LinkedIn Job Data Analysis

A dashboard analysis on the state of 2022's Data Analysis Job market

Introduction

This presentation is made by Harris Zheng using PowerPoint.

The **goal** of this presentation is to show the insights I have derived from the LinkedIn Job data, as well as any relevant context.

All data and visualizations in this presentation are produced from Tableau and Python Plotly. Links to all data sources are provided in footnotes.

Table of Contents (ToC)

- Origin of Data (Data Source)
- Insights
  - Interesting Findings
- Technical Details
- Next Steps
- Resources (Codebase/Dashboards)
- Reference
  - General Dataset Details

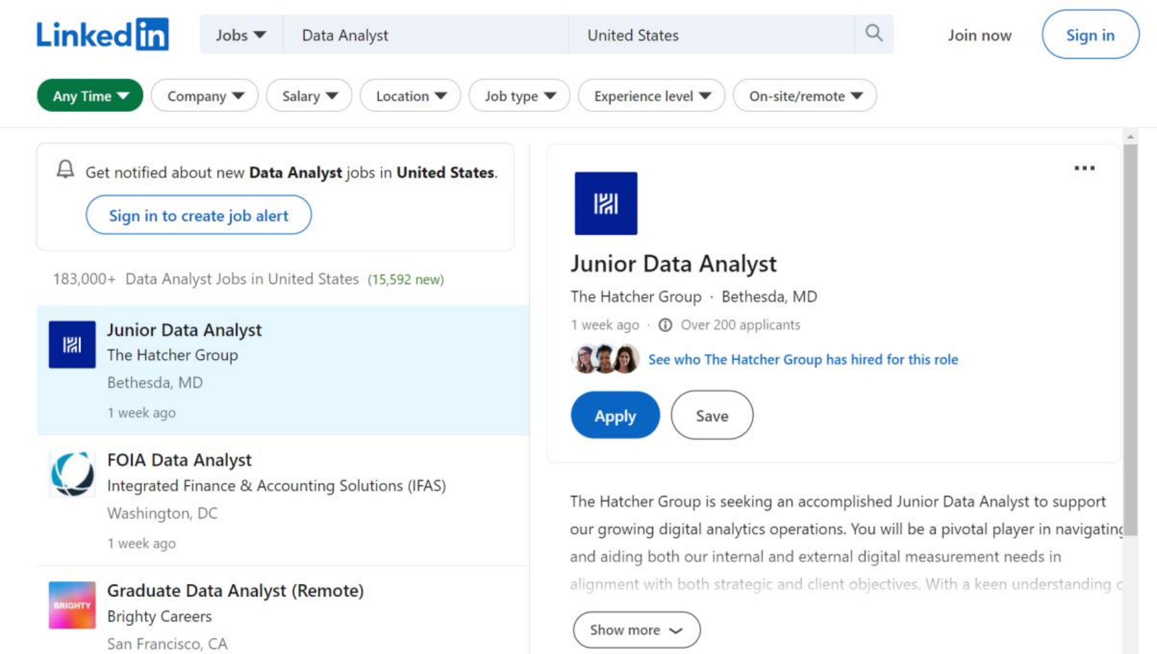
Origin of Data (Data Source)

- LinkedIn Job Data was downloaded from this Kaggle dataset [1], which is a collection of jobs scraped from LinkedIn using BeautifulSoup. These jobs were posted from three locations: Africa, Canada and the USA
- A subset of the data was sliced from the dataset for analysis, consisting of **only Canadian and US jobs** totalling in a CSV of **5718 rows**

[1] https://www.kaggle.com/datasets/cedricaubin/linkedin-data-analyst-jobs-listings

Origin of Data (Data Source)

- The jobs are scraped from the same location as where a normal non-robot user would access the jobs: By querying "Data Analyst" through the LinkedIn job search panel.



Get notified about new **Data Analyst** jobs in **United States**.

Sign in to create job alert

183,000+ Data Analyst Jobs in United States (15,592 new)

Junior Data Analyst
The Hatcher Group
Bethesda, MD
1 week ago

FOIA Data Analyst
Integrated Finance & Accounting Solutions (IFAS)
Washington, DC
1 week ago

Graduate Data Analyst (Remote)
Brighty Careers
San Francisco, CA

Junior Data Analyst
The Hatcher Group · Bethesda, MD
1 week ago · Over 200 applicants
See who The Hatcher Group has hired for this role

The Hatcher Group is seeking an accomplished Junior Data Analyst to support our growing digital analytics operations. You will be a pivotal player in navigating and aiding both our internal and external digital measurement needs in alignment with both strategic and client objectives. With a keen understanding c

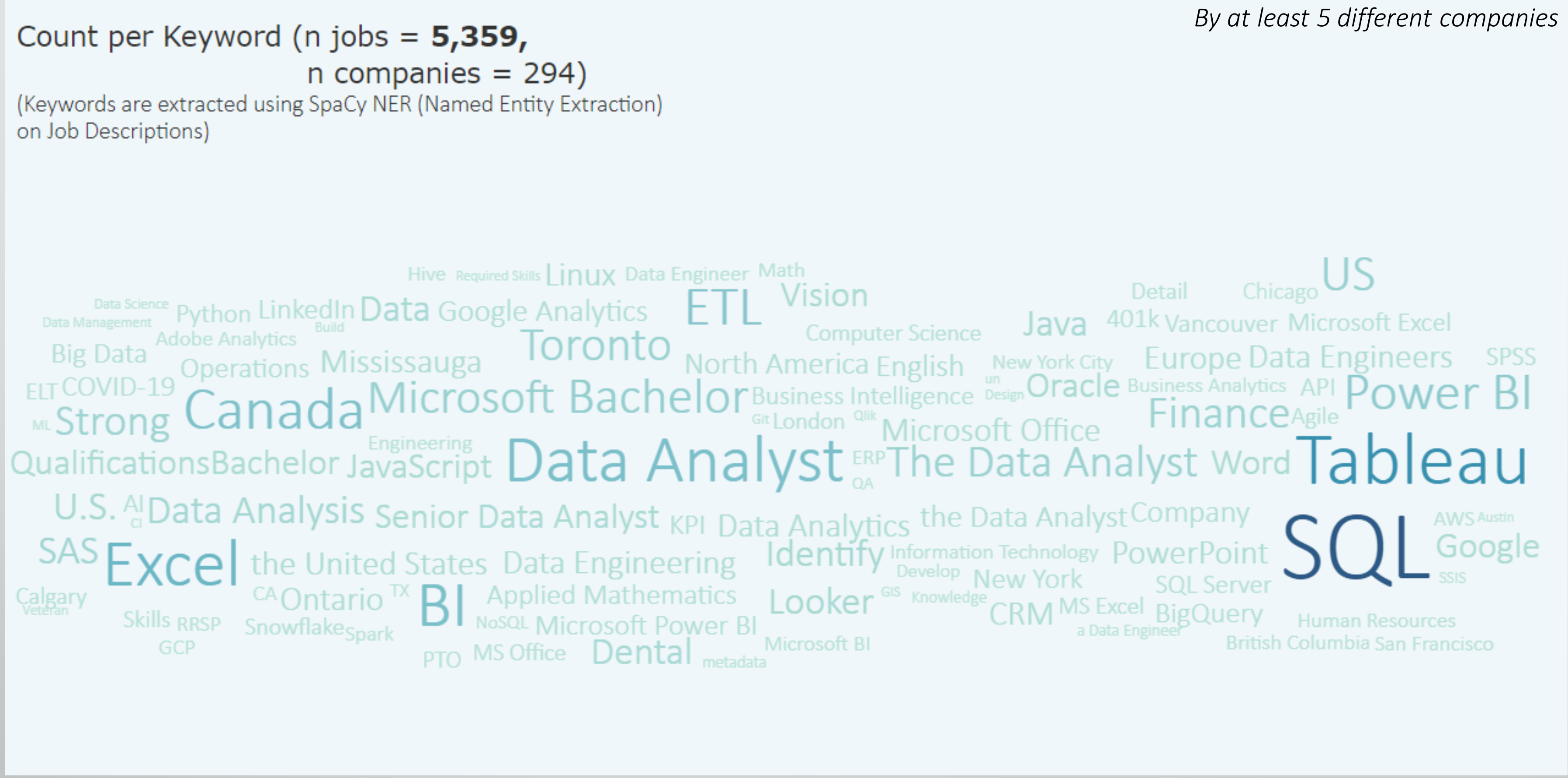
https://www.linkedin.com/jobs/search?keywords=data%20analyst&location=United%20States&geoId=103644278&trk=public_jobs_jobs-search-bar_search-submit&position=1&pageNum=0

Insights

Interesting Findings

Top Skill Words for Data Analyst Positions

Each skill is reported
By at least 5 different companies



Count per Keyword (n jobs = **5,359,**
n companies = 294)

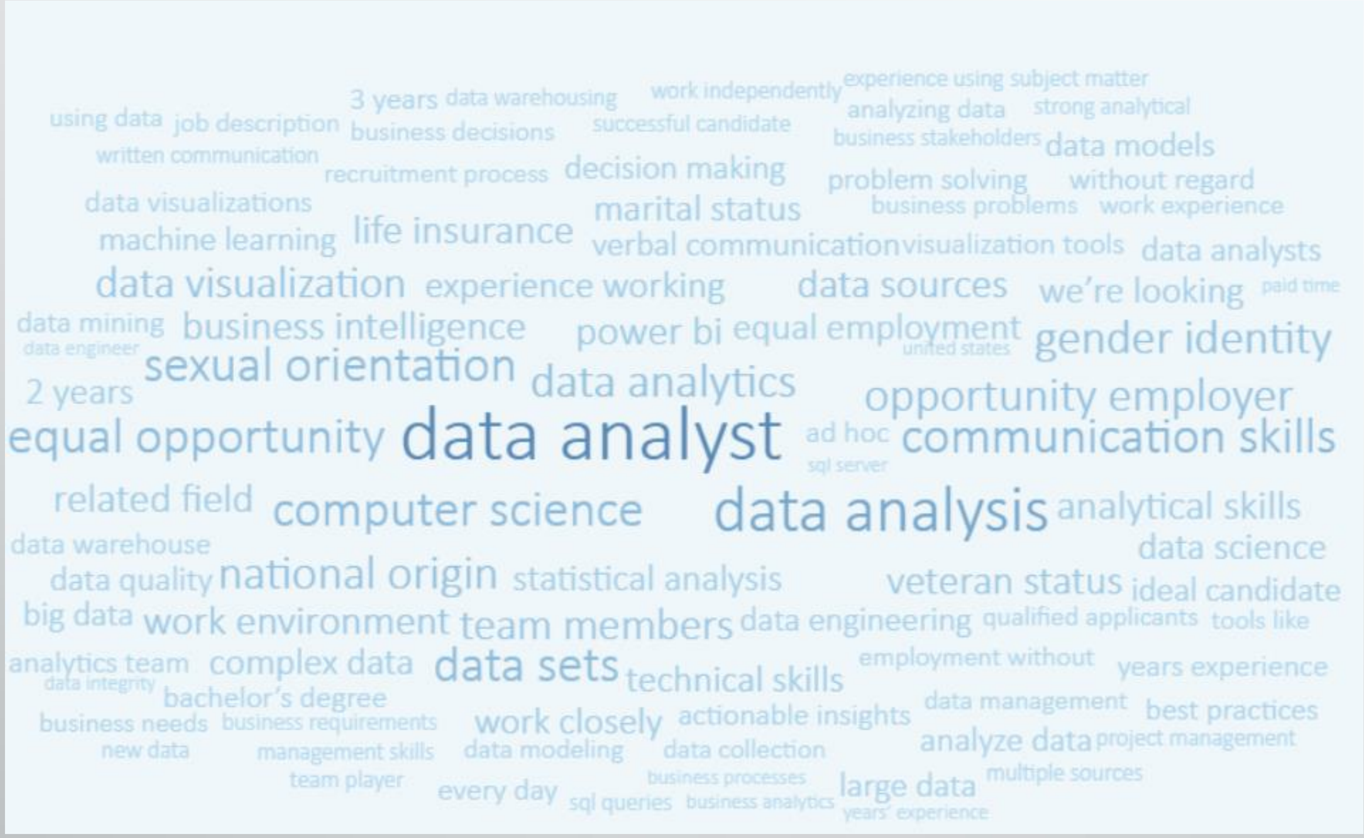
(Keywords are extracted using SpaCy NER (Named Entity Extraction) on Job Descriptions)

SQL, Tableau, BI, Excel and ETL are amongst the most prevalent skill words to appear in Data Analyst job postings

Other slightly less prevalent keywords include Python, Linux, Java, SAS etc..

Bigrams

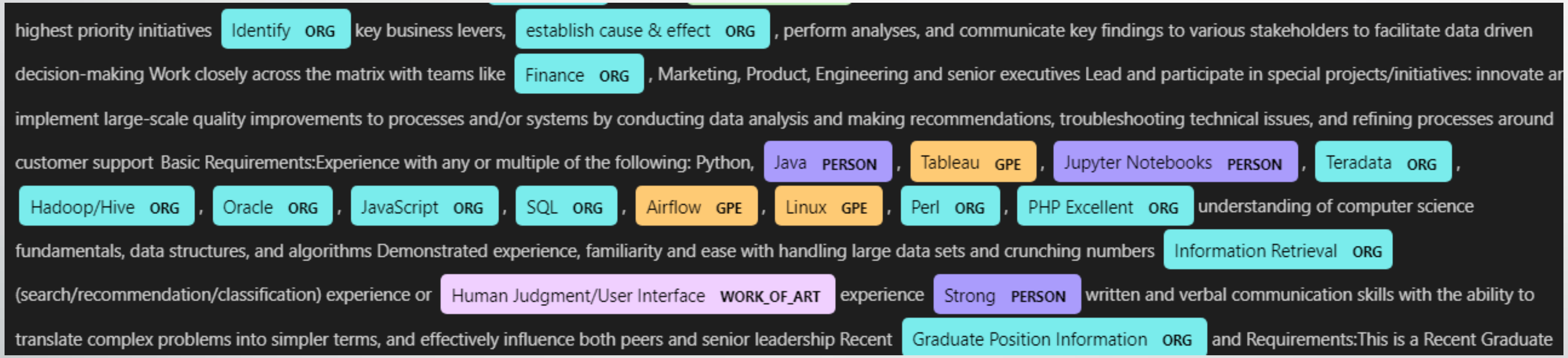
Each Bigram is reported
By at least 5 different companies



Apart from every company wanting to present themselves as an "equal opportunity" employer, there appears to be interest in *"business intelligence"*, *"statistical analysis"*, *"data visualization"*, *"data mining"*, *"data science"*, *"machine learning"* etc., showing that Data Analysts are expected to be well-versed in both **modern and traditional methods of analysis and insight extraction**

How are these Skill Words extracted?

- By using SpaCy’s pretrained *en_core_web_sm* NLP Model to extract skill words from unstructured job descriptions.



highest priority initiatives Identify ORG key business levers, establish cause & effect ORG , perform analyses, and communicate key findings to various stakeholders to facilitate data driven
decision-making Work closely across the matrix with teams like Finance ORG , Marketing, Product, Engineering and senior executives Lead and participate in special projects/initiatives: innovate ar
implement large-scale quality improvements to processes and/or systems by conducting data analysis and making recommendations, troubleshooting technical issues, and refining processes around
customer support Basic Requirements:Experience with any or multiple of the following: Python, Java PERSON , Tableau GPE , Jupyter Notebooks PERSON , Teradata ORG ,
Hadoop/Hive ORG , Oracle ORG , JavaScript ORG , SQL ORG , Airflow GPE , Linux GPE , Perl ORG , PHP Excellent ORG understanding of computer science
fundamentals, data structures, and algorithms Demonstrated experience, familiarity and ease with handling large data sets and crunching numbers Information Retrieval ORG
(search/recommendation/classification) experience or Human Judgment/User Interface WORK_OF_ART experience Strong PERSON written and verbal communication skills with the ability to
translate complex problems into simpler terms, and effectively influence both peers and senior leadership Recent Graduate Position Information ORG and Requirements:This is a Recent Graduate

SpaCy is not trained to extract job skills specifically. However, it is still able to extract skills through other entity types such as GPE (Locations), ORG (Organizations), and PERSON (People). As shown in the above image of an example extraction from job descriptions, we can see many programming skills are identified.

Note that this extraction method isn't perfect. One problem I noticed with this method is that SpaCy seems to recognize Python as an animal instead of a job skill… We need ML training and job descriptions labelled with skills to ensure accuracy

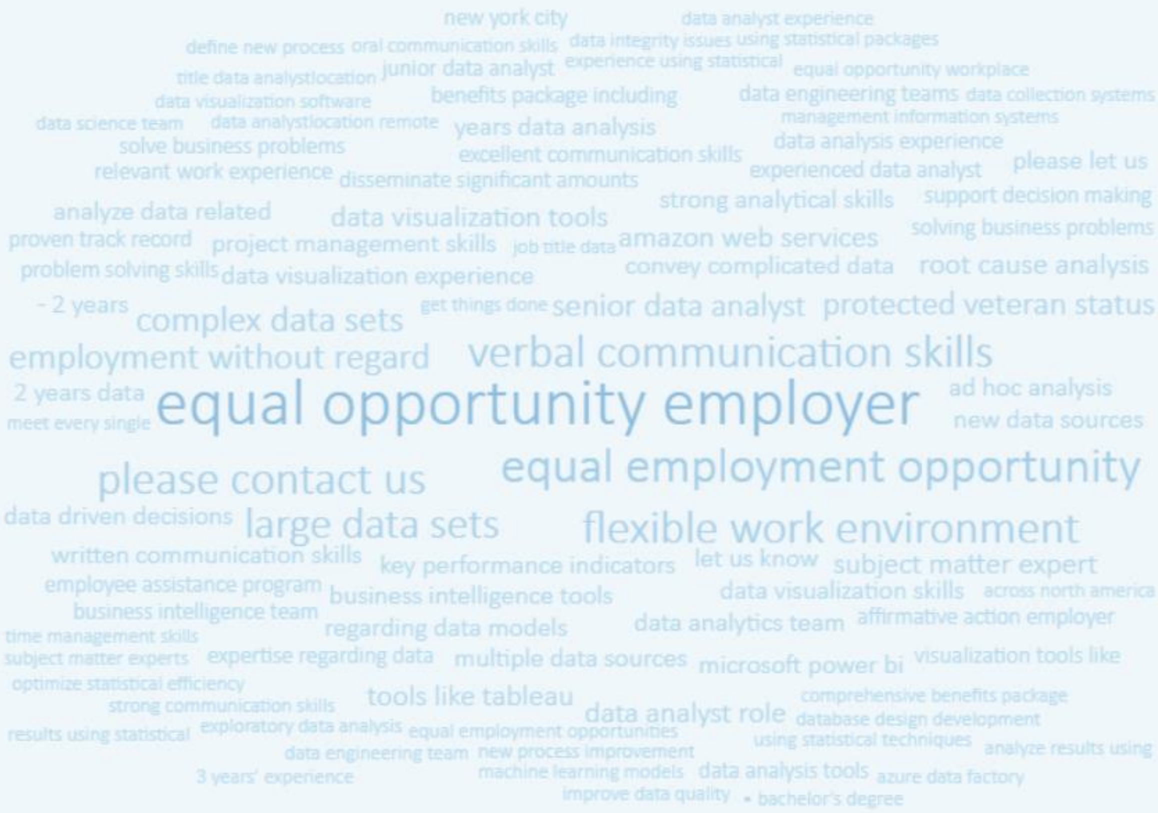
Trigrams

From trigram analysis, we can see that Data Analysts are generally expected to work with “large/complex data sets”, as well as perform “ad hoc analysis” **based on stakeholder need**.

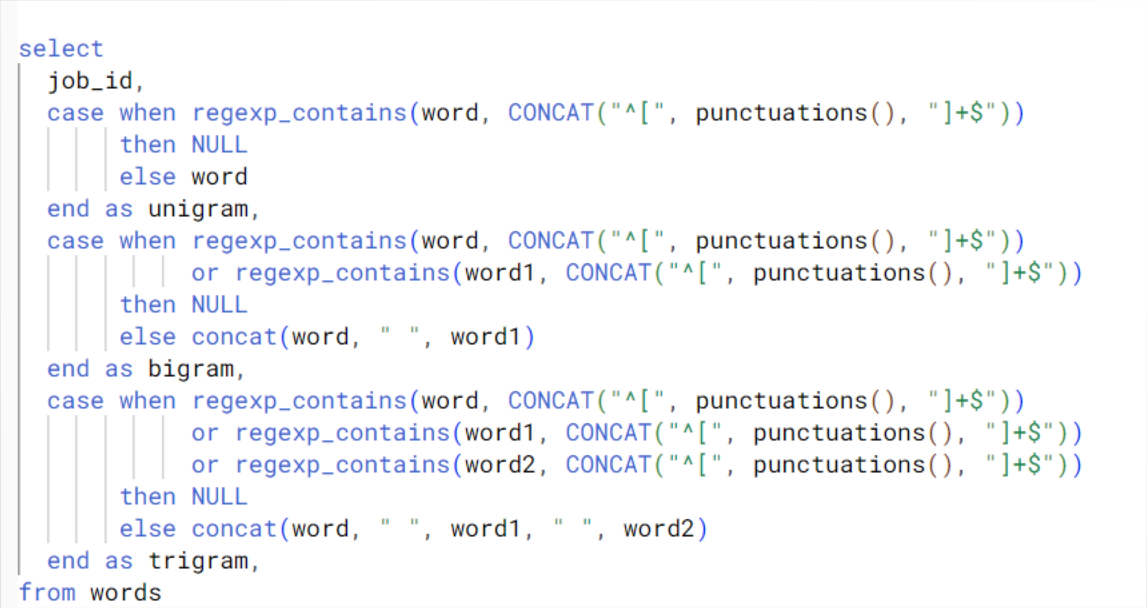
We can see that analysts are generally expected to meet KPIs and perform on-call but are not micromanaged as they enjoy “**flexible work environment**”.

Trigram (n jobs = 5,618)

Each Trigram is reported
By at least 5 different companies



Bigram and Trigram Preprocessing

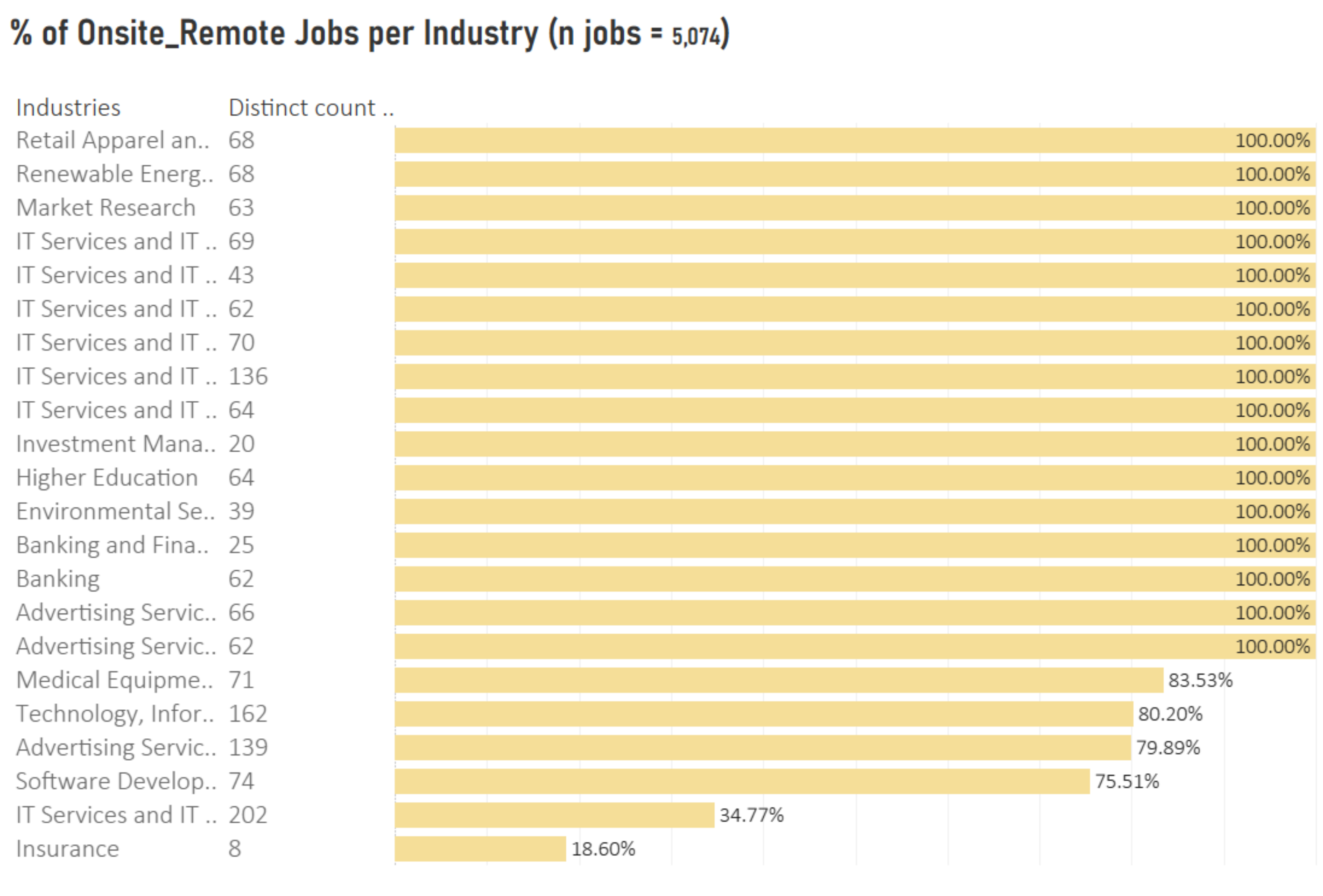


```
select
  job_id,
  case when regexp_contains(word, CONCAT("^[", punctuations(), "]+$"))
       then NULL
       else word
  end as unigram,
  case when regexp_contains(word, CONCAT("^[", punctuations(), "]+$"))
            or regexp_contains(word1, CONCAT("^[", punctuations(), "]+$"))
       then NULL
       else concat(word, " ", word1)
  end as bigram,
  case when regexp_contains(word, CONCAT("^[", punctuations(), "]+$"))
            or regexp_contains(word1, CONCAT("^[", punctuations(), "]+$"))
            or regexp_contains(word2, CONCAT("^[", punctuations(), "]+$"))
       then NULL
       else concat(word, " ", word1, " ", word2)
  end as trigram,
from words
```

Take a set of tokenized words (including punctuation!) from LinkedIn job descriptions **and nullify any bigrams/trigrams that contain punctuations**.

(e.g. if we did not include punctuation in our original set of words, “I felt stupid. Boy I could’ve done better.” would’ve falsely produced a bigram “stupid Boy”).

Remoteness by Sector

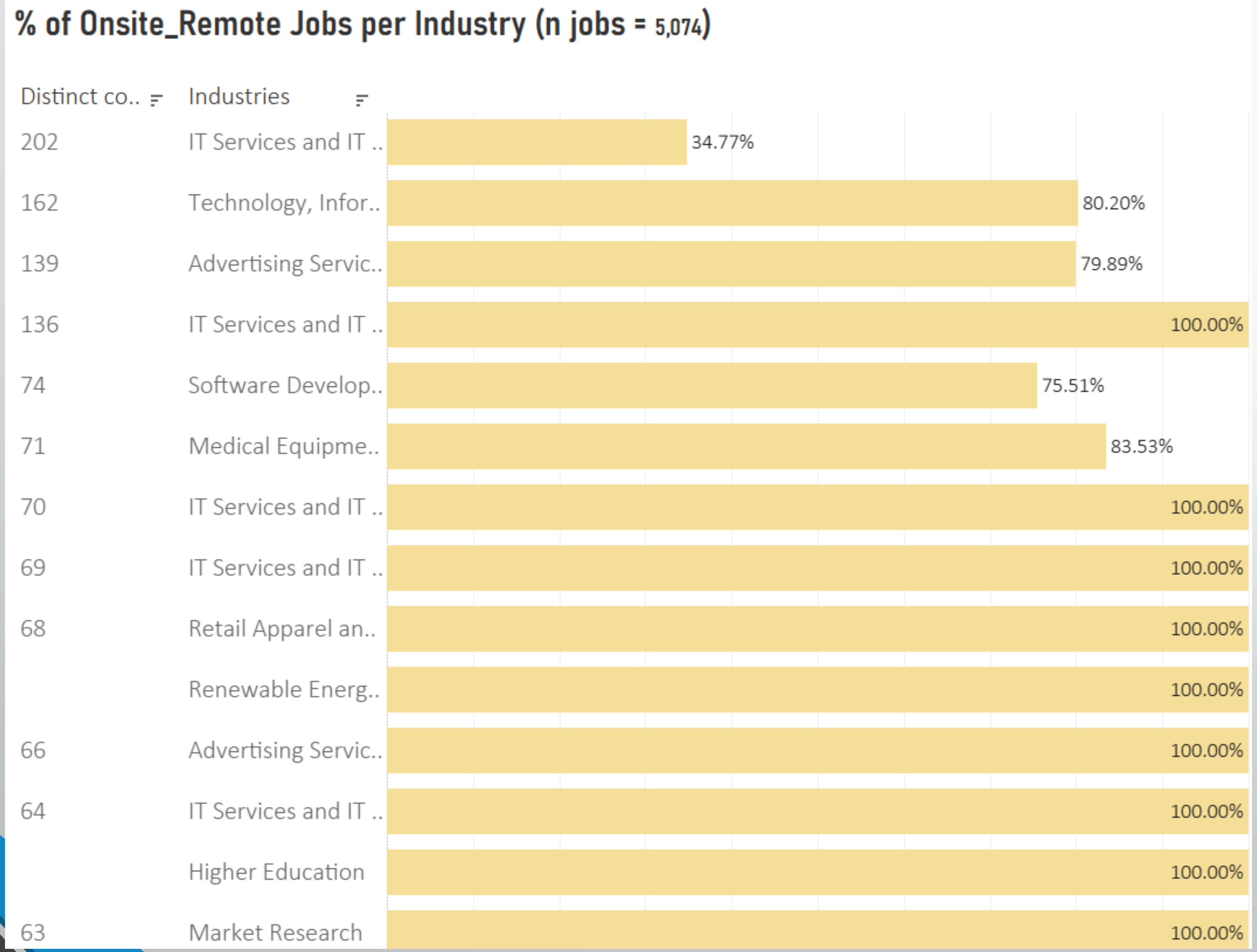


% of Onsite_Remote Jobs per Industry (n jobs = 5,074)

Industries	Distinct count ..	
Retail Apparel an..	68	100.00%
Renewable Energ..	68	100.00%
Market Research	63	100.00%
IT Services and IT ..	69	100.00%
IT Services and IT ..	43	100.00%
IT Services and IT ..	62	100.00%
IT Services and IT ..	70	100.00%
IT Services and IT ..	136	100.00%
IT Services and IT ..	64	100.00%
Investment Mana..	20	100.00%
Higher Education	64	100.00%
Environmental Se..	39	100.00%
Banking and Fina..	25	100.00%
Banking	62	100.00%
Advertising Servic..	66	100.00%
Advertising Servic..	62	100.00%
Medical Equipme..	71	83.53%
Technology, Infor..	162	80.20%
Advertising Servic..	139	79.89%
Software Develop..	74	75.51%
IT Services and IT ..	202	34.77%
Insurance	8	18.60%

A large proportion of IT Service/Consultancy-related jobs are fully remote, as well as banking jobs, retail jobs, renewable energy and market research jobs.

Remoteness by Sector

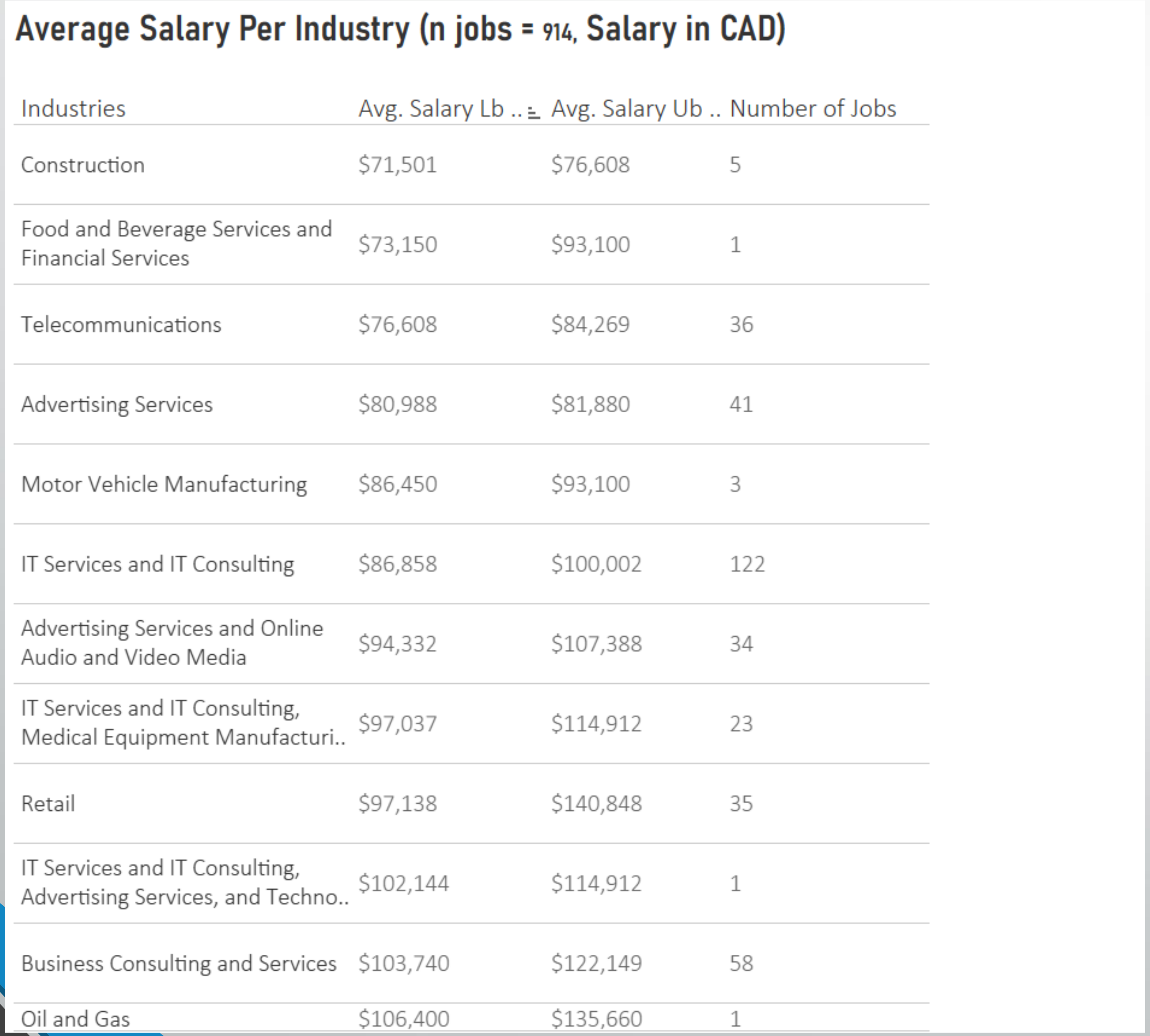


% of Onsite_Remote Jobs per Industry (n jobs = 5,074)

Distinct co..	Industries	
202	IT Services and IT ..	34.77%
162	Technology, Infor..	80.20%
139	Advertising Servic..	79.89%
136	IT Services and IT ..	100.00%
74	Software Develop..	75.51%
71	Medical Equipme..	83.53%
70	IT Services and IT ..	100.00%
69	IT Services and IT ..	100.00%
68	Retail Apparel an..	100.00%
	Renewable Energ..	100.00%
66	Advertising Servic..	100.00%
64	IT Services and IT ..	100.00%
	Higher Education	100.00%
63	Market Research	100.00%

If we just rank industries by number of remote job offerings instead of proportions, then we see Technology/Information jobs as well as Advertising and Services jobs are often remote

Lowest Paid Industries on Average

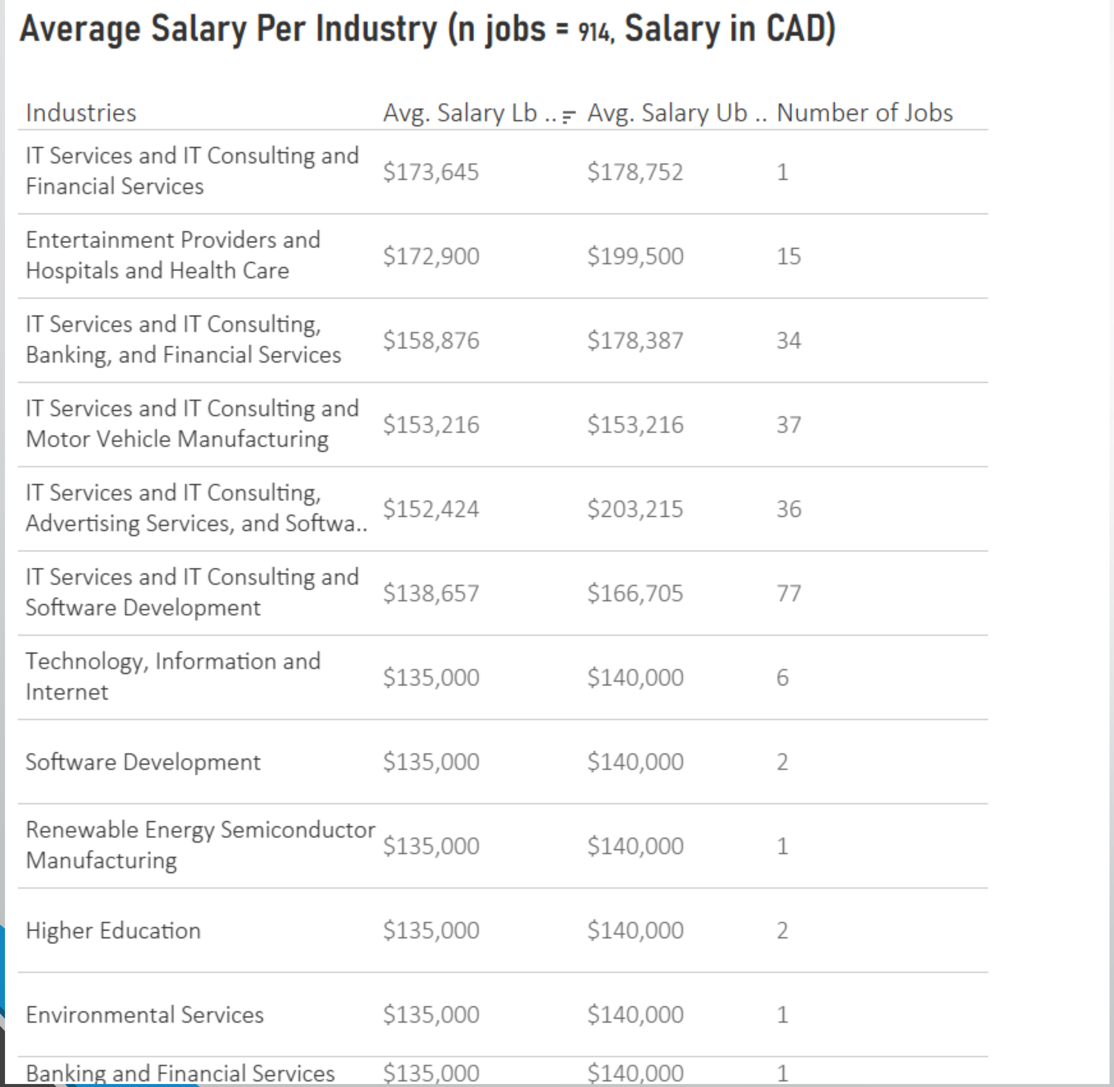


Average Salary Per Industry (n jobs = 914, Salary in CAD)

Industries	Avg. Salary Lb ..	Avg. Salary Ub ..	Number of Jobs
Construction	$71,501	$76,608	5
Food and Beverage Services and Financial Services	$73,150	$93,100	1
Telecommunications	$76,608	$84,269	36
Advertising Services	$80,988	$81,880	41
Motor Vehicle Manufacturing	$86,450	$93,100	3
IT Services and IT Consulting	$86,858	$100,002	122
Advertising Services and Online Audio and Video Media	$94,332	$107,388	34
IT Services and IT Consulting, Medical Equipment Manufacturi..	$97,037	$114,912	23
Retail	$97,138	$140,848	35
IT Services and IT Consulting, Advertising Services, and Techno..	$102,144	$114,912	1
Business Consulting and Services	$103,740	$122,149	58
Oil and Gas	$106,400	$135,660	1

Lowest Paid Industries on average include “Construction”, “Telecommunications”, "Advertising Services" etc.

Highest Paid Industries on Average



Average Salary Per Industry (n jobs = 914, Salary in CAD)

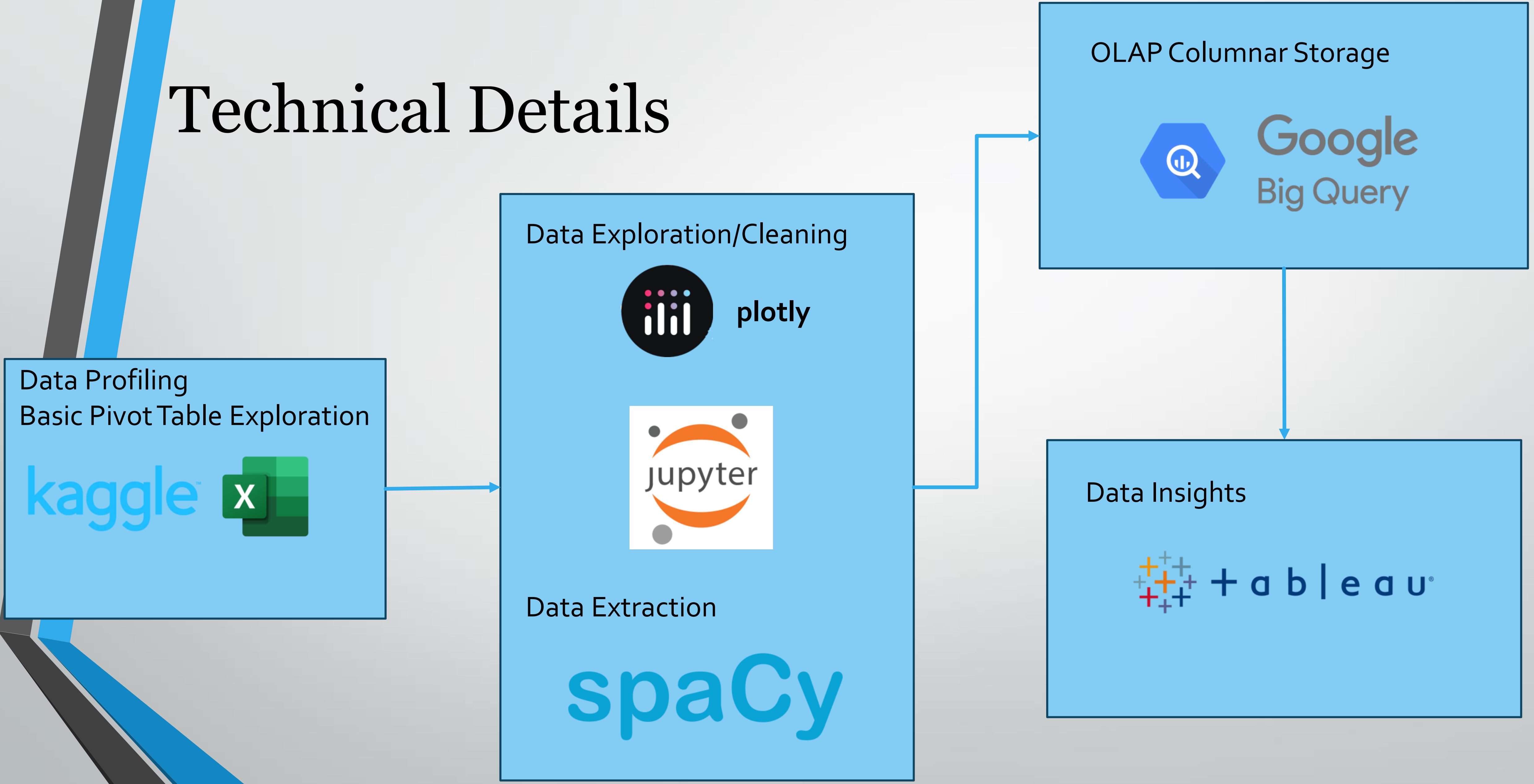
Industries	Avg. Salary Lb ..	Avg. Salary Ub ..	Number of Jobs
IT Services and IT Consulting and Financial Services	$173,645	$178,752	1
Entertainment Providers and Hospitals and Health Care	$172,900	$199,500	15
IT Services and IT Consulting, Banking, and Financial Services	$158,876	$178,387	34
IT Services and IT Consulting and Motor Vehicle Manufacturing	$153,216	$153,216	37
IT Services and IT Consulting, Advertising Services, and Softwa..	$152,424	$203,215	36
IT Services and IT Consulting and Software Development	$138,657	$166,705	77
Technology, Information and Internet	$135,000	$140,000	6
Software Development	$135,000	$140,000	2
Renewable Energy Semiconductor Manufacturing	$135,000	$140,000	1
Higher Education	$135,000	$140,000	2
Environmental Services	$135,000	$140,000	1
Banking and Financial Services	$135,000	$140,000	1

Highest Paid Industries are IT Service and IT Consulting companies that specialize in a niche, such as Financial Services/Health Care etc..

Summary of Learnings

- **SQL, Tableau, ETL, Excel, Business Intelligence, and BI** are amongst the most required skills in LinkedIn job postings.
- From our bigram analysis, we learned that job postings require **a strong familiarity with data analysis methods**, from traditional methods such as statistics to modern solutions like machine learning.
- From our trigram analysis, we learned that Data Analysts boast a **flexible work environment**, at the cost of being able to stay on-call for **ad-hoc analysis**.
- A large proportion of IT Services and Consulting jobs were **remote**, as well as Banking jobs and Advertising jobs
- IT Services and Consulting for Specific Industries (e.g. Healthcare, Financial Services, Banking) are **paid highest** for Data Analysts (~$130,000 - $170,000 CAD/year), while **lowest paid** industries include Telecommunications/Food Beverages Services/Construction (~$70,000-$80,000 CAD/year)

Technical Details



Data Profiling
Basic Pivot Table Exploration

Data Exploration/Cleaning

Data Extraction

OLAP Columnar Storage

Data Insights

Next Steps:

- Build automated job scheduler (ex. Cron or Airflow) to scrape LinkedIn job data daily, so that we can keep up with Data Analyst trends (e.g. required skills, best salaries)
- Annotate job dataset using a commercial annotation tool like LabelBox and apply transfer learning on various pre-trained NLP models (like SpaCy) to learn skill word entity extraction. This gives us more confidence in our skill extraction model, since we can calculate precision and recall based on our annotations and choose the best model based on those metrics.
- Implement ML Data Cleaning models and data extraction techniques to make data more accurate.
- Utilize various tools such as data processing tools like Hadoop or extensions to Python like Cython to make analysis pipeline more performant

Resources (Codebase and Dashboards)

- Tableau Dashboards:
  - Job Trend Dashboard (includes Job Descriptions analysis)
  - Bigram Trigram Dashboard
- Github
- Blog

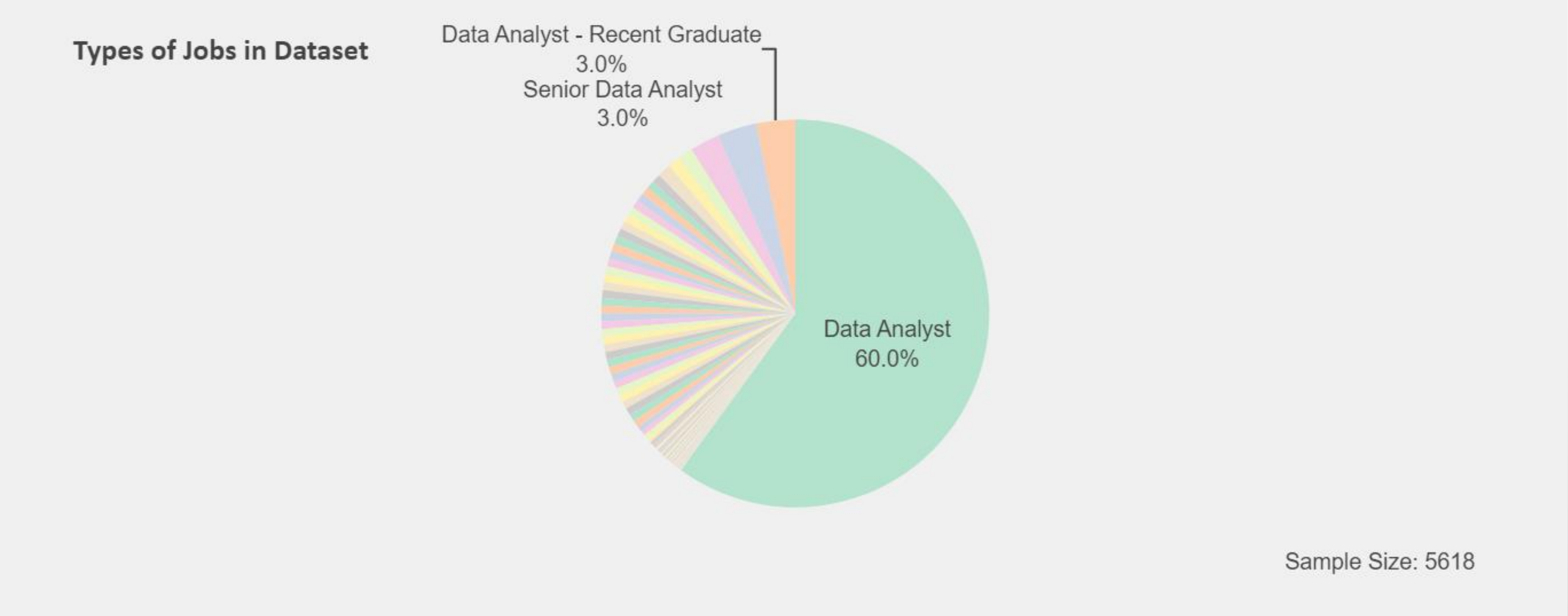
Thank you for Going through this Presentation!

Appreciate your time and interest, always

Reference

General Dataset Details

Top Job Types



Most prevalent Job Title found is “Data Analyst”, representing 60% of the dataset.

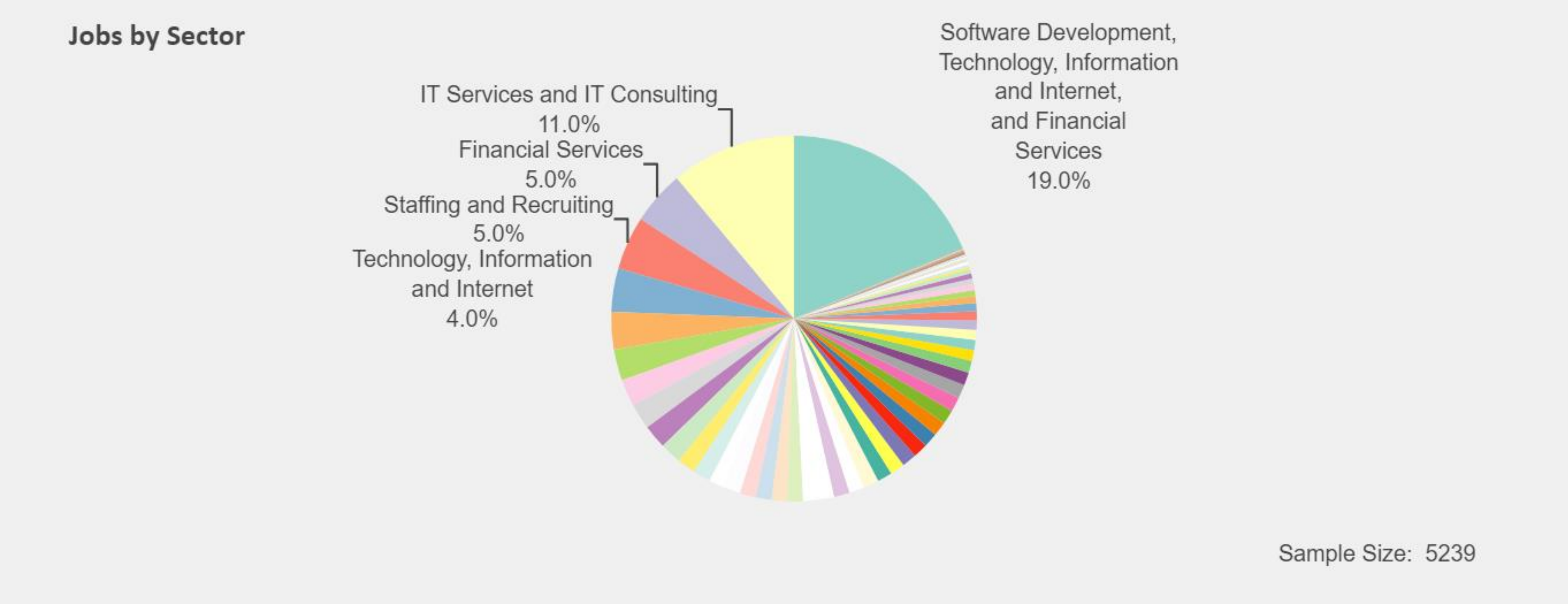
Other job titles include “Data Analyst – Recent Graduate”, “Senior Data Analyst”, “Junior Data Analyst”

General Dataset Details

Top Job Sectors

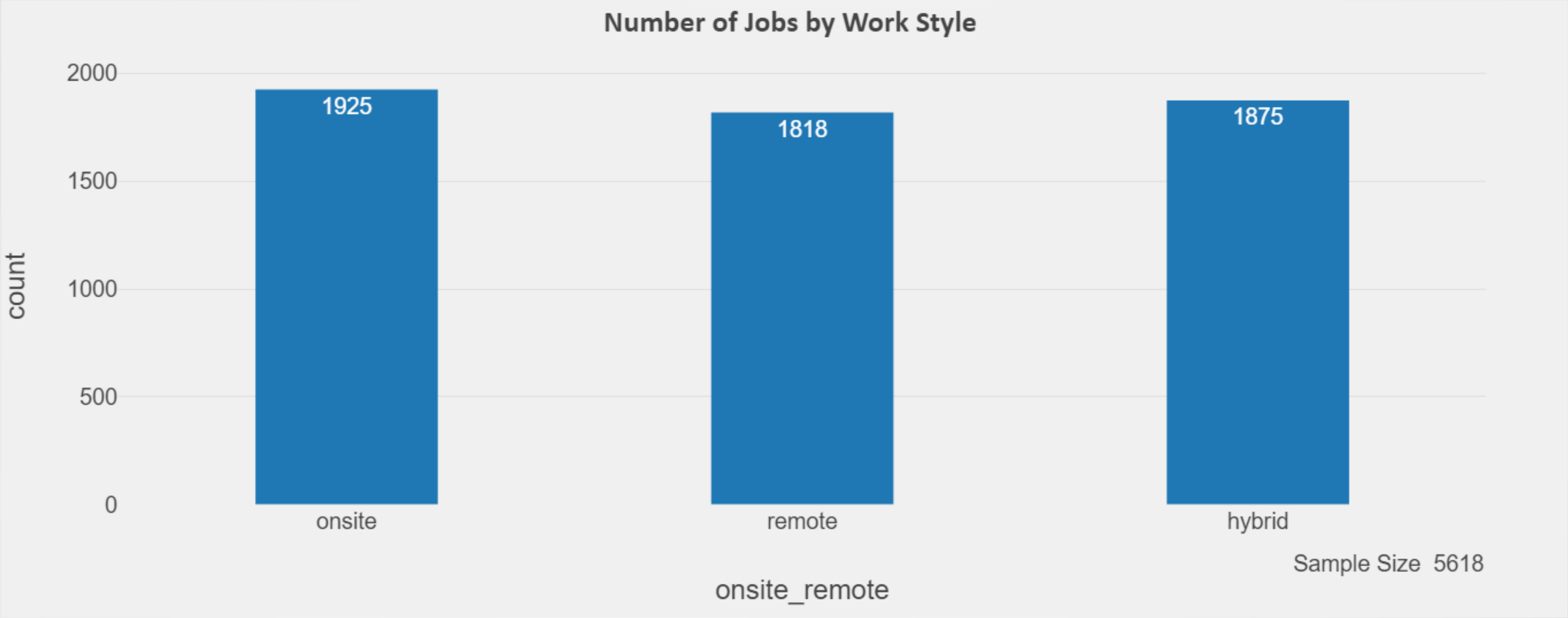
Most prevalent Job Sector is Software Development/Financial Services at 19% of jobs in dataset

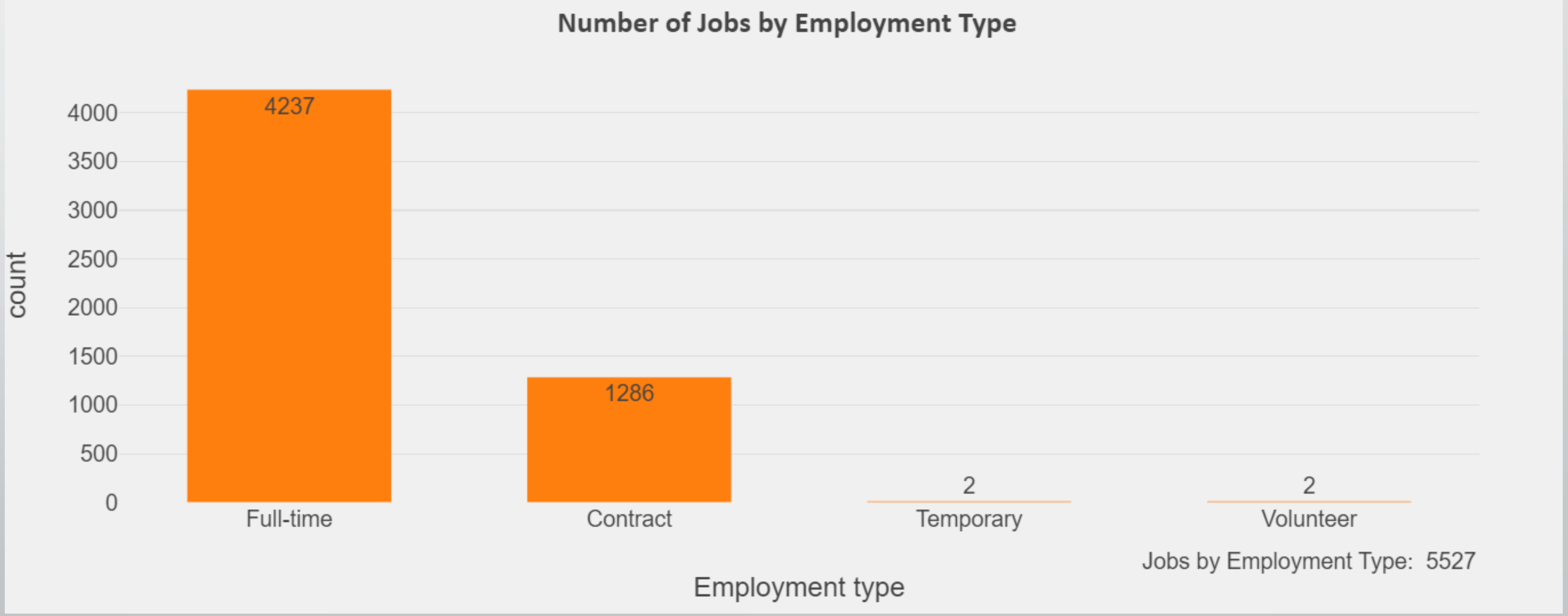
Alongside with other top sectors such as IT Consulting, Staffing and Recruiting etc..



General Dataset Details

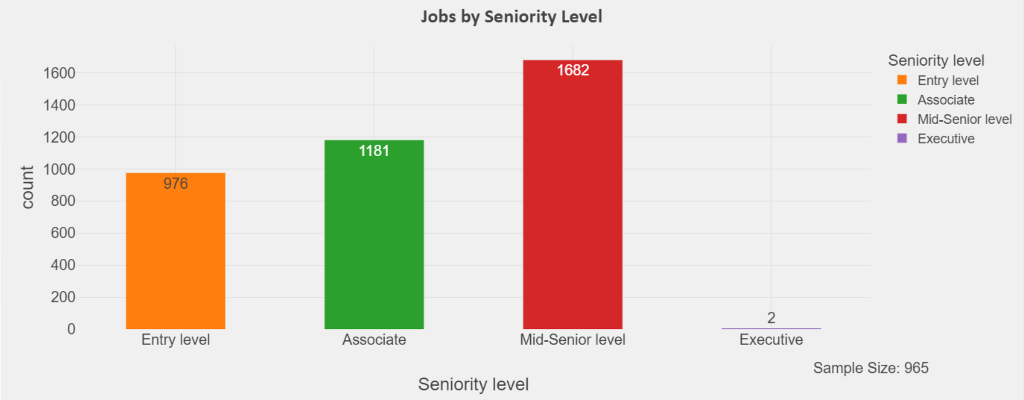
Most jobs are either Full-time or Contract jobs. Work styles are evenly distributed.

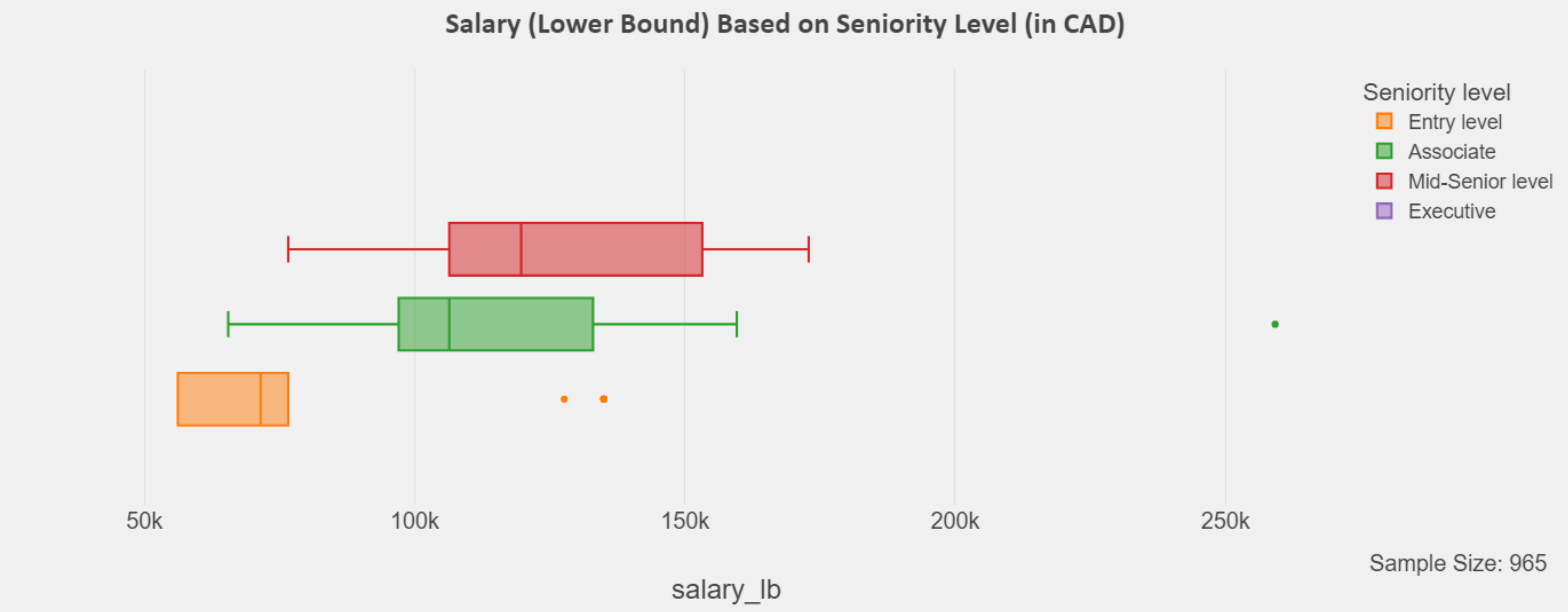




General Dataset Details

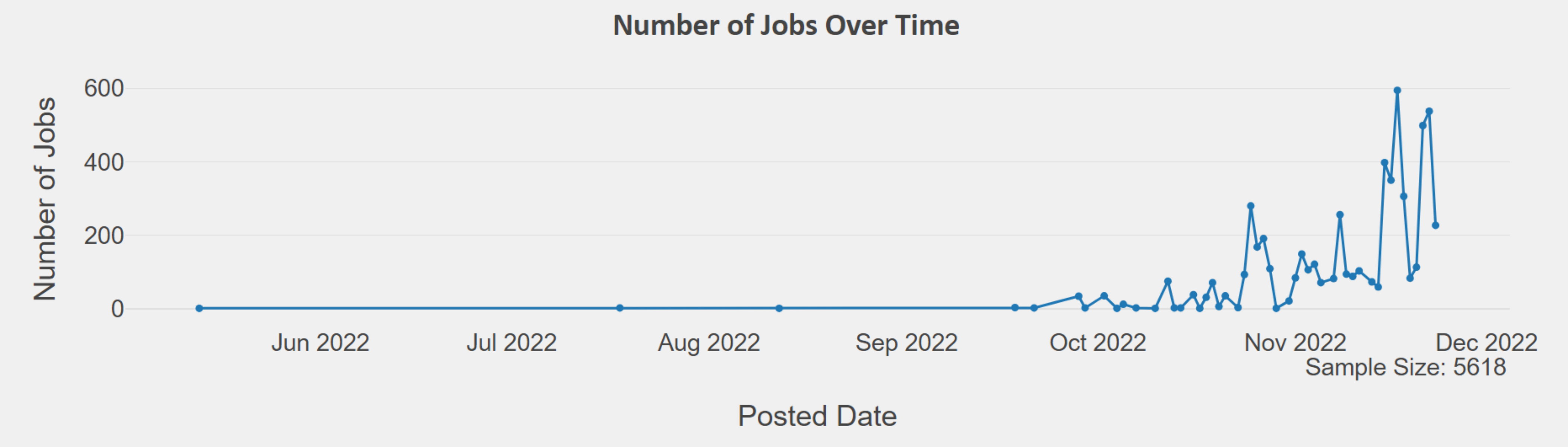
Salary by Seniority Level





Note that salaries posted in LinkedIn job posting have upper and lower bounds (e.g. $25 - $27/hr). The above visualization shows a box plot of lower bound values. Salaries are adjusted based on December 2022 currency conversion rates (1 USD = 1.33 CAD)

General Dataset Details



Posted Dates have “Day” resolution

Job postings in this dataset were data scraped mostly from Sept – Dec 2022