Data Science Survey

Section I: PURPOSE & OUTCOMES

Section II: TECHNOLOGY, TOOLS & PLATFORMS

Section III: DATA SIZE

Section IV: DATA CHARACTERISTICS, STORAGE & MANAGEMENT

Section V: ANALYSIS TECHNIQUES

I. PURPOSE & OUTCOMES

For questions Q1-Q6, please rate your experience on a scale of 0–4, where:

- *0: have no knowledge or experience*
- *1: aware user; limited practical experience*
- *2: basic user; some past performance (clients, tasks, prior personal projects)*
- *3: skilled user; significant past performance (clients, tasks, prior personal projects)*
- *4: power user; expert knowledge; significant past performance*

Q1-Q6: Please rate your experience with performing:

Data Science Concepts	0	1	2	3	4
Q1: *Exploratory data analysis* - the practice of analyzing data sets to summarize their main characteristics, including use of statistical and visual methods, for preliminary data discoveries.	○	○	○	◉	○
Q2: *Diagnostic data analysis* - the practice of looking at past performance to determine what happened and why.	○	○	○	◉	○
Q3: *Predictive data analysis* - the practice of using many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future events.	○	○	◉	○	○
Q4: *Time series analysis* - the practice of analyzing time series data to extract meaningful statistics and other characteristics of the data.	○	○	◉	○	○
Q5: *Longitudinal data analysis* - the practice of deriving insights from longitudinal data that captures changes in individual entities over time.	○	○	◉	○	○

Q6: *Data analysis insights as inputs to another analysis* - for example, using insights into key variables to inform the development of simulation models.	○	○	◉	○	○

II. TECHNOLOGY, TOOLS & PLATFORMS

For questions Q7-Q19, please rate your experience on a scale of 0–4, where:

- *0: have no knowledge or experience*
- *1: aware user; limited practical experience*
- *2: basic user; some past performance (clients, tasks, prior personal projects)*
- *3: skilled user; significant past performance (clients, tasks, prior personal projects)*
- *4: power user; expert knowledge; significant past performance*

Q7: Please rate your experience with the following programming languages:

Technology, Tools, & Platforms	0	1	2	3	4
C	○	◉	○	○	○
C++	○	○	◉	○	○
C#	○	◉	○	○	○
Java	○	○	◉	○	○
JavaScript	○	◉	○	○	○
Julia	○	◉	○	○	○
Lisp/Clojure	○	◉	○	○	○
.NET	○	◉	○	○	○
Perl	○	◉	○	○	○
PHP	○	◉	○	○	○

Python	○	○	◉	○	○
R	○	○	○	◉	○
Ruby/Rails	○	◉	○	○	○
Scala	○	◉	○	○	○
SQL	○	○	○	◉	○
Visual Basic	○	○	◉	○	○
(Free Text) **Other languages** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q8: Please rate your experience with the following categories for database management systems:

Technology, Tools, & Platforms	0	1	2	3	4
Desktop relational databases (*e.g., MS Access*)	○	○	○	◉	○
Client-server relational databases (*e.g. MS SQL Server, Oracle, MySQL*)	○	○	○	◉	○

Non-Relational / NoSQL databases (*e.g. Apache Cassandra, HBase, Bigtable/MapReduce, DataStax, Dynamo, MarkLogic, MongoDB, Redis*)	○	◉	○	○	○
Distributed databases (*e.g. blockchain*)	○	◉	○	○	○
(Free Text) **Other languages** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q9: Please rate your experience with data reshaping and analysis tools:

Technology, Tools, & Platforms	0	1	2	3	4
Open-source (*e.g. Python—for example pandas, R— for example dplyr*)	○	○	○	○	◉
(Free Text) **Other open-source** data reshaping and analysis tools you are skilled in (Rating Level 2 or higher), separated by commas.					

Proprietary (*e.g. MATLAB, SAS, SPSS, Stata, Tableau*)	○	○	◉	○	○
(Free Text) **Other proprietary** data reshaping and analysis tools you are skilled in (Rating Level 2 or higher), separated by commas.					

Q10: Please rate your experience with machine learning tools:

Technology, Tools, & Platforms	0	1	2	3	4
Open-source (*e.g. Python— for example scikit-learn, Java— for example WEKA, R— for example caret*)	○	○	◉	○	○
(Free Text) **Other open-source** machine learning tools you are skilled in (Rating Level 2 or higher), separated by commas.					
Proprietary *(e.g. Amazon Machine Learning, Mathematica, MATLAB, Splunk*)	○	◉	○	○	○

(Free Text) **Other proprietary** machine learning tools you are skilled in (Rating Level 2 or higher), separated by commas.	

Q11: Please rate your experience with network analysis tools:

Technology, Tools, & Platforms	0	1	2	3	4
Open-source (*e.g. Gephi, pajek, Python — for example networkx, Java — for example JUNG, R — for example sna*)	○	◉	○	○	○
(Free Text) **Other open-source** network analysis tools you are skilled in (Rating Level 2 or higher), separated by commas.					
Proprietary (*e.g. ORA, UCINET*)	◉	○	○	○	○
(Free Text) **Other proprietary** network analysis tools you are skilled in (Rating Level 2 or higher), separated by commas.					

Q12: Please rate your experience with using the following:

Technology, Tools, & Platforms	0	1	2	3	4
Deep Learning frameworks (*e.g. Caffe, TensorFlow, Theano, Torch*).	○	◉	○	○	○
(Free Text) **Other deep learning** tools you are skilled in (Rating Level 2 or higher), separated by commas.					

Q13: Please rate your experience with using the following:

Technology, Tools, & Platforms	0	1	2	3	4
Source code management and control tools (*e.g. Git, Mercurial, Subversion (SVN)*	○	○	◉	○	○
(Free Text) **Other source code management and control tools** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q14: Please rate your experience with the following:

Technology, Tools, & Platforms	0	1	2	3	4
Generating data visualizations (*e.g. Chart.js, D3.js, Gephi, Processing, Tableau*).	○	○	◉	○	○
(Free Text) **Other data visualization tools** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q15: Please rate your experience with the following:

Technology, Tools, & Platforms	0	1	2	3	4
Generating data dashboards (*e.g., MicroStrategy, MS Excel, Tableau, Qlik*).	○	○	◉	○	○
(Free Text) **Other data dashboard tools** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q16: Please rate your experience with the following:

Technology, Tools, & Platforms	0	1	2	3	4
Tools for distributed computing	○	◉	○	○	○

(*e.g., Flink, Hive/Impala, Pig, Spark, Storm, Vertica*).					
(Free Text) **Other distributed computing tools** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q17: Please rate your experience with the following:

Technology, Tools, & Platforms	0	1	2	3	4
Cloud-based resources for big-data computing (*e.g. Amazon AWS, Cloudera, Google Cloud, IBM Bluemix/SoftLayer, Microsoft Azure, Oracle Cloud, Salesforce*).	○	○	◉	○	○
(Free Text) **Other distributed computing tools** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q18: Please rate your experience with geographical information system (GIS) tools:

Technology, Tools, & Platforms	0	1	2	3	4
Open-source GIS tools (*e.g. GRASS GIS, QGIS, uDig*)	○	◉	○	○	○

(Free Text) **Other open-source GIS tools** you are skilled in (Rating Level 2 or higher)					
Proprietary GIS tools (e.g. *Esri products such as ArcGIS*)	○	○	○	◉	○
(Free Text) **Other proprietary GIS tools** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q19: Please rate your experience with leveraging the following technologies for enabling more efficient data analytics:

Technology, Tools, & Platforms	0	1	2	3	4
GPU processing	○	◉	○	○	○
In-memory processing	○	◉	○	○	○
Microservices	○	◉	○	○	○
Multi-threading	○	◉	○	○	○

III. DATA SIZE

For questions Q20-Q22, please rate your experience on a scale of 0–4, where:

- *0: have no knowledge or experience*
- *1: aware user; limited practical experience*
- *2: basic user; some past performance (clients, tasks, prior personal projects)*
- *3: skilled user; significant past performance (clients, tasks, prior personal projects)*
- *4: power user; expert knowledge; significant past performance*

Q20: Please rate your experience in using the following data set sizes for analysis:

Data Size	0	1	2	3	4
Less than 10GB	○	○	○	○	◉
More than 10GB, up to 100GB	○	◉	○	○	○
More than 100GB, up to 1TB	○	◉	○	○	○
More than 1TB, up to 100TB	○	◉	○	○	○
More than 100TB, up to 1PB	○	◉	○	○	○
More than 1PB	○	◉	○	○	○

Q21-Q22: Please rate your experience with performing the following:

Data Size	0	1	2	3	4
Q21: ***Feature selection*** - the process of *selecting* a subset of *relevant features* (variables, predictors) for use in model construction. Feature selection returns a subset of the features.	○	○	○	◉	○

Q22: *Feature extraction* – the process of *transforming* data in the high-dimensional space to a space of fewer dimensions. Feature extraction creates *new features* from functions of the original features. (*e.g. principal component analysis* (*PCA*), *multiple correspondence analysis* (*MCA*)).	○	○	◉	○	○

IV: DATA CHARACTERISTICS, STORAGE & MANAGEMENT

For questions Q23-Q26, please rate your experience on a scale of 0–4, where:

- *0: have no knowledge or experience*
- *1: aware user; limited practical experience*
- *2: basic user; some past performance (clients, tasks, prior personal projects)*
- *3: skilled user; significant past performance (clients, tasks, prior personal projects)*
- *4: power user; expert knowledge; significant past performance*

Q23: Please rate your experience with processing and handling the following data formats:

Data Characteristics, Storage, & Management – Data Characteristics	0	1	2	3	4
Audio formats	○	○	◉	○	○
Geospatial formats (*e.g. Esri, GeoJSON, GML, LandXML*)	○	○	◉	○	○
Hierarchical Data Format (HDF)	◉	○	○	○	○
Image formats	○	◉	○	○	○

JavaScript Object Notation (JSON)	○	○	◉	○	○
Tabular CSV/ASCII	○	○	○	○	◉
Video formats	○	◉	○	○	○
XML	○	○	◉	○	○
Unstructured text	○	○	◉	○	○
(Free Text) **Other data formats** you are skilled in (Rating Level 2 or higher), separated by commas.					

Q24-Q25: Please rate your experience with the following:

Data Characteristics, Storage, & Management – Data Storage	0	1	2	3	4
Q24: Data warehouses - central repositories of integrated data from one or more disparate sources, but the data (as opposed to data lakes) is *predominantly structured and processed*. In other words, before data may be loaded into a data warehouse, it first needs some shape and structure (also known as *schema-on-write*). The end users are typically professionals interested in business intelligence insights.	○	○	◉	○	○

Q25: Data lakes - storage repositories that hold a vast amount of *raw data in its native format*. As opposed to data warehouses, data lakes may include structured, semi-structured, and unstructured data, as well as formats like video and audio. Increasingly, the term is being accepted as a way to describe any large data pool in which the schema and data requirements are not defined until the data is queried. In other words, with a data lake, you load in the raw data, as-is; when it is needed, you give it shape and structure (also known as *schema-on-read*). The end users are typically data science professionals.	○	○	◉	○	○

Q26: Please rate your experience with the following:

Data Characteristics, Storage, Management – Data Management	0	1	2	3	4
Q26: Developing policies for data governance - the practice of managing the availability, usability, integrity, ownership, stewardship, and security of the data employed in an enterprise.	○	○	◉	○	○

V. ANALYSIS TECHNIQUES

For questions Q27-Q34, please rate your experience on a scale of 0–4, where:

- *0: have no knowledge or experience*
- *1: aware user; limited practical experience*
- *2: basic user; some past performance (clients, tasks, prior personal projects)*
- *3: skilled user; significant past performance (clients, tasks, prior personal projects)*
- *4: power user; expert knowledge; significant past performance*

Q27-Q28: Please rate your experience with performing the following:

Analysis Techniques – Data Engineering and Processing	0	1	2	3	4
Q27: Data engineering (includes what some companies might call data infrastructure or data architecture) - Data engineering involves gathering and collecting the data, storing the data, running batch processing or real-time processing on the data, and serving the data via an API to data scientists who can easily query it.	○	○	◉	○	○
Q28: Extract, Transform, Load (ETL) - a data integration process for transferring raw data from a source server to a data warehouse on a target server and then preparing the information for downstream use.	○	○	◉	○	○

Q29-31: Please rate your experience with performing the following:

Analysis Techniques – Data Science	0	1	2	3	4
Q29: Feature engineering - practice of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on data not explicit in the set.	○	○	◉	○	○

Q30: Supervised feature learning – a process in which features are learned with labeled input data (e.g. *supervised neural networks, multilayer perceptron, and (supervised) dictionary learning*).	○	○	◉	○	○
Q31: Unsupervised feature learning – a process in which features are learned with unlabeled input data (*e.g. dictionary learning, independent component analysis, auto-encoders, matrix factorization, and various forms of clustering such as k-means*).	○	○	◉	○	○

Q32: Please rate your experience with performing the following machine learning techniques and algorithms:

Analysis Techniques – Data Science	0	1	2	3	4
Convolutional Neural Network (CNN)	○	◉	○	○	○
Decision Tree	○	○	○	○	◉
Deep Learning	○	◉	○	○	○
Deep Neural Network	○	◉	○	○	○
K-means Clustering	○	○	○	◉	○
Support Vector Machine (SVM)	○	○	◉	○	○
Text Mining/Natural Language Processing (NLP)	○	○	◉	○	○
(Free Text) Other machine learning techniques and algorithms you are skilled in (Rating Level 2 or higher), separated by commas.					

Q33: Please rate your experience with the following simulation modeling methods:

Analysis Techniques – Simulation Modeling	0	1	2	3	4
Agent-based	○	◉	○	○	○
Discrete-event	○	○	◉	○	○
Microsimulation	◉	○	○	○	○
Monte Carlo	○	○	○	◉	○
Systems dynamics	○	○	○	◉	○

Q34: Please rate your experience with the following:

Analysis Techniques – Simulation Modeling	0	1	2	3	4
Q34: Using simulation modeling (e.g. *agent-based modeling, Monte Carlo*) to **generate pseudo–data sets** for study in scenarios where no real-world data exists.	○	○	◉	○	○

END SURVEY