## Week 1 – Introduction

1. This course is part of a specialisation on Coursera designed to train data analysts to become data scientists.
   1. True.
2. Databricks Community Edition is available for free.
   1. True.
3. What tabular data structures are available in Spark?
   1. Spark SQL Table.
   2. Spark DataFrame.
4. What can be used to describe Delta Lake?
   1. Data lakehouse.
5. What filetype can be used to save a collection of Databricks Notebooks as a single file?
   1. Databricks Archive (DBC).

## Week 2a – The Scientific Method

1. The scientific method is important to performing quality data science.
   1. True.
2. What contributes to an effective science question?
   1. Objectively measurable.
   2. Relevancy to the field.
3. The alternative hypothesis is usually associated with no impact or outcome.
   1. False.
4. For what step of the scientific method is data collection imperative?
   1. Experimenting.
5. Which step/s are important to the scientific process?
   1. Question asking.
   2. Hypothesis construction.
   3. Experimenting.

## Week 2b – Defining Skills of Data Science

1. Applied statistics is only practiced by trained statisticians.
   1. False.
2. What can be classified under the field of applied statistics?
   1. Hypothesis testing.
   2. Sampling methods.
   3. Statistical modelling.
3. Machine learning is at the intersection of applied statistics and computer science.
   1. True.
4. What skill is an application of computer science.
   1. Programming.
5. What are two similarities between applied statistics and computer science?
   1. Vital to data science.
   2. Rooted in mathematics.

## Week 3a – Statistics

1. The field of statistics provides tools to find patterns and relationships in data using mathematics.
   1. True.
2. Which of the following can be considered “data”?
   1. Numbers.
   2. Words describing something.
   3. Information that is represented or coded in some form that is suitable for processing.
3. What is correct?
   1. Qualitative data is descriptive.
   2. Quantitative data is numerical.
   3. Quantitative data can be discrete or continuous.
4. What are the two main branches of statistics?
   1. Descriptive and inferential.
5. What are some measures of descriptive statistics?
   1. Mean.
   2. Median.
   3. Mode.
   4. Standard deviation.
   5. Range.
6. The mean is the middle point of a dataset and the median is the average.
   1. False.
7. What are the uses for descriptive and inferential statistics?
   1. Descriptive statistics is used to understand and define the sample data.
   2. Inferential statistics is used to make generalisations.
   3. Inferential statistics is used to make predictions.
8. What must be true about a sample in order to be able to use it to make accurate inferences?
   1. The sample has to be representative of the actual population.
   2. Each and every member of the population as an equal and mutually exclusive chance of being selected in the sample.
   3. The sample must be unbiased.
9. A good sample will be perfectly representative of the entire population with no errors.
   1. False.
10. Why do we use inferential statistics?
    1. To determine whether the patterns observed in the data actually generalise to the population.
    2. To determine whether sample statistics adequately support our hypotheses about the population.

## Week 3b – Probability Distributions

1. A distribution of the probability of there being 7 or more rainy days in July: Is this a continuous or discrete probability distribution?
   1. Discrete.
2. A distribution of the probability of there being 7 or more rainy days in July: Is this a continuous or discrete probability distribution?
   1. Continuous.
3. A probability distribution represents continuous variables.
   1. True.

## Week 3c – Hypothesis Testing

1. The first step of hypothesis testing is to formulate a null and alternative hypothesis.
   1. True.
2. Which decisions can we make based on the results of a hypothesis test?
   1. Reject the null hypothesis.
   2. Fail to reject the null hypothesis.
3. Which statements are true?
   1. The null hypothesis us usually associated with an expectation that something will have no impact or effect.
   2. The alternative hypothesis is usually associated with an expectation that something will have an impact or effect.

## Week 3d – Introductory Statistics for Data Science

1. Descriptive statistics is used to understand and define the sample data, while inferential statistics is used to make generalisations about a population.
   1. True.
2. What are some measures of central tendency?
   1. Mean.
   2. Median.
   3. Mode.
3. The probability of a specific outcome of an event is equal to the number of times the outcome occurred divided by the total number of events.
   1. True.
4. What statements are true?
   1. Binomial probability distributions model binary outcomes.
   2. Multinomial distributions model probability when there are more than two possible outcomes.
   3. Poisson distributions model the probability of an event happening based on the mean number of occurrences of the event.
   4. Bernoulli distributions model single trials of random variables that have only two possible outcomes.
5. Random variables can only be discrete, not continuous.
   1. False.
6. Continuous probability distributions model scenarios where the outcomes of experiments can only take on single, specific values.
   1. False.
7. Some continuous probability distributions include:
   1. Uniform.
   2. Normal.
   3. Logistic.
8. Hypothesis testing is a way to determine whether the patterns we observe in a sample could be due to sampling error rather than being representative of a true effect in the population.
   1. True.
9. What is one key assumption that hypothesis tests make about the data?
   1. It is normally distributed.
10. Outliers should always be removed from the dataset before training a machine learning model.
    1. False.

## Week 4a – Basics of Machine Learning

1. Machine learning is the process of computers learning to accomplish tasks without being explicitly programmed to do so.
   1. True.
2. What are the advantages of using machine learning to solve problems?
   1. More accurate predictions.
   2. Consider ore data observations.
   3. Consider more data features.
3. Unsupervised learning creates a function mapping a series of inputs to an output based on true values known as labels.
   1. False.
4. What are examples of a supervised learning problem?
   1. Predicting whether or not a football team will win its upcoming match.
   2. Estimating the full-grown height of recently planted trees.
   3. Predicting whether the next day’s weather will be sunny.
5. What are examples of an unsupervised learning problem?
   1. Grouping recipes together based on their ingredients.
   2. Clustering customers into groups based on their past shopping behaviour.

## Week 4b – Classification, Regression and Clustering

1. What are examples of supervised learning?
   1. Regression.
   2. Classification.
2. What is an example of unsupervised learning?
   1. Clustering.
3. Regression is concerned with predicting one of a predefined set of classes.
   1. False.
4. What are examples of a clustering problem?
   1. Grouping recipes together based on their ingredients.
   2. Clustering customers into groups based on their past shopping behaviour.
5. What is an example of a regression problem?
   1. Estimating the full-grown height of recently planted trees.
6. What is an example of a classification problem?
   1. Predicting whether or not a football team will win its upcoming match.

## Week 5a – Regression Evaluation

1. The residuals for a regression model indicate what percentage of the predicted values the model got right.
   1. False.
2. Because it squares the difference between the predicted and actual values, mean squared error (MSE) penalises larger errors.
   1. True.
3. The root mean squared error (RMSE) penalises smaller errors because it takes the square root of the MSE.
   1. False.
4. Which of the following regression evaluation metrics is in the same units as the dependent variable?
   1. RMSE.
5. What is true about Mean Absolute Error?
   1. It is more robust to outliers.
   2. It doesn’t the errors as extremely as MSE.
   3. It isn’t suitable for applications where you want to pay more attention to the outliers.

## Week 5b – Bias-Variance Trade-off

1. You trained a model and got 99% accuracy on the training set and 71% on the test set. What’s true?
   1. The model has high variance.
2. You trained a model and got 65% accuracy on the training set. What’s true?
   1. The model has high bias.
3. You trained a model and got 97% accuracy on the training set and 94% accuracy on the test set. What’s true?
   1. None of the above.

## Week 5c – Logistic Regression

1. What are some of the reasons that logistic regression is better suited for classification tasks than linear regression?
   1. Linear regression predicts numbers that could range from negative to positive infinity, while logistic regression predicts probabilities.
   2. A straight line doesn’t fit a probabilistic, binary outcome as well as a sigmoid curve.
   3. Linear regression predicts continuous outputs and logistic regression predicts probabilities of belonging to a particular class.
2. The outputs of a logistic regression model are always between 0 and 1.
   1. True.
3. How do you interpret the coefficients of logistic regression?
   1. The amount of evidence for belonging to a particular class, given a change in the predictor variable.
4. How can a data scientist change which class label is applied, given a predicted probability value?
   1. Lower or raise the threshold value for assigning a sample to a given class.

## Week 5d – Classification Evaluation

1. Which are true?
   1. A Type I error means a false positive.
   2. A Type II error means a false negative.
2. Classification accuracy is calculated as the number of correct predictions divided by the total number of predictions.
   1. True.
3. Of all the cases that the model predicted to be positive, what percentage were actually positive?
   1. Precision.
4. Of all the cases that were actually positive, what percentage did the model correctly identify as positive?
   1. Recall.
5. The F1 score is the harmonic mean between precision and recall.
   1. True.

## Week 5e – Decision Trees Part 1

1. You have a trained classification decision tree that determines what kind of animal you are looking at: a crow, a penguin, a dolphin, or an elephant. The series of yes/no questions that the tree makes are as follows:  
   1a) Does it live only on land? -> 2a) Does it fly?  
   1b) Does it spend its whole life in the water?  
   Use this decision tree to classify an animal that lives only on land and doesn’t fly.
   1. Elephant.
2. Using the same decision tree, classify an animal that doesn’t live only on land and doesn’t spend its whole life in the water.
   1. Penguin.
3. You now have a trained regression decision tree that determines the temperature range outside based on the weather. The series of decisions it makes are as follows:  
   1a) Is the time of day between 1PM and 6PM? -> 2a) Is the cloud cover less than 20%? (If both are yes, the temperature is predicted to be >60 degrees Fahrenheit. If the first is yes, but the second is no, the temperature is predicted to be between 40 and 60 degrees Fahrenheit.)  
   1b) Is it night time? -> 2b) Are you in the Arctic? (If both are yes, the temperature is predicted to be < 0 degrees Fahrenheit. If the first is yes but the second is no, the temperature is predicted to be between 0 and 40 degrees Fahrenheit.)  
   Use this decision tree to determine the temperature when it’s 4:45PM and the cloud cover is 60%.
   1. Between 40 and 60 degrees Fahrenheit.
4. Using the same decision tree, determine the temperature when it’s 11PM and you aren’t in the Arctic.
   1. Between 0 and 40 degrees Fahrenheit.

## Week 5f – Decision Trees Part 2

1. What is a terminal node in a decision tree called?
   1. A leaf node.
2. The decision tree algorithm uses the information gained to determine how to split the data.
   1. True.
3. The decision tree algorithm seeks for each branch to be as heterogeneous as possible.
   1. False.
4. What hyperparameters can be adjusted to control the training of a decision tree.
   1. Tree depth.
   2. Node size.
   3. Leaf size.
   4. The number of features to consider.

## Week 6a – Measuring Successes and Constraints

1. Objectives of data science projects should be measurable.
   1. True.
2. It’s advantageous to have an objective that’s measurable against a machine learning evaluation metric.
   1. True.
3. This is a measurable objective: “We would like to decrease the number of out-of-stock items in each store by 50% at the end of the day.”
   1. True.
4. This is a measurable objective: “We would like to use the results of our clustering analysis to learn about weather days.”
   1. False.
5. This is a measurable objective: “We would like to increase the likelihood of recovery for infected patients after administering this medicine.”
   1. True.
6. What are common constraints that apply to data science projects.
   1. Time.
   2. Data availability.
   3. Technology availability.
   4. Regulatory requirements.
7. What’s an example of a technological constraint?
   1. All of the relevant data doesn’t fit on the computer.

## Week 6b – Machine Learning Solutions

1. Data science solutions should be assessed in the real-world production environment.
   1. True.
2. What are reasons for assessing data science solutions after they have been deployed?
   1. Data collection issues.
   2. Data cleansing issues.
   3. Model serving issues.
   4. Model drift.
3. Machine learning solutions are fully complete after they’ve been deployed.
   1. False.
4. Machine learning models drift when the relationship between the feature variances and the label variable changes from the model’s understanding of the relationships.
   1. True.
5. What is a strategy for combating machine learning model drift?
   1. Retrain the model on more recent observations to learn new relationships.