Data Analysis of Data Science Job Ads with suggestions

Liuhua Peng
Lecturer, School of Mathematics and Statistics
Thanks!

Fiona Simpson
Careers & Industry Consultant
Academic Engagement
Faculty of Science
The University of Melbourne

Peter Karutz
Senior Advisor Experiential Learning, Industry
Global Leadership and Employability, Academic Services
University Services
The University of Melbourne

Kathy Ryan
Academic Services and Registrar
University Services
The University of Melbourne

Copyright: burningglass®
An Increasing Market !!!!
## Trend in Advertised Occupations

<table>
<thead>
<tr>
<th>Occupation (Job Title)</th>
<th># Job Ads in 2016</th>
<th># Job Ads in 2017</th>
<th># Job Ads in Last 365 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science</td>
<td>616</td>
<td>1030</td>
<td>1165</td>
</tr>
<tr>
<td>Data Analyst</td>
<td>13</td>
<td>38</td>
<td>32</td>
</tr>
<tr>
<td>Software Programmers</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Database Administrator</td>
<td>5</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>University Lecturer</td>
<td>5</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Software Engineer</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Business Intelligence Analyst</td>
<td>2</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Business Manager</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Employers

Ernst and Young

Swinburne University of Technology
Monash University

General Electric Company

General Assembly

Woolworths

Accenture

Suncrop Group Limited

Amazon

Big Wednesday Digital

BHP Billiton Limited

Gold Coast City Council

Hatch Limited

Sas Institute

Teradata Australia

The University of Sydney

Australian Department of Defence

Science New Zealand Inc

The Australian Security Investments Commission

Queensland University of Technology

Bureau of Meteorology

Australian Taxation Office

Expedia Incorporated

Readify Pty Ltd

Tabcorp Holdings Limited
Degree and Experience

Levels of Education and Levels of Experience Requested

- Postgraduate Degree
  - No Experience Required: 9
  - 0 to 2 Years of Experience: 49
  - 3 to 5 Years of Experience: 4
  - 6 to 8 Years of Experience: 6

- Bachelors and Honours Degrees
  - No Experience Required: 51
  - 0 to 2 Years of Experience: 101
  - 3 to 5 Years of Experience: 12

Legend:
- No Experience Required
- 0 to 2 Years of Experience
- 3 to 5 Years of Experience
- 6 to 8 Years of Experience
- 9+ Years of Experience
Top Skills: Baseline and Specialized Skills

Baseline Skills

- Mathematics: 40.55%
- Communication Skills: 32.96%
- Research: 32.36%
- Problem Solving: 16.82%
- Organisational Skills: 16.31%
- Creativity: 15.19%
- Writing: 11.48%
- Presentation Skills: 10.27%

Specialized Skills

- Machine Learning: 62.38%
- Data Science: 61.95%
- Predictive Models: 34.86%
- Big Data: 29.68%
- Data Analysis: 22.35%
- Data Mining: 21.4%
- Statistics: 11.82%
- Optimization: 9.84%
- Neural Network: 8.89%
- Business Modeling: 8.8%
- Data Visualization: 8.71%

Percent of Jobs Requested
Top Skills: Software and Programming Skills

Percent of Job Ads with specific software and programming skills requested

- Python: 59.28\% (2016), 58.09\% (2017), 54.15\% (Last 365 Days)
- SQL: 55.09\% (2016), 43.58\% (2017), 43.38\% (Last 365 Days)
- R: 41.33\% (2016), 27.14\% (2017), 25.11\% (Last 365 Days)
- SAS: 38.05\% (2016), 22.26\% (2017), 23.82\% (Last 365 Days)
- Apache Hadoop: 27.14\% (2016), 22.26\% (2017), 23.82\% (Last 365 Days)
- Tableau: 19.35\% (2016), 14.8\% (2017), 14.63\% (Last 365 Days)
- JAVA: 19.35\% (2016), 14.8\% (2017), 14.63\% (Last 365 Days)
- MATLAB: 17.56\% (2016), 14.67\% (2017), 11.86\% (Last 365 Days)
- NoSQL: 17.56\% (2016), 14.67\% (2017), 11.86\% (Last 365 Days)
- C++: 17.56\% (2016), 14.67\% (2017), 11.86\% (Last 365 Days)
- Microsoft Excel: 17.56\% (2016), 14.67\% (2017), 11.86\% (Last 365 Days)
- SPSS: 17.56\% (2016), 14.67\% (2017), 11.86\% (Last 365 Days)
- Teradata: 17.56\% (2016), 14.67\% (2017), 11.86\% (Last 365 Days)
Salaries

2016

Salary Range

Mean salary: $110,000
Median salary: $100,000

2017

Salary Range

Mean salary: $109,000
Median salary: $100,000

Last 365 Days

Salary Range

Mean salary: $112,000
Median salary: $108,000
Salaries ! ! ! ! ! ! ! !

2016

Salary Range

- $35,000 to $49,999: 16, 19%
- $50,000 to $74,999: 28, 33%
- $75,000 to $99,999: 31, 37%
- $100,000 to $149,999: 9, 11%
- More than $150,000: 0%

Mean salary: $110,000
Median salary: $100,000

2017

Salary Range

- $35,000 to $49,999: 57, 37%
- $50,000 to $74,999: 47, 31%
- $75,000 to $99,999: 18, 12%
- $100,000 to $149,999: 27, 18%
- More than $150,000: 2%

Mean salary: $109,000
Median salary: $100,000

Last 365 Days

Salary Range

- $35,000 to $49,999: 84, 44%
- $50,000 to $74,999: 36, 19%
- $75,000 to $99,999: 20, 10%
- $100,000 to $149,999: 50, 26%
- More than $150,000: 1%

Mean salary: $112,000
Median salary: $108,000
The Data Science Process

Useful Resources

2. Coursera  [https://www.coursera.org/](https://www.coursera.org/)
3. Amazon Web Services  [https://aws.amazon.com/](https://aws.amazon.com/)

The Elements of Statistical Learning

An Introduction to Statistical Learning with Applications in R
[http://www-bcf.usc.edu/~gareth/ISL/](http://www-bcf.usc.edu/~gareth/ISL/)
Top Python Data Analysis Libraries

1. Numpy  http://www.numpy.org/
2. Scipy   https://www.scipy.org/
3. Pandas  https://pandas.pydata.org/
4. Matplotlib https://matplotlib.org/
5. Seaborn https://seaborn.pydata.org/
7. Scikit-learn http://scikit-learn.org/
8. Theano  http://deeplearning.net/software/theano/
10. BeautifulSoup4, Urlib2, Selenium, Scrapy
1. **Anaconda** is a free and open source distribution of the **Python** and **R** for data science.

2. It aims to simplify package management and deployment.

3. Use **Jupyter Notebook** as IDE

4. Anaconda Distribution is used by over 6 million users, and it includes more than 250 popular data science packages suitable for Windows, Linux, and MacOS.

5. [https://www.anaconda.com/download/](https://www.anaconda.com/download/)
High Performance Computing in R

1. **Rcpp, RcppArmadillo, RcppEigen**: integrate R with C++

2. Parallel Computing in R:
   - ✓ *parallel, foreach*: execute the for loop in parallel
   - ✓ *SNOW, doSNOW*
   - ✓ *H2o*: facilitates machine learning (e.g. random forests, GBM) in a parallel environment
   - ✓ *Rhadoop, RHIPE*: interface between R and Hadoop
   - ✓ *Rmpi, pbdMPI*: interface MPI in R
   - ✓ *rgpu, gcbd, OpenCl*: GPU programming
   - ✓ *data.table, ff, bigmemory*: large memory and out-of-memory data

https://cran.r-project.org/web/views/HighPerformanceComputing.html
Useful Packages in R/Rstudio

1. tidyverse  [https://www.tidyverse.org/](https://www.tidyverse.org/)
2. rvest
3. dplyr, tidyr, stringr, purrr
4. lubridate
5. ggplot2, maptools, rgdal
6. leaflet
7. caret, e1071, kernlab, nnet, rpart, xgboost, tensorflow
   [https://cran.r-project.org/web/views/MachineLearning.html](https://cran.r-project.org/web/views/MachineLearning.html)
9. knitr
Thank you

Questions?

___

Liuhua Peng
Lecturer, School of Mathematics and Statistics