oxylabs

Report on

# APPENDIX 2

## Preparing the Data Set

Prepared by

Laima Lukoseviciute

February 11, 2026

# Table of contents

## Introduction

This report presents a data preparation for the EDA step for Coresignal jobs data. The primary goal of the further analyses will be to identify trends, relationships, and interesting angles within the Tech Stacks landscape of 2025.

The initial raw data was gained from the Coresignal Multi-Source Jobs Dataset, which aggregates listings from major global job boards. Since the Job listings for 2025 have more than 60 mln job postings I have decided to extract the job titles in the US, only where the job title contains selected keywords (plural of the words was also accepted):

`developer` OR `analyst` OR `programmer` OR `programming` OR `scientists` OR `data` OR `researcher` OR `engineer` OR `engineering`.

The data was extracted using this SQL query that can be seen below.

```
SELECT
  title, description, company_name, company_industry, state, created_at,
FROM
  `oxy-analytics.raw_external_cosi_core.multisource_job`
WHERE
  created_at > "2024-12-31" AND
  country = "United States" AND
  REGEXP_CONTAINS(title, r"(?i)\b(developers?|analysts?|programmers?|programming|
  scientists?|data|researchers?|engineers?|engineering)\b")
```

The analyzed dataset has 6 primary features and over 2 062 382 observations, ranging from January 1, 2025 to December 31, 2025. For a detailed breakdown of the features, please refer to the Table 1.

Table 1: The description of variables for data.

| Variable Name | Type | Description |
|---|---|---|
| title | STRING | The professional title of the job listing. |
| description | STRING | The full text of the job post, used for keyword extraction. |
| company_name | STRING | The name of the hiring organization. |
| company_industry | STRING | The sector the company operates in (e.g., Tech, Finance). |
| state | STRING | The US state of the job location. |
| created_at | TIMESTAMP | The date when the job listing was added to the database. |

A preview of the analysed dataset is presented below in Table 2.

Table 2: Raw data pre-view first 5 rows. Note: if the table is not visible in pdf, please see html version of the report.

| | title | description | company_name | company_industry | state | created_at |
|---|---|---|---|---|---|---|
| 0 | Scientist (non-P... | Why Patients Nee... | NaN | NaN | California | 2025-09-23 18:17... |
| 1 | Sr. Scientist, C... | About Loyal Loya... | Loyal | Biotechnology Re... | California | 2025-09-19 19:39... |
| 2 | Transit Coordina... | Posted: Oct 1, 2... | National Grants ... | Non-profit Organ... | Texas | 2025-09-29 09:04... |
| 3 | Senior Manager, ... | Description As t... | Amazon | Software Develop... | Tennessee | 2025-09-10 20:07... |
| 4 | Business Intelli... | Job Title: Busin... | IntelliSavvy | IT Services and ... | Washington | 2025-09-22 12:07... |

## Data Preparation

The extracted data totals approximately 10 GB. To ensure the system processes this volume of information efficiently, I have partitioned the data into 11 separate files. After the programming language data is extracted, the "description" column will be removed. This adjustment allows the information from all files to be combined into a single dataset for the analysis and saved to the file `data/processed/jobs_proc_2025_no_desc.csv`.

The function `extract_tech_tools` extracts these 23 tools: "Excel", "Google_Sheets", "Fivetran", "Airbyte", "dbt", "Snowflake", "BigQuery", "Airflow", "Prefect", "Power BI", "Tableau", "Looker", "Git", "Docker", "Kubernetes", "Terraform", "AWS", "Azure", "GCP", "Databricks", "Kafka", "Spark", "Monte Carlo". The detection of all the tools are han-

daled with regular expressions. To process the tools like “Excel”, “Airflow”, and “Prefect” I will be using LLM to interpret the meaning of these words in the description.

Below you can see the list of all the tools which presence was evaluated.

```
['Excel',
 'Google_Sheets',
 'Fivetran',
 'Airbyte',
 'dbt',
 'Snowflake',
 'BigQuery',
 'Airflow',
 'Prefect',
 'Power_BI',
 'Tableau',
 'Looker',
 'Git',
 'Docker',
 'Kubernetes',
 'Terraform',
 'AWS',
 'Azure',
 'GCP',
 'Databricks',
 'Kafka',
 'Spark',
 'Monte_Carlo']
```

After verification process we have 3 additional columns in the data set: “Excel_verified”, “Airflow_verified”, and “Prefect_verified”. Below Table 3 you can see the preview of data with these new columns.

```
(2062297, 31)
```

Table 3: Verified data pre-view first 5 rows. Note: if the table is not visible in pdf, please see html version of the report.

| | title | Excel | Excel_verified | Airflow_verified | Prefect_verified |
|---|---|---|---|---|---|
| 0 | Scientist (non-P... | 0 | 0 | 0 | 0 |
| 1 | Sr. Scientist, C... | 0 | 0 | 0 | 0 |
| 2 | Transit Coordina... | 0 | 0 | 0 | 0 |
| 3 | Senior Manager, ... | 1 | 1 | 0 | 0 |
| 4 | Business Intelli... | 0 | 0 | 0 | 0 |

NOTE:

The `run_full_verification` function utilizes the Ollama Large Language Model (LLM) to identify the meaning of the words “Excel”, “Airflow”, and “Prefect”. To ensure the most consistency of these results and minimize variability in model output, the temperature parameter is set to lower value.

Below I will evaluate the error rate for the tech tool detection precision. Which of them mean the tech tool and which of them do not.

Below at Table 4 you can see the preciton evaluation for tools “Excel”, “Airflow”, and “Prefect”.

Table 4: Precision evaluation for each tech tool

| | tool | precision |
|---|---|---|
| 0 | Excel | 0.87 |
| 1 | Airflow | 0.94 |
| 2 | Prefect | 0.64 |

## Observations and Features

To make sure the function did what it was suppose to do let's do the exploration of the dataset's structure. I will examine the characteristics of each column to ensure data integrity and understand the available information.

Below you can see the list of all the columns in the processed dataframe. All the programming languages were included.

```
Index(['title', 'company_name', 'company_industry', 'state', 'created_at',
       'Excel', 'Google_Sheets', 'Fivetran', 'Airbyte', 'dbt', 'Snowflake',
       'BigQuery', 'Airflow', 'Prefect', 'Power_BI', 'Tableau', 'Looker',
       'Git', 'Docker', 'Kubernetes', 'Terraform', 'AWS', 'Azure', 'GCP',
       'Databricks', 'Kafka', 'Spark', 'Monte_Carlo', 'Excel_verified',
       'Airflow_verified', 'Prefect_verified'],
      dtype='object')
```

Below Table 5 you can see the description of categorical data. We can see that there are 598373 unique title in the data and 107161 unique companies. Top industry is Software Development, and top state is California. At Table 6 you can see that the data covers 2025.

Table 5: Description of the categorical data

| | title | company_name | company_industry | state |
|---|---|---|---|---|
| count | 2062297 | 2039563 | 1782048 | 1459598 |
| unique | 598373 | 107161 | 427 | 138 |
| top | Financial Analyst | Jobs via Dice | Software Develop... | California |
| freq | 6863 | 68134 | 240228 | 215729 |

Table 6: Description of the date data

| | created_at |
|---|---|
| count | 2062297 |
| mean | 2025-06-22 23:52... |
| min | 2025-01-01 00:14... |
| 25% | 2025-03-25 02:28... |
| 50% | 2025-06-25 07:36... |
| 75% | 2025-09-12 19:41... |
| max | 2025-12-19 09:18... |

## Duplicate and Missing Values

In this section I will analyse if the data set has any duplicated observations or missing values. From the outputs below (Table 7) we can see that data have some missing values in column `state` 29% of values are missing, 14% in column `company_industry`, and 1% in `company_name`. The full breakdown can be seen below in Table 7.

Table 7: Missing values in the data set by column.

| | column_name | no_values_missing | percentage_values_missing |
|---|---|---|---|
| 3 | state | 602699 | 29.22 |
| 2 | company_industry | 280249 | 13.59 |
| 1 | company_name | 22734 | 1.10 |
| 0 | title | 0 | 0.00 |
| 23 | GCP | 0 | 0.00 |
| 19 | Kubernetes | 0 | 0.00 |
| 20 | Terraform | 0 | 0.00 |
| 21 | AWS | 0 | 0.00 |
| 22 | Azure | 0 | 0.00 |
| 25 | Kafka | 0 | 0.00 |
| 24 | Databricks | 0 | 0.00 |
| 17 | Git | 0 | 0.00 |

Table 7: Missing values in the data set by column.

| | column_name | no_values_missing | percentage_values_missing |
|---|---|---|---|
| 26 | Spark | 0 | 0.00 |
| 27 | Monte_Carlo | 0 | 0.00 |
| 28 | Excel_verified | 0 | 0.00 |
| 29 | Airflow_verified | 0 | 0.00 |
| 18 | Docker | 0 | 0.00 |
| 15 | Tableau | 0 | 0.00 |
| 16 | Looker | 0 | 0.00 |
| 14 | Power_BI | 0 | 0.00 |
| 13 | Prefect | 0 | 0.00 |
| 12 | Airflow | 0 | 0.00 |
| 11 | BigQuery | 0 | 0.00 |
| 10 | Snowflake | 0 | 0.00 |
| 9 | dbt | 0 | 0.00 |
| 8 | Airbyte | 0 | 0.00 |
| 7 | Fivetran | 0 | 0.00 |
| 6 | Google_Sheets | 0 | 0.00 |
| 5 | Excel | 0 | 0.00 |
| 4 | created_at | 0 | 0.00 |
| 30 | Prefect_verified | 0 | 0.00 |

720756 were duplicated values. I will keep missing values, and will do the analyses with them in mind, and I will remove the duplicated values, since it is the same job posting.

```
Number of duplicated values:  720756
```

## Outliers

In this section let's look for some obvious outliers or other descrepencies in the data. The job `title` contains some obvious outliers, like resercher that is not related to data, but rather the academic enviroment, and egnineering manager mignt not need any knowlage of the programming languages, thus I will remove all the rows that have no mentions of any programming languages. I will also filter out the job

```
Number of rows and columns after filtering jobs (rows) that have no mentions of tech tools:
(596962, 31)
```

Below you can see the list of all USA states that are in the data and needs to be cleaned and unified.

```
array(['Tennessee', 'Washington', 'California', 'District of Columbia',
       'Massachusetts', 'Indiana', 'Florida', nan, 'Iowa', 'Texas',
       'New Jersey', 'Illinois', 'Utah', 'New York', 'Maryland',
       'Arizona', 'North Carolina', 'Virginia', 'Georgia', 'Ohio',
       'Rhode Island', 'Connecticut', 'Oklahoma', 'Kansas', 'Mississippi',
       'New Hampshire', 'North Dakota', 'Maine', 'Alabama', 'Wisconsin',
       'Colorado', 'WI', 'Arkansas', 'Pennsylvania', 'South Carolina',
       'Nebraska', 'Minnesota', 'Nevada', 'Missouri', 'Michigan',
       'Hawaii', 'Kentucky', 'New Mexico', 'Wyoming', 'Oregon',
       'United States', 'Delaware', 'Vermont', 'MN', 'Idaho', 'MA',
       'Montana', 'South Dakota', 'Puerto Rico', 'Louisiana', 'MD',
       'West Virginia', 'Alaska', 'TX', 'SC', 'Metropolitan Area', 'DC',
       'GA', '   ', 'San Juan', 'ND', 'US Virgin Islands',
       'Gurabo Municipio', 'Carolina', 'Dededo Municipality',
       'Eastern District', 'Barrigada Municipality', 'Sarasota Area',
       'WA', 'Yigo Municipality', 'County', 'Guam', 'North', 'Young',
       'Grants Pass Area', 'American Samoa', 'Virgin Islands',
       'Guayanilla Municipio', 'Provincia de Las Palmas',
       'Cayey Municipio', 'DE', 'indiana', 'Guaynabo', 'Chicago',
       'U.S. Virgin Islands', '   ', 'Washington D.C.', '    ',
       'Floride', '      ', '   ', '    ',
       'Doddridge County', 'Nowy Jork', 'Nueva York', 'Cidra Municipio',
       'Municipality', 'FL', '         ', '   ', 'states',
       '  ', 'Menomonie Area', 'Northern Mariana Islands'], dtype=object)
```

These are states and US territories, which have left after the cleaning.

```
Number of unique states after cleaning:  55
```

```
array(['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California',
       'Colorado', 'Connecticut', 'Delaware', 'District of Columbia',
       'Florida', 'Georgia', 'Guam', 'Hawaii', 'Idaho', 'Illinois',
       'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Louisiana', 'Maine',
       'Maryland', 'Massachusetts', 'Michigan', 'Minnesota',
       'Mississippi', 'Missouri', 'Montana', 'Nebraska', 'Nevada',
       'New Hampshire', 'New Jersey', 'New Mexico', 'New York',
       'North Carolina', 'North Dakota', 'Ohio', 'Oklahoma', 'Oregon',
       'Pennsylvania', 'Puerto Rico', 'Rhode Island', 'South Carolina',
       'South Dakota', 'Tennessee', 'Texas', 'Utah', 'Vermont',
       'Virgin Islands', 'Virginia', 'Washington', 'West Virginia',
       'Wisconsin', 'Wyoming'], dtype=object)
```

## Feature Engineering

To improve the clarity of the recruitment landscape, I am standardizing the industry classifications within the dataset. Currently, the data includes 427 unique industries, many of which overlap or share similar characteristics. This level of granularity can make it difficult to identify broader market trends.

I have grouped these industries into 11 primary categories to make the analysis more accessible and highlight high-level patterns. The distribution of job postings across these groups is as follows:

| broad_industry_group | count |
|---|---|
| Tech, Data & Telecom | 226571 |
| Professional, Legal & Business Services | 87252 |
| Miscellaneous | 76292 |
| Finance, Insurance & Real Estate | 61138 |
| Manufacturing, Industrial & Defense | 55533 |
| Healthcare, Pharma & Wellness | 23832 |
| Logistics, Travel & Construction | 19385 |
| Education, Government & Non-profit | 15065 |
| Consumer, Retail & Agriculture | 12883 |
| Energy, Utilities & Environment | 12372 |
| Media, Entertainment & Arts | 6639 |

To provide a more structured view of the recruitment landscape, I am standardizing the job titles within the dataset. Currently, the data contains approximately 600 000 unique job titles, which is a level of detail that can obscure broader market trends.

By grouping these titles into five primary categories, the analysis becomes more accessible for identifying high-level patterns. These categories include:

- Manager: Roles focused on leadership and project oversight.
- Engineer: Positions centered on building and maintaining technical systems.
- Analyst: Roles dedicated to interpreting data and providing insights.
- Scientist: Research-oriented positions, including Data Scientists and Researchers.
- Developer: Traditional software creation and programming roles.

This categorization simplifies the comparison of programming language requirements across different professional functions. The final dataframe can be found in file `data/processed/tools/jobs_filtered_2025_no_desc.csv`. The Table 9 shows the preview to the final data set.

```
(596962, 37)
```

Table 9: Final data set pre-view first 5 rows and 7 columns. Note: if the table is not visible in pdf, please see html version of the report.

| | title | broad_industry_group | analyst | engineer | Excel | Airflow | Excel_verified |
|---|---|---|---|---|---|---|---|
| 3 | Senior Manager, ... | Tech, Data & Tel... | 0 | 1 | 1 | 0 | 1 |
| 4 | Business Intelli... | Tech, Data & Tel... | 0 | 1 | 0 | 0 | 0 |
| 5 | HAZARDOUS SUBSTA... | Energy, Utilitie... | 0 | 1 | 1 | 0 | 1 |
| 8 | PMS 378 Senior T... | Tech, Data & Tel... | 1 | 0 | 1 | 0 | 1 |
| 11 | COMPENSATION ANA... | Miscellaneous | 1 | 0 | 1 | 0 | 1 |

## Summary

This stage of the project focused on transforming around 10 GB of raw job posting data from the Coresignal Multi-Source Jobs Dataset into a structured, analysis-ready format. To ensure the data remained manageable on local hardware while maintaining depth, the following steps were completed:

- Targeted Extraction: I have narrowed the scope to over 2 000 000 observations from 2025 specifically within the United States, filtering for key technical roles such as developers, analysts, and engineers.
- Data Integrity: The dataset was refined by removing more than 720 000 duplicate entries and filtering out job titles that lacked any tech tools mentions (e.g., academic researchers or pure management roles).
- Technical Standardization: I have identified mentions of 23 tech tools. Ambiguous terms like “Excel” were processed using a local Large Language Model (Ollama).
- Categorization: To understand market trends, I have engineered five high-level job categories: Manager, Engineer, Analyst, Scientist, and Developer.

The resulting processed dataset contains 602 359 high-quality job postings, significantly reduced from the initial data, allowing for efficient and high-impact EDA.

## Suggestions for Further Improvements

The error rate calculations for the LLM used for the interpretation tech tools could be done. I think this will be marginal improvement and will not change the results that much, but just for the sake of being precice and redusing the error this should be done.

There could be some other jobs that require these tools and are in the tech position but the Coresignal data fields were lacking this information in the job title or description, because the fields were not present in the data of mixed.

There are also other tools that could be added specifically talking about the developers, but the additional analyses, should be conducted.