# Unit Assessment: Analytical Modeling and Big Data

**Due** Sunday by 11:59pm **Points** 100 **Submitting** an external tool

You are about to complete the Analytical Modeling and Big Data Unit Assessment! This Unit Assessment allows you to check your knowledge, as well as demonstrate your competency in key concepts from Modules 15 through 19.

After submitting the assessment, you will see a summary of your performance. While you will not be able to see your performance on individual questions, you are allowed unlimited attempts to complete the assessment.

## Question 1

Using R, how would you import a CSV file, “data_file.csv” into your environment? Select all R statements that could import a CSV file without error.

- [x] `read.csv(file=data_file.csv',check.names=F,stringsAsFactors = F)` ✔
- [ ] `read(file=data_file.csv')`
- [ ] `csv(file=data_file.csv',header=T)`
- [x] `read.csv(check.names=F,stringsAsFactors = F)` ✘
- [ ] `read(file=data_file.csv',check.names=F,stringsAsFactors = F)`
- [x] `read.csv(file=data_file.csv',stringsAsFactors = F)` ✔

## Question 2

You are trying to summarize the heights of your students using the tidyverse in R. Within your R environment you have a dataframe called “test_df” that contains the names of each student and their heights in inches.

Select the correct R statement that produces a dataframe containing summary statistics of the student’s heights:

- [ ] `sum(test_df,Mean_Height=mean(height),SD_Height=sd(height))`

- `test_df %>% summarize(Mean_Height=mean(height),SD_Height=sd(height))` ✔
- `test_df %>% sum(Mean_Height=mean(height),SD_Height=sd(height))`
- `test_df %>% summary(Mean_Height=mean(height),SD_Height=sd(height))`

## Question 3

Consider that you have imported a dataframe into R with three columns - `model`, `year`, `Mean_Hwy`. You wish to create a heatmap that compares the mean highway fuel-efficiency across vehicle models and years.

Complete the following code below so that it will do the following tasks:

1. Create a ggplot2 object.
2. Add a heatmap plot to the ggplot2 visualization.
3. Rotate the x-axis labels 45 degrees.
4. Add labels to x-axis, y-axis, and legend.

```
plt <- ggplot(data,aes(x=model ✔ ,
                       y=factor(year) ✔ ,
                       fill=Mean_Hwy ✔ ))

plt + geom_tile() ✔ + labs(
                      x="Model" ✔ ,
                      y="Vehicle Year" ✔ ,
                      fill="Mean Highway (MPG)" ✔ ) + theme(
axis.text.x = element_text(angle=45,hjust=1) ✔ )
```

## Question 4

Match the following definitions of statistical concepts with the correct term in the word bank:

1. Multiple linear regression ✔ : builds a linear regression model with two or more independent variables.
2. Alternate hypothesis ✔ : known as Ha and is generally the hypothesis that is influenced by non-random events.
3. Continuous data ✔ : a data type that can be subdivided infinitely.
4. Pearson correlation coefficient ✔ : denoted as “r” in mathematics and is used to quantify a linear relationship between two numeric variables.
5. Chi-squared test ✔ : a statistical test used to compare the distribution of categorical frequencies across two groups.
6. Dichotomous data ✔ : a data type that is either one of two categories.
7. One-way ANOVA ✔ : a statistical test used to test the means of a single dependent variable across a single independent variable with multiple groups.
8. Shapiro-Wilk test ✔ : a statistical test to quantify the probability of whether or not the test data came from a normally distributed dataset.
9. P-value ✔ : tells us the likelihood that we would see similar results if we tested our data again.
10. Student’s t-test ✔ : a statistical test used to compare the mean of one dataset to another under a few assumptions.

## Question 5

As lead data engineer, you are tasked with handling the website shop orders. Every minute a new batch of orders comes in that needs to be processed. Which of the four V’s would best describe this data?

- Variety

- Volume
- Velocity ✔
- Veracity

## Question 6

You are working with a DataFrame that can contain the sales for each day for the past 5 years. The DataFrame has already been loaded into your notebook. In order to process the data, you first group the data by the date, then sum all the total sales for the day. After this you took a look at the resulting DataFrame to make sure it is correct, then filtered for dates in the past year before displaying the final result.

Which is the correct order of actions and transformations you called on the DataFrame?

- Transformation -> Transformation -> Transformation -> Transformation -> Action
- Action -> Transformation -> Transformation -> Transformation -> Action
- Transformation -> Transformation -> Action -> Transformation -> Action ✔
- Action -> Transformation -> Transformation -> Action -> Transformation

## Question 7

You have been tasked with analyzing the common voice commands used in your company's new software for home voice activation software. You have been given the data set, and you have processed it and separated the text into a list of words.

What would be the next stage in the NLP pipeline to perform?

- Tokenization
- Term Frequency-Inverse Document Frequency
- Normalization
- Stop Word Removal ✔

## Question 8

Which of the following is not true about using AWS's S3 to store your data?

- S3 allows for massive storage of data without any normalization required.
- S3 allows for data files to be accessed easily across multiple people.

- [x] S3 are public by default but can easily be made private. ✔
- [ ] S3 can be directly read into a notebook and stored into a DataFrame.

## Question 9

You're a data analyst for a school district. Based on data of high school students, such as average grade, number of missed school days, number of detentions and suspensions, and reduced lunch rate status, your task is to identify at-risk students who are predicted to drop out in the upcoming academic year.

Which of the following libraries is most directly relevant in performing your task?

- [ ] `sklearn.decomposition.FactorAnalysis`
- [ ] `sklearn.exceptions`
- [ ] `sklearn.linear_model.LinearRegression`
- [x] `sklearn.linear_model.LogisticRegression` ✔

## Question 10

How will `sklearn.model_selection.train_test_split` help you accomplish your task? Choose the best answer.

- [ ] Splitting the dependent and independent variables allows parametrization.
- [x] Splitting a dataset into training and testing sets allows validation of the machine learning model used. ✔
- [ ] Splitting the dependent variables is necessary to instantiate a machine learning model.
- [ ] Splitting the independent variables is necessary to instantiate a machine learning model.

## Question 11

In your discussion with the superintendent, she informs you that her aim is to identify as many potential drop-out candidates as possible, and to monitor them during the year. Which of the following is the most relevant validation metric for your model?

- [ ] RMSE
- [ ] Accuracy
- [ ] Precision
- [x] Recall ✔

## Question 12

You also decide to explore using other models. Which of the following observations is incorrect?

- Decision tree models can be prone to overfitting.
- A benefit of the random forest classifier is the ability to rank the importance of features.
- The choice of the kernel in SVM can yield very different results.
- SVM, unlike decision trees, does not benefit from feature scaling. ✔

## Question 13

A large amount of raw financial data has just been uploaded to your data storage. You have been tasked with priming the raw data so unsupervised learning models can be applied and the results analyzed. Which one of the following is not something you should consider?

- How can I get this data to be used to create great visualizations? ✔
- Will I be able to easily hand off this data set to other teams?
- Does the data contain excess data that we don't really need?
- Are there different types of data that ?

## Question 14

How does the K-Means algorithm determine how many clusters are made and which data points belong to them?

- The amount of clusters is given to the algorithm which then assigns data points to the cluster based on similarity or distance. ✔
- The algorithm first finds the ideal amount of clusters then randomly assigns the data points to a cluster.
- The algorithm looks at the data points based on similarity or distance then finds the optimal amount of clusters.
- The data points are broken into groups and the algorithm determines if those points should be placed in more or less clusters.

## Question 15

Which form of dimensionality reductions does Principal Component Analysis perform?

Which form of dimensionality reductions does Principal Component Analysis perform?

- (selected) Feature Extraction ✔
- Feature Elimination
- Feature Compilation
- Feature Reduction

## Question 16

Your company is looking to open a new location and you are given data sets with all of it's competitors information and are asked to cluster it to see where your demographic can best be served. Which would be the best reason to use hierarchical clustering over K-Means clustering for this data set.

- The data is not super large and does not need to be analyzed with super high performance.
- (selected) The data set lends itself to be grouped based off distance. ✔
- The data set does not make it clear how many different clusters should be known ahead of time.
- The data set lends itself to be better visualized with a dendogram instead of an Elbow Curve

## Question 17

Match the following definitions of neural network and modeling concepts with the correct term in the word bank:

1. Linearly separable ✔ : a dataset that can be separated perfectly using a straight line
2. Bucketing ✔ : reduces the number of unique categorical variables in a dataset
3. Loss metric ✔ : the model performance score based on the inaccuracy of a single input
4. ReLU activation function ✔ : returns a value from 0 to infinity, so any negative input through the function is 0
5. Perceptron ✔ : a single neuron within a neural network model

6. Sigmoid activation function ✔ : transforms the output to a range between 0 and 1. It is the same curve used in logistic regression
7. Epoch ✔ : a training iteration in Tensorflow machine learning
8. Random Forest Classifier ✔ : an ensemble learning technique that resembles neural network models in terms of structure and performance
9. Keras ✔ : module within Tensorflow used to build neural networks
10. Hidden Layer ✔ : component of a sequential model. Using one in the Sequential model creates a neural network, while using more than one creates a deep learning model.

## Question 18

Complete the following code below so that it will do the following tasks:

1. Define the deep learning classification model
2. Add the first hidden layer with a non-linear activation function
3. Add the second hidden layer with less neurons than the first hidden layer
4. Add an output layer
5. Compile the model
6. Check the model structure

```
nn model = tf.keras.models.
                          Sequential()                          ✔

nn model.add(tf.keras.layers.
        Dense(units=16, activation="relu", input_dim=8)         ✔ )

nn model.add(tf.keras.layers.
               Dense(units=8, activation="relu")                ✔ )

nn model.add(tf.keras.layers.
```

```
            Dense(units=1, activation="sigmoid")                ✔ )

nn_model.compile(loss="binary_crossentropy", optimizer="adam", metrics=
["accuracy"])                                                   ✔

nn_model.                          summary()                    ✔
```

## Question 19

The following steps are used to preprocess categorical and numerical data for use in a neural network model. Arrange the steps in the correct order:

- ≡ Check if any categorical variables require bucketing ✔
- ≡ Perform bucketing transformation on any applicable categorical variable ✔
- ≡ Encode categorical variables using `OneHotEncoder` instance ✔
- ≡ Merge the encoded categorical dataframe with the input dataframe ✔
- ≡ Scale training and testing data using fitted `StandardScalar` instance ✘
- ≡ Fit `StandardScalar` instance on training data ✔
- ≡ Split input dataframe into training and testing datasets ✘

## Question 20

The following code builds, trains, and evaluates a neural network model. Which of the following code changes is the most likely to optimize the model's performance?

```
# Define the basic neural network model
nn_model = tf.keras.models.Sequential()
nn_model.add(tf.keras.layers.Dense(units=16, activation="relu", input_dim=8))
nn_model.add(tf.keras.layers.Dense(units=1, activation="sigmoid"))
```

```
# Compile the Sequential model together and customize metrics
nn_model.compile(loss="binary_crossentropy", optimizer="adam", metrics=["accuracy"])

# Train the model
fit_model = nn_model.fit(X_train_scaled, y_train, epochs=100)
```

- [x] Change the number of neurons in the hidden layer
- [ ] Swap the training and testing datasets
- [ ] Change the number of neurons in the input layer
- [ ] Remove the output layer

Retake