

LABOR MARKET RESEARCHER

TECHNICAL TEST



CARLOS GOMEZ HERNANDEZ  
https://www.carlosgomezhernandez.com/

Table of Contents

[DATA DICTIONARY: 2](#_Toc56272957)

[ASSUMPTIONS: 2](#_Toc56272958)

[QUESTIONS TO THE DATA SCIENCE TEAM: 3](#_Toc56272958)

[MOST RELEVANT DATA 4](#_Toc56272959)

[SKILLS INVENTORY NEEDED IN 2020 4](#_Toc56272959)

[SKILLS INVENTORY OFFERED IN 2020 5](#_Toc56272960)

[WRAPPING THE DATA - INDUSTRY FOCUSED: 6](#_Toc56272961)

# DATA DICTIONARY:

|  |  |  |  |
| --- | --- | --- | --- |
| **Vector Location** | **Vector** | **Data type** | **Content** |
| 0 | skill\_title | object | Skill required or offered |
| 1 | DocumentId | int64 | ID of the event document: |
| 2 | Type | object | Definition of the request: resume = offer job = demand |
| 3 | OriginalTitle | object | - |
| 4 | CreationDate | datetime64[ns] | Date of document creation. |
| 5 | OriginalCompanyName | object | Company Name |
| 6 | RangeFrom | float64 | - |
| 7 | RangeTo | float64 | - |
| 8 | ClusteredTitle | object | Cluster aggrupation for job titles. |
| 9 | OriginalLocation | object | - |
| 10 | OriginalSalary | object | - |
| 11 | CountryCode | object | location0 |
| 12 | City | object | location2 |
| 13 | SalaryParsed | object | - |
| 14 | State | object | location1 |
| 15 | Complexity | float64 | - |
| 16 | importance | float64 | - |

**ROWS = 231713  
VECTORS = 16**

ASSUMPTIONS:  
  
The following series is an event-based tabular set. Each row represents a skill requested or offered for each document (job or resume) submitted to the web application.  
  
Because the same job or resume can have multiple skills offered or requested, **this data cannot give insights as to the job counts themselves**.  
  
However, this data can be aggregated by skills and filtered to find insights of the time, location, type of document, salary, complexity and importance.  
  
The vectors : skill title, and ClusteredTitle seem to have very high-dimensional space, which will cause the plots to be meaningless unless multiple filters are applied.   
  
QUESTIONS TO THE DATA SCIENCE DEPARTMENT:  
  
I would ask them to clarify the column “CONTENT” from the data dictionary and confirm that the assumptions from said column are correct.  
  
MOST RELEVANT DATA:The following table contains the relevant columns that could provide relevant information to the client.  
  
- Skill title is the main vector to been analysed.

- Type is necessary to split the dataset between offer and demand.  
- City, State, Country Code, provide location information.  
- Creation Date allows to build trend plots to track specific skills, how the demand for them has change.- Complexity and Importance flags provide information about the skill complexity and importance.

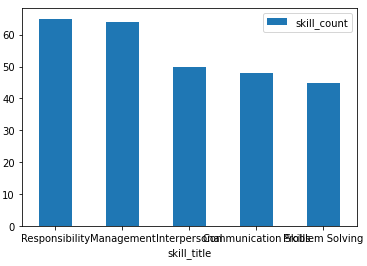
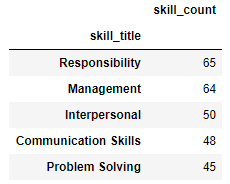
|  |  |
| --- | --- |
| **Vector Location** | **Vector** |
| 0 | skill\_title |
| 2 | Type |
| 4 | CreationDate |
| 5 | OriginalCompanyName |
| 8 | ClusteredTitle |
| 10 | OriginalSalary |
| 11 | CountryCode |
| 12 | City |
| 14 | State |
| 15 | Complexity |
| 16 | importance |

# 

# SKILLS INVENTORY NEEDED IN 2020

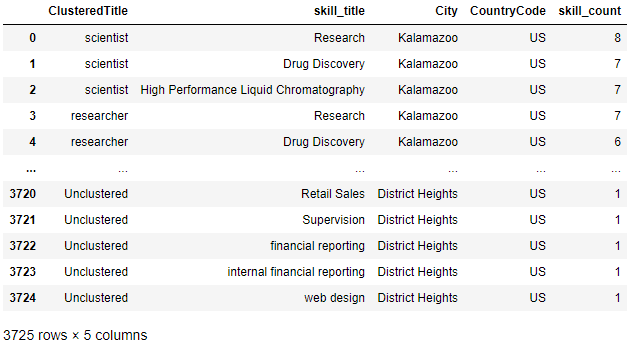
The skill\_inventory.xlsx report contains demanded aggregations of job titles, skills, and locations for the year 2020 and the total count of each slice. Note that the following table shows that “regulatory affairs” - located in the first column - contains multiple records (many to many relationship).



  
With this inventory we can find meaningful insights into business questions such as:  
  
WHAT ARE THE 5 MOST DEMANDED SKILLS IN TORONTO IN THE YEAR 2020:

# SKILLS INVENTORY OFFERED IN 2020

The skill\_inventory.xlsx report also contains offered aggregations of job titles, skills, and locations for the year 2020 and the total count of each slice.



# WRAPPING THE DATA - INDUSTRY FOCUSED:

In order to be able to wrap this data as industry focused, I would need an additional lookup table or column – or industry vector – to provide a more detailed picture as to how the data could be used.  
  
This could be achieved by adding an input field in the web application called “INDUSTRY” or by having a dimensional table in the database that provides information about the company.  
  
This report would include the following information:

|  |
| --- |
| **Vector** |
| skill\_title |
| Type |
| CreationDate |
| INDUSTRY |
| ClusteredTitle |
| Complexity |
| importance |

With INDUSTRY defined, I would be able to include aggregations about the top skills that each industry required.  
  
Multiline charts can be performed with the date vector, the industry and the skill title.  
Slicers, filters and plots can be applied using INDUSTRY, location, complexity and importance.