

Out [58]: [Click here to toggle on/off the raw code.](#)

Data Mining & Wrangling | Final Project



**IF DATA IS THE  
NEW OIL,  
WHERE ARE THE  
OIL FIELDS?**

**Identifying In-Demand  
Job-Skill Segments In The  
Philippine Data Industry**

## Abstract

The Data Science, Analytics, and Engineering job sector has experienced substantial growth, creating opportunities for both companies and specialists. Navigating this dynamic field is a continual challenge as professionals adapt their skills to evolving industry demands. This study focuses on identifying pivotal skills for roles in the Philippine data industry, providing aspiring professionals with insights for strategic skill development and targeted career growth. Data from JobStreet PH, a prominent online hiring platform, was used, and dimensionality reduction techniques revealed key job segments and their associated skills. In the result, the study shows the importance of not only technical proficiency (SQL, Excel) but also soft skills and core competencies, such as strong communication, client interaction, and customer-centricity. It can be concluded that for professionals to thrive across industries in the Philippines, it is not only important to be technically knowledgeable, but also have soft skills such as communication and client management.

## Import Libraries and Functions

## Import Libraries

The following libraries and functions were imported:

- **Pandas** for managing dataframes and tabular data
- **sklearn's TfIdfVectorizer** for vectorizing the tokens and normalizing them into TF-IDF

- **Numpy** for faster manipulation of matrices
- **matplotlib's pyplot** for plotting some of the relevant visualizations in the paper
- **SQLite3** for accessing and querying created databases
- **Regular Expression** for acquiring the skills from the job descriptions
- **Collections' Counter** for collecting frequencies for the word cloud
- **Sklearn's English Stop Words** for additional stop words to add to the vectorizer
- **Warnings** for hiding the warnings and making the notebook cleaner
- **Seaborn** for plotting the other relevant visualizations in the paper
- **WordCloud** for getting the stop words, generating the word cloud, and generating image color
- **IPython** for displaying purposes
- **json** for parsing json-like columns

## Define Functions

The following functions were defined:

- **truncated\_svd** for performing Singular Value Decomposition on the given design matrix
- **project\_svd** for projecting the relevant singular values on the design matrix
- **plot\_svd** for plotting the transformed data and features on to the first two singular vectors
- **grey\_color\_func** for generating color in varying shades of gray for a given word in the word cloud
- **parse\_classification** for parsing classifications from the given list of classifications
- **parse\_location** for parsing locations from the given list of locations
- **remove\_empty\_strs** for removing empty strings in the `skill` column
- **word\_cloud\_svd** for creating word clouds for the features of the singular values
- **nssd\_cum** for plotting the cumulative nssd and finding the significant number of singular values at a given threshold
- **color** for deciding the color of each word in each singular value's word cloud

## Table of Contents

[ABSTRACT](#)

[IMPORT LIBRARIES AND DEFINE FUNCTIONS](#)

[I. PROBLEM STATEMENT](#)

[II. MOTIVATION](#)

- [III. DATA SOURCE](#)
- [IV. DATA EXPLORATION](#)
- [V. METHODOLOGY](#)
- [VI. RESULTS AND DISCUSSION](#)
- [VII. CONCLUSION](#)
- [VIII. RECOMMENDATION](#)
- [ACKNOWLEDGMENT](#)
- [REFERENCES](#)

## I. Problem Statement

The Data Science, Analytics, and Engineering job industry has been growing rapidly in the past decade. This industry presents a huge opportunity for growth and development for companies and specialists alike, but can serve as a challenge for data professionals to know what companies need and for companies to know what data professionals do. While there may be indeed unicorns in the industry who may assume every possible role in the industry, they are a rarity, and it has been considered counterproductive to assume that every data professional is a unicorn, who possesses all relevant data skills and "can do everything needed by the company" (Davenport & Patil, 2022).

Hence, by identifying the current, in-demand job clusters, we may know how the data industry is developing to make practical use of the data that they have and how they are making the best use of existing expertise. Furthermore, we can direct our own development to include the critical skills for the specific job segments we are entering in the industry to remain both competitive and on the leading edge of where this technology is developing.

Ultimately, this study aims to address the question: **What are the current, in-demand job segments of the Philippine Data Industry, and what corresponding skills are most needed in each job segment?**

[Back to TOC](#)

## II. Motivation

The evolving landscape of the job market, particularly in the aftermath of the global pandemic, presents a compelling context for our study. In the Philippines, the job market has undergone significant transformations. Phenomena such as the Great Resignation and Quiet Quitting have led to what can be described as an 'Invisible Revolution' in employment trends (Page, 2023). These shifts not only underscore the urgency to understand the current state of job markets but also highlight the unique dynamics within sectors like data science and analytics. **Our study aims to delve into this evolving landscape, focusing on the in-demand job segments and the**

**requisite skills in the data industry, a domain that remains partially explored in existing literature.**

This necessity is further amplified by the changing attitudes and motivations of employees across industries. The decline in job loyalty and the continuous search for better opportunities signal a shift in employee mindset, validating the importance of our study in identifying sought-after roles and skills in the data science domain (Page, 2023). Understanding these elements is paramount for professionals navigating this new era of employment, where flexibility and career growth are key drivers. Our study bridges these gaps, providing insights that are not only academic in nature but also practical for career planning and organizational development.

Furthermore, the growing emphasis on work-life balance and flexibility in job roles has become a critical aspect of employee satisfaction and retention. This shift in preferences underscores the need to comprehend the evolving nature of job roles, particularly in the data science sector, and their offerings in terms of flexibility and work-life balance (Page, 2023). **Our study directly responds to these trends, examining how these factors influence the data science job market in the Philippines.**

The increased demand for data science roles reinforces the relevance of our research. As industries become more data-driven, the need for skilled data scientists and analysts is more pronounced than ever. These professionals are essential in unlocking insights from vast, unprocessed data streams, making their roles critical in contemporary business operations (Page, 2023). **By identifying specific skills and roles in demand, our study not only contributes to academic discourse but also provides a practical guide for aligning educational and professional development with market needs.**

**Our study is positioned at a critical juncture, where understanding the dynamics of the data science job market is essential for both professionals and organizations.** It is uniquely poised to offer valuable insights into the development of the data industry in the Philippines, guiding individuals and organizations in navigating and strategizing their career paths and talent management in this new era of work. The implications of our findings extend beyond academia, offering practical applications for a workforce in transition.

[Back to TOC](#)

### **III. Data Source**

## **Data Source**

JobStreet PH (<https://www.jobstreet.com.ph/>) is a leading hiring platform in the Philippines, which caters to job seekers and employers. JobStreet is known for its

comprehensive listings across various industries and sectors, making it a valuable resource for labor market analysis.

Job postings were scraped from this hiring platform for data-related job posts. The scraped data comprised of 10,608 job listings, with 37 features, including the job title, job details, company details, and job location. After cleanup and preprocessing, and removing unnecessary columns, the final data was reduced to 2,681 rows with 15 features.

## Data Description

Our first set of scraped data consists of 37 features which are: `advertiser`, `automaticInclusion`, `branding`, `bulletPoints`, `classification`, `companyName`, `companyProfileStructuredDataId`, `displayStyle`, `displayType`, `listingDateDisplay`, `location`, `locationId`, `locationWhereValue`, `id`, `isPremium`, `isStandOut`, `jobLocation`, `listingDate`, `logo`, `roleId`, `salary`, `solMetadata`, `subClassification`, `teaser`, `title`, `tracking`, `workType`, `isPrivateAdvertiser`, `tags`, `currencyLabel`, `suburb`, `suburbId`, `suburbWhereValue`, `searchInsights`, `area`, `areaId`, `areaWhereValue`.

We refined our initial dataset of 37 features obtained from the first scraping, narrowing it down to a more focused set of 15 key features. This selection process was guided by the relevance of each feature to our study's objectives, ensuring that the refined dataset is both comprehensive and directly aligned with our research goals. The features retained are:

- `id` - The unique identifier for each job listing
- `jobTitle` - The specific title of the job
- `jobDetails` - An in-depth description of the job
- `teaser` - A concise summary of the job\_details
- `roleId` - Categories defining the job's nature and domain
- `classification` - Categories defining the job's nature and domain
- `subClassification` - Subcategories defining the job's nature and domain
- `workType` - Denoting whether the job is full-time, part-time, or contractual.
- `companyName` - Name of the company
- `location` - Location of the job
- `suburb` - District of the location
- `jobLocation` - Address or the region of the job
- `averageSalary` - Average salary of the job, derived from the range of 'salary'
- `currency` - Currency of the salary
- `payScheme` - Payment scheme if monthly, daily

In the end, these were reduced further to two columns, a TF-IDF vectorized `skill` column derived from the `jobDetails` column and columns from the one-hot

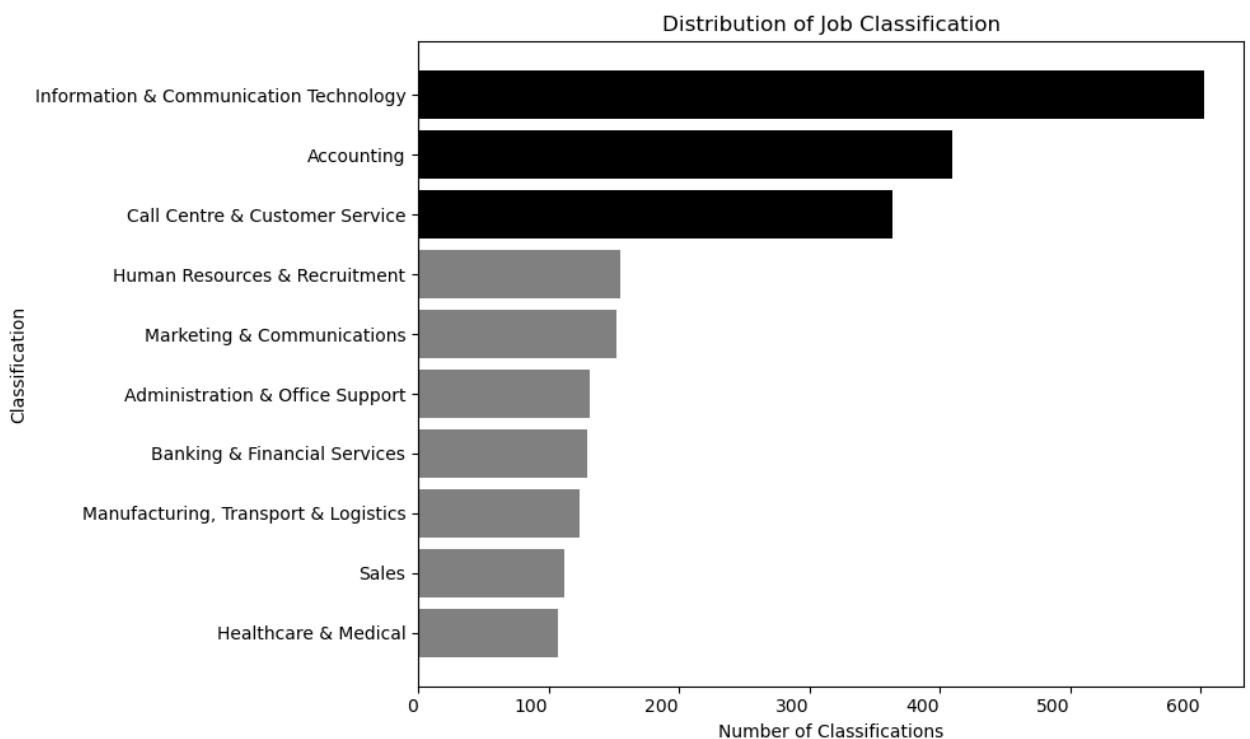
encoded `classification` column.

## IV. Data Exploration

# Data Exploration

This phase of our study is foundational to achieving these objectives. Here, we meticulously analyze the raw data collected from JobStreet PH, a leading job platform in the Philippines. This phase is dedicated to unraveling the layers of our dataset, which includes information on thousands of job listings across the data field. By scrutinizing each aspect of the data, from job titles and locations to detailed job descriptions, we aim to paint a comprehensive picture of the current job market in the data industry.

Through this exploration, we strive to lay a solid foundation for answering our central research question: What are the current, in-demand job segments in the Philippine Data Industry, and what specific skills are most needed in each of these segments? By meticulously dissecting our dataset, we prepare the groundwork for more sophisticated analyses that will follow, ultimately guiding individuals and organizations towards more informed decisions in the ever-evolving realm of data science and analytics.



**Figure 1.** Distribution of Job Classifications

As seen above, most of the jobs are comprised of Information & Communication Technology, Accounting, and Call Centre and Customer Service. Hence, it can be inferred from this that these groups would have a large effect on the grouping done

by SVD.

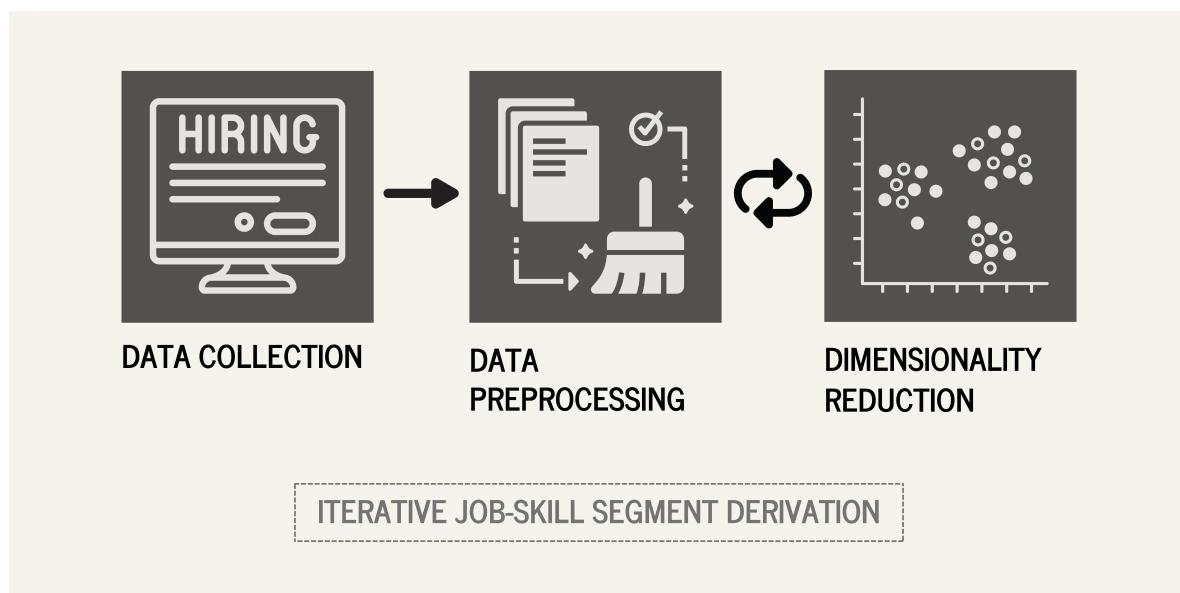


**Figure 2.** Word Cloud of `jobDescription` Column

Moreover, in Figure 2, most "skills" are general words such as management, experience, work, and skill, and so, domain-specific stop words should be set in place when vectorizing this column, as these would not have much value when analyzing the groups produced by SVDs which contain these words.

[Back to TOC](#)

## **V. Methodology**



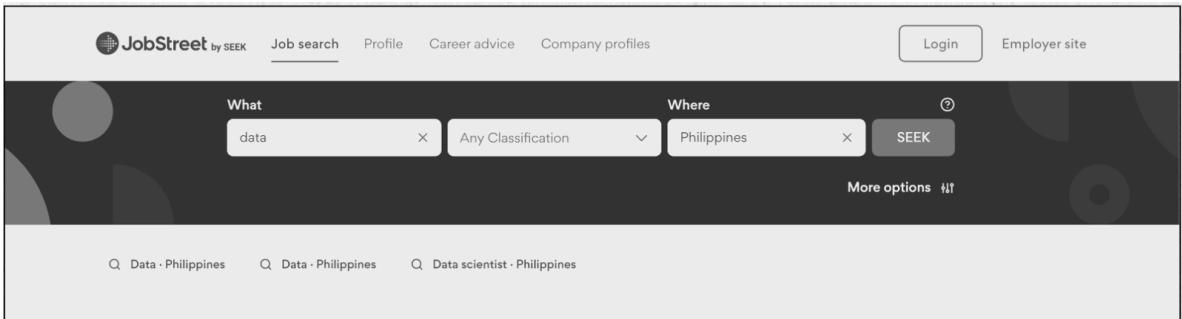
The methodology consisted of **3 phases**:

- ## 1. Data Collection

2. Data Preprocessing
3. Dimensionality Reduction & Analysis

## Data Collection

The Data Collection phase consisted of 2 parts: webscraping of data-related jobs from JobStreet (1), and webscraping the `job\_details` or job description of those jobs.



**Figure 3.** Distribution of Job Classifications

For the data, the group utilized JobStreet PH, a leading hiring platform in the Philippines, catering to job seekers and employers. JobStreet is known for its comprehensive listings across various industries and sectors, making it a valuable resource for labor market analysis.

The data collection process began with a targeted search on JobStreet PH. The group initiated a query using the keyword 'data' coupled with a geographical filter set to 'Philippines'. This search strategy was designed to yield a wide range of job opportunities related to 'data' across various industries in the Philippines. The results from this query were extensive, reflecting the diverse and dynamic nature of the data-related job market in the country.

```
js_params = {
  'siteKey': 'PH-Main',
  'sourcesystem': 'houston',
  'userqueryid': '29f535e70ed6837e41115b411d0e4f34-0245335',
  'userid': '1430f579-f0af-4b9a-b969-77535c91cc72',
  'usersessionid': '1430f579-f0af-4b9a-b969-77535c91cc72',
  'eventCaptureSessionId': '1430f579-f0af-4b9a-b969-77535c91cc72&',
  'where': 'Philippines',
  'page': '1',
  'seekSelectAllPages': 'true',
  'keywords': 'data',
  'pageSize': '100',
  'include': 'seodata',
  'locale': 'en-PH',
  'solId': 'cf9086ae-797a-4772-8f88-8dd47c63fad0'
}
```

```

js_joblist = []

for pagenum in range(0, 105):
    js_params['page'] = str(pagenum)

    response = requests.get(
        'https://www.jobstreet.com.ph/api/chalice-
        search/v4/search',
        headers=js_headers,
        params=js_params
    )

    time.sleep(1)

    js_joblist.extend(response.json()['data'])

```

**Figure 4.** Web Scraping Code Using JobStreet API

**Table 1.**

*Preview of the Preliminary DataFrame*

	advertiser	automaticInclusion	branding	bulletPoints	classification	com
0	{'id': '60350712', 'description': 'ISTA Soluti...}	0	{'id': 'f460edef-5f55-4281-949f-cc047938f9ba.1...}	[]	{"id": "6281", "description": "Information & C..."}	
1	{'id': '60282266', 'description': 'Satellite O...}	0	{'id': '11d87da4-efee-4415-8f94-58fecb763ffa.1...}	[]	{"id": "1223", "description": "Science & Techn..."}	
2	{'id': '60241355', 'description': 'Chempro QLD...}	0	{'id': '9b5fcf2b-15a5-4c75-9a26-940c3484f762.1...}	['HMO for you and one dependent', 'Up to 23 le...']	{"id": "6281", "description": "Information & C..."}	

3 rows × 37 columns

Subsequently, the group employed web scraping techniques to extract job details from these search results. This initial scraping effort successfully captured a dataset comprising 10,608 job listings, each described by 37 distinct features. These features encompass a wide array of information, including job title, company name, location, salary range, and more, offering a rich insight into the job market landscape. However, it is important to note that this initial dataset did not include the 'job\_details' or job descriptions, which are crucial for understanding the specifics of each role, including responsibilities, required skills, and potential benefits.

To address this gap, a second phase of scraping was conducted. The team used the unique job IDs ('id') from the first dataset to individually visit each job listing page and extract the 'job\_details'. This targeted scraping effort resulted in a smaller, more focused dataset of 2,681 jobs, each now including its detailed job description.

The second phase of scraping was streamlined to focus on three key features:

- 'title': The job title, providing a concise identifier of the job role.
- 'job\_id': A unique identifier for each job listing, facilitating data management and analysis.
- 'job\_details': Detailed descriptions of each job, including role responsibilities, requirements, benefits, and necessary skills.

```
job_data = []

for job_id in todo_ids_list:
    try:
        content =
requests.get(f'https://www.jobstreet.com.ph/job/%7Bjob_id%7D').text
        soup = BeautifulSoup(content, 'lxml')

        jobtitle = soup.find('h1', class='_1wkzzau0 a1msqi4y
_lnocuo0 _lnocuo1 _1d0g9qk4 _lnocuo21').text
        jobdetails = soup.find('div', class='_1wkzzau0
_1pehz540').text

        job_details = re.sub(r'[\xa0-\xff]', '', job_details)
        job_details = re.sub(r'[\u2000-\u200f]', '', job_details)
        job_details = job_details.replace('·', '')
        job_details = job_details.replace('•', '')

        job_data.append({'id': job_id, 'job_title': job_title,
'job_details': job_details})
    except Exception as e:
        print(f"Error processing job ID {job_id}: {e}")
```

**Figure 5.** Web Scraping Code To Extract 'Job Details'

**Table 2.**

*Job Details DataFrame*

	jobTitle	jobId	jobDetails
0	Data Analyst	71648955	Description ISTA Solutions an outsourcing offs...
1	NIGHTSHIFT   DATA ANALYST (POWER BI)   WFH	71644101	DATA ANALYST Work for our global clients and i...
2	Data Analyst	71571071	We are Gold Coast's leading pharmacy group wit...

The 'job\_details' would be integrated with the original dataset during the data preprocessing. This integration would enrich the initial dataset, creating a more detailed and informative collection for analysis. The enhanced dataset would now feature comprehensive information on each job listing, including its title, unique job ID, and a detailed description, offering valuable insights for a nuanced analysis of the data-related job market in the Philippines.

This two-stage data collection process ensured the acquisition of a dataset that is not only extensive in its coverage but also detailed in its content, making it an ideal foundation for in-depth analysis of the data-related job market in the Philippines.

## Data Preprocessing

The Data Preprocessing consisted of:

- 1. Removing unrelated columns from the jobs data**

The original jobs data consisted of different features about the job, which included the company's details, the display of job post or job ad, and different columns about the job's location. These columns were removed, and only columns directly related to the job was kept.

**Table 3.**

*DataFrame with Several Dropped Columns*

Out [223...]	classification	companyName	location	id	jobLocation	roleId	salary
0	{'id': '6281', 'description': 'Information & C...', 'label': 'Information & Communication Technology', 'countryCode': 'PH', 'city': 'Makati City', 'lat': 14.600000000000002, 'lon': 121.00000000000001}	ISTA Solutions, Inc.	Makati City	71648955	{'label': 'Makati City, Metro Manila', 'countryCode': 'PH', 'city': 'Makati City', 'lat': 14.600000000000002, 'lon': 121.00000000000001}	data-analyst	₱30,000 per month
1	{'id': '1223', 'description': 'Science & Technology', 'label': 'Information & Communication Technology', 'countryCode': 'PH', 'city': 'Taguig City', 'lat': 14.600000000000002, 'lon': 121.00000000000001}	Satellite Office	Taguig City	71644101	{'label': 'Taguig City, Metro Manila', 'countryCode': 'PH', 'city': 'Taguig City', 'lat': 14.600000000000002, 'lon': 121.00000000000001}	data-analyst	Normal
2	{'id': '6281', 'description': 'Information & Communication Technology', 'label': 'Information & Communication Technology', 'countryCode': 'PH', 'city': 'Metro Manila', 'lat': 14.600000000000002, 'lon': 121.00000000000001}	Chempro	Metro Manila	71571071	{'label': 'Metro Manila', 'countryCode': 'PH', 'city': 'Metro Manila', 'lat': 14.600000000000002, 'lon': 121.00000000000001}	data-analyst	₱35,000 per month

The classifications and subclassifications were broken down from the `classification` and from the `subClassification` columns using the

functions `parse_classification` and `parse_location`, respectively. Moreover, the `jobLocation` column was broken down as well. The other columns aside from the resulting description columns, and these description columns were then renamed as `classification`, `subClassification`, and `jobLocation`.

**Table 4.**

*DataFrame with Several Transformed and Renamed Columns*

Out [230...]	companyName	location	id	roleId	salary	teaser	title
0	ISTA Solutions, Inc.	Makati City	71648955	data-analyst	₱30,000 – ₱35,000 per month	Data Analytics	Data Analyst
1	Satellite Office	Taguig City	71644101	data-analyst	None	DATA ANALYST Work for our global clients and im...	NIGHTSHIFT   DATA ANALYST (POWER BI)   WFH
2	Chempro	Metro Manila	71571071	data-analyst	₱35,000 – ₱45,000 per month	The successful candidate must have a good unde...	Data Analyst

Moving on, the previously shown dataframe which contains the information about the jobs such as `jobDetails`,  `jobId`, and  `jobTitle` are then merged to the worked-on dataframe. Duplicate entries based on the column  `id` are then dropped, and the redundant column  `title` is dropped as well.

**Table 5.**

*Merged Data Frame*

	companyName	location				salary	teaser	workType
0	ISTA Solutions, Inc.	Makati City	71648955	data-analyst	₱30,000 – ₱35,000 per month	DATA ANALYST	Data Analytics	Full time
1	Satellite Office	Taguig City	71644101	data-analyst	None	DATA ANALYST	Work for our global clients and im...	Full time
2	Chempro	Metro Manila	71571071	data-analyst	₱35,000 – ₱45,000 per month	The successful candidate must have a good unde...		Full time

Columns such as `classification_description`, `subClassification_description`, and `jobLocation_label` are cleaned further, where the words after the underscore are removed, and the camel case is retained. Moreover, the columns are rearranged.

**Table 6.**

*"Finalized" Data Frame*

	id	jobTitle	jobDetails	companyName	location	roleId	salary
0	71648955	Data Analyst	Description ISTA Solutions an outsourcing off...	ISTA Solutions, Inc.	Makati City	data-analyst	₱30,000 – ₱35,000 per month
1	71644101	NIGHTSHIFT   DATA ANALYST (POWER BI)   WFH	DATA ANALYST Work for our global clients and i...	Satellite Office	Taguig City	data-analyst	None
2	71571071	Data Analyst	We are Gold Coast's leading pharmacy group wit...	Chempro	Metro Manila	data-analyst	₱35,000 – ₱45,000 per month

The "finalized" dataframe is placed into a database named as `jobs_final.db`, as a table named `jobs`. This is done through the following code:

```
conn = sqlite3.connect('jobs_final.db')
df_combined.to_sql('jobs', conn, if_exists='replace',
index=False)
```

**Figure 6.** Code for Saving the DataFrame As An SQL Database

- **2. Cleaning up the `job\_details` text column**

The job details column consisted of text that had Unicode characters and combined words. RegEx was used to separate words and replace Unicode characters with spaces.

First, the table `jobs` is retrieved from the newly created `jobs_final.db` database. This is done by connecting to the database, creating a dataframe using the pandas' `read_sql_query` and entering the query:

```
'SELECT * FROM jobs'
```

**Figure 7.** Code for Retrieving the Dataframe

Looking at the preview of the dataframe,

**Table 5.**

*DataFrame From the Database*

Out [548...]		id	jobTitle	jobDetails	companyName	location	roleId
	0	70301897	URGENT!   Real Time Analyst   Pampanga Site	JOB SUMMARY:Intraday specialist is a critical ...	Tata Consultancy Services	Central Luzon	real-time-analyst
	1	70307629	Accountant	Candidate must possess at least Bachelor's/Col...	None	Quezon City	accountant
	2	70315694	Financial Reporting Analyst	QualificationsBachelor of Science in Accountan...	Knowles Electronics (Philippines) Corporation	Central Visayas	financial-reporting-analyst

Seeing first that there are words which are conjoined such as `directedQUALIFICATIONS`, words which are directly preceded by a period or

colon without any space separating them, and cryptic unicode characters or punctuations such as \u202f and the floating dot, the team decided to clean the job\_details by separating the conjoined words and punctuation-word pairs and by converting the cryptic words or punctuations to space. These regex patterns were used:

```
patterns = [
    (r'([a-z])([A-Z])', r'\1 \2'),
    (r'([.])([A-Z])', r'\1 \2'),
    (r'([:])([A-Z])', r'\1 \2'),
    (r'\u202f', ' '),
    ('..', '')
]
```

**Figure 7.** Patterns in Cleaning the `job\_details`

Out [549...]: JOB SUMMARY: Intraday specialist is a critical role holding responsibilities to support and manage the Real Time monitoring activities of contact center operations. Intraday specialist must ensure the execution of workforce plans with focus on adherence and achievement of Service Level objectives through real time interventions. KEY RESPONSIBILITIES AND DUTIES: Good understanding of Real Time Management and its deliverables Knowledge of WFM tool like, Verint, Genesys, Aspect, IEX. Creating a game plan for the day and ensuring right people in the right place at the right time Working knowledge of the operation SLAs, GOS, Productivity, Utilization & other KPIs Performing Real Time SL Escalation Procedures Ensuring Service Levels are met daily and if they are not met, complete root cause analysis Ensure timely delivery of reports with proper insights Perform Skill Audits Managing Intraday Plan and communication of intraday performance Tracking documentation and communication of System Issues Communication, tracking and entering overtime (OT) and VTO availability on an intraday basis Perform other duties and assignments as directed QUALIFICATIONS: 3 to 4 years of relevant experience Proficient in MS Word, MS Excel, MS Power Point, VBA, Advance Excel, etc. Proficient in at least one WFM tool (Genesys, NICE, Aspect, Calabrio, Teleopti, Verint etc.) Hands on experience in working with Telephony databases Strong mathematical, analytical, communication, and organization skills Self-motivated and must excel in a highly dynamic work environment Willing to work onsite'

**Figure 8.** Cleaned `job\_details`

- **3. Creating a new feature `skills` to isolate skills from the `job\_details`**

The skills feature was made to isolate the skills from the job\_description.

By retaining as much as possible the rows with job titles, job details, or teasers which contain the terms data or analyst, the number of rows are reduced from 2681 to 2663.

Moreover, given that job\_details may contain terms such as skills; the following code was done to check if the term skill exists in the job\_details column.

```

# Find the index of the first row where 'job_details' column
# contains 'skill'
index_with_skill =
df.index[df['job_details'].str.contains('skill')].tolist()

# If there is at least one row with 'skill', get the first row
if index_with_skill:
    first_row_with_skill = df.loc[index_with_skill[0]]
    print(first_row_with_skill)
else:
    print("No row contains 'skill' in the 'job_details' column.")

```

**Figure 9.** Code for Checking Whether The Term `skill` Exists In The `job\_details` Column

Checking the printed output, it returned the first row with skill, and checking the first row with skill,

Out[552...]

```
'JOB SUMMARY: Intraday specialist is a critical role holding responsibilities to support and manage the Real Time monitoring activities of contact center operations. Intraday specialist must ensure the execution of workforce plans with focus on adherence and achievement of Service Level objectives through real time interventions. KEY RESPONSIBILITIES AND DUTIES: Good understanding of Real Time Management and its deliverables Knowledge of WFM tool like, Verint, Genesys, Aspect, IEX. Creating a game plan for the day and ensuring right people in the right place at the right time Working knowledge of the operation SLAs, GOS, Productivity, Utilization & other KPIs Performing Real Time SL Escalation Procedures Ensuring Service Levels are met daily and if they are not met, complete root cause analysis Ensure timely delivery of reports with proper insights Perform Skill Audits Managing Intraday Plan and communication of intraday performance Tracking documentation and communication of System Issues Communication, tracking and entering overtime (OT) and VTO availability on an intraday basis Perform other duties and assignments as directed QUALIFICATIONS: 3 to 4 years of relevant experience Proficient in MS Word, MS Excel, MS Power Point, VBA, Advance Excel, etc. Proficient in at least one WFM tool (Genesys, NICE, Aspect, Calabrio, Teleopti, Verint etc.) Hands on experience in working with Telephony databases Strong mathematical, analytical, communication, and organization skills Self-motivated and must excel in a highly dynamic work environment Willing to work onsite'
```

**Figure 10.** First Row with `skill`

The skills were then retrieved using this regex pattern:

```
skill_pattern = r'(?P<skill>.*?)(?P<rest>\s(.*)?)'
```

**Figure 11.** Pattern for finding the Skills

This then led to a column named `skill` where the text below shows one of the content of its rows. As one can see in the excerpt below, there are skills such as MS

Word, MS Excel, MS Powerpoint, Communication, and Organization skills.

Audits Managing Intraday Plan and communication of intraday performance Tracking documentation and communication of System Issues Communication, tracking and entering overtime (OT) and VTO availability on an intraday basis Perform other duties and assignments as directed **QUALIFICATIONS:** 3 to 4 years of relevant experience Proficient in MS Word, MS Excel, MS Power Point, VBA, Advance Excel, etc. Proficient in at least one WFM tool (Genesys, NICE, Aspect, Calabrio, Teleopti, Verint etc.) Hands on experience in working with Telephony databases Strong mathematical, analytical, communication, and organization skills Self-motivated and must excel in a highly dynamic work environment Willing to work onsite

**Figure 12.** `skill` Column Excerpt

Given that the `job_title`, `subClassification`, and `roleId` is very specific and thus, not helpful to the goal of finding general job-skill segments and that the `teaser` is essentially an excerpt of the `job_description`, it was removed from the columns. Moreover, other columns such as `suburb`, and `currencyLabel` were removed because of the null values in their data points. Furthermore, `salary` and `currency` were removed because of their lack of relevance, retrospectively, in the past iterations of dimensionality reduction. `jobLocation`, `location`, and `companyName` were removed because of their lack of significance in finding the job-skill segments. Lastly, `workType` was removed because although it may be interesting to add it as an additional feature, it tends to divert the focus from job-skill segments to job-workType segments, which detracts from the original purpose of the study, and `id` was removed as it does not provide any significant information at all. Hence, the resulting columns are the `classification` and `skills` column.

**Table 6.**

*Skill and Classification Dataframe*

Out [555...]	classification	skill
0	Call Centre & Customer Service	Audits Managing Intraday Plan and communicatio...
1	Accounting	Honest, Hardworking, Dependable, Preferably CP...
2	Accounting	and both oral and written communication skills...

The `classification` column is then one-hot encoded (where one column was not dropped in order to maintain interpretability for the job-skill segments), while the `skill` column would be TFIDF vectorized for dimensionality reduction.

**Table 8.**

*One-hot encoded Dataframe*

Out [556...]

	skill	Accounting	Administration & Office Support	Advertising, Arts & Media	Banking & Financial Services	CEO & General Management
0	Audits Managing Intraday Plan and communicatio...	0	0	0	0	C
1	Honest, Hardworking, Dependable, Preferably CP...	1	0	0	0	C
2	and both oral and written communication skills...	1	0	0	0	C

3 rows × 29 columns

Lastly, the columns are checked for null values and for empty strings through the `isna()` method of Pandas where none is found for all columns and through the `remove_empty_strs` function.

```

Out[558]: skill          0
          Accounting      0
          Administration & Office Support 0
          Advertising, Arts & Media    0
          Banking & Financial Services 0
          CEO & General Management   0
          Call Centre & Customer Service 0
          Community Services & Development 0
          Construction        0
          Consulting & Strategy      0
          Design & Architecture     0
          Education & Training      0
          Engineering           0
          Farming, Animals & Conservation 0
          Government & Defence      0
          Healthcare & Medical       0
          Hospitality & Tourism     0
          Human Resources & Recruitment 0
          Information & Communication Technology 0
          Insurance & Superannuation   0
          Legal                  0
          Manufacturing, Transport & Logistics 0
          Marketing & Communications   0
          Mining, Resources & Energy     0
          Real Estate & Property       0
          Retail & Consumer Products    0
          Sales                  0
          Science & Technology        0
          Trades & Services          0
          dtype: int64

```

**Figure 12.** Number of Null Values For Each Column

- **4. TF-IDF Vectorizing the `skills` Column**

The skills are vectorized and then normalized using the TF-IDF vectorizer.

A list of stop words were created in order to remove unrelated words from the LSA analysis in the latter part.

```

# English Stop Words
with open('stopwords.txt') as f:
    en_words = f.readlines()
    en_words = [word.replace('\n', '') for word in en_words]

df_stop = pd.read_sql('SELECT * FROM jobs', conn)

# Company Stop Words
company_names = df_stop['companyName'].dropna().tolist()
all_phrases = ' '.join(company_names)
words_without_punctuation = re.findall(r'\b\w+\b', all_phrases)
comp_stopwords = [word.lower() for word in
words_without_punctuation]

```

```

# SKLearn English Stop Words
sklearn_en = list(ENGLISH_STOP_WORDS)

# Business Words
business_words = ['citi', 'workday', 'bonifacio', 'senior',
'specialist', 'officer', 'associate', 'assistant', 'hmo',
'work', 'plaza', 'officers', 'opening', 'openings', 'us', 'll',
'join', 'joins', 'pasig', 'azure', 'japanese', 'role',
'apply', 'day', 'benefits', 'job', 'time',
'description', 'experience', 'taguig', 'workday', 'complete',
'application', 'reminder', 'skills', 'required']

# Warning Words
warn_words = ['ek', 'five', 'fr', 'inkjets', 'millennium', 'r',
'sec', 'x']

```

**Figure 13.** Code for Consolidating The Stop Words

At first, only English words were included through the sklearn's English Stop Words and the `stopword.txt` acquired online (<https://gist.github.com/larsyencken/1440509>). However, through iteration, we also find words related to the job that does not necessarily add to the analysis of the job's needed skills.

Among these unrelated words were the Company Names (eg. Accenture, Emapta). Then, business words about the job was added (eg. Pasig, workday, benefits).

The following is the code used for the TFIDFVectorizer:

```

tfidf_vectorizer = TfidfVectorizer(token_pattern=r'[a-z]+',
                                    stop_words=combined_stopwords,
                                    min_df=.05, max_df=.95,
                                    ngram_range=(1,2))

details_bow = tfidf_vectorizer.fit_transform(df['skill'])

```

**Figure 14.** TFIDF Vectorizer Settings

As seen above, the minimum document frequency is .05 and the max document frequency is 0.95, to remove the extreme outliers in the most common and rarest side of the corpus. Moreover, bigrams were also captured in order to capture interesting skills such as "SQL Proficiency" and such. Lastly, the said stop words were used and only alphabetical tokens were captured from the documents. As a side note, the

`skill` column was first extracted for TFIDF vectorizing then concatenated in the first axis back to a copy of the original dataframe by converting it into a dataframe through the `sparse.from_spmatrix` method which is used for sparse matrices (where the `skill` column is first removed in the original dataframe before concatenation as to prevent removing the word `skill` which may have been captured by the vectorizer). The copy of the original dataframe's and the skill dataframe's indices are reset as to prevent mismatching during concatenation).

**Table 9.**

*TFIDFVectorized 'skill' Column*

Out [562...]

	abilities	able	accessible	accessible locations	accounting	accuracy	achieve	activities
0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000
1	0.0	0.0	0.0	0.0	0.288326	0.0	0.0	0.000000
2	0.0	0.0	0.0	0.0	0.287454	0.0	0.0	0.135782

3 rows × 334 columns

**Table 10.**

*Final Dataframe for SVD*

Out [569...]

	Accounting	Administration & Office Support	Advertising, Arts & Media	Banking & Financial Services	CEO & General Management	Call Centre & Customer Service	De
0	0	0	0	0	0	0	1
1	1	0	0	0	0	0	0
2	1	0	0	0	0	0	0

3 rows × 362 columns

## Dimensionality Reduction & Analysis

The dimensionality reduction technique used was Truncated SVD or Singular Value Decomposition with Latent Semantic Analysis. This was used in order to define the varying features of job. Two parts of the job was explored: the vectorized `skills` column and the `classification` column.

These dimensionality reduction techniques were chosen because they can work well

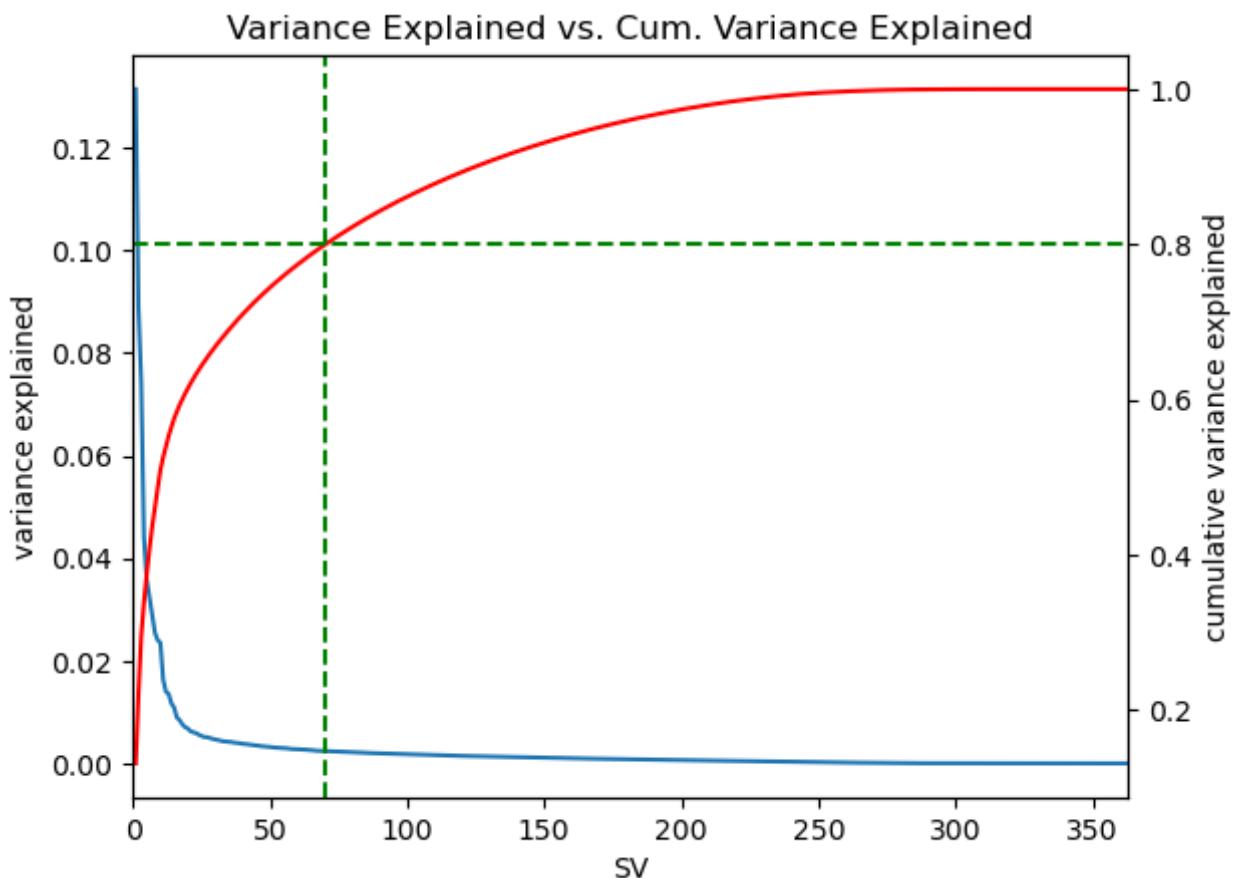
with sparse data, and the data mostly consisted of varying text about the job (job title, job description). Furthermore, with these methods, the segmentation of job posts and important features about them will surface.

After getting the decomposition of the design matrix through the function `truncated_svd`, the SVD-projected matrix is acquired by matrix multiplying the original design matrix with the acquired, truncated `p` based on the results given by the cumulative sum of the normalized sum of squared distances (the variance explained in SVD) where the set threshold (based on the standard rule-of-thumb is 80%) (UCLA Institute for Digital Research and Education, n.d.). The number of singular values to be used is derived from the plot given by `nssd_cum` which is further explained in the Results and Discussion section.

Afterwards, using the `plot_svd`, the first two singular values are plotted, and using codes from `matplotlib.pyplot` and the defined function `word_cloud_svd`, the bar plots for the first 10 singular values are accompanied by a word cloud which is produced by the `word_cloud_svd` function.

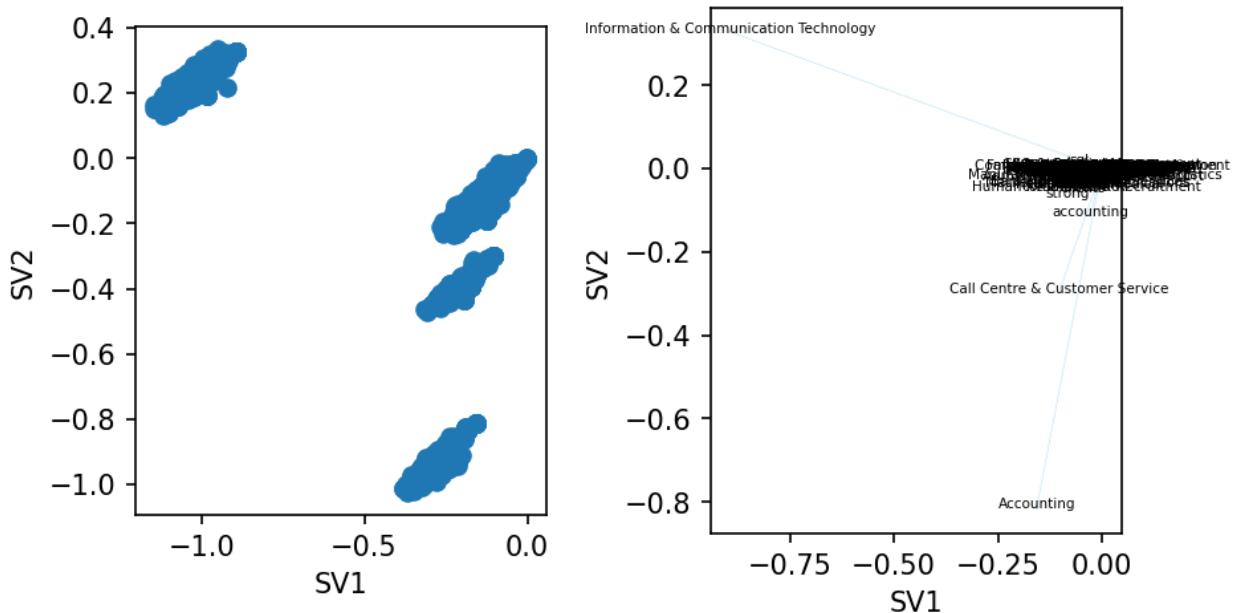
[Back to TOC](#)

## VI. Results and Discussion



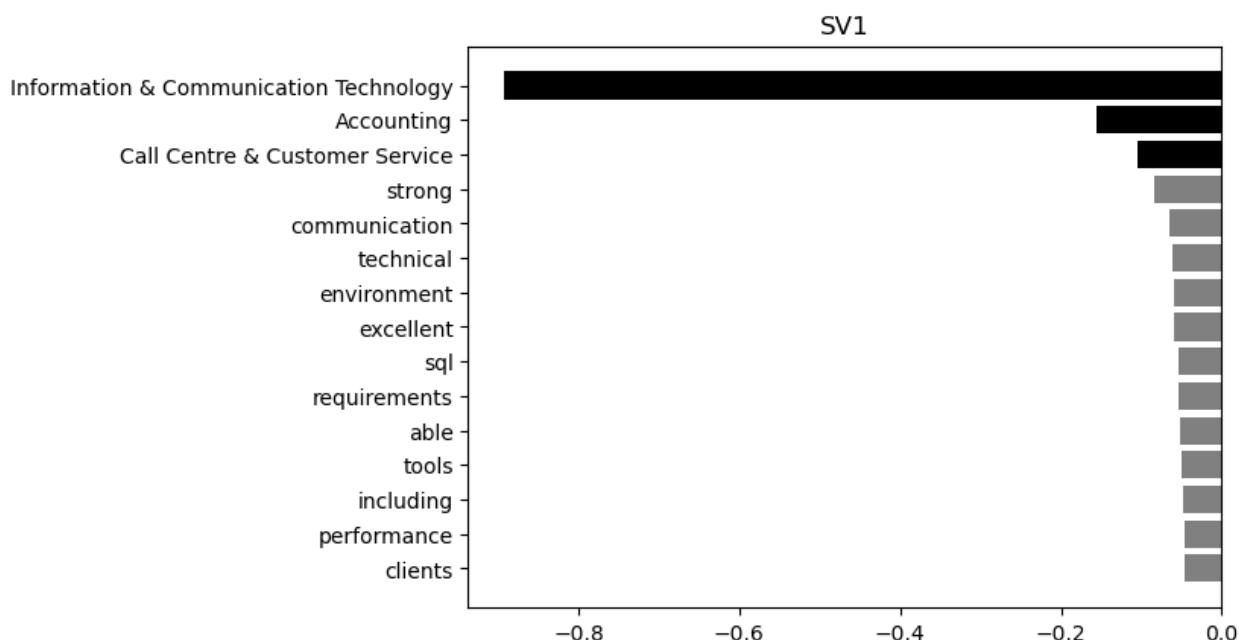
**Figure 15.** Variance Explained vs. Cum Variance Explained

Given a threshold of 80%, the needed singular values to explain most of the variance is **71 singular values**.

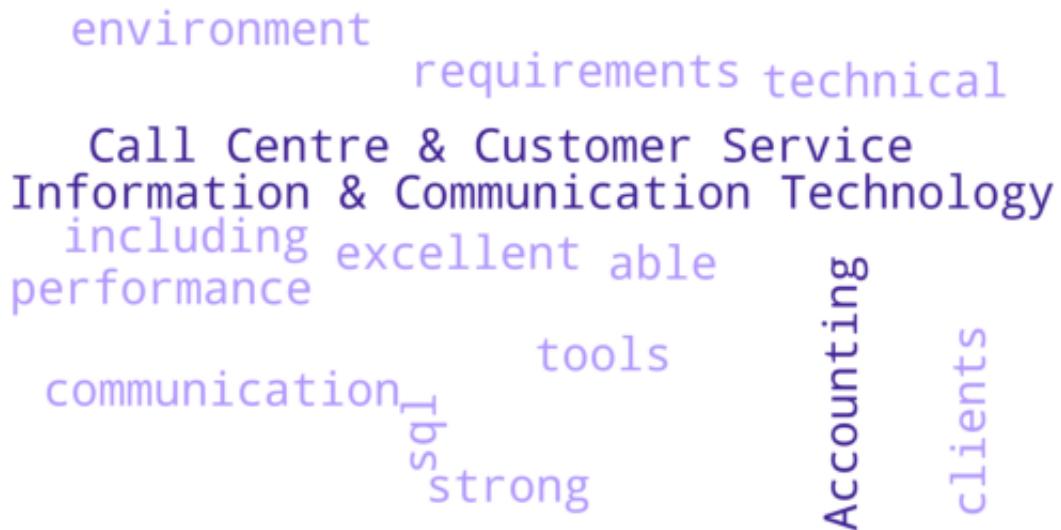


**Figure 16.** SVD Plot (Scatter and Feature Vector Arrows)

Given this plot, while clusters can be seen across the plot wherein there are two clusters in the middle of SV2 and lower end of SV1 (zero), one cluster in the lower ends of both SVs (negative for SV2 and zero for SV1, respectively), and another cluster in the positive upper end of SV2 and negative "upper" end of SV1, it would be hard to interpret the singular values using this plot alone given the numerous features, and so, to interpret each singular value, the first 10 singular values would be plotted using bar graphs and word clouds.



**Figure 17.** SV1 Horizontal Bar Chart



**Figure 18.** SV1 Word Cloud

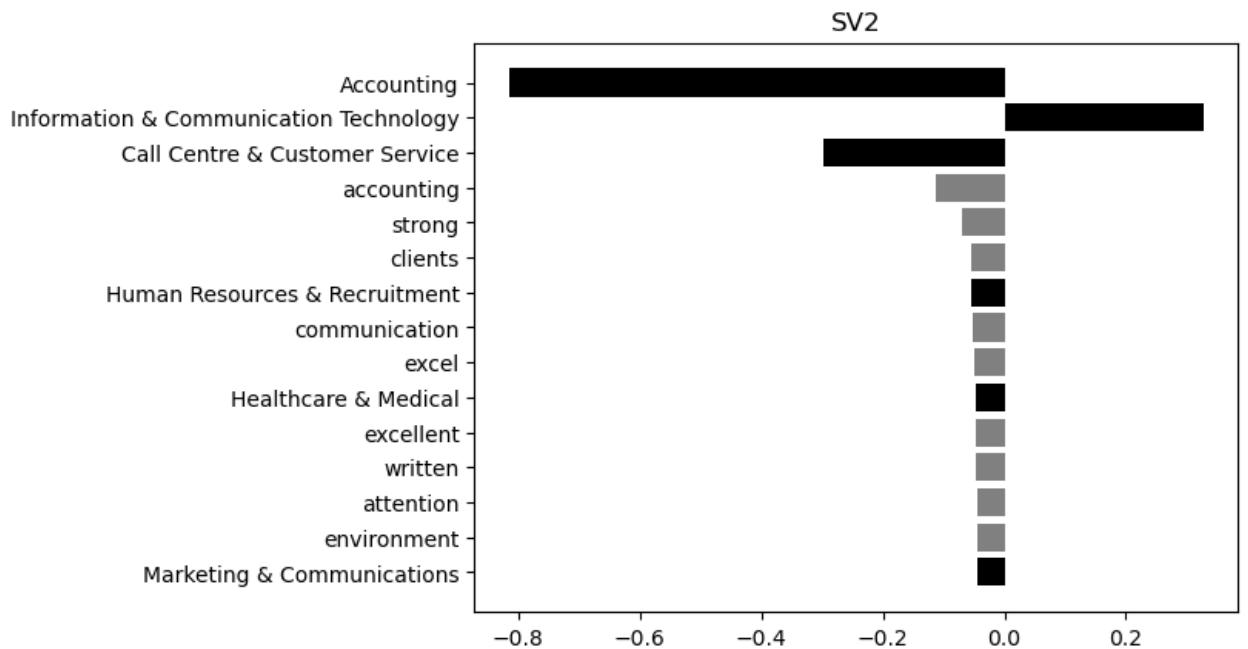
### SV1 - Technical Client Support Roles

The top segments which are hiring data-related roles are from the Information & Communication Technology, Accounting, and Call Centre & Customer Service.

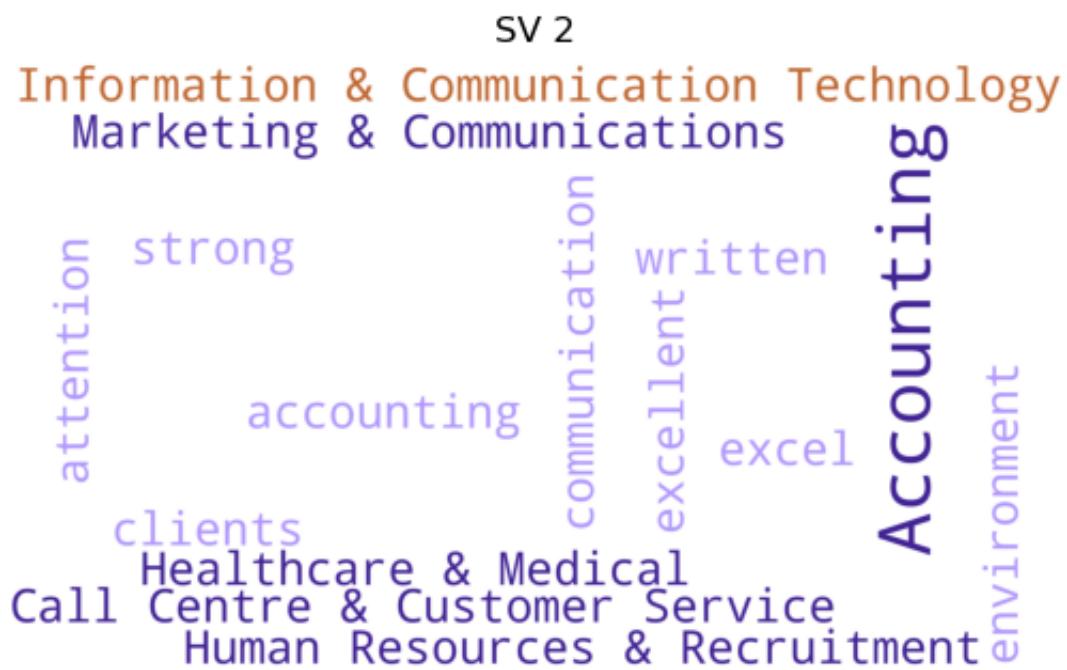
The needed skills for these group of segments are:

- Should be a strong communicator
- Should have technical skills and knowledge of relevant tools
- Must know SQL
- Should be able to deal with clients

The desired hiree for this should both be technically knowledgeable and a strong communicator, especially with clients.



**Figure 19.** SV2 Horizontal Bar Chart



**Figure 20.** SV2 Word Cloud

#### SV2 - Client-Centric Business Services Roles

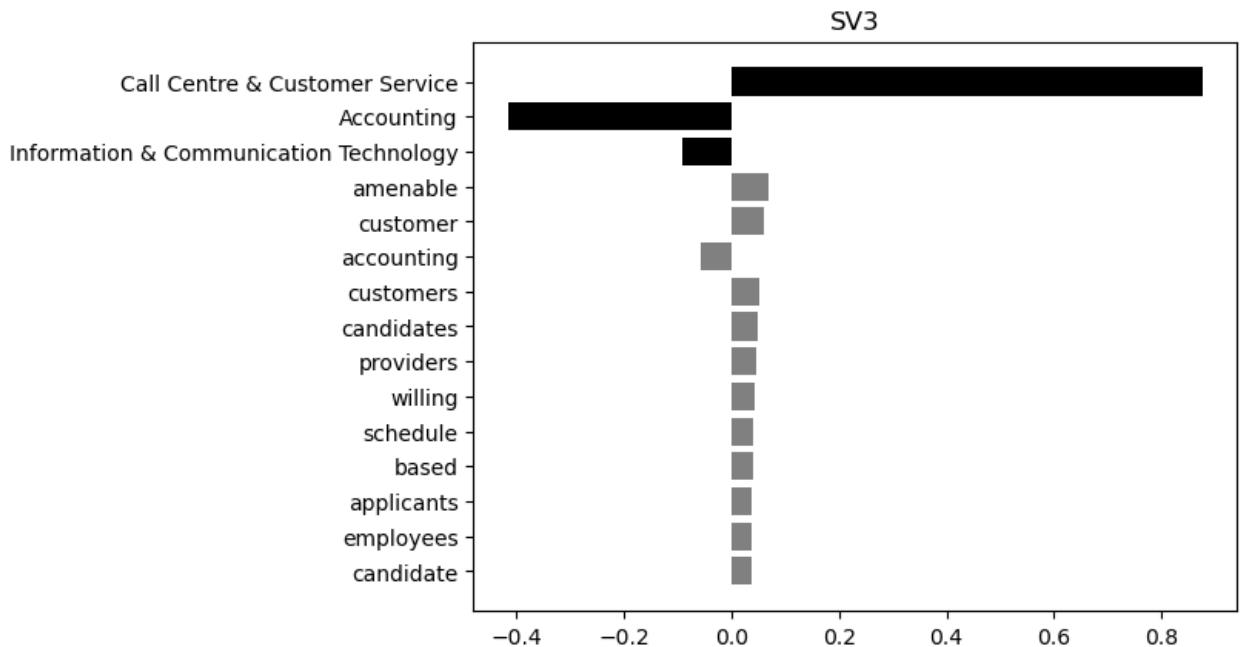
On the left side of the axis, the top segments include Accounting, Call Centre & Customer Service, Human Resources & Recruitment and Healthcare & Medical.

The needed skills for these group of segments are:

- Have Accounting or financial knowledge
- Should be a strong communicator
- Should be able to deal with clients
- Must know Excel

- Have attention to detail
- Must be good in written communication

The desired hiree for this should also be technically knowledgeable and be a strong communicator, on top of having knowledge of a more common tool, Excel.



**Figure 21.** SV3 Horizontal Bar Chart



**Figure 22.** SV3 Word Cloud

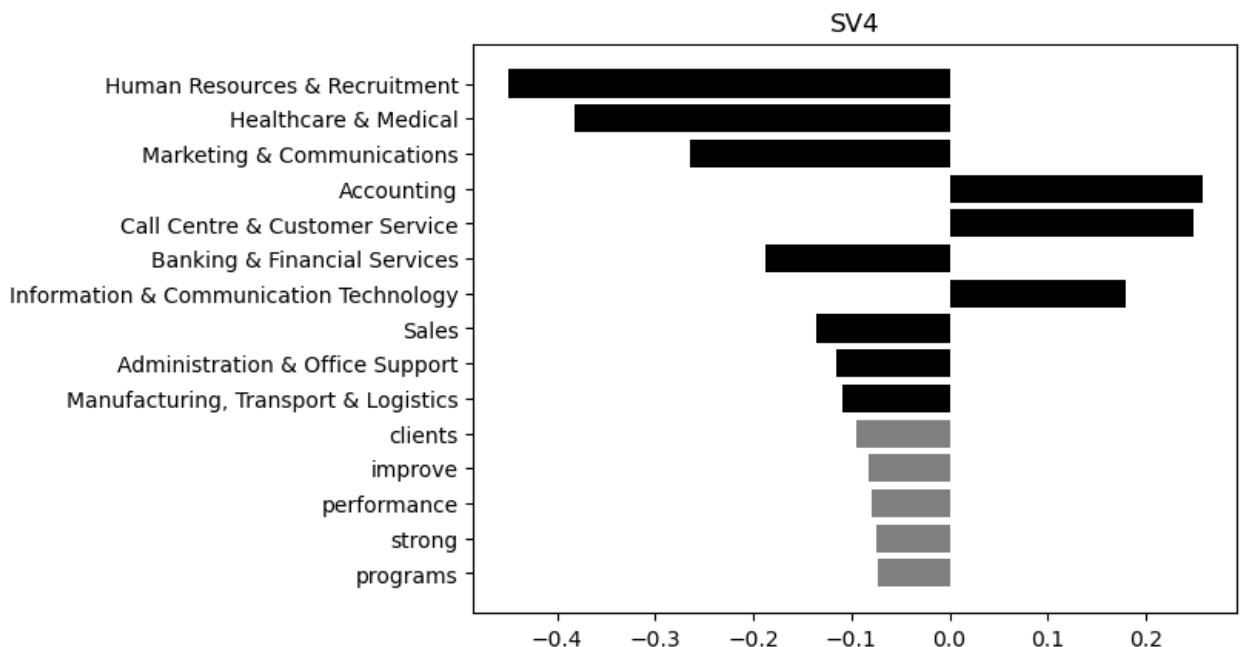
### SV3 - Customer-centric Professions

On the right side of the axis, the top segment includes Call Centre & Customer Service.

The needed skills for this segment are:

- Be amenable or friendly
- Be customer-centric
- Be organized

Hirees in the Call Centre & Customer Service segment deal a lot with customers, and so their needed hire has to have the capacity to deal with people.



**Figure 23.** SV4 Horizontal Bar Chart



**Figure 24.** SV4 Word Cloud

#### SV4 - Client-centric Professions

On the left side of the axis, the top segments include Human Resources & Recruitment, Healthcare & Medical, Marketing & Communications, Banking & Financial Services, Sales, Administration & Office Support, Manufacturing, Transport

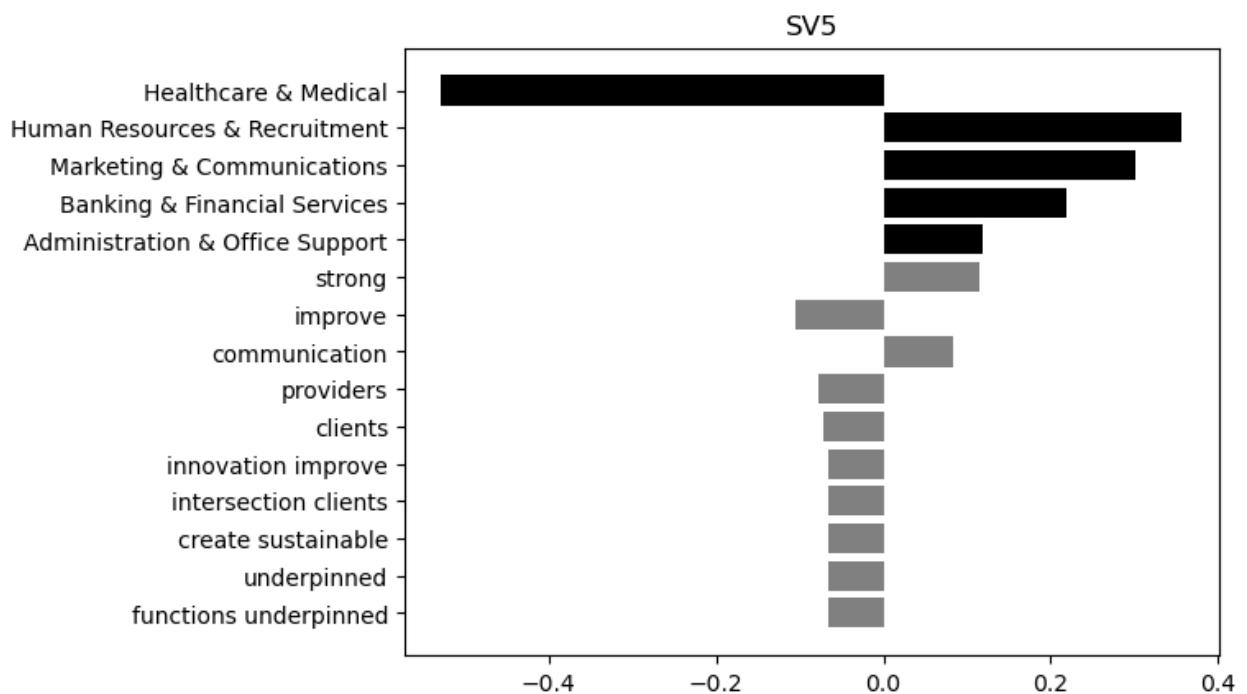
& Logistics.

The skills associated with this group is:

- Must be able to deal with clients
- Must be able to improve performance that satisfies clients

Note that in the opposing axis here, the top segments include: Accounting, Call Centre & Customer Service, and Information & Communication Technology. These segments were previously seen in SV1, which may mean that those in these segments are opposed to that of SV1.

In SV4, the hirees in these segments give a strong emphasis to working with clients, while those segments in SV1 give a balance to being technically-inclined while still communicating with clients.



**Figure 25.** SV5 Horizontal Bar Chart



**Figure 26.** SV5 Word Cloud

#### **SV5 - Empathy-Communication Spectrum**

The left side of the axis includes the segment Healthcare & Medical with the skills:

- Must be willing to improve lives

On the other side of the axis includes the segment Human Resources & Recruitment, Marketing & Communications, Banking & Financial Services, and Administration & Office Support with the skill:

- Must have strong communication skills

The spectrum represents the continuum between the desire to make a positive impact on people's lives and the importance of effectively conveying ideas and messages.

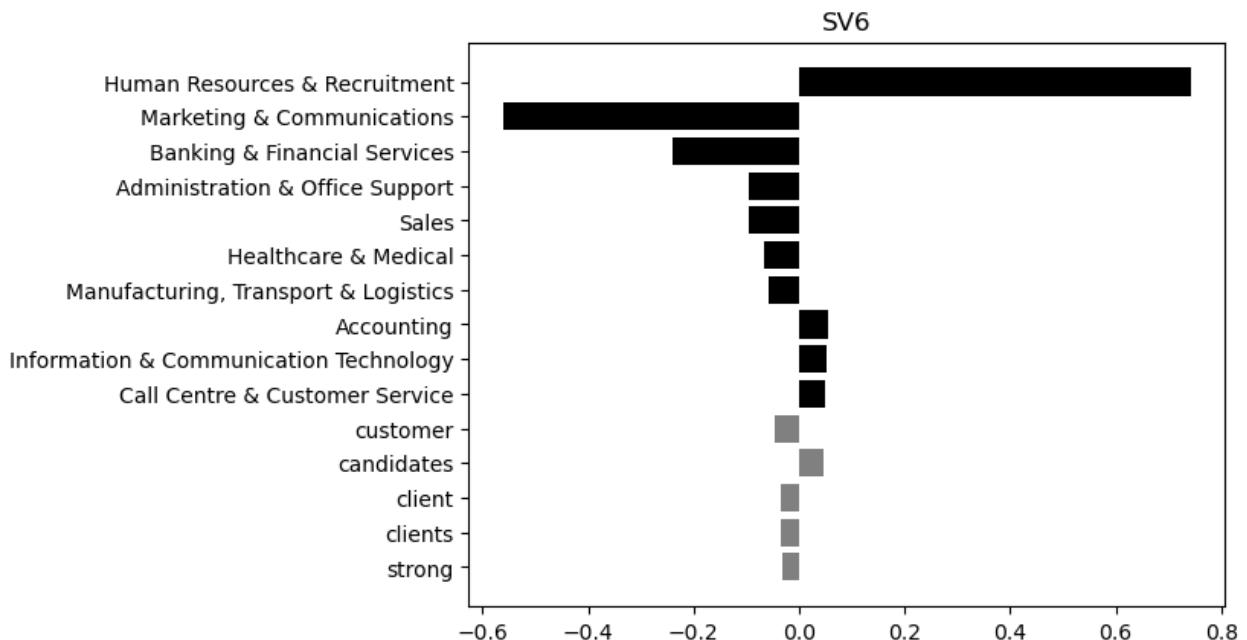


Figure 27. SV6 Horizontal Bar Chart

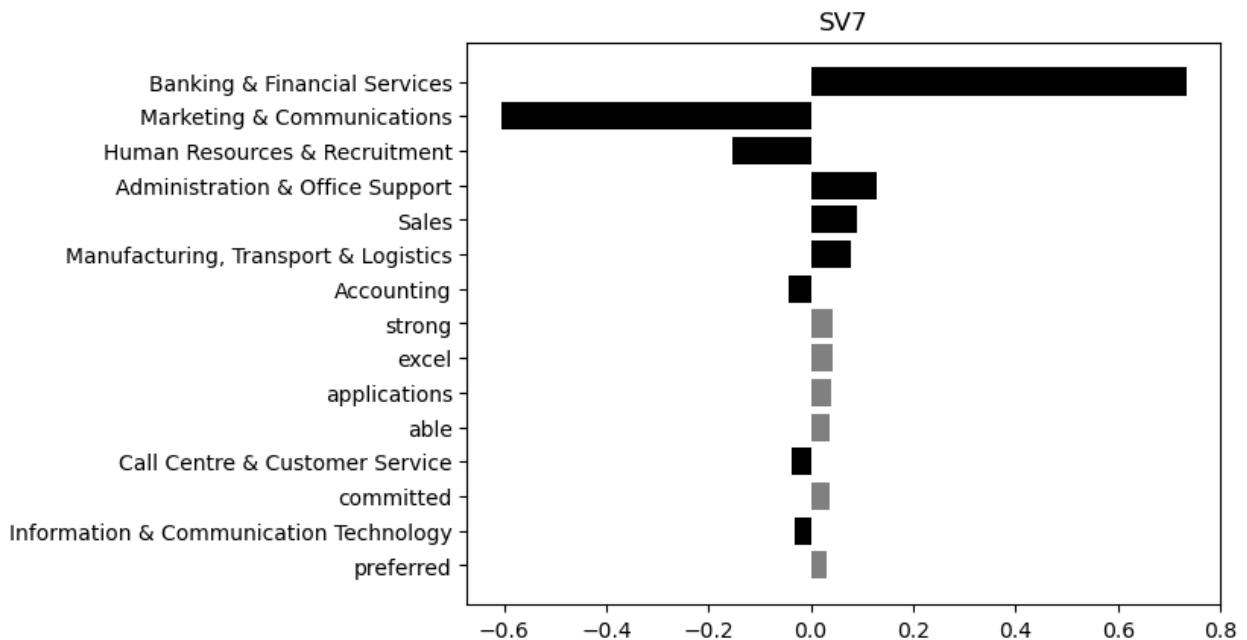


Figure 28. SV6 Word Cloud

### SV6 - Serving Customers vs. Candidates

The left side of the axis includes the following segments: Marketing & Communications, Banking & Financial Services, Administration & Office Support, Sales, Healthcare & Medical, Manufacturing, Transport & Logistics, who will handle mostly customers and clients.

The right side of the axis includes the following segments: Human Resources & Recruitment, Accounting, Information & Communication Technology, Call Centre & Customer Service, who will handle mostly candidates.



**Figure 29.** SV7 Horizontal Bar Chart



**Figure 30.** SV7 Word Cloud

#### **SV7 - Service and Communication to Technical Expertise Spectrum**

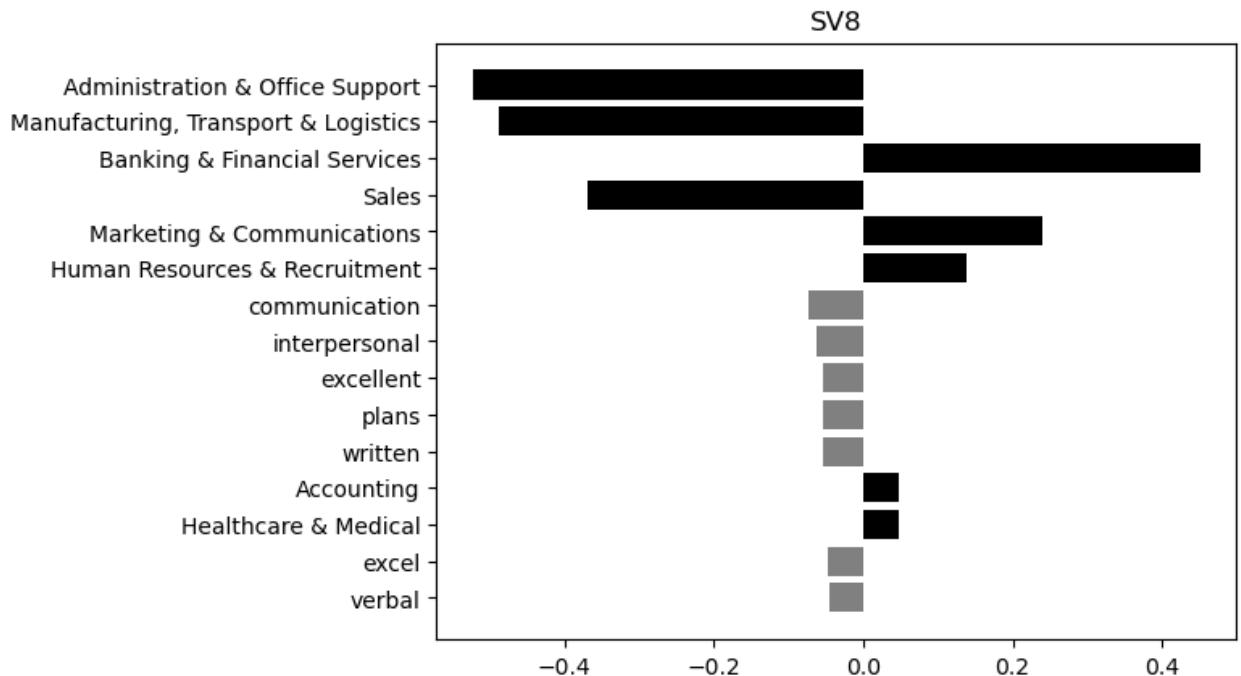
The left side of the axis includes the following segments: Marketing & Communications, Human Resources & Recruitment, Accounting and Call Centre & Customer Service.

The right side of the axis includes the following segments: Banking & Financial Services, Administration & Office Support, Sales, Manufacturing, Transport & Logistics, which needs the skills:

- Must know Excel

- Must know industry-specific applications
- Must have financial knowledge
- Must be committed to improvement

These two sides of the spectrum emphasize the shifting focus from communication and service-oriented roles to those that require technical skills and industry-specific knowledge.



**Figure 31.** SV8 Horizontal Bar Chart



**Figure 32.** SV8 Word Cloud

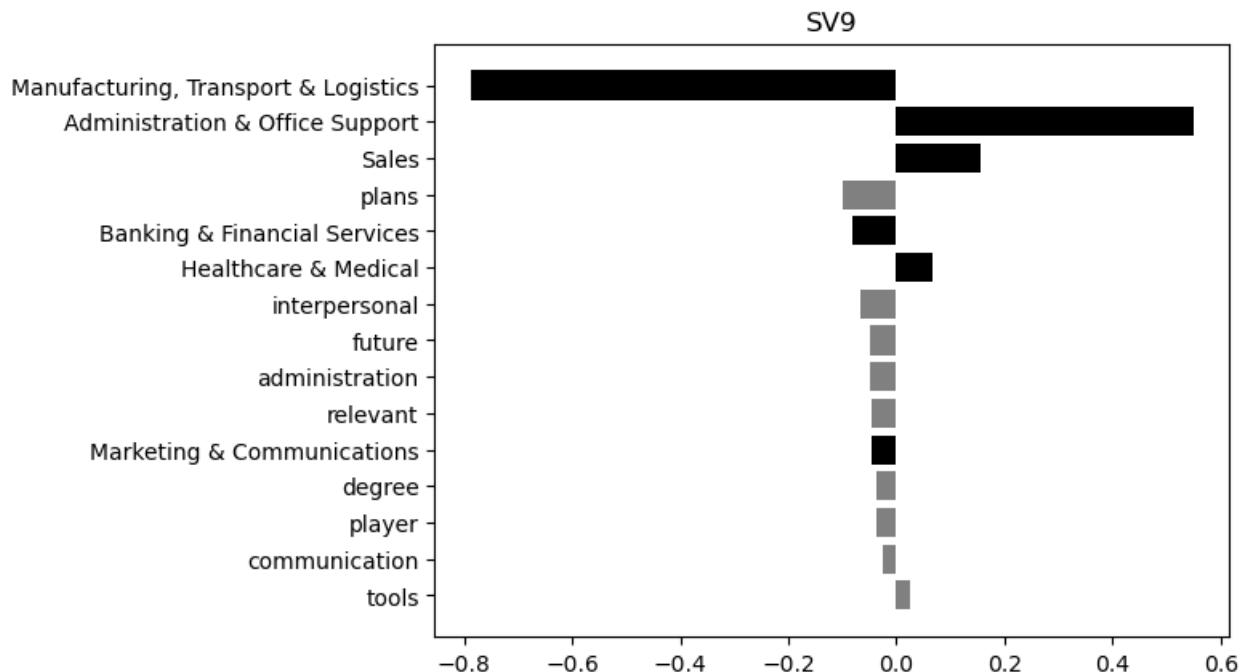
#### SV8 - Technical Operations to Professional Services Continuum

The left side of the axis includes the following segments: Administration & Office Support, Manufacturing, Transport & Logistics, Sales, and Engineering. The skills required here are:

- Have good interpersonal skills
- Have strong written and verbal communication skills
- Must know Excel
- Must be a good planner

The right side of the axis includes the following segments: Banking & Financial Services, Marketing & Communications, Human Resources & Recruitment, and Healthcare & Medical.

The axis highlights the spectrum from roles that emphasize operational and interpersonal skills to those focused on professional services in finance, marketing, HR, and healthcare.



**Figure 33.** SV9 Horizontal Bar Chart

**Figure 34.** SV9 Word Cloud**SV9 - Strategic Leadership to Technical Expertise Continuum**

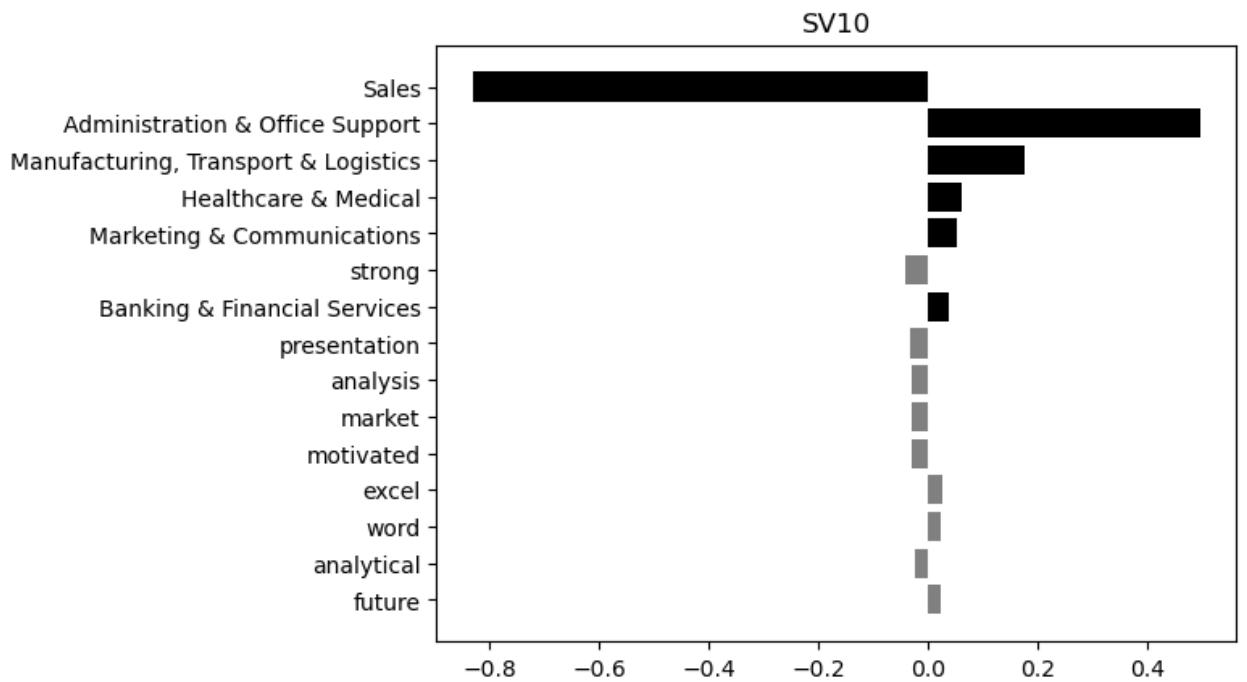
The left side of the axis includes the following segments: Manufacturing, Transport & Logistics, Banking & Financial Services, and Marketing & Communications. The skills required here are:

- Must be a good planner
- Must have strong interpersonal skills
- Must be future-thinking
- Must be a team player
- Must have strong communication skills

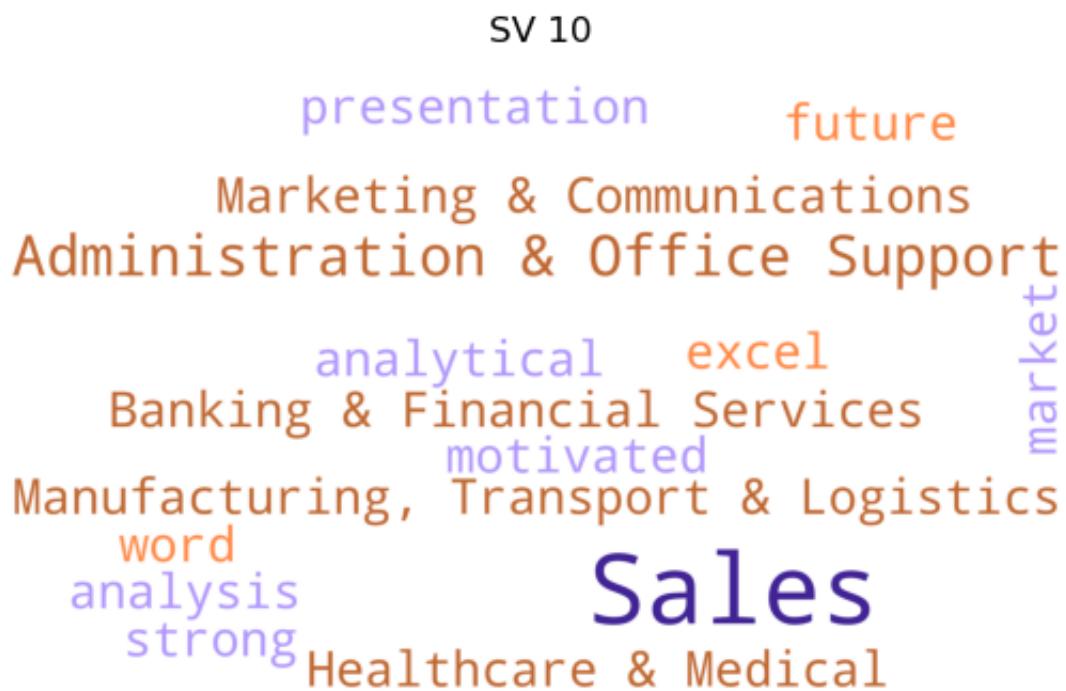
The right side of the axis includes the following segments: Administration & Office Support, Sales, and Healthcare & Medical, with the required skills:

- Must have knowledge with industry-specific tools

This spectrum highlight the range from roles that prioritize strategic planning and interpersonal skills to those that emphasize technical proficiency with industry-specific tools.



**Figure 35.** SV10 Horizontal Bar Chart



**Figure 36.** SV10 Word Cloud

#### SV10 - Analytical vs Administrative Professions

The left side of the axis includes the segment Sales, with the required skills:

- Be motivated
- Have strong analytical skills
- Must have presentation skills

The right side of the axis includes the segments: Administration & Office Support, Manufacturing, Transport & Logistics, Healthcare & Medical, Marketing &

Communications, Banking & Financial Services, with the required skills:

- Must know Excel and Word
- Must be future-thinking

This continuum aims to capture the range from roles that require motivation and analytical skills to those that prioritize administrative proficiency and future-thinking.

[Back to TOC](#)

## VII. Conclusion

# Summary and Insights

Shown here is a summary table of the results and discussion:

**Table 11.**

*Summary of Job-Skill Segments*

SV	Job Segment	Required Skills	Summary
SV1	Technical Client Support Roles	<ul style="list-style-type: none"><li>- Strong communicator</li><li>- Technical skills and knowledge of relevant tools</li><li>- SQL proficiency</li><li>- Ability to deal with clients</li></ul>	Hirees need to be technically knowledgeable and strong communicators, especially with clients.
SV2	Client-Centric Business Services Roles	<ul style="list-style-type: none"><li>- Accounting or financial knowledge</li><li>- Strong communicator</li><li>- Ability to deal with clients</li><li>- Excel proficiency</li><li>- Attention to detail</li><li>- Good in written communication</li></ul>	Desired hirees should be technically knowledgeable, strong communicators, and proficient in Excel.
SV3	Customer-Centric Professions	<ul style="list-style-type: none"><li>- Amenable and friendly</li><li>- Customer-centric</li><li>- Organized</li></ul>	The focus is on dealing with customers, requiring the ability to engage with people effectively.
SV4	Client-Centric Professions	<ul style="list-style-type: none"><li>- Ability to deal with clients</li><li>- Improve performance that satisfies clients</li><li>- Willingness to improve lives</li></ul>	Emphasis on working with clients, providing a strong focus on client interaction.

<b>SV5</b>	Empathy-Communication Spectrum	(Healthcare & Medical) - Strong communication skills (HR, Marketing, Banking, Admin)	Represents the continuum between making a positive impact on lives and effective communication.
<b>SV6</b>	Serving Customers vs. Candidates	- Customer-centric roles (Marketing, Sales, Healthcare, etc.) - Candidate-centric roles (HR, Accounting, IT, Call Centre)	Highlights the distinction between roles focusing on customers/clients and those on candidates.
<b>SV7</b>	Service and Communication to Technical Expertise Spectrum	- Communication-focused roles (Marketing, HR, Accounting, etc.) - Technical roles (Banking, Manufacturing, Logistics, etc.)	Emphasizes the shift from communication/service-oriented roles to those requiring technical expertise.
<b>SV8</b>	Technical Operations to Professional Services Continuum	- Interpersonal skills - Strong communication skills - Excel proficiency - Good planner (Admin, Manufacturing, Logistics, etc.)	Represents the spectrum from roles emphasizing operational skills to those in professional services.
<b>SV9</b>	Strategic Leadership to Technical Expertise Continuum	- Good planner - Strong interpersonal skills - Future-thinking - Team player - Strong communication skills (Manufacturing, Banking, Marketing, etc.)	Captures the range from roles prioritizing strategic planning to those emphasizing technical expertise.
<b>SV10</b>	Analytical vs Administrative Professions	- Motivation - Analytical skills - Presentation skills (Sales) - Excel and Word proficiency - Future-thinking (Admin, Manufacturing, Healthcare, etc.)	Represents the continuum from roles requiring motivation and analytics to those emphasizing administrative proficiency.

These are some insights gleaned from the results of the study:

- The primary job segments in data-related positions emerge from the Accounting, Call Centre & Customer Service, and Information & Communication Technology

domains.

- Working with clients and communication skills emerged as the persistent and prevailing skill wanted across many job segments.
- While career shifters and students focus on hard skills, it was shown here that employers value core competency and soft skills that help them solve their problems, rather than the specific technical skills.

To fully conclude the study, at the beginning, the team was expecting mostly technical skills to be the source of variance for needed job skills per segment. However, as shown here, most job-skill segments vary also by soft skills and core competencies such as strong communication skills and having an analytical mind. This further necessitates how future data professionals should not only focus on their technical know-how, but make sure that they also grow in terms of their soft skills and core competencies. This should also guide companies to know which soft skills, aside from hard skills, should be expected of a data professional entering a specific department of their company.

[Back to TOC](#)

## VIII. Recommendation

For the next set of studies, we recommend:

### 1. Focus on hard skills

In this study, we found that soft skills and core competency are important skills to have for anyone wanting to go in this industry. Hard skills or the specific tools were not mentioned as much in the results, although the words "technical", "Excel" and "SQL" did come on the latter part of the results. Many will be curious to know what specific tools or fields of study in "data" will trend and be in-demand in the years to come.

### 2. Focus on specific job titles

Data is all around us, and so are data-related jobs. The job titles here consisted of analysts, consultants, and data encoders and even accountants and customer service representatives who work with data. To make a more focused study serving a specific group in the industry, we recommend focusing on specific job titles of interest. Furthermore, not only the job titles can be filtered, but also varying features about the job (eg. salary, location).

### 3. Use another data source aside from job posts

Job posts are free form text, which needs much data cleaning in order to generate specific insights of interest. In particular, the job posts, while enumerates skills and tools, also enumerate needed characteristics of the applicant (eg. being detail-oriented, amicable). To find the specific hard skills or toolset needed, it's possible to use a different set of data that would be more descriptive and niched to the data-related jobs of interest.

[Back to TOC](#)

## Acknowledgment

We would like to extend our heartfelt gratitude to everyone who supported us throughout this project. Special thanks go to our mentor, Prof. Christian Alis for his invaluable guidance and encouragement. We are also grateful to our colleagues and peers for their insights and contributions, which greatly enriched our work. Additionally, we appreciate the understanding and support from our families and friends during this endeavor. This project would not have been possible without the collective effort and encouragement from all who were involved.

[Back to TOC](#)

## References

### References

1. Davenport, T., Patil, D. (2022, July 15). Is Data Scientist Still the Sexiest Job of the 21st Century?. Harvard Business Review. [https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century?fbclid=IwAR08iUssthOE\\_mF4VkJYZ2njPfkY6Aj6fCPYtwYp9\\_kLYsdVhMqi4mK\\_LQ](https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century?fbclid=IwAR08iUssthOE_mF4VkJYZ2njPfkY6Aj6fCPYtwYp9_kLYsdVhMqi4mK_LQ)
2. Michael Page. (2023). The Invisible Revolution. Retrieved from <https://www.michaelpage.com.ph/talent-trends/the-invisible-revolution>
3. Michael Page. (2023). 5 most in-demand tech roles for 2023. Retrieved from <https://www.michaelpage.com.ph/advice/market-insights/market-updates/5-most-demand-tech-roles-2023>
4. UCLA Institute for Digital Research and Education. (n.d.). Principal Components (PCA) and Exploratory Factor Analysis (EFA) with SPSS. Retrieved from <https://stats.oarc.ucla.edu/spss/seminars/efa-spss/>
5. Yencken, L. (2011, December 6). stopwords.txt. Retrieved from <https://gist.github.com/larsyencken/1440509>

[Back to TOC](#)