DATA SCIENCE COMPETENCY CHECKLISTS 2017

Reviewed by Malaysian Employers

Produced by:

DataViz My
Your Data Education

In collaboration with:

MDEC®

BIG DATA Malaysia

CONTENTS

Enquires: Dr Sandra Hanchard: sandra@dataviz.my

OBJECTIVES

Investing in data science capabilities enables businesses to gain a competitive advantage by deriving value from big data. While the industry is well-served with commoditised and open-source tools, there remains a large gap in the available candidates with the necessary skill sets to implement data science initiatives. The lack of precision in defining a data scientist, and supporting roles, is one hindrance to sourcing and developing talent that is in demand by Malaysian employers.

The objective of this report is to help job-seekers, employers and educators identify requirements for data science teams in Malaysia. We have focused on 3 core roles: data scientist, data engineer and data analyst; representing natural areas of focus and division of labour. Data scientists have specialised skills in developing models; engineers build platforms to overcome big data constraints; while analysts are responsible for communicating actionable insights.

We offer a checklist of core and desired skills and tools grouped into functional areas for each role. These checklists have been validated by leading Malaysian employers. We have supplemented these checklists with insights on preferred qualifications, industry experience, salary expectations and capacity. Job-seekers can use the checklists to evaluate their strengths to develop deeper knowledge for the role that specifically interests them. Employers can apply the checklists for building job descriptions and benchmarking capabilities across their teams. Educators may also find the checklists helpful for developing industry-relevant curriculum.

OVERVIEW

Data-centric industries including business services, information technology, banking and telecommunications are leading the way with data science recruitment in Malaysia. Common business applications with predictive analytics include personalised marketing, managing the customer lifetime value, identifying risk and forecasting supply. Deep learning and natural language understanding technologies are topical; while communications, geospatial and biomedical data are of high interest.

Job-seekers for data science roles require baseline higher education, holding a Bachelor's Degree at a minimum. Employers however value short courses and MOOCs in resumes as they reflect active lifelong learning and commitment. While a PhD is not a prerequisite to becoming a data scientist, advanced education is valued. Soft skills such as critical and creative thinking are sought after. Finding an individual that is strong in all the competencies for a data scientist is very rare: the formation of teams with complementary skill sets can address this challenge. Ultimately, employers desire team members who will add value to the 'bottom line' of a business through delivering actionable insights.

Employers indicated expansion of hiring for data science teams, although recruitment is still at an early stage in Malaysia. The size of data science teams, requirements and salary expectations varies with the maturity and size of organization and industry. Many employers are building their teams from 'scratch', accepting candidates with entry-level industry experience. The large salary range reported by employers for a data scientist from more than RM 15,000 per month to less than RM 5,000 per month may reflect a lack of precision in defining the role of a data scientist. While compensation will differ by candidate experience and performance, it is important that this role is not undervalued. Data science competencies such as statistical modelling and machine learning require a high education investment which should be recognised.

REVIEWERS

There are 31 companies represented in our pool of 35 reviewers. All reviewers, who are based in Malaysia, are senior subject matter experts with 28 out of 35 directly responsible for hiring. The primary location for hiring was Kuala Lumpur and Selangor, with one hirer in Penang and two in Singapore. The industries represented by our reviewers, in descending order, included:

- Business Services
- Information Technology
- Fintech
- Banking
- Human Resources
- Marketing Services
- Telecommunications
- Agribusiness
- Aviation
- Digital Marketing
- Insurance
- Lifestyle
- Manufacturing
- Oil and Gas
- Property
- Transportation

Indicative job titles of reviewers included: CEO, Managing Director, CIO, CTO, Director, Analytics Head, Engineering Head, Consultant (Principal); Data Scientist, Data Engineer, Data Analyst and Data Modeller (Senior and Intermediate).

ACKNOWLEDGEMENTS

We are extremely grateful for the time and expertise shared by our reviewers. The following companies have given permission to publish their logo:

KEY ROLES

The 3 key roles identified for data science teams represent natural areas of focus and division of labour:

DATA SCIENTIST

Use analytical techniques combined with data skills to develop scalable and robust analytical models

DATA ENGINEER

Design and develop high-performance infrastructure and tools to enable users to consume and understand data more effectively

DATA ANALYST

Communicate insights that deliver business value based on exploratory analysis

The most common job titles used by our reviewers were data analyst, data scientist and data engineer. The amount of demand for data analysts in the job market should not be under-estimated. Data analysts may play an important role in deriving value from data assets while companies build out their predictive capabilities.

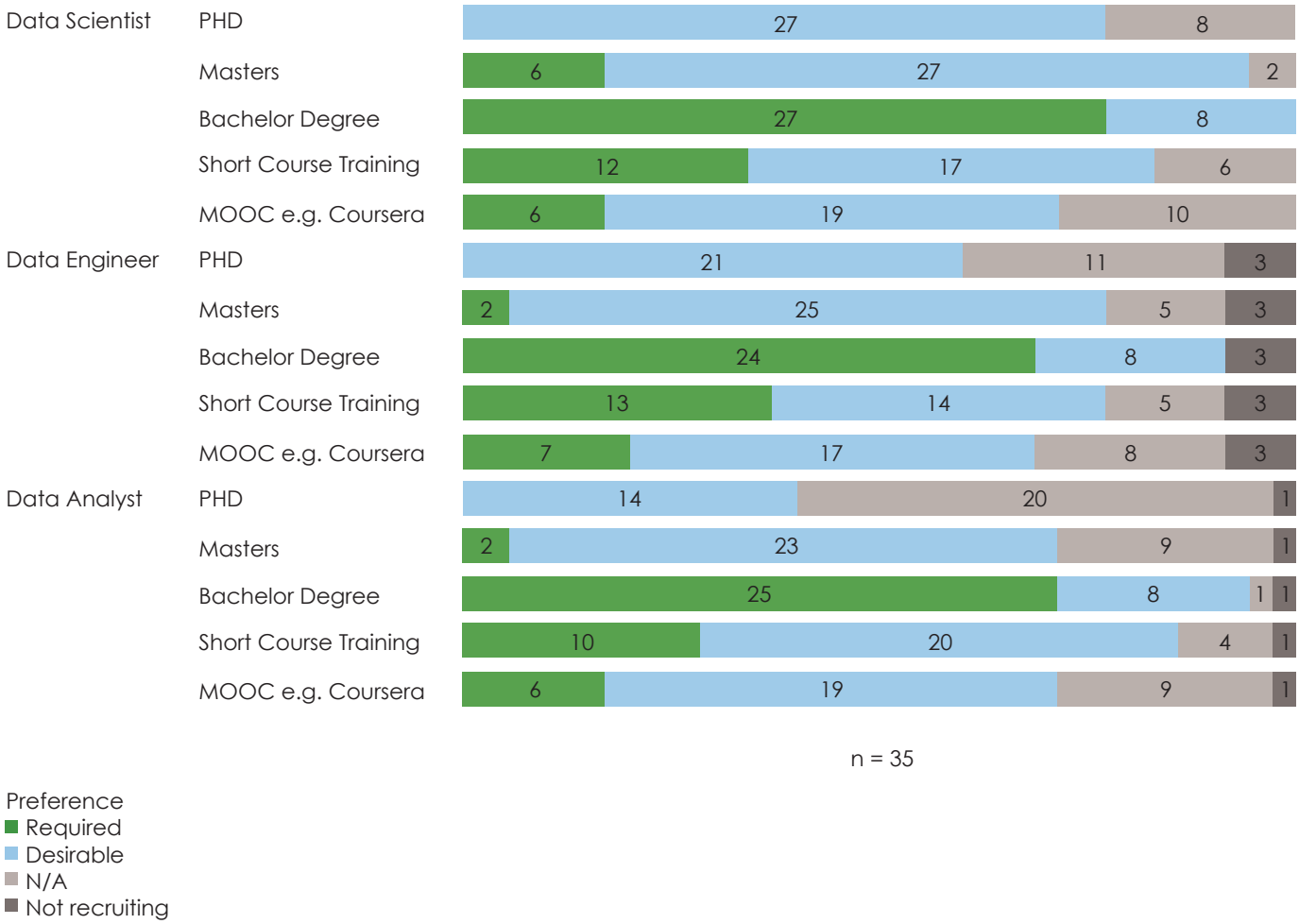
While some employers are still recruiting for data modellers and data miners, we anticipate that these titles will be eventually encompassed and replaced by 'data scientist'; skills such as programming and data wrangling are likely to be enhanced by these professionals. Other titles that employers used included data architect, data visualization engineer, software research engineer, research analyst, researcher, statistician and actuary. A data science team may also be supported by other roles such as chief data analytics officer, database administrator and data governance officer. It should be noted that several job titles may refer to a similar role. For example, a data engineer may be called a software engineer or big data engineer.

RECRUITMENT INSIGHTS

Qualifications

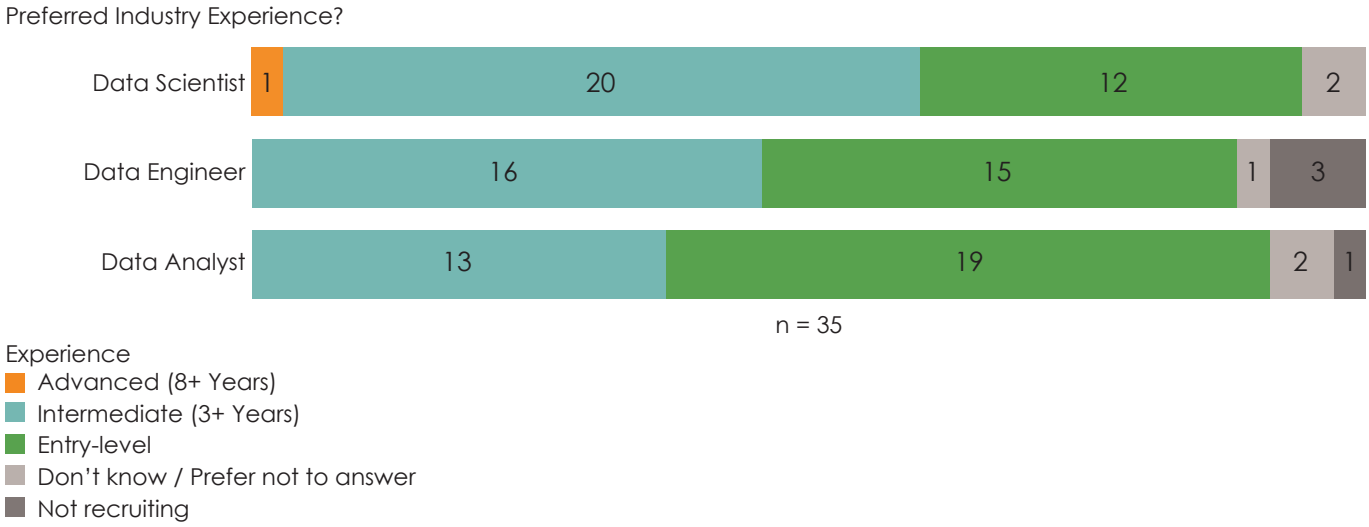
A Bachelor Degree was a minimum requirement by most employers for each role. Notably, some employers also required short course training and MOOCs, highlighting the importance and value of extra-curricular learning activities. While a PhD was not regarded as mandatory for a Data scientist role, some employers wanted at least a Master's degree, indicating the advanced knowledge desired.

Preferred Qualifications?



Industry Experience

Many employers were willing to recruit data scientists with entry-level industry experience. These employers are building their teams from the ground up, a reflection that the data science recruitment landscape in Malaysia is still at an early stage. Strong education backgrounds of candidates would be vital for building these types of teams.



Salary Expectations

While the highest bracket a data scientist could expect was more than RM15,000 per month, the highest bracket for data engineers and data analysts was RM8,001 – RM15,000. The skills that differentiate data scientists, namely, statistics and modelling and machine learning competencies, can add high value to a graduate's earning potential.

Some employers offered less than RM5,000 per month for a data scientist. The large range in salaries could be due to the maturity and size of organizations. Employers are also likely to be compensating based on experience and past performance of candidates. However, some organizations could be using 'data scientist' as a catch-all job title for roles with lower skill requirements, thereby undervaluing the role.

Capacity

Most employers reported increased capacity for data science teams for 2017 – 2020. While many employers indicated the current and future capacity for each type of role was less than 5 people, some organizations said they were willing to recruit up to 30 data scientists between 2017 – 2020. Employers should be careful to distinguish the functionality between data science roles to ensure successful recruitment.

CORE FUNCTION AREAS

The functions of a data science team can be divided into 7 core areas:

Business analysis, approach and management

Problem framing, experimentation and benefits realization

Insights, storytelling and data visualization

Communication of insights that can guide actionable results

Programming

Data sourcing, integration, scaling and optimization

Data wrangling and database concepts

Ensure provenance and hygiene of data for trusted insights

Statistics and data modelling

Identifying factors and relationships to make predictions for competitive advantage

Machine learning algorithms

Classification and prediction for intuitive results at scale

Big data

Remove constraints on volume, variety and velocity of data

Baseline Competencies for Roles

The table below highlights the core function areas for each role, with varying degrees of competency expected across roles. It should be noted that in some teams, one hire may be responsible for competencies beyond each role scope, depending on the budget and size of the team.

	DATA SCIENTIST	DATA ENGINEER	DATA ANALYST
Business analysis, approach and management	✓	✓	✓
Insights, storytelling and data visualization	✓		✓
Programming	✓	✓	✓
Data wrangling and database concepts	✓	✓	✓
Statistics and data modelling	✓		✓
Machine learning algorithms	✓		
Big Data	✓	✓	

DATA SCIENTIST CHECKLIST

CORE FUNCTION

Use analytical techniques combined with data skills to develop scalable and robust analytical models

BUSINESS ANALYSIS, APPROACH AND MANAGEMENT

CORE SKILLS

DESIRED SKILLS

- Demonstrate curiosity and an analytical mindset
- Forming and testing a hypothesis
- Use subject matter knowledge to ask the right questions and identify areas of business impact
- Design an experimental project and apply the workflow of data science
- Curate data relevant to business outcomes
- Project manage a data science project to successful completion
- Present a data science project to stakeholders
- Break new ground in terms of research and new algorithm development

CORE TOOLS

DESIRED TOOLS

- Sharpies and paper
- Presentation e.g. Excel, Tableau, Powerpoint, Keynote, Prezi
- Project Mangements e.g. Trello, Jira

INSIGHTS, STORYTELLING AND DATA VISUALIZATION

CORE SKILLS

DESIRED SKILLS

- Explore and analyse data for insights using visualization
- Explain insights with clear and compelling written, verbal and visual communication
- Curate visualizations that assist stakeholder understanding
- Present storybooks using dashboards
- Illustrate entity relationships graphically

CORE TOOLS

DESIRED TOOLS

- Analytics e.g. PowerBI, Qlik, SAS, Microstrategy, Tableau
- Notebooks e.g. Jupyter, Zeppellin, Databricks
- Data Visualization e.g. D3.js, Matplotlib, ggplot
- Entity Relationship Diagramming

PROGRAMMING

CORE SKILLS

DESIRED SKILLS

- Apply the fundamentals of languages for effective analysis
- Write code to read data, access packages, apply logic
- Debugging, profiling and optimization
- Organizing and commenting code
- Map data sources to data visualization libraries
- Write reusable libraries
- Working with GPUs

CORE TOOLS

DESIRED TOOLS

- Scripting and environments e.g. Python, R, Matlab, Julia
- IDEs, Interpreters and Notebooks
- Other languages e.g. C/C++, Java, Scala, Go
- Source control e.g. git, Mercurial, Github, Bitbucket

DATA WRANGLING AND DATABASE CONCEPTS

CORE SKILLS

DESIRED SKILLS

- Clean data through statistical approaches, such as identifying outliers
- Transform data into machine-readable formats
- Create, read, update and delete on databases and apply data normalization
- Extract, Transform and Load data
- Collect data from variety of sources including APIs and web
- Optimize performance of database queries/schema (basics such as adding indices etc.)
- Processing unstructured data (text, voice, image)

CORE TOOLS

DESIRED TOOLS

- ETL and Data Integration e.g. Talend, Kettle, Alteryx
- Data cleaning e.g. Google Refine, Monarch, Orange, KNIME
- Databases and data stores e.g. MySQL, Postgres, Couch, MongoDB, Neo4j
- Scraping e.g. Scrapy, Apache Tika, BeautifulSoup

STATISTICS AND DATA MODELLING

CORE SKILLS

DESIRED SKILLS

- Build, monitor and measure quality of models over time
- Deep knowledge of statistical and mathematical concepts
- Identify trends and behaviours with descriptive statistics
- Apply inferential statistics on randomised samples to understand characteristics of a population
- Apply common models such as linear and logistic regression for prediction
- Run A/B tests

CORE TOOLS

DESIRED TOOLS

- R, R Studio, Pandas, SAS, SPSS, Excel
- A/B testing e.g. Adobe Target, Convert, Optimizely, Maxymizer, Unbounce, VWO
- Data Mining e.g. Weka, RapidMiner, KNIME, KXEN, Rattle

MACHINE LEARNING ALGORITHMS

CORE SKILLS

DESIRED SKILLS

- Select appropriate algorithms for prediction
- Test and assess goodness of fit
- Apply clustering and classification algorithms e.g. K-Means, Decision Tree, Naïve Bayes
- Apply Natural Language Processing algorithms
- Apply neural networks and support vector machines
- Apply dimensionality reduction techniques

CORE TOOLS

DESIRED TOOLS

- Scikit learn, Tensorflow
- Amazon Machine Learning, Azure, and IBM Watson

CORE SKILLS

DESIRED SKILLS

- Recommend appropriate searching and indexing methods
- Use big data tools and platforms to access data and run models
- Manage the deployment of models lifecycle
- Develop technical documentation for overall architecture
- Scale data science projects with architectural components
- Architect high performance frameworks to process variety of data
- Productionize code structure by data scientist
- Make data easily accessible for analysis

CORE TOOLS

DESIRED TOOLS

- Data storage and processing platforms e.g. MongoDB, Hadoop and Spark ecosystems
- Cloud e.g. AWS, Azure, Google Cloud
- Search e.g. Solr, ELK
- Web Services e.g. Google Big Query

SOFT SKILLS

- Agile Problem-Solving
- Business acumen
- Creative
- Critical thinking
- Data Inquisitive
- Data Intuition
- Self-starting Motivation
- Strong Communication
- Team building

RELATED JOB ROLES

- Data Modeller
- Junior Data Scientist
- Quantitative Analyst
- Statistical Analyst
- Statistician

CORE FUNCTION

Design and develop high-performance infrastructure and tools to enable users to consume and understand d ata more effectively

BUSINESS ANALYSIS, APPROACH AND MANAGEMENT

CORE SKILLS

DESIRED SKILLS

- Demonstrate curiosity and an analytical mindset
- Work with stakeholders to identify data sources
- Communicate solution to team members
- Use subject matter knowledge to ask the right questions and identify areas of business impact
- Manage expectations through effective project management

CORE TOOLS

DESIRED TOOLS

- Sharpies and paper
- Project Mangements e.g. Trello, Jira

PROGRAMMING

CORE SKILLS

DESIRED SKILLS

- Apply the fundamentals of languages for effective analysis
- Write code to read data, access packages, apply logic
- Debugging, profiling and optimization
- Organizing and commenting code
- Write reusable libraries
- Package software and / or models for deployment
- Map data sources to data visualization libraries
- Working with GPUs

CORE TOOLS

DESIRED TOOLS

- Scripting and environments e.g. Python, R, Matlab, Julia
- IDEs, Interpreters and Notebooks
- Other languages e.g. C/C++, Java, Scala, Go
- Source control e.g. git, Mercurial, Github, Bitbucket
- Build e.g. Jenkins, TravisCI, CircleCI, Team City, Bamboo

DATA WRANGLING AND DATABASE CONCEPTS

CORE SKILLS

DESIRED SKILLS

- Manage database schema
- Transform data into machine-readable formats
- Create, read, update and delete on databases and apply data normalization
- Extract, Transform and Load data
- Optimize performance of database queries/schema
- Collect data from variety of sources including APIs and web
- Clean data through statistical approaches, such as identifying outliers
- Develop software for adding data collection points or integrating to sensors
- Use data pipelining to channel streams of data from a variety of sources

CORE TOOLS

DESIRED TOOLS

- ETL e.g. Talend, Kettle, Alteryx
- Data cleaning e.g. Google Refine, Monarch, Orange, KNIME
- Databases and data stores e.g. MySQL, Postgres, Couch, MongoDB, Neo4j
- Scraping e.g. Scrapy, Apache Tika, BeautifulSoup
- Data pipeling e.g. Apace Nifi, Kafka, Amazon Kinesis

BIG DATA

CORE SKILLS

DESIRED SKILLS

- Scale data science projects with architectural components
- Architect high performance frameworks to process variety of data
- Productionize code structure by data scientist
- Make data easily accessible for analysis
- Manage the deployment of models lifecycle
- Develop technical documentation for overall architecture

CORE TOOLS

DESIRED TOOLS

- Data storage and processing platforms e.g. MongoDB, Hadoop and Spark ecosystems
- Cloud e.g. AWS, Azure, Google Cloud
- Search e.g. Solr, ELK
- Web Services e.g. Google Big Query

SOFT SKILLS

- Agile Problem-Solving
- Business acumen
- Creative
- Critical thinking
- Data Inquisitive
- Data Intuition
- Self-starting Motivation
- Strong Communication
- Team building

RELATED JOB ROLES

- Big Data Solution Architect
- Big Data Engineer
- Data Architect
- Systems Engineer

DATA ANALYST CHECKLIST

CORE FUNCTION

Communicate insights that deliver business value based on exploratory analysis

BUSINESS ANALYSIS, APPROACH AND MANAGEMENT

CORE SKILLS

DESIRED SKILLS

- Demonstrate curiosity and an analytical mindset
- Forming and testing a hypothesis
- Use subject matter knowledge to ask the right questions and identify areas of business impact
- Design an experimental project and apply the workflow of data science
- Curate data relevant to business outcomes
- Project manage a data science project to successful completion
- Present a data science project to stakeholders

CORE TOOLS

DESIRED TOOLS

- Sharpies and paper
- Presentation e.g. Excel, Tableau, Powerpoint, Keynote, Prezi
- Project Mangements e.g. Trello, Jira

INSIGHTS, STORYTELLING AND DATA VISUALIZATION

CORE SKILLS

DESIRED SKILLS

- Explore and analyse data for insights using visualization
- Explain insights with clear and compelling written, verbal and visual communication
- Curate visualizations that assist stakeholder understanding
- Present storybooks using dashboards
- Illustrate entity relationships graphically

CORE TOOLS

DESIRED TOOLS

- Analytics e.g. PowerBI, Qlik, Microstrategy, SAS, Tableau
- Notebooks e.g. Jupyter, Zeppellin, Databricks
- Data Visualization e.g. D3.js, Matplotlib, ggplot
- Entity Relationship Diagramming

PROGRAMMING

CORE SKILLS

DESIRED SKILLS

- Apply the fundamentals of languages for effective analysis
- Write code to read data, access packages, apply logic
- Debugging, profiling and optimization
- Organizing and commenting code
- Write reusable libraries
- Map data sources to data visualization libraries

CORE TOOLS

DESIRED TOOLS

- Scripting and environments
  e.g. Python, R, Matlab, Julia
- IDEs, Interpreters and Notebooks
- Other languages e.g. C/C++, Java, Scala, Go
- Source control e.g. git, Mercurial, Github, Bitbucket

DATA WRANGLING AND DATABASE CONCEPTS

CORE SKILLS

DESIRED SKILLS

- Clean data through statistical approaches, such as identifying outliers
- Create, read, update and delete on databases
- Extract, Transform and Load data
- Collect data from variety of sources including APIs and web
- Apply data normalization
- Transform data into machine-readable formats
- Optimize performance of database queries/schema (basics such as adding indices etc.)

CORE TOOLS

DESIRED TOOLS

- ETL e.g. Talend, Kettle, Alteryx
- Data cleaning e.g. Google Refine, Monarch, Orange, KNIME
- Databases and data stores e.g. MySQL, Postgres, Couch, MongoDB, Neo4j
- Scraping e.g. Scrapy, Apache Tika, BeautifulSoup
- Data Integration

STATISTICS AND DATA MODELLING

CORE SKILLS

DESIRED SKILLS

- Identify trends and behaviours with descriptive statistics
- Apply inferential statistics on randomised samples to understand characteristics of a population
- Identify, monitor, and measure quality of models over time
- Apply common models such as linear and logistic regression for prediction
- Run A/B tests

CORE TOOLS

DESIRED TOOLS

- R, R Studio, Pandas, SAS, SPSS, Excel
- A/B testing e.g. Adobe Target, Convert, Optimizely, Maxymizer, Unbounce, VWO
- Data Mining e.g. Weka, RapidMiner, KNIME, Rattle

SOFT SKILLS

- Agile Problem-Solving
- Business acumen
- Creative
- Critical thinking
- Data Inquisitive
- Data Intuition
- Self-starting Motivation
- Strong Communication
- Team building

RELATED JOB ROLES

- Business Data Analyst
- Business Intelligence Analyst
- Data Miner
- Data Wrangler
- Product Data Analyst

METHODOLOGY

These checklists were compiled by reviewing job advertisements on leading online portals and global course curriculum; stakeholder engagement with leading data scientists in Malaysia and regionally through the Big Data Malaysia network; and distributing a survey targeted to employers who are actively recruiting for data science teams. Further research is required on estimating the size and growth of the Malaysian data science job market as it expands between 2017 – 2020.

ABOUT US

This report was compiled by Dr Sandra Hanchard and Jessie Wong. We welcome your feedback on the usefulness of the checklists. Contact: sandra@dataviz.my

DataViz My
Your Data Education

The mission of DataViz My is to promote a culture of active learning around data. We help users of varying technical skillsets access big data technologies to apply insights for real world outcomes. DataViz My has helped more than 100 entrepreneurs and professionals through our data analytics and data science programmes. The network of instructors for DataViz My have a comprehensive track record of big data analytics industry experience.

BIG DATA
Malaysia

Big Data Malaysia is a community for professionals with an interest in all things data. Since 2012, the group has run regular meetups, produced reports on the state of Big Data in Malaysia, collaborated with industry, education and government sectors, and provided content curation. Our activities are conducted for helping stakeholders benefit from advances in technology and data ubiquity.

MDEC®

Malaysia Digital Economy Corporation Sdn. Bhd. (MDEC) is the holistic, government-owned agency launched in 1996 to pioneer the transformation of Malaysia's digital economy. Its roots stemmed from Vision 2020, the plan to develop Malaysia into a fully-developed nation by 2020. MDEC pursues a digital future that unlocks significant economic, environmental, and social value within the nation.

DataViz My

Your Data Education