COURSE ONE:

MODULE ONE:

* Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
* There are many paths to a career in data science; most, but not all, involve a little math, a little science, and a lot of curiosity about data.
* New data scientists need to be curious, judgmental and argumentative.
* Why data science is considered the sexiest job in the 21st century, paying high salaries for skilled workers.

QUIZ:

Q1:

Harvard Business Review called data science the sexiest job in the 21st century: TRUE

Q2:

According to the report by the McKinsey Global Institute, by 2018, it is projected that there will be a shortage of people with deep analytical skills in the United States. What is the size of this shortage? 140,000-190,000 people

Q3:

According to this Module's reading assignment, how did Walmart address its analytical needs? CROWDSOURCING

Q4:

The New York Times reported that the average base salary of a data scientist is $85,000 + competitive bonus. FALSE

Q5:

According to professor Haider, the three important qualities to possess in order to succeed as a data scientist are:

CURIOUS, JUDGEMENTAL, GOOD STORY TELLER (ARGUMENTATIVE)

MODULE TWO:

LEARNING OBJECTIVES

* Learn how organizations are using data science to solve problems.
* Learn about some key concepts, tools and algorithms used in data science shared by data science professionals.
* Hear from Professor Murtaza Haider on how the cloud has expanded the role of the data scientist.
* The typical workday for a Data Scientist varies depending on what type of project they are working on.
* Many algorithms are used to bring out insights from data.
* Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily

QUIZ:

Q1:

According to this Module's reading assignment, Hal Varian, the chief economist at Google, declared that "the sexy job in the next ten years will be: STATISTICIANS

Q2:

According to this Module's reading assignment, the author defines a data scientist as someone who uses complicated machine learning to build predictive models, publishes their research results in online blogs, and works from home. FALSE

Q3:

According to this Module's reading assignment, the author defines data science as the art of uncovering the hidden secrets in data. FALSE

Q4:

According to this Module's reading assignment, what is admirable about Dr. Patil’s definition of a data scientist? HIS DEFINITION IS INCLUSIVE OF INDIVIDUALS FROM VARIOUS ACADEMIC BACKGROUNDS AND TRAINING.

Q5:

According to this Module's reading assignment, what characteristics are said to be exhibited by the **best** data scientists? CURIOUS INDIVIDUALS WHO ASK GOOD QUESTIONS AND ARE O.K. DEALING WITH UNSTRUCTURED SITUATIONS AND TRYING TO FIND STRUCTURE IN THEM.

MODULE THREE:

LEARNING OBJECTIVES

* Learn about the 5 Vs of Big Data.
* Learn about how Hadoop and other tools are handling the demands of big data.
* Hear from Norman White, Professor at New York University on data science and big data.
* Learn about data mining and the steps that comprise the process of mining a given data set.

In this module, you have learned:

* How Big Data is defined by the Vs: Velocity, Volume, Variety, Veracity, and Value.
* How Hadoop and other tools, combined with distributed computing power,  are used to handle the demands of Big Data.
* What skills are required to analyse Big Data.
* About the process of Data Mining, and how it produces results.

QUIZ:

Q1:

According to this Module's reading assignment, the output of a data mining exercise largely depends on: THE QUALITY OF THE DATA

Q2:

According to this Module's reading assignment, what should you do when data is missing in a systematic way? DETERMINE THE IMPACT OF MISSING DATA ON THE RESULTS AND WHETHER MISSING DATA CAN BE EXCLUDED FROM THE ANALYSIS

Q3:

According to this Module's reading assignment, what is an example of a data reduction algorithm? PRINCIPAL COMPONENT ANALYSIS

Q4:

According to this Module's reading assignment, after the data is appropriately processed, transformed, and stored, what is a good starting point for data mining? DATA VISUALIZATION

Q5:

"Formal evaluation could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data." This is known as: IN-SAMPLE FORECAST

MODULE FOUR:

LEARNING OBJECTIVES

* Learn the difference between Machine Learning and Deep Learning.
* Learn about some of the many applications of Machine Learning.
* Learn about regression and what questions can be put to regression analysis.

In this module, you have learned:

* The differences between some common Data Science terms, including Deep Learning and Machine Learning.
* Deep Learning is a type of Machine Learning that simulates human decision-making using neural networks.
* Machine Learning has many applications, from recommender systems that provide relevant choices for customers on commercial websites, to detailed analysis of financial markets.
* How to use regression to analyze data.

QUIZ:

Q1:

According to this Module's reading assignment, which of the following best describes the real added value of the author's research on residential real estate properties? QUANTIFYING THE MAGNITUDE OF RELATIONSHIPS BETWEEN HOUSING PRICES AND DIFFERENT DETERMINANTS. THE RESEARCH CONFIRMED MANY PERCEPTIONS THAT PEOPLE HAVE ABOUT REAL ESTATE PROPERTIES BUT IT MAJOR CONTRIBUTION IS QUANTIFYING THE MAGNITUDE OF THE RELATIONSHIPS BETWEEN THE HOUSING PRICES AND DIFFERENT DETERMINANTS.

Q2:

Regression is a statistical technique developed by Blaise Pascal. FALSE, REGRESSION IS A STATISTICAL TECHNIQUE DEVELOPED BY SIR FRANCES GALTON.

Q3:

According to this Module's reading assignment, what did the author's research reveal about the impact of an additional washroom on the price of a housing unit? THE AUTHOR FOUND THAT AN ADDITIONAL WASHROOM ADDS MORE TO THE HOUSING PRICES THAN AN ADDITIONAL BEDROOM.

Q4:

What did the author's research reveal about proximity to large shopping centres on the price of a housing unit? THE AUTHOR DISCOVERED THAT PROXIMITY TO LARGE SHOPPING CENTRES HAD A NONLINEAR IMPACT ON THE HOUSING PRICES.

Q5:

According to this Module's reading assignment, which of the following are questions that can be put to regression analysis? WHAT IS THE IMPACT OF LOT SIZE ON HOUSING PRICE? DO HOMES WITH BRICK EXTERIOR SELL FOR LESS THAN HOMES WITH STONE EXTERIOR?

MODULE 5:

IBM CLOUD ACCOUNT

1cf128005052714054f5192973b6720d

IBM Watson Studio is a service from IBM, that provides a suite of tools and a collaborative environment for data scientists, developers and domain experts. In this lab, you will use Watson Studio and explore different datasets. As we have learnt in the course, the data is not only about numbers, it can be anything such as numeric data, text data, images, videos, audios etc. You will work on three samples.

**Sample 1** in which you will learn about the dataset in which only numeric attributes are present.

**Sample 2** in which you will learn about the dataset in which numeric & text attributes are present.

**Sample 3** in which you will analyze how the Jupyter Notebooks look like. How a Data Scientist create the models?

OBJECTIVES:

* Launch Watson Studio for accessing Data Science Problems
* Evaluate Numeric dataset
* Evaluate dataset with Non-Numeric attributes
* Evaluate Jupyter Notebook

EXERCISE 2: EVALUATE NUMERIC DATASET

The data is related to forest fires where the aim is to predict the burned area of forest fires, in the northeast region of Portugal, by using meterological and other data.

**Attribute Information:**

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

EXERCISE 2: EVALUATE NON-NUMERIC DATASET

The dataset comprises of three main tables:

* listings - Detailed listings data showing 96 attributes for each of the listings. Some of the attributes used in the analysis are price(continuous), longitude (continuous), latitude (continuous), listing\_type (categorical), is\_superhost (categorical), neighbourhood (categorical), ratings (continuous) among others.
* reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (datetime), listing\_id (discrete), reviewer\_id (discrete) and comment (textual).
* calendar - Provides details about booking for the next year by listing. Four attributes in total including listing\_id (discrete), date(datetime), available (categorical) and price (continuous).

### EXERCISE 3: EVALUATE JUPYTER NOTEBOOK

This notebook shows you how Decision Optimization can help to prescribe decisions for a complex constrained problem using Python to help determine the optimal location for a new store.

The objective is to minimize the total distance from libraries to coffee shops so that a book reader always gets to our coffee shop easily. It can be done by analyzing and displaying the location of the coffee shops on a map.

#### Summary

In this lab, you have learnt about how different datasets are available and how a data scientist create and predict the models using the model building in IBM Watson Jupyter Notebook.

**LEARNING OBJECTIVES**

* Learn about what companies need to do in order to start with data science.
* Learn about some of the qualities that differentiate data scientists from other professionals.
* Learn about some applications of data science.
* Learn about analytics and what important role data scientists play in this process.
* Learn about story-telling and the importance of an effective final deliverable.
* Learn about the main components of an effective final deliverable.
* Apply what you learned about data science to answer open-ended questions.
* Demonstrate your understanding of the readings to define what data science and data scientist mean.
* Demonstrate your understanding of the readings to answer a question about the final deliverable of data science project.

**IN THIS MODULE, YOU HAVE LEARNED:**

* Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
* That companies can start on their data science journey by capturing data. Once they have data, they can begin analysing it.
* Some ways that data is generated by consumers.
* How businesses like Netflix, Amazon, UPs, Google, and Apple use the data generated by their consumers and employees.
* The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.

QUIZ:

Q1:

According to this Module's reading assignment, what is the ultimate purpose of analytics? TO COMMUNICATE FINDINGS TO STAKEHOLDERS TO FORMULATE POLICY OR STRATEGY

Q2:

In this Module's reading assignment, what role of a data scientist is discussed?

USING INSIGHTS TO BUILD A NARRATIVE TO COMMUNICATE FINDINGS.

Q3:

The United States Economic Forecast is a publication by McKinsey University Press. FALSE, IT IS BY DELOITTE UNIVERSITY PRESS.

Q4:

The report, discussed in this Module's reading assignment, successfully did the job of: USING DATA AND ANALYTICS TO GENERATE THE LIKELY ECONOMIC SCENARIOS

Q5:

According to this Module's reading assignment, it is recommended that a team waits until the results of analytics are out before they can decide on the final deliverable. FALSE, IN ORDER TO PRODUCE A COMPELLING NARRATIVE, INITIAL PLANNING AND CONCEPTUALIZING OF THE FINAL DELIVERABLE IS OF EXTREME IMPORTANCE.

MODULE SIX:

LEARNING OBJECTIVES

* Learn about what companies need to do in order to start with data science.
* Learn about some of the qualities that differentiate data scientists from other professionals.
* Learn about some applications of data science.
* Learn about analytics and what important role data scientists play in this process.
* Learn about story-telling and the importance of an effective final deliverable.
* Learn about the main components of an effective final deliverable.
* Apply what you learned about data science to answer open-ended questions.
* Demonstrate your understanding of the readings to define what data science and data scientist mean.
* Demonstrate your understanding of the readings to answer a question about the final deliverable of data science project.

IN THIS MODULE, YOU HAVE LEARNED:

* The length and content of the final report will vary depending on the needs of the project.
* The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgments, references, and appendices.
* The report should present a thorough analysis of the data and communicate the project findings.

QUIZ:

Q1:

According to this Module's reading assignment, regardless of the length of the final deliverable, which of the following does the author recommend that you include: A COVER PAGE, APPENDICES, TABLE OF CONTENTS

Q2:

According to this Module's reading assignment, an introductory section is always helpful in: SETTING UP THE PROBLEM FOR THE READER WHO MIGHT BE NEW TO THE TOPIC

Q3:

According to this Module's reading assignment, The results section is where you craft your main arguments and present your conclusion. FALSE, THE RESULTS SECTION IS WHERE YOU PRESENT THE EMPIRICAL FINDINGS.

Q4:

According to this Module's reading assignment, the discussion section is where you: RELY ON THE POWER OF NARRATIVE TO ENABLE NUMBERS TO COMMUNICATE YOUR IMPORTANT FINDINGS TO THE READERS, REFER THE READER TO THE RESEARCH QUESTION AND THE KNOWLEDGE GAPS YOU IDENTIFIED EARLIER, HIGHLIGHT HOW YOUR FINDINGS PROVIDE THE ULTIMATE MISSING PIECE TO THE PUZZLE

Q5:

According to this Module's reading assignment, what is an example of housekeeping? ADDING A LIST OF REFERENCES

COURSE TWO:

Much of the world's data resides in databases. SQL (or Structured Query Language) is a powerful language which is used for communicating with and extracting data from databases. A working knowledge of databases and SQL is a must if you want to become a data scientist.

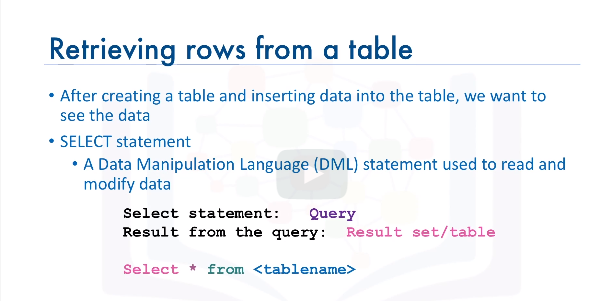
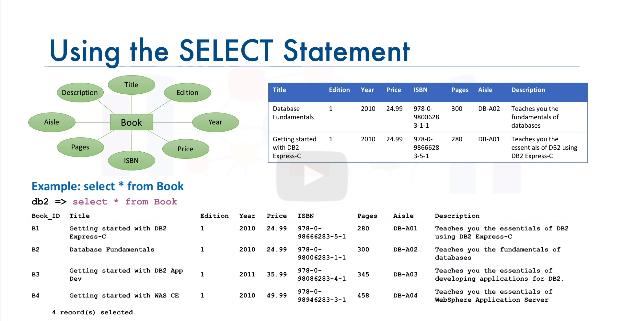
The purpose of this course is to introduce relational database concepts and help you learn and apply foundational knowledge of the SQL language. It is also intended to get you started with performing SQL access in a data science environment.

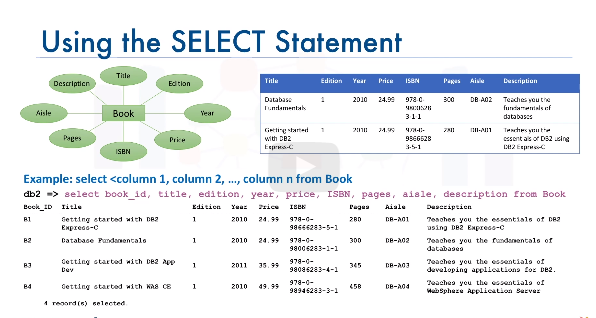
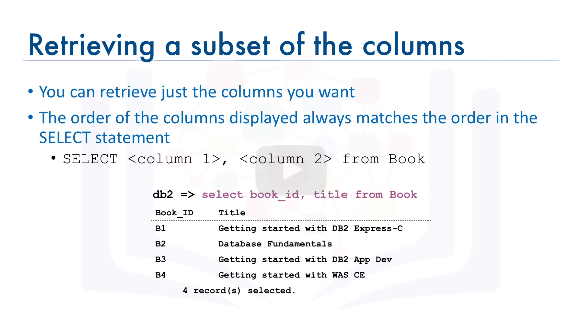
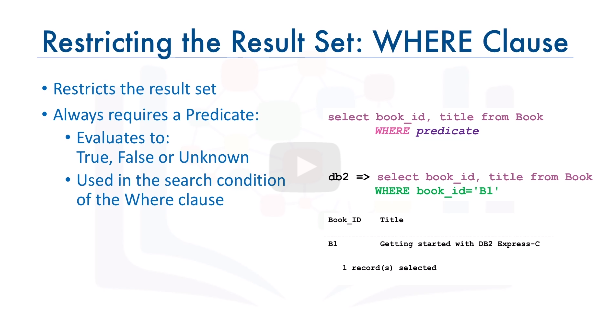
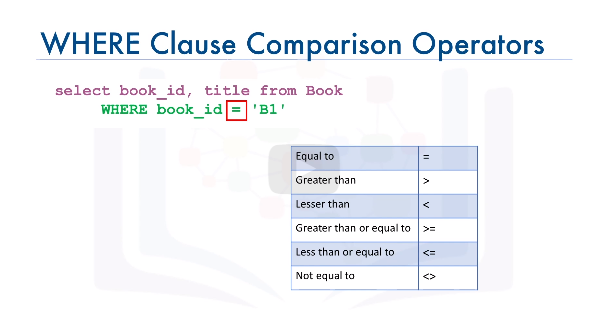
The emphasis in this course is on hands-on and practical learning. As such, you will work with real databases, real data science tools, and real-world datasets. You will create a database instance in the cloud. Through a series of hands-on labs, you will practice building and running SQL queries. You will also learn how to access databases from Jupyter notebooks using SQL and Python.

MODULE ONE:

### ****LEARNING OBJECTIVES****

* Describe SQL and Databases
* Explain the syntax of basic SQL statements - Select, Insert, Update, Delete
* Compose and execute basic SQL statements hands-on on a live database
* Demonstrate how to write basic SQL statements

## **Objectives**

At the end of this lab you will be able to:

* use SELECT queries to retrieve data from the database

**Effort:** 2 min

The general syntax of SELECT statments is:

**select COLUMN1, COLUMN2, ... from TABLE1 ;**

To retrieve all columns from the COUNTRY table we could use "\*" instead of specifying individual column names:

**select \* from COUNTRY ;**

The WHERE clause can be added to your query to filter results or get specific rows of data. To retrieve data for all rows in the COUNTRY table where the ID is less than 5:

**select \* from COUNTRY where ID < 5 ;**

In case of character based columns the values of the predicates in the where clause need to be enclosed in single quotes. To retrieve the data for the country with country code "CA" we would issue:

\*\*select \* from COUNTRY where CCODE = 'CA'; \*\*

* Retrieve the number of rows that match a query criteria
* Remove duplicate values from a result set and return the unique values
* Restrict the number of rows retrieved from a table

