Indeed Web Scraper: Job Market Trends for Data Careers

Taesun Yoo

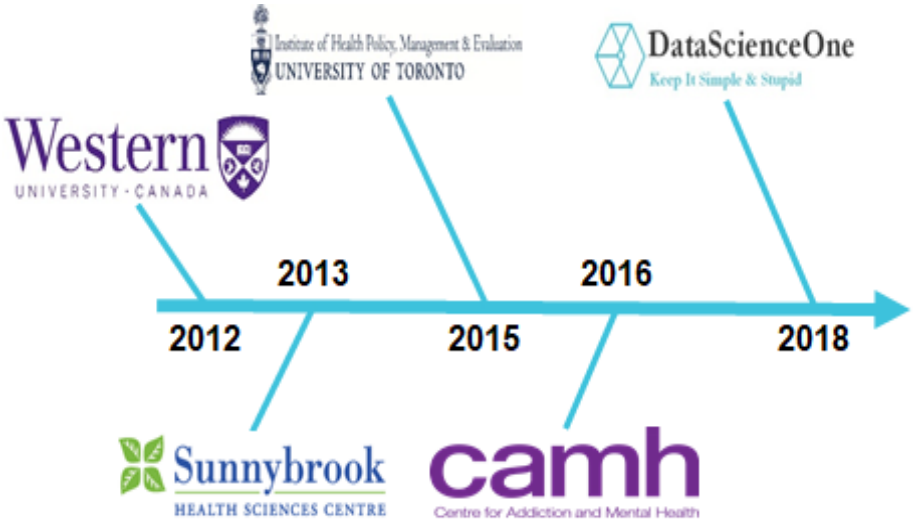
- Aug 30, 2018 -

About Myself: Taesun Yoo

- Former BI QA Analyst, ML Enthusiast
- Founder of DataScienceOne (Youtube Channel)
- Completed Master's in Health Informatics

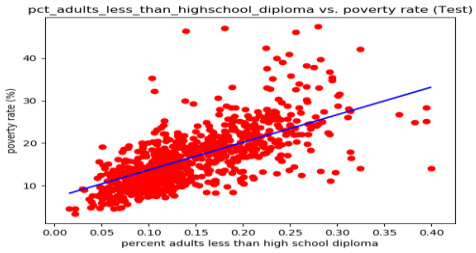
- Research experience: Sunnybrook
  - ❑ Medical Imaging (image processing)
  - ❑ Radiation Physics (cancer treatment)

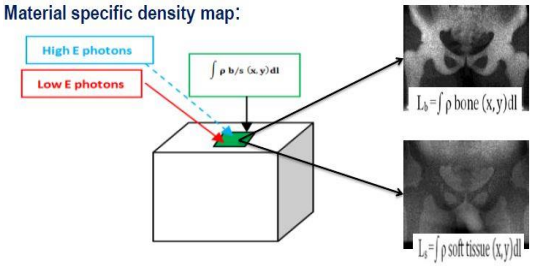
- Work experience: CAMH
  - ❑ Business Intelligence
  - ❑ QA data warehousing
  - ❑ Data visualization/reporting



Kicking some side machine learning projects …

MajorityVote Classifier		
	Predicted Class	
Actual Class	Stroke	Non-stroke
Stroke	41%	9%
Non-stroke	11%	39%





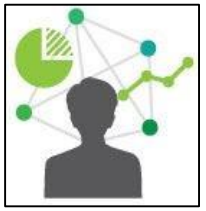
Agenda

1. • Problem Overview
2. • Data Wrangling: Indeed.com
3. • Exploratory Data Analysis
4. • Future Work and Recommendations

Problem Overview

Indeed Job Search Results in Toronto

On Aug 26, 2018

95 jobs	Data Scientist	
90 jobs	Data Engineer	
136 jobs	Data Analyst	
647 jobs	Business Intelligence	

Indeed Job Market: Trend Analysis

Why should we care?

- Improve % of getting an interview
- Strategize job application submissions

Stakeholders:

- Job applicants
- Hiring managers and recruiters

Goal:

- Provide overall data career market insights in Toronto!

Objectives:

- Save time job application submission vs. keyword-based search
- Prioritize and learn top required skills/tools

Indeed Web Scraped: Dataset Overview

Dataset contains **10** variables for exploratory data analysis:

- 7 categorical & 3 numerical features
- Demographics: required education, tools/skills, salary and company
- Sample size: x < 1,000 jobs

Observations (rows)

Required Skills			Required Education Level				Salary & Company Demographics				
Job Title	Tools	Count %	Job Title	Education Level	Majors	Count %	Job Title	Company Name	Salary	City	Province
Data Scientist	Python	15	Data Scientist	Bachelor	Engineering	29	Data Analyst	York University	87109	Toronto	Ontario
Data Scientist	Hadoop	9	Data Scientist	Master	Computer Science	29	Business Intelligence	Humber River Hospital	65733.5	Toronto	Ontario
Data Scientist	R	8.5	Data Scientist	PhD	Math	42	Data Scientist	ChefHero	95000	Toronto	Ontario
Data Engineer	Python	10	Data Engineer	Bachelor	Engineering	61	Data Engineer	JB Micro Inc	95000	Brampton	Ontario
Data Engineer	Spark	9	Data Engineer	Master	Computer Science	29	Data Engineer	Workbridge Associates	110000	Toronto	Ontario
Data Engineer	Java	8.5	Data Engineer	PhD	Math	10	Data Engineer	Jobspring Partners	105000	Toronto	Ontario

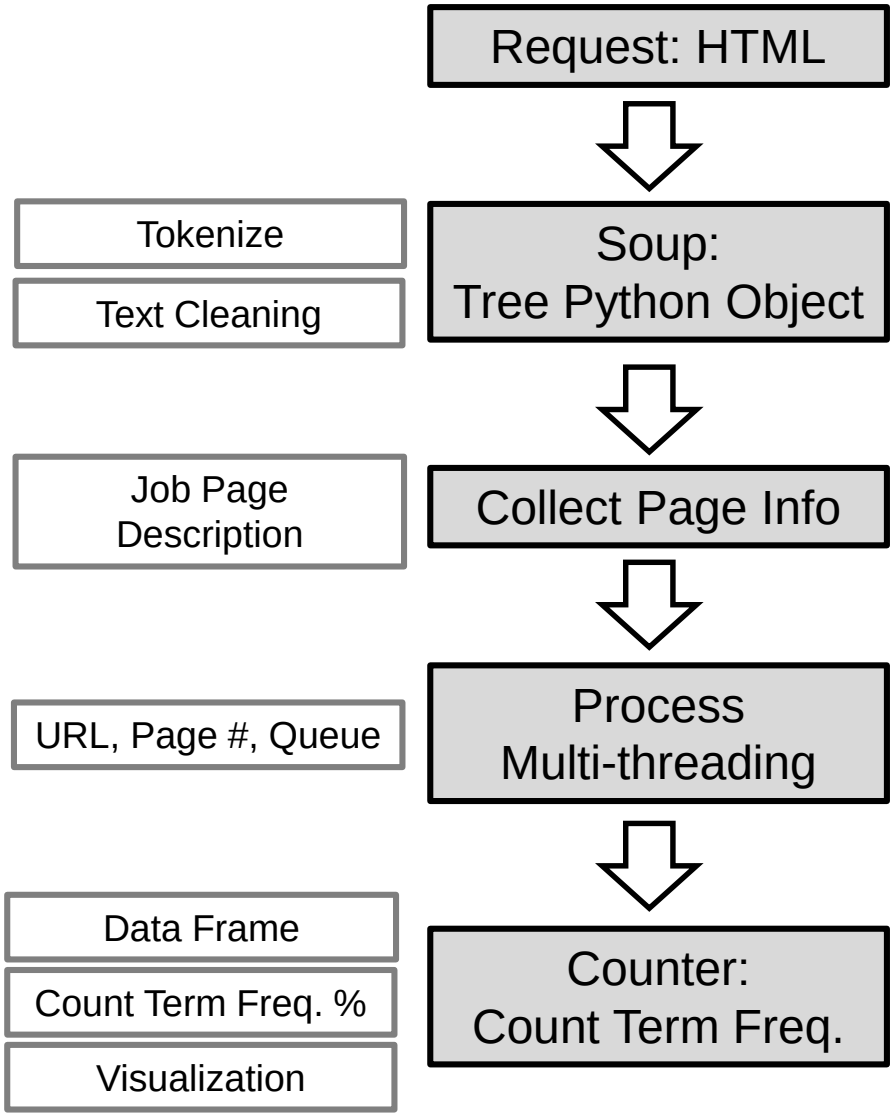
Features (attributes)

Challenges:

- Missing salary information (job posting)
- Inconsistent format (i.e., hourly/weekly salary)
- Duplication (i.e., same job postings)

Data Wrangling: Indeed.com

Data Wrangling Pipeline



$$\text{Term Freq.} = \frac{Term}{\sum_1^n Term} \text{ x } 100\%$$

Sample Code: Getting Page Info

Input Parameters

```
# Function 7: job scraper and visualize data #
def indeed_job_scraper(city=None, state=None, job_title=None):
    city_copy = city[:]

    if city is not None:
        # For city name like 'San Francisco', we want to convert it into 'San+Francisco'
        city_list = city.split()
        city = '+'.join(city_list)
        url_list = ['http://www.indeed.ca/jobs?q=', job_title, '&l=', city, '%2C+', state]
    else:
        url_list = ['http://www.indeed.ca/jobs?q=', job_title]

    url = ''.join(url_list)    # URL for job search
    print("Using URL " + url)

    try:
        html = requests.get(url).text
    except:
        print('The location ' + city_copy + ', ' + state + ' could not be found.')
        return

    soup = make_soup(html)

    num = soup.find(id = 'searchCount')
    if num == None:
        num = soup.find(id = 'searchCount')
    num_jobs = num.string.encode('utf-8')

    # Total number of jobs found
    job_numbers = re.findall('\d+', str(num_jobs))

    # Process commas in large number representations
    if len(job_numbers) > 3:
        total_num_jobs = (int(job_numbers[2]) * 1000) + int(job_numbers[3])
    else:
        total_num_jobs = int(job_numbers[2])

    if city is None:
        print(str(total_num_jobs) + ' jobs found nationwide')
    print(str(total_num_jobs) + ' jobs found in ' + city_copy + ', ' + state)

    num_pages = int(total_num_jobs / 10)

    # Obtain company list
    page_company = []
    for element in soup.find_all('span', class_='company'):
        page_company.append(element.get_text().strip()) # append company list
```

```
# Multi-threading #
total_job_descriptions = process_url(num_pages, url)    # Four-dimensional list
# Convert four-dimensional list into two-dimensional list
total_job_descriptions = sum(sum(total_job_descriptions, []), [])
total_jobs_found = len(total_job_descriptions)

print('Done with collecting the job ads!')
print('There were ' + str(total_jobs_found) + ' jobs successfully found.')
```

Term Frequency Counter

```
# Counter for terms within job description of each page url
doc_freq = Counter()
[doc_freq.update(item) for item in total_job_descriptions]

## List of company names: numjob openings ##
##########################################
comp_dict = Counter(page_company)
# Convert results into a dataframe
df_comp = pd.DataFrame.from_dict(comp_dict, orient='index').reset_index()
df_comp.columns = ['CompName', 'NumJobs']
# sort data for plotting
df_comp.sort_values(by = 'NumJobs', ascending=False, inplace=True)
pd.set_option('display.width', 1000)
```

Output #1: Education Level

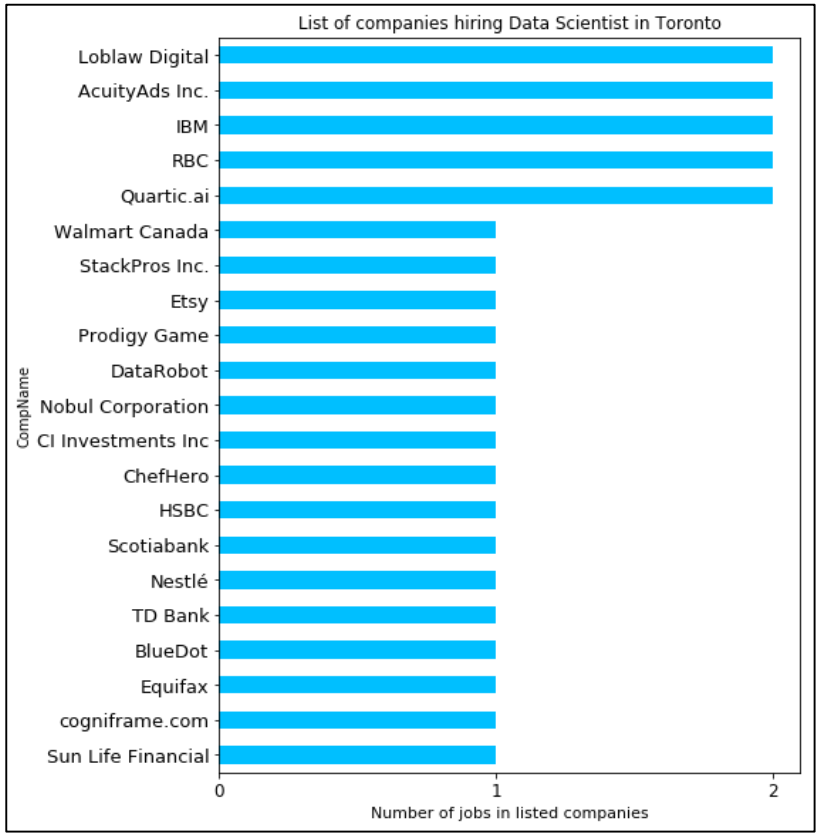
```
## Requirement by education level: calculate % apperance in job ads ##
######################################################################
educ_dict = Counter({'PhD': doc_freq['phd']+doc_freq['doctor']+doc_freq['doctorate'],
                     'Master': doc_freq['msc']+doc_freq['master'],
                     'Bachelor': doc_freq['bsc']+doc_freq['bachelor']+doc_freq['university']})

# Convert results into a dataframe
df_educ = pd.DataFrame(list(educ_dict.items()), columns = ['Educ_Level', 'NumJobs'])
# Percentage of job ads having a education requirement
df_educ['Percentage'] = (df_educ.NumJobs / df_educ.NumJobs.sum()) * 100
# sort data for plotting
df_educ.sort_values(by='Percentage', ascending=False, inplace=True)
pd.set_option('display.width', 1000)
######################################################################
```

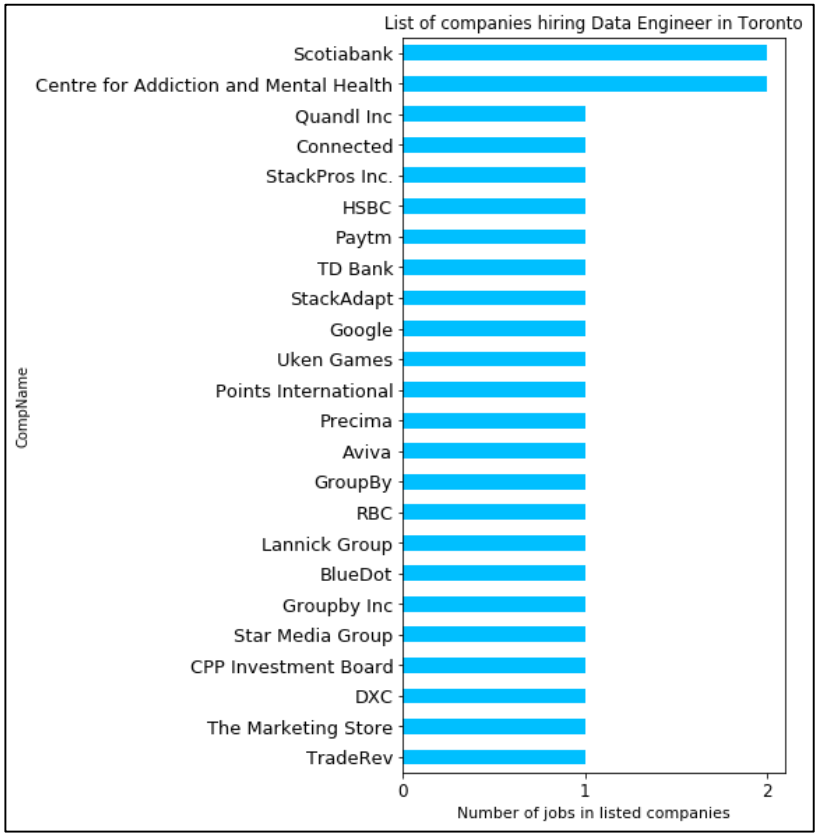
Exploratory Data Analysis

Hiring Companies: Data Scientist | Engineer

Data Scientist



Data Engineer

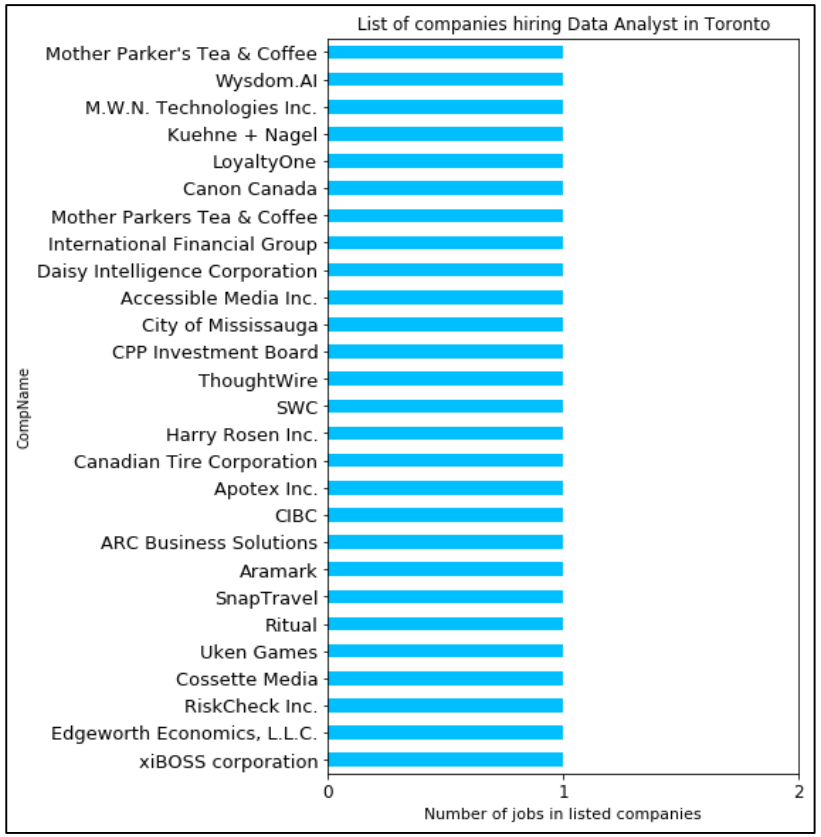


Companies hiring ML talents in Toronto

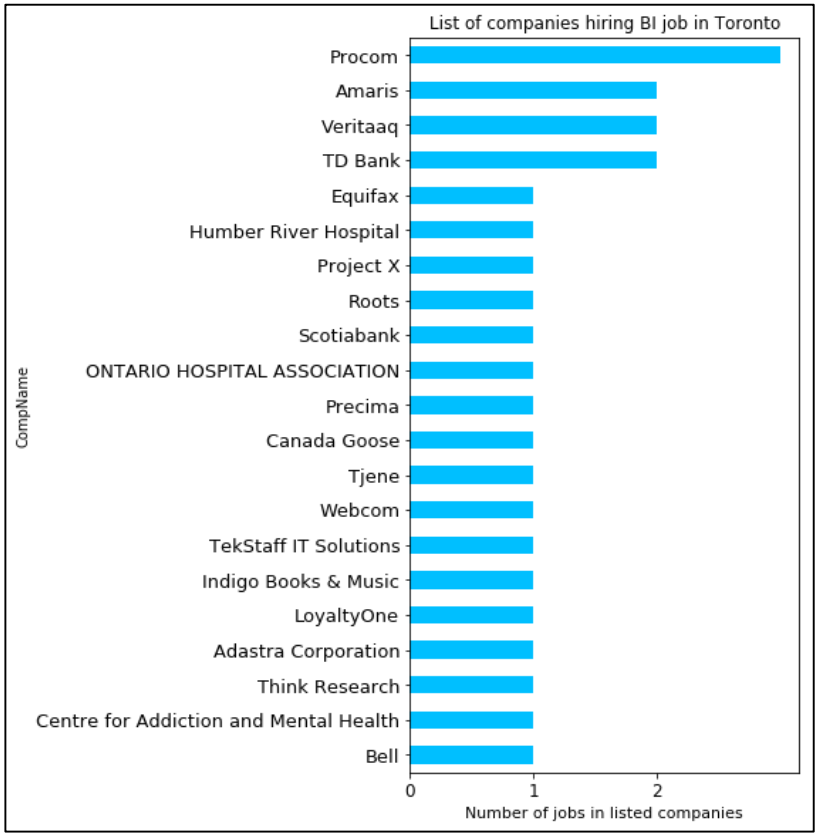
- Retail: Loblaw, Walmart, …
- Finance: TD Bank, HSBC, …
- Insurance: Sun Life, Aviva
- Start-Ups: LoyaltyOne, Precima
- Healthcare: CAMH

Hiring Companies: Data Analyst | BI Job

Data Analyst



Business Intelligence

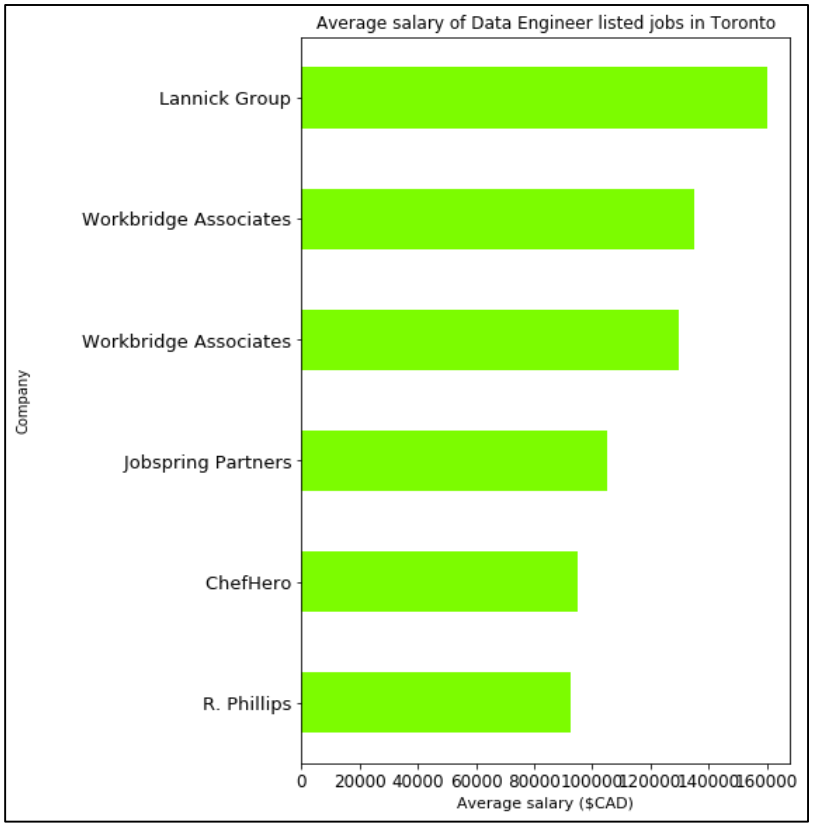


Companies hiring BI/analytics talents in Toronto

- Retail: Canada Goose, Harry Rosen, …
- Finance: TD Bank, HSBC, …
- Consulting Firm: Adastra, …
- Tech: Amazon, …
- Government: City of Mississauga, …

Salary: Data Scientist | Engineer

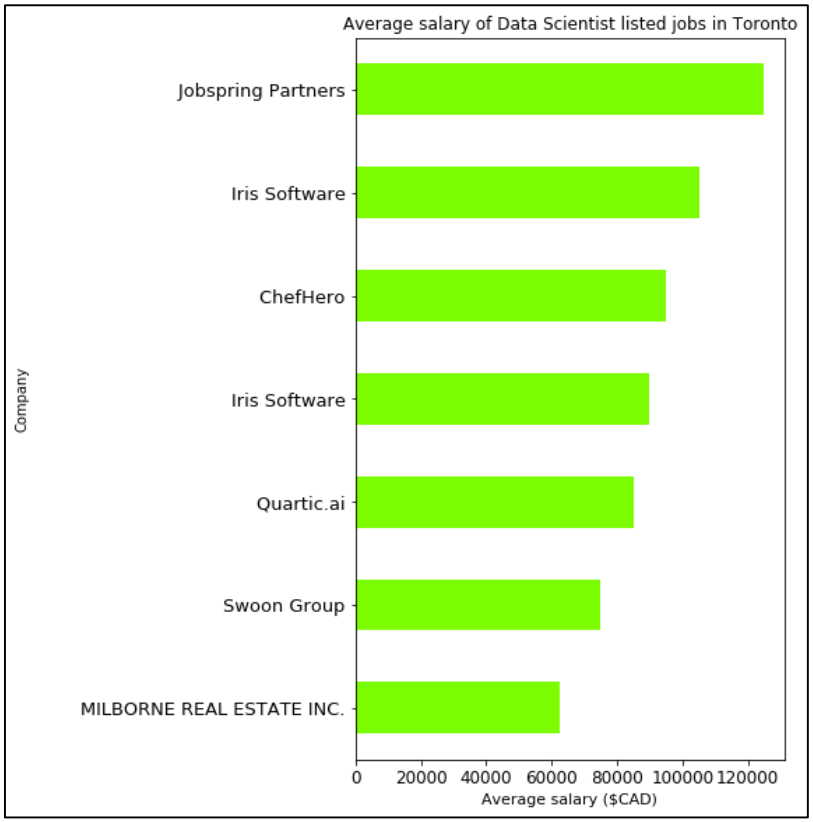
Data Engineer



Average salary

$ 119,500 per year

Data Scientist

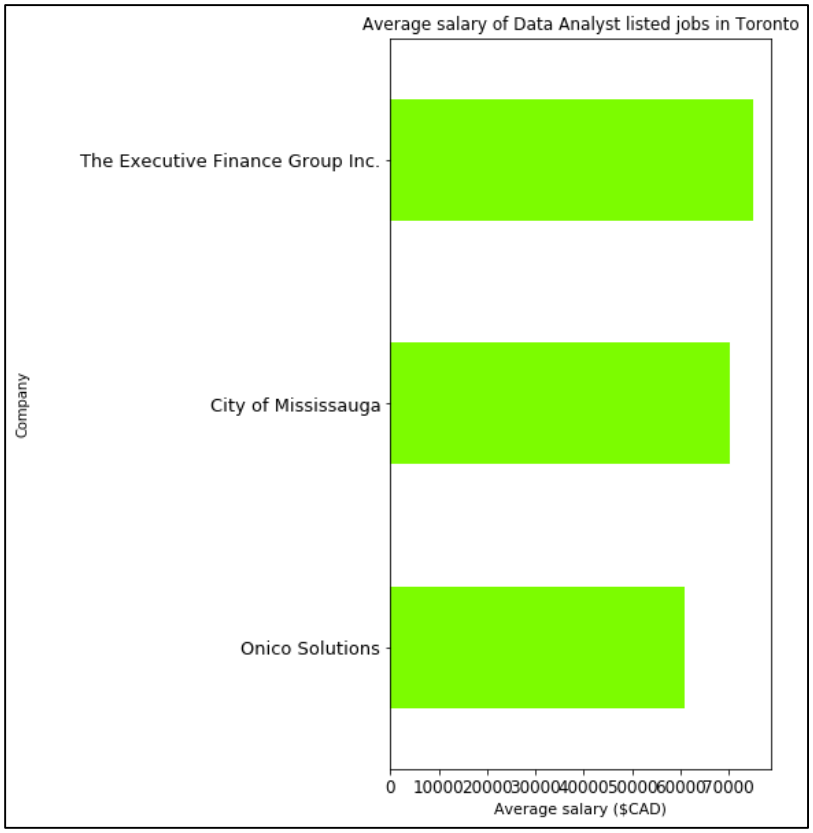


Average salary

$91,071 per year

Salary: Data Analyst | BI Jobs

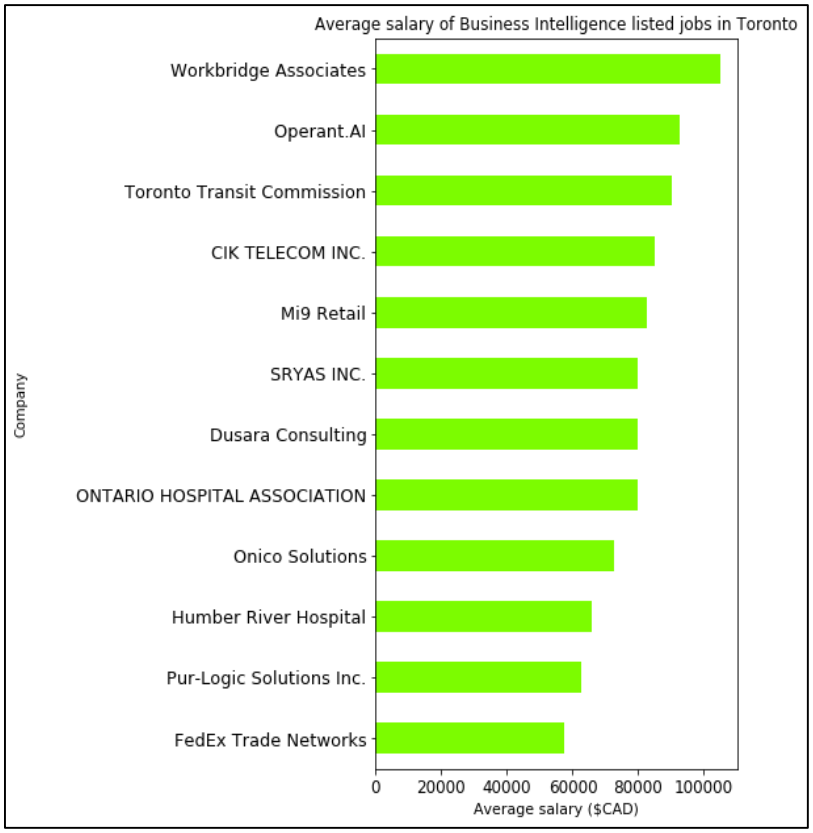
Business Intelligence



Average salary

$79,426 per year

Data Analyst



Average salary

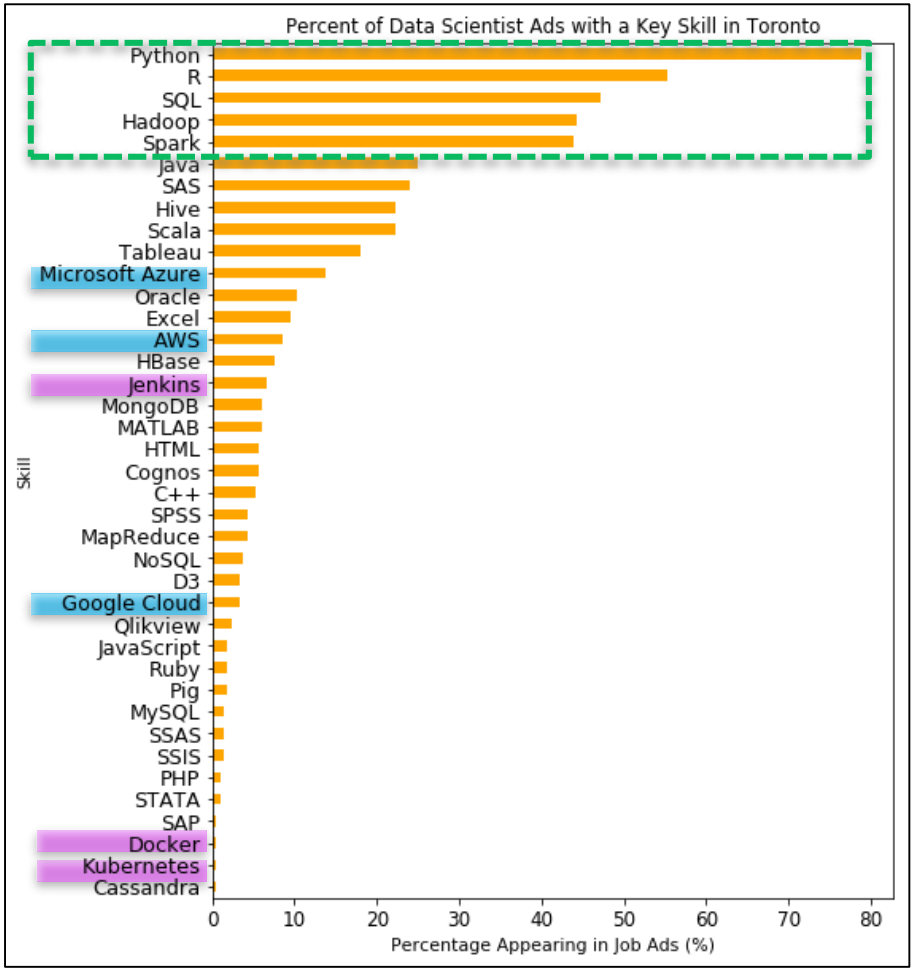
$68,722 per year

DataScienceOne
Keep It Simple & Stupid

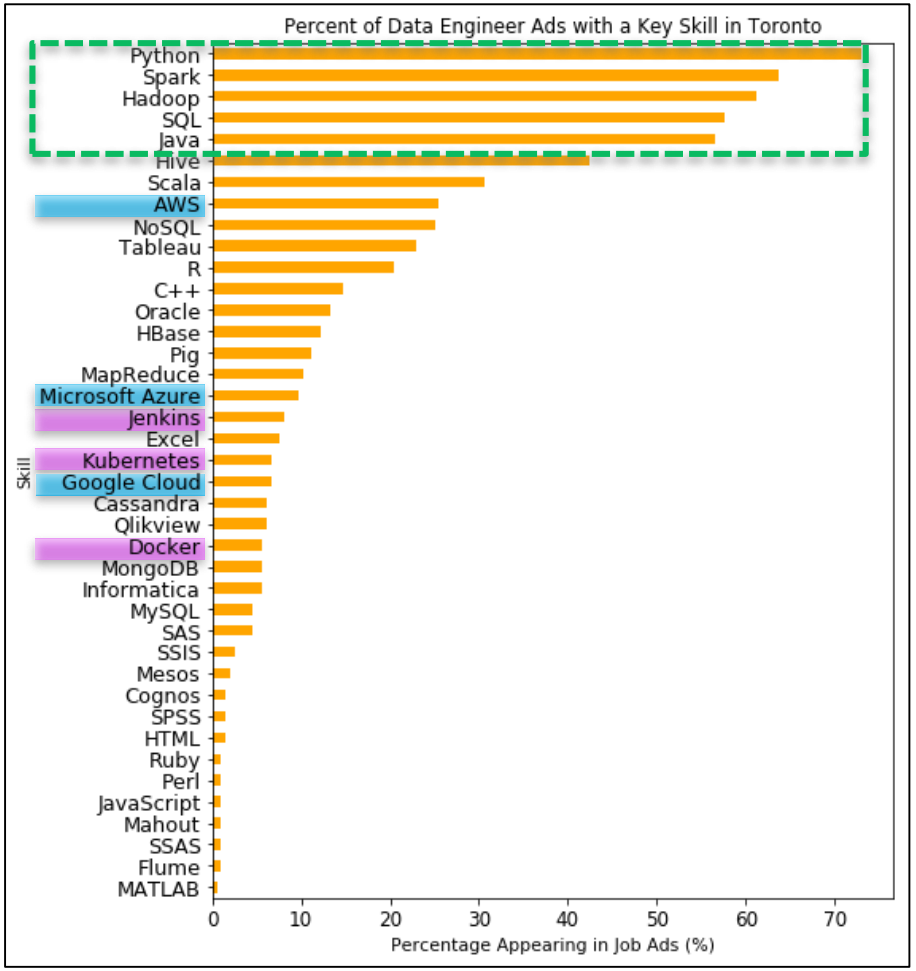
Skill Requirement: Data Scientist | Engineer

Top 5 Skills DevOps Clouds

Data Scientist



Data Engineer

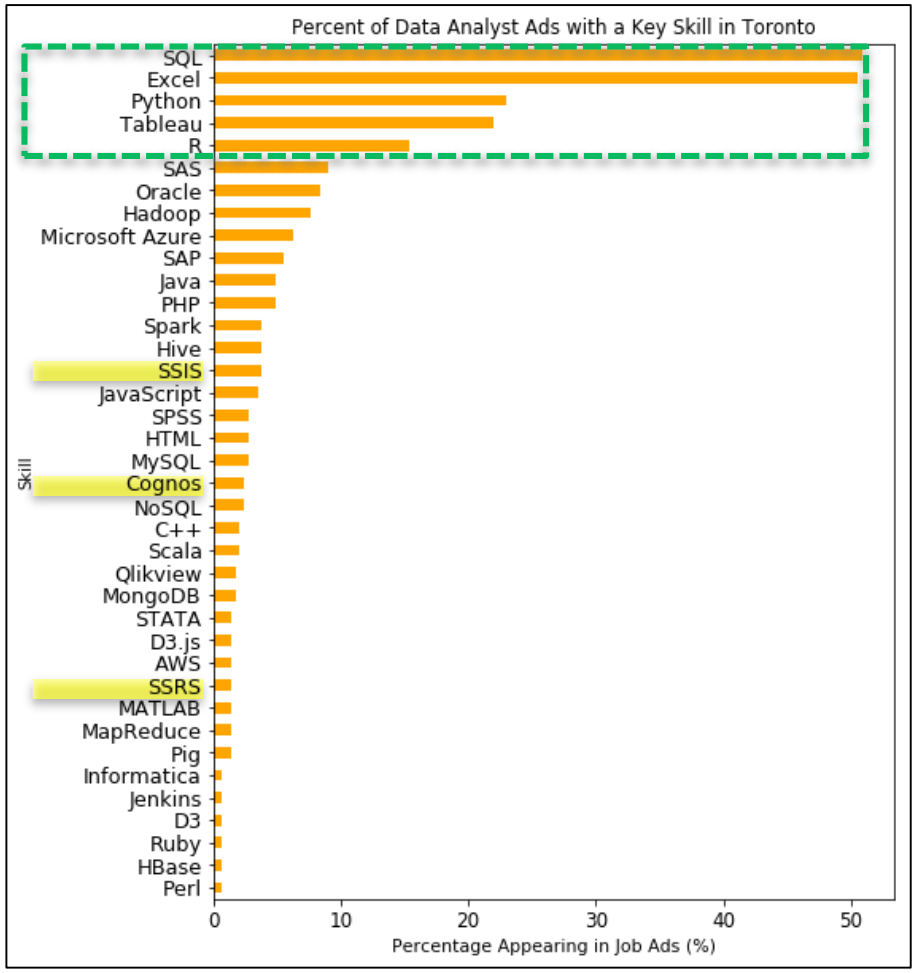


DataScienceOne
Keep It Simple & Stupid

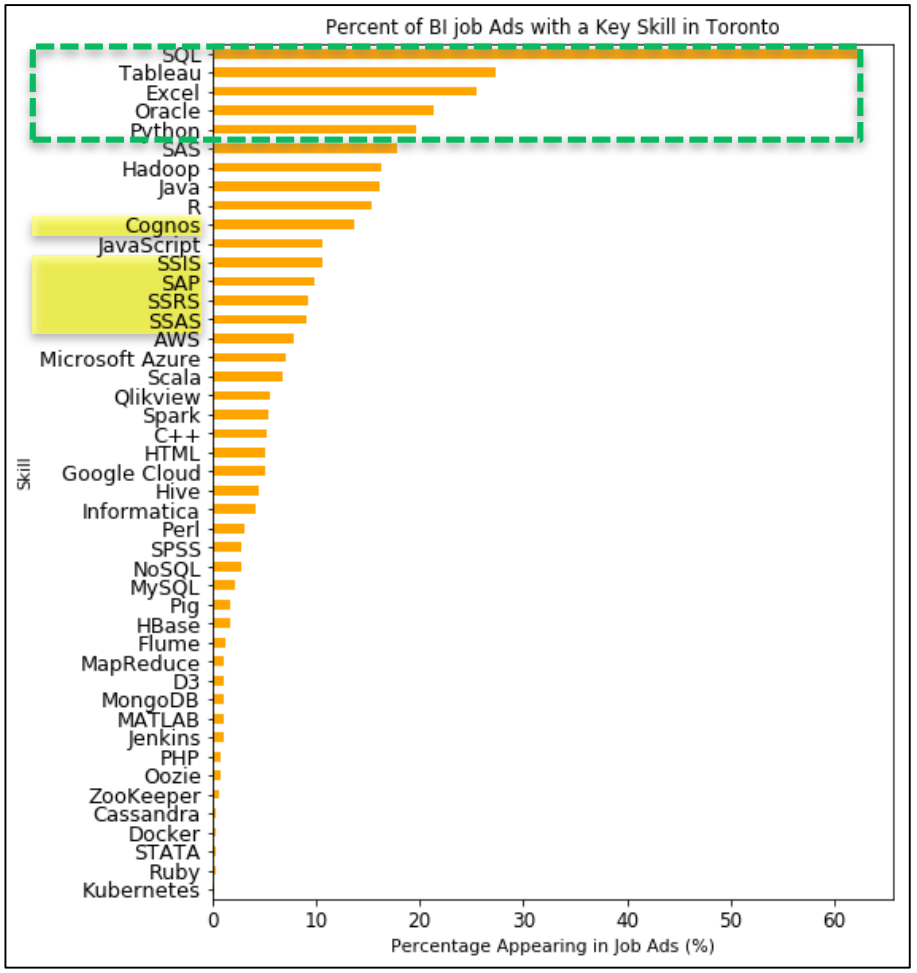
Skill Requirement: Data Analyst | BI Jobs

Top 5 Skills ETL

Data Analyst

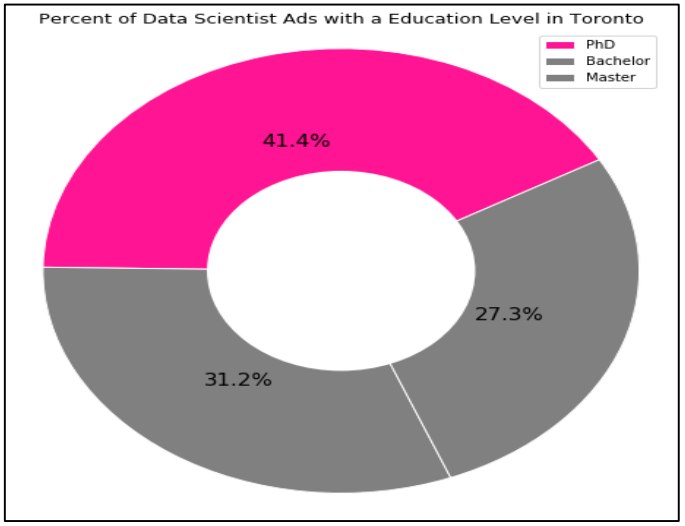


Business Intelligence

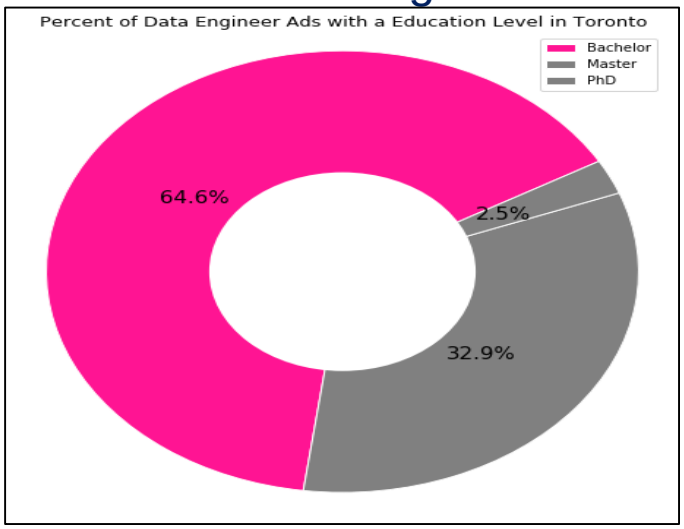


Education Requirement: Levels

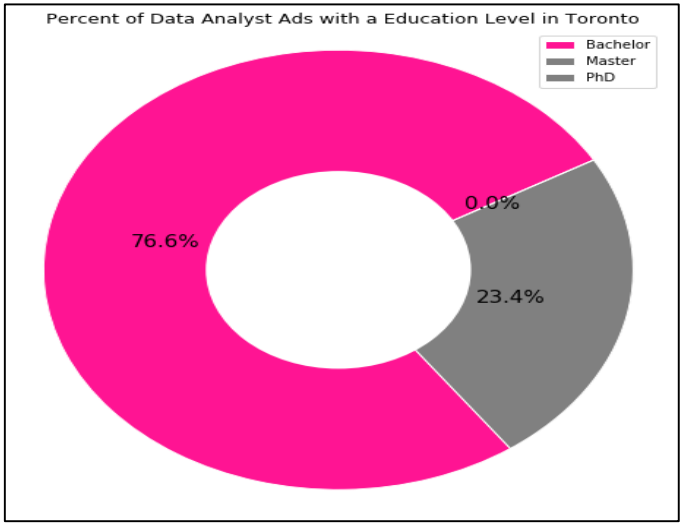
Data Scientist



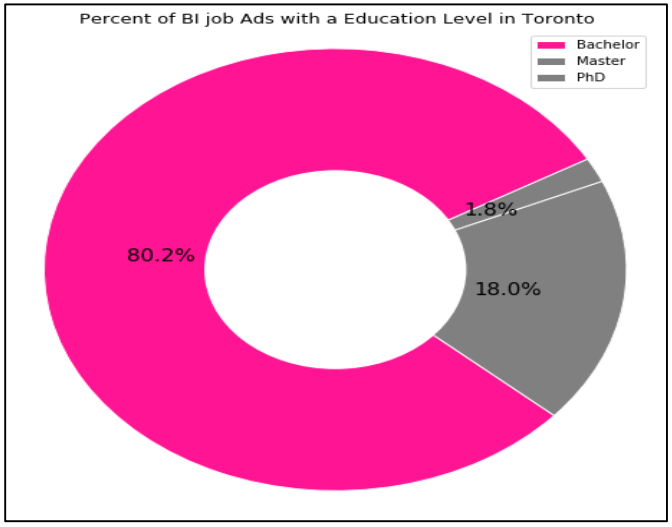
Data Engineer



Data Analyst

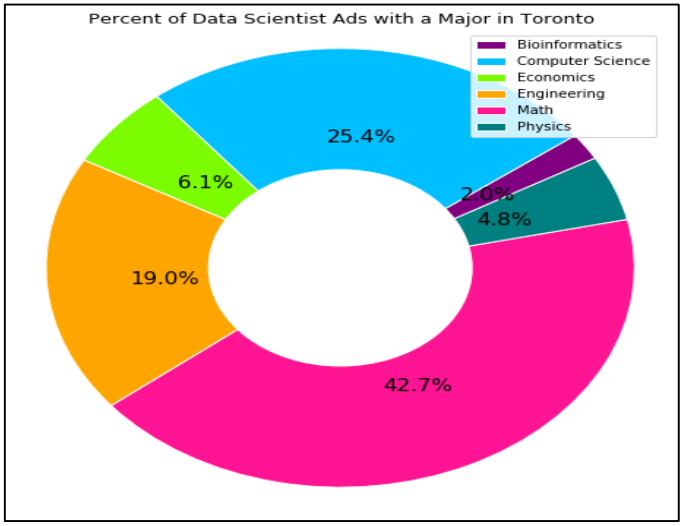


Business Intelligence

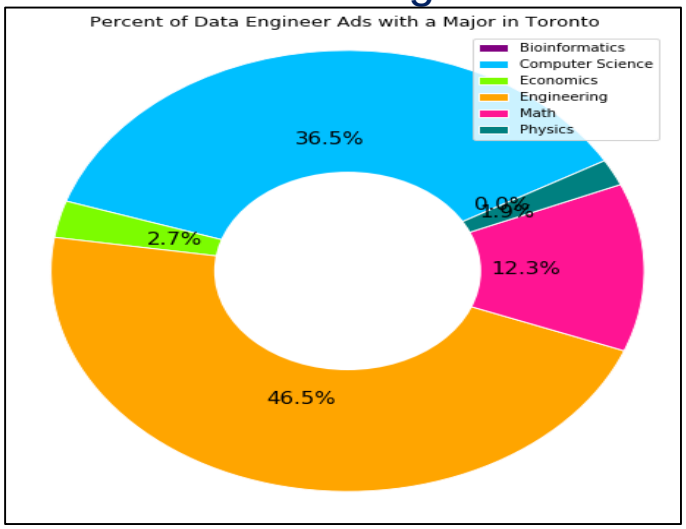


Education Requirement: Majors

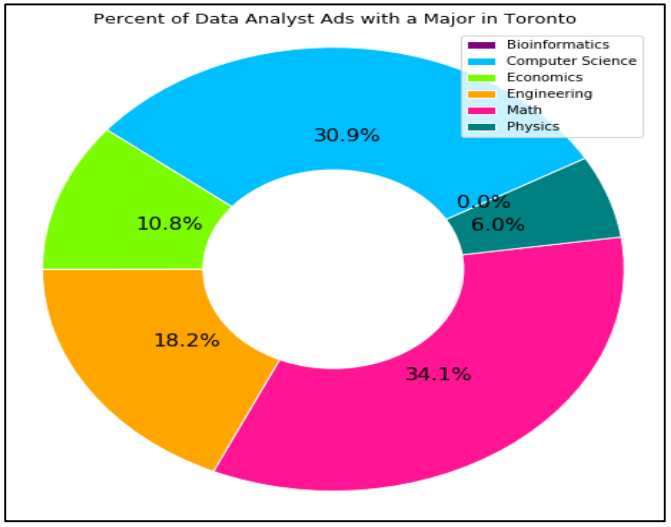
Data Scientist



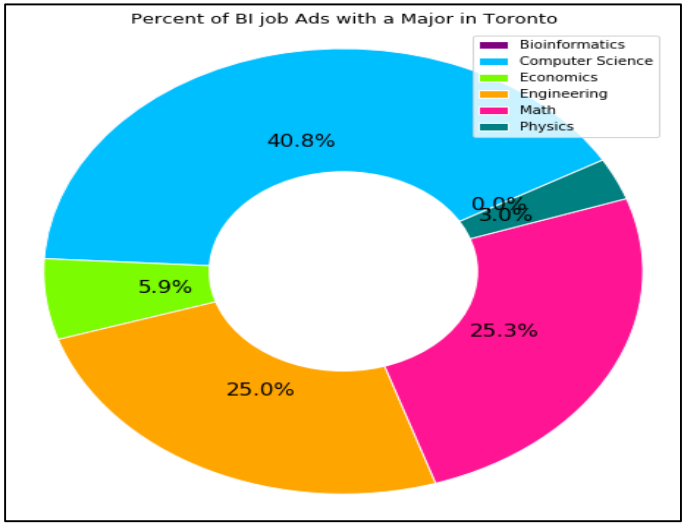
Data Engineer



Data Analyst



Business Intelligence



Summary: Indeed Job Scraper

Goal

- Provide overall data career market insights in Toronto

Key Insights

- Most of sectors hiring for ML/BI talents are from finance, insurance, retail, etc.
- Avg. annual salary: 1st rank - data engineer ($119K) vs. 4th rank - data analyst ($69K).
- Required skill set:
  - Data engineer focused more on DevOps and cloud experience
  - BI job focused more on ETL tools (i.e., Oracle, Microsoft BI stack)
- Required education level & majors:
  - Data scientist was only job with preferred qualification at PhD level
  - Most of data career jobs preferred majors were from S.T.E.M field

Future Work & Recommendations

Limitations & Future Work

Limitation:

- Limited # of job ads with salary info.
- Sample size is not at large scale.

Future Work:

- Data sources: consider web scraping other sites such as
  - MonsterJob, GlassDoor and LinkedinJob

- Data collection:
  - Company's review comments by employees
  - Company's roles and environment ratings by employees

- ML modeling:
  - Sentiment analysis on positive and negative comments
  - LDA topic modeling and word cloud by job categories
  - Data career salary regression model by job categories

Recommendations

Strategic job application submissions

- Understand education requirement before applying
- Learn required tools by a job (ranked importance)
- Compare average salary of your targeted companies

Design ML model portfolios

- Salary regression model
- Sentiment analysis on review comments
- Cluster analysis by job categories

Data collection

- Consider other data sources: Linkedin, Glassdoor, etc.
- Collection of data on company's review comments/ratings

Next Episode

Boston Housing Price Prediction Model with Regression