Digital Data for Development:Industry Jobs and Skills Trends - About
2) The World Bank: Industry Skills Needs Dataset(3500 X 7), Skill Penetration Dataset(20780 X 7)
3) The World Bank: Talent Migration Dataset(Industry Migration-5295 X 13)
4) The World Bank: Industry Employment Shifts Dataset(7335 X 13)
5) The World Bank: Terms of Use for Datasets(CC BY 4.0)

"License- The aggregated datasets and visuals are available to all for the public good under the Creative Commons Attribution 3.0 IGO license with attribution to both LinkedIn Corporation and the World Bank. The World Bank Group and LinkedIn Corporation (including its affiliates) do not take responsibility and are not liable for any damage caused through use of data and insights through this website, including any indirect, special, incidental or consequential damages."

Country – countries with 100,000+ LinkedIn members.

World Bank Region – countries as classified given the most recent 6 regional World Bank country categories.

World Bank Income Group – countries are classified given the most recent World Bank country classification by GNI into 4 categories: Low Income, Lower Middle Income, Upper Middle Income, and High Income.

Industry – Detailed economic activity defined through the LinkedIn industry classification (approximately ISIC Rev. 4 2 digit level), covering approximately 140 industries (industries may be excluded based on data quality considerations) which compose the six ISIC Rev. 4 tradable sectors (ISIC Index: B, C, K, J, M, R). Please see LinkedIn – ISIC industry mapping file https://datacatalog.worldbank.org/node/144635

ISIC Section – The LinkedIn industry taxonomy is mapped to ISIC Rev. 4 Sector (1 digit) categories. Data is limited to 6 tradable sectors (ISIC Index: B, C, K, J, M, R). Please see LinkedIn – ISIC industry mapping file. https://datacatalog.worldbank.org/node/144635
Tradable and Knowledge-Intensive Sectors – Six knowledge-intensive and tradable sectors, using ISIC Rev. 4 classification, are: B-mining and quarrying; C-manufacturing; J-information and communication; K-financial and insurance activities; M-professional, scientific, and technical activities; and R-arts, entertainment and recreation.

Skill Group – Skill groups categorize the 50,000 detailed individual skills into approximately 250 skills groups (some skill groups may be excluded based data quality considerations). Skill related metrics are presented at the skill group rather than detailed skill level.

Industry Skills Needs – Captures the most-distinctive, most-represented skills of LinkedIn members working in a particular industry. Based on the skills section of the LinkedIn profile. It's calculated using an adapted version of a text mining technique called Term Frequency - Inverse Document Frequency (TF-IDF).

Skill Penetration – Measures the time trend of a skill across all occupations within an industry. Based on skill addition rates, and the number of times a particular skill appears in the top 30 skills added across all of the occupations within an industry. For example, if 3 of 30 skills for Data Scientists in the Information Services industry fall into the Artificial Intelligence skill group, Artificial Intelligence has a 10% penetration for Data Scientists in Information Services. These penetration rates are averaged across occupations to derive the industry averages reported.

Migration Overview – All the metrics are based on net migration (arrivals minus departures). These net migration figures are each normalized differently to enable fairer comparisons across samples. We calculate all on an annual basis, and report an average of the last three years.
Industry Migration – Industries gained and lost. Based on the industry associated with a member's company at the time of migration. The net gain or loss of members from another country working in a given industry divided by the number of LinkedIn members working in that industry in the target (or selected) country, multiplied by 10,000.

Industry Employment Shifts – Captures the transitions among industries over time by LinkedIn members as a proxy for industry employment growth. Based on the industries declared by the companies in a member's work history.