When developing a course, it is first important to be clear on the direction you are taking with the course, it is recommended that you start with an outline that clearly outlines the modules, topics, objectives, assessments, activities, etc.

Aspects to consider when developing the content:

* Introductions and summaries in each module/section are required
* Examples such as modeling and demonstration(s) support skill and knowledge
* Additional Resources need to be copyright cleared (if you are unsure please contact the Program Owner or Delegate)
* Feedback strategies for assessments and activities are required

**NOTES TO SME:**

* In the following pages, anything highlighted in yellow is content that should be inputted.
* If you have not already received the Course Description and Course Learning Outcomes from the Program Owner or Delegate, please contact them for this, as you will need to adhere to these for the development of course content.

**Reminder: APA Citation**

Please, in accordance to the policy outlined in the Continuing Education course outline template, remember to cite sources you are using in APA format.

Any work that you submit in a Continuing Education course must be your own and must be created specifically for the course for which it is being submitted.

Go to the Continuing Education web page on Academic Honesty to learn more about plagiarism and the consequences of plagiarizing: <http://conted.ucalgary.ca/info/honesty.jsp>

Course work must make use of the American Psychological Association (APA) citation style to reference the work of others (or yourself). Learn more about APA here: <https://owl.english.purdue.edu/owl/resource/560/01/>

Here is a narrated presentation on referencing and how to avoid plagiarism: <http://www.ucalgary.ca/cted/breeze/referencing>

**Template Legend**

* Highlighted text – sections for SME completion
* Gray text – instructions for SME

**Course Description**

Every field of computing is being impacted by Machine Learning. This includes software engineering, data analysis, and artificial intelligence.

This course is for individuals who would like to apply hands-on experience in solving complex and simple real-world problems across different industries, including; retail, finance, tech, healthcare, etc.

In this course, you will discover how Machine Learning helps to create models that understand large amounts of data. In this course, you will learn how to use Python libraries for predictive problems (also called ‘supervised learning’) and data clustering problems (also called ‘unsupervised learning’).

You will also study fundamental and advanced machine learning techniques such as multiple linear regressions (Ridge and Lasso), generalized linear models and classification, and clustering and dimensionality reduction methods.

This course culminates with an experiential learning project called a ‘capstone project’. In this project, students will learn how to create freelance coding projects and will complete their own small freelance coding assignments

**Course Learning Outcomes**

By the completion of this course, successful students will be able to:

1. Describe how Machine Learning helps to create models that understand a large amount of data.
2. Learn how to acquire and prepare data for Machine Learning models.
3. Explore Machine Learning oriented practical applications of scientific libraries such as SciKit-Learn and Keras.
4. Identify learning applications, formulation of learning tasks as computational problems, and methods designed to solve these problems.
5. Design and implement methods for problems in pattern recognition, system identification or predictive analysis.
6. Apply Machine Learning methods and algorithms in the context of real-world problems.
7. Complete a capstone project to show proficiency with the Python programming language.

**Introduction to the Course**

Write an introduction to the course for the learners. The purpose of this section is to provide the learner with an overview of what to expect from the course in general terms. This should be different from the course outline and more informal. This is the place to explain to students the structure of the course and how the resources will be used.

In addition to this written introduction, we ask SMEs to please provide a script for a short one minute video introduction to welcome students and introduce the course. You can use your imagination in this section. If you have alternative ideas to the one-minute introduction video, please reach out to the Program Owner or Delegate.

Write Here:

In this course, you will be introduced to the concepts of machine learning. You will learn the python libraries for machine learning. You will go through classical models of machine learning. A complete application of the recommendation system will be introduced as an example of the classical machine learning. An introduction to the neural network and deep learning is given at the end of the course to give you an overview of the modern techniques of machine learning. You will complete a capstone project, in which you will apply all what you learn in the course.

In this course, you will:

* Be introduced to the concepts of machine learning.
* Be introduced the python libraries for machine learning.
* Go through classical models of machine learning.
* Be introduced to a complete application of the recommendation system as an example of the classical machine learning.
* Be given an introduction to the neural network and deep learning at the end of the course to give an overview of the modern techniques of machine learning.
* Complete a capstone project, in which you will apply all what you learn in the course.

**Course Modules**

Number and name the weeks/modules in the course (please check with the Program Owner or Delegate to determine if there are pre-defined modules).

Describe what is covered in each week/module. All course materials, learning activities, and assignments will relate to each specific week/module.

| **Date** | **Missions and Descriptions** | **Readings, Learning Activities, Assignments, etc.** |
| --- | --- | --- |
| Pre-Course Activities | Set-up your computer | Example:  Read Course Outline  View Course Orientation video |
| Module 0  (1 week) | Course Orientation | * Course Orientation – Interactive Lesson Mission * Course Outline – Interactive Lesson Mission |
| Module 1 (1 week) | Introduction to Data | * Module Overview – Interactive Lesson Mission * Introduction to Data – Interactive Lesson Mission * Read and Modify a dataset – Editor Mission * Extract information from a dataset – Editor Mission * Read and Modify a dataset – Editor Mission Score Booster * Read, modify and extract data – Editor Mission Score Booster * Introduction to Machine Learning – Quiz Mission * Introduction to Machine Learning – Quiz Mission Score Booster * Data Exploration – Interactive Lesson Mission * Extracting Information – Editor Mission * Apply Statistics to the Iris dataset – Editor Mission * Count and Extract Information – Editor Mission Score Booster * Apply more statistics on the Iris dataset – Editor Mission Score Booster * Data Exploration – Quiz Mission * Data Exploration – Quiz Mission Score Booster * Data Pre-Processing – Interactive Lesson Mission * Data Encoding – Editor Mission * Data Scaling – Editor Mission * Data Encoding – Editor Mission Score Booster * Data Scaling – Editor Mission Score Booster * Data Pre-Processing – Quiz Mission * Data Pre-Processing – Quiz Mission Score Booster * Data Visualization – Interactive Lesson Mission * Plot a Parabola – Emulator Mission * Plot independent points – Emulator Mission * Plot a line – Emulator Mission Score Booster * Scattered Points – Emulator Mission Score Booster * Data Visualization – Quiz Mission Score Booster * Data Visualization – Quiz Mission Score Booster * Module Recap– Interactive Lesson Mission * Breast Cancer Campaign! Part 1 –Accumulative Project Mission * Chocolate! – Mini-project Mission * Football! – Mini project Mission |
| Module 2 (1 week) | Classification | * Module Overview – Interactive Lesson Mission * Introduction to Classification – Interactive Lesson Mission * Classifying Iris using Logistic Regression – Editor Mission * Classifying Mushrooms using Logistic Regression – Editor Mission * Classifying Iris using Logistic Regression – Editor Mission Score Booster * Classifying Mushrooms using Logistic Regression – Editor Mission Score Booster * Introduction to Classification – Quiz Mission * Introduction to Classification – Quiz Mission Score Booster * K-Nearest Neighbors Classification – Interactive Lesson Mission * Classifying Iris using KNN – Editor Mission * Classifying Mushrooms using KNN – Editor Mission * Classifying Iris using KNN – Editor Mission Score Booster * Classifying Mushrooms using KNN – Editor Mission Score Booster * K-Nearest Neighbors using KNN – Quiz Mission * K- Nearest Neighbors using KNN – Quiz Mission Score Booster * Support Vector Machine for Classification – Interactive Lesson Mission * Classifying Iris using SVMs – Editor Mission * Classifying Mushrooms using SVMs – Editor Mission * Classifying Iris using SVMs – Editor Mission Score Booster Mission * Classifying Mushrooms using SVMs – Editor Mission Score Booster Mission * Support Vector Machine for Classification – Quiz Mission * Support Vector Machine for Classification – Quiz Mission Score Booster * Tree-based Algorithms – Interactive Lesson Mission * Classifying Iris using Decision Trees – Editor Mission * Classifying Mushrooms using Random Forest – Editor Mission * Classifying Iris using Random Forest – Editor Mission Score Booster * Classifying Mushrooms using Decision Trees – Editor Mission Score Booster * Tree-based Algorithms – Quiz Mission * Tree-based Algorithms – Quiz Mission Score Booster * Module Recap– Interactive Lesson Mission * Breast Cancer Campaign! Part 2 –Accumulative Project Mission * Credit Risks – Mini-project Mission * Medial Appointments – Mini project Mission |
| Module 3 (1 week) | Regression | * Module Overview – Interactive Lesson Mission * Introduction to Regression – Interactive Lesson Mission * Estimating house prices using Linear Regression – Editor Mission * Regression Iris – Editor Mission * Estimating house prices using Linear Regression – Editor Mission Score Booster * Regression Iris – Editor Mission Score Booster * Introduction to Regression – Quiz Mission * Introduction to Regression – Quiz Mission Score Booster * Regression using SVMs – Interactive Lesson Mission * Estimating house prices using SVMs – Editor Mission * Iris regression using SVMs – Editor Mission * Estimating house prices using SVMs – Editor Mission Score Booster Mission * Iris regression using SVMs – Editor Mission Score Booster Mission * Regression using SVMs – Quiz Mission * Regression using SVMs – Quiz Mission Score Booster * Decision Tree Regression – Interactive Lesson Mission * Predicting house prices using Decision Trees – Editor Mission * Iris Regression using Random Forests – Editor Mission * Predicting house Prices using Random forests – Editor Mission Score Booster * Iris regression using Decision trees – Editor Mission Score Booster * Decision Tree Regression – Quiz Mission * Decision Tree Regression – Quiz Mission Score Booster * K-Nearest Neighbors regression – Interactive Lesson Mission * Predicting house prices using KNN – Editor Mission * Iris regression – using KNN – Editor Mission * Predicting house prices using KNN – Editor Mission Score Booster * Iris regression – using KNN – Editor Mission Score Booster * K-Nearest Neighbors Regression – Quiz Mission * K-Nearest Neighbors Regression – Quiz Mission Score Booster * Module Recap– Interactive Lesson Mission * Breast Cancer Campaign! Part 3 –Accumulative Project Mission * Avocados – Mini-project Mission * Product Demand Forecasts – Mini project Mission |
| Module 4 (1 week) | Clustering | * Module Overview – Interactive Lesson Mission * Introduction to Clustering – Interactive Lesson Mission * Introduction to Clustering – Quiz Mission * Introduction to Clustering – Quiz Mission Score Booster * K-Means Clustering – Interactive Lesson Mission * Clustering Iris using K-Means – Editor Mission * Clustering Mushrooms using K-Means – Editor Mission * Clustering Iris using K-Means – Editor Mission Score Booster * Clustering Mushrooms using K-Means – Editor Mission Score Booster * K-Means Clustering – Quiz Mission * K-Means Clustering – Quiz Mission Score Booster * Mean Shift Clustering – Interactive Lesson Mission * Clustering Iris using Mean Shift – Editor Mission * Clustering Mushrooms using Mean Shift – Editor Mission * Clustering Iris using Mean Shift – Editor Mission Score Booster * Clustering Mushrooms using Mean Shift – Editor Mission Score Booster * Mean Shift Clustering – Quiz Mission * Mean Shift Clustering – Quiz Mission Score Booster * Module Recap– Interactive Lesson Mission * Breast Cancer Campaign! Part 4 –Accumulative Project Mission * Wholesales – Mini-project Mission * Image Compression – Mini project Mission |
| Module 5 (1 week) | Recommendation System | * Introduction to Recommendation Systems – Interactive Lesson Mission * Recommender System – Interactive Lesson Mission * Popularity Recommender System – Editor Mission * Content based Recommender System – Editor Mission * Popularity Recommender System – Editor Mission Scale Booster * Content based Recommender System – Editor Mission Score Booster * Collaborative Filtering – Interactive Lesson Mission * Collaborative based Recommender System – Editor Mission * Collaborative based Recommender System – Editor Mission Score Booster * Good Reads Recommender System – Mini-project Mission |
| Module 6 (1 week) | Introduction to Deep Learning | * Module Overview – Interactive Lesson Mission * Understanding Artificial Neural Networks – Interactive Lesson Mission * Understanding Artificial Neural Networks – Quiz Mission * Understanding Artificial Neural Networks – Quiz Mission Score Booster * Classification and Regression using ANN – Interactive Lesson Mission * Build an ANN Classifier – Editor Mission * Artificial Neural Network – Editor Mission * Build an ANN – Editor Mission Score Booster * ANN Regressor – Editor Mission Score Booster * Convolutional Neural Networks – Interactive Lesson Mission * Convolutional Neural Networks – Quiz Mission * Convolutional Neural Networks – Quiz Mission Score Booster * Image Recognition – Interactive Lesson Mission * Digits Identification – Editor Mission * Categorical Identification – Editor Mission Score Booster * Understanding Recurrent Neural Networks – Interactive Lesson Mission * RNN – Quiz Mission * RNN – Quiz Mission Score Booster * Natural Language Processing – Interactive Lesson Mission * Shakespeare – Editor Mission * IMDB Reviews – Editor Mission Score Booster * Module Recap– Interactive Lesson Mission * Breast Cancer Campaign! Part 5 –Accumulative Project Mission * Fruit Identification – Mini-project Mission * Tweet Sentiment Analysis– Mini project Mission |
| Module 7 (1 week) | Capstone Project | * Diamonds Predictor Application – Capstone Project |
| Final Course Task(s) | Submit the Course Evaluation Survey | Take 15 minutes to fill out the anonymous course evaluation survey that has been emailed to you.  Note: Instructors do not see the results of the survey until final grades have been submitted. |

IMPORTANT NOTE FROM LS

**Course Content – Module 1 (Name: Introduction to Data)**

**Course Objective(s) addressed in Module 1**

Please indicate which course objective(s) this module is addressing. See page 2 for details.

1. Describe how Machine Learning helps to create models that understand a large amount of data.
2. Learn how to acquire and prepare data for Machine Learning models.
3. Explore Machine learning oriented practical applications of scientific libraries such as SciKit-Learn and Keras.

**Module 1 Objective(s) Are the module objectives the “missons”**

Please indicate which module objectives (missions) that this module addresses. These go a step further in detail than the course objectives. Ie. Think of a tree, the trunk are the course objectives and the branches are the module objectives.

* Explain the basic concepts and terminology used in machine learning.
* Identify the different types of machine learning problems.
* Identify the steps required to design a machine learning model.
* Acquire and read various types of datasets from different sources.
* Analyze and visualize data.
* Prepare and preprocess data to be ready for machine learning models.

**Module 1 Introduction**

Write an overview statement here and indicate how many missions are in this module, as well as the structure of the module.

Write Here:

Machine Learning is an application of Artificial Intelligence that enables systems to learn existing patterns in the data automatically and improve themselves based on experiences that they have gained. In this module, you will learn how to use machine learning to obtain new and hidden information from your dataset.

This module contains six submodules and three mini-projects. Each submodule is comprised of a reading mission and three application missions. In each section of the reading mission (also referred to as an interactive lesson mission) you will encounter some knowledge checks. The knowledge checks are short quizzes based on the content which you read. You are expected to provide an answer to each knowledge check question in order to proceed to the next section of the reading mission. The knowledge checks are assessed automatically, and an explanation is provided for the answers.

There are three types of application mission which you will encounter in this module: Quiz, Emulator and Editor.

A quiz mission is used to test your knowledge of a lesson. You will be asked an average of four questions. For each question, you are expected to select the correct answer from several options. In order to evaluate your understanding of the lesson, you are expected to also specify your confidence level about the answer you selected.

An editor mission will allow you to practically apply the knowledge acquired from a lesson. The editor is divided into 3 sections. The first section contains the question you are expected to solve. The second section contains a box where you can type-in your codes. Finally, the last section is where the result of the codes typed will be displayed.

An emulator mission is very similar to quiz mission. An emulator is used to test your understanding of the lesson through a feature which allows you to view the result of your chosen answer in comparison to that of the correct answer.

Based on your performance in these application missions, you might expect to take another set of application missions called Score Booster Missions. The score booster missions are there primarily to help you boost your overall performance for that module.

Finally, there are three mini-projects required for this module: two Application and one Through Mini-project. The Application Mini-project provides an exercise based on the topics covered in this module. In the Through Mini-project, you will be solving a piece of a bigger project which relates to this module.

For more information on the missions and mini-projects and their grading, please refer to the course Orientation through this link. Also, for information about the GIT naming convention for the mini-projects, please refer to the GIT page through this link. You will be required to set up a Github account and install it on your computer. The link provided will walk you through the installation.

**Before starting any coding please install Anaconda. Anaconda** is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.).

Anaconda is the distribution you will be using throughout this course for development. See The installation Manual Here:

For any information regarding the legends please refer to the Orientation of this course.

Box blue: is for defining a term

Box yellow: is for giving notes, hint or important information.

Box green: is for giving knowledge and regular information.

# Submodule 1. Introduction to Machine Learning

## **Submodule 1 Activities Summary:**

List the activities learners will be completing during the module, and in the sequence they will be the stated module objectives listed above. The following list displays examples of activities that may be included. If you have questions about the activities you would like to include in the course, please speak to Learning Services.

Missions in Submodule 1:

* Interactive Lesson mission
* Quiz Mission
* Editor Mission

**Submodule 1 Essential Question (Purpose):**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

What is Machine learning? What applications machine learning has on our day to day lives?

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Develop Submodule 1 Content**

Repeat this section for every mission.

Start explaining, based on the course structure, the objectives to be covered. You may do this in a table if you find it beneficial. In this section, you will develop content. This section is where you use your knowledge to build connection pieces between theory, resources, and assessments. Use the module objectives as the base for the structure of your content. This section is not focused on curation of resources; it is focused on you, as the SME, providing your expertise to create content. However, if you are referencing any type of source, please cite it in APA format (including open source materials and creative commons material).

Write Here:

Mission 1: Interactive lesson mission: Introduction to Machine learning

Section 0: Introduction

|  |
| --- |
| What will you learn in this lesson? |
| In this lesson, You will be introduced to Machine Learning. By the end of this lesson, you understand:   1. What is Machine learning? 2. How did Machine learning start? 3. Types of Learning. 4. Modern real-life applications of Machine learning. |

Section 1: Siri as an example of machine learning

|  |
| --- |
| Siri |
| You are probably familiar with Siri, the virtual assistant in any IPhone. You can verbally ask Siri questions and she understands what you are saying, responds with an action and sends feedback. How does that happen? |
| Citation:  Siri [Digital Image]. Retrieved from https://support.apple.com/ |











Section 02: Machine Learning Overview

In this section, you will be introduced to machine learning and its history.

|  |
| --- |
| Machine Learning |
| Machine Learning is one of the fields falling under the idea of Artificial intelligence which aims directly at understanding different types of data including that data’s structure and distinct features in order to fit this data to models that could be easily understood by people.  Machine learning is a subset to Artificial Intelligence aiming at automating the process of decision making using input features and attributes and mapping them to the target dependent variable.  Machine learning’s main aim is creating machines that can take directions and learn from their mistakes; additionally, human beings using statistical methods as an approximation for the actual decision-making process that takes place in our minds all the time. |
|  |

**Citation: RoboGarden Course Resources** Copyright 2019

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Fill in the Missing Blank | | | | | | |
| Machine learning is considered a | | Dropdown | | | To Artificial intelligence. | |
| Sub-set | | |
| Super-set | | |
| Alternative | | |
| Opposite | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Machine learning is considered a sub-set of Artificial intelligence. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Revise the definition of Machine learning. | |  |
|  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| History of Machine Learning  Since the early 1940s, scientists started to introduce the idea of using statistical analysis in creating a decision-making model which mimics the human brain.  In this section, you will get acquainted with the steps of the evolution of machine learning through the years. Click through the carrousel below to learn more about the evolution of machine learining. | | | | |
|  |  | 1940s  In the early 1940s, a mathematician named Walter Pitts  and a neurophysiologist named Warren McCulloch  introduced the idea of creating a mathematical model for  what they termed a neuron and described the working of such.  The idea was simple, an electric circuit that fires  when a condition is satisfied. |  |  |
|  |  |     |  |  |

Text Carousel Content Table

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | 1940s  In the early 1940s, a mathematician named Walter Pitts  and a neurophysiologist named Warren McCulloch  introduced the idea of creating a mathematical model for  what they termed a neuron and described the working of such.  The idea was simple, an electric circuit that fires  when a condition is satisfied. |
|  | 1950s  During the 1950s, a scientist by the name of Alan Turing created a test where a machine tried to convince a human that the machine was not a machine.  Later that year, Arthur Samuel created a program  which knew how to play a game named checkers.  These two scientists have helped Frank Rosenblatt to create  the first artificial neural network in 1958.  Parallel to these efforts, different scientists had been working on decision making algorithms based on different statistical methods. One of the most important algorithms developed was the tree based algorithms. Developed in 1959, a tree based algorithm  was developed to use entropy and information theory to make decisions. |
|  | 1980s – 1990s  In the early years of the 1980s, a scientist by the name of Hopfield suggested creating a bidirectional neural network in a manner that is considered quite similar to how human brain works. This helped Stanford researchers in creating an algorithm  that allowed creating a neural network like architecture which they called the slow learners.  In the 1990s, the field had started to prosper when an idea that helped in making a machine learn patterns faster called the backpropagation was first proposed.  In 1997, an IBM computer was able to beat the world champion in chess. |
|  | Currently, machine learning has become one of the most promising fields of science. Computers are able to identify objects, sounds, videos and process structured data faster and better predict patterns more accurately than a human being.  You will learn more about the applications of machine learning later in this lesson. |
|  | citation   * Keith, D. (2019, March 26). A brief history of machine learning. Retrieved from [https://www.dataversity.net/a-brief-history-of-machine-learning](https://www.dataversity.net/a-brief-history-of-machine-learning/) * Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial intelligence: a modern approach (2nd ed.). Prentice Hall. ISBN 978-0137903955 * Sohrob K. (2018, March 15). Alan Turing and The birth of machine learning. Retrieved from https://www.vectra.ai/blogpost/alan-turing-and-the-birth-of-machine-intelligence |

Section 03: Machine Learning Problem Types

|  |  |
| --- | --- |
| Supervised, Unsupervised and Reinforcement | |
| The problems tackled by machine learning are mainly divided into three main definitions of problems.  Scientists have utilized machine learning after characterizing the difference between:   1. Supervised learning problems 2. Unsupervised learning problems. 3. Reinforcement learning problems. |  |

**Citation: RoboGarden Course Resources** Copyright 2019

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *The Learning Phase in Machine Learning*  In this section, you will learn about different learning techniques.  Click through the carrousel below to learn more about Learning approaches in Machine learning | | | | |
|  |  | Citation: Supervised-Learning [Digital Image]. (2018). Retrieved from https://medium.com/@canburaktumer/machine-learning-basics-with-examples-part-2-supervised-learning-e2b740ff014c |  |  |
| *Supervised learning*  Supervised learning is learning when a student is provided with the labels of the data samples. Consider a problem where the category of an article is needed to be predicted using machine learning models. Supervised learning is training your model on data samples with known target responses that are target classes or even numeric values.  The learning algorithm dictates that training on samples of labeled data will help the model fit the data-points after training. The model is then used to predict the target responses of new Unlabeled data.  This method helps the model during training to **learn** by comparing the correct outputs it already knows to the outputs predicted to modify model and eventually be able to induce correct values for outputs.  The purpose of this method is for the algorithm to be able to “learn” by comparing its actual output with the “taught” outputs to find errors, and modify the model accordingly.  The Supervised learning was called supervised learning since it almost mimics the concept of human learners learning a certain subject under the supervision of some more knowledgeable learners. The knowledgeable learners can give the students examples for him to formulate general rules for predicted new examples.  Examples:   * Classification * Regression   Citation: Supervised-Learning [Digital Image]. (2018). https://medium.com/@canburaktumer/machine-learning-basics-with-examples-part-2-supervised-learning-e2b740ff014c | | | | |
|  |  |    |  |  |

Click through the carrousel below to learn more about different learning methods.

Image Carousel Content Table

|  |  |  |
| --- | --- | --- |
| **#** | **Image Description** | **Image** |
|  | Supervised learning  Supervised learning is learning when a student is provided with the labels of the data samples. Consider a problem where the category of an article is needed to be predicted using machine learning models. Supervised learning is training your model on data samples with known target responses that are target classes or even numeric values.  Assume you have images of a definite type of flowers, A supervised approach is to train a machine learning model on these images and then if a new image of these species of flower, the machine learning model should be able to identify based on the labeled data it has seen in the learning phase.  The learning algorithm dictates that training on samples of labeled data will help the model fit the data-points after training. The model is then used to predict the target responses of new Unlabeled data.  The Supervised learning was called supervised learning since it almost mimics the concept of human learners learning a certain subject under the supervision of some more knowledgeable learners. The knowledgeable learners can give the students examples for him to formulate general rules for predicted new examples.  Examples:   * Classification * Regression | Citation: Supervised-Learning [Digital Image]. (2018). Retrieved from <https://medium>.com/@canburaktumer/machine-learning-basics-with-examples-part-2-supervised-learning-e2b740fff014c |
|  | **Unsupervised learning**  Unsupervised learning is the type of learning where the labels of the training data are unknown. No target responses are added to the examples added or the data samples before training.  Unsupervised learning aims at recognizing different patterns in data and use this information to construct an output different from the input data.  The algorithm aims at determining patterns in data and different measures of similarities between data samples. It mimics the idea of a slow learner trying to notice main differences between two different classes of objects with no supervisor to give him interpretation for the different classes of examples the learner sees.  Examples:   * Clustering | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Reinforcement learning  Reinforcement learning is the type of learning which is mainly accompanied by making decisions based on a cost and reward system. The data in reinforcement learning is not labeled but accompanied with some feedback.  Some of the decisions the training model would make have negative feedback – known as the cost of the action or the loss- which helps the model in avoiding doing the same mistakes multiple times. In the same manner, a reward system could be used. Some of the decisions the training model would take will be accompanied by a reward pushing the model to learn the best patterns which is more likely to succeed and avoid patterns that would most likely fail. | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  A machine learning model is aimed at studying 20000 examples of cat images then decide on new images whether it is a cat image or not. What type of learning problems is that example? | | | | | | |
|  |  | a. | Unsupervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Semi-supervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Supervised learning. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Reinforcement learning. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The initial cat images are already labeled which means this is an example of supervised learning problems. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Notice that the initial cat images’ labels are known in advance. | |  |
|  |  |  |  |

Section 04: Applications

In this section, you will see some application of the machine learning in real life.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Machine learning real-life applications:  Machine learning has become one of the most promising technologies and one of the closest to the public as a result of the wide variety of applications which use artificial intelligence currently.  Machine learning has created an opportunity to create applications in which the computers could predict outcomes of a certain problem automatically with accuracy higher than that of a human expert.  In this section, you will get acquainted with some of the major applications using machine learning  Click through the carrousel below to learn more about real life applications | | | | |
|  |  |  |  |  |
| Citation: RoboGarden Course Resources Copyright 2019  Personal Assistants  You are probably familiar with Siri, Alexa or Google by now. These three are called virtual personal assistants controlled via voice commands to do different types of mobile activities such as display schedules, update information or even call a friend.  These personal assistants are one example of voice-recognition using machine learning and then using the words said by the person to create an action. | | | | |
|  |  |    |  |  |

Image Carousel Content Table

|  |  |  |
| --- | --- | --- |
| **#** | **Image Description** | **Image** |
|  | Image Recognition  Image recognition is considered one of the main fields that are strongly supported by artificial intelligence and machine learning techniques. Detecting different objects in images or videos, Tracking the movement of objects in visual data or predicting the next image or shot in video are all examples of things you could do with machine learning.  Examples:  Face Detection: Machine learning can be used to detect faces in images and predict who these images belong to.  Hand writing recognition is another application where the user could use machine learning to let the program understand different hand-written digits and letters and eventually read the human handwriting. | Citation: RoboGarden Resources Copyright 2019 |
|  | Speech Recognition  Speech recognition is the recognition of voiced out words in audio data and the ability to turn this data into text. Speech signals are interpreted to some set of numbers which are accordingly used to create datasets of audio data.  Machine learning helps in classifying the spoken words, clustering the same voices together or text recognition.  Speech recognition applications include:   * Speech operated locks. * Automatic speech translators. | Citation: RoboGarden Resources Copyright 2019 |
|  | Medical Diagnosis  Machine learning has conquered the field of medical research and the biomedical engineering since its applications are countless. Machine learning is not just the best analysis tool to be used in the medical field but also Machine learning can detect patterns in symptoms to actively diagnose a patient based on some of information about the patient.  Machine learning is also researched to be of help to the experts in the medical field as well. Creating smart machines to effectively monitor, alarm and diagnose patients would improve the efficiency of the medical healthcare system. | Citation: RoboGarden Resources Copyright 2019 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Which of these are machine learning applications?  Select all that apply. | | | | | | |
|  |  | a. | Camera that tracks the pattern of movement of cars in a certain area. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Predicting stock market price movements. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Creating database to retrieve data of students. | | |  |
|  |  | | |  |  |  |
|  |  | d. | A Matrix Calculator which calculates the determinant of a given matrix. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

MCQ Responses Table

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Correct | Great! This is an example of a machine learning application. The machine needs to learn since we cannot write a sequential program to solve the problem without the study of the data. |
| b. | Correct | Great! This is an example of a machine learning application. The machine needs to learn since we cannot write a mathematic equation to give the predication without the study of the data. |
| c. | Incorrect | Unfortunately, this is incorrect. This is not a machine learning application. The machine does not need to learn since it is a programmatically solvable problem without the study of the data. |
| d. | Incorrect | Unfortunately, this is incorrect. This is not a machine learning application. The machine does not need to learn since it is a mathematically solvable problem without the study of the data. |

Section 06: Lesson Summary

|  |
| --- |
| What did we cover in this lesson? |
| Great work! You were introduced to the definition of Machine learning, History of machine learning and the Real-life applications tackled by machine learning.  Machine learning is a subset to Artificial Intelligence aiming at automating the process of decision making using input features and attributes and mapping them to the target dependent variable. Machine learning problem types include supervised, unsupervised and reinforcement learning and is used in various fields including medicine, technology, marketing, sales, etc. |



## **Submodule 1 Application Mission(s) Details**

Repeat this section for every mission. The activities listed below are examples. See page 7 for the activities summary.

Below you will see the details of all the application missions within the submodule 1.

Mission 2 – graded Knowledge check: Quiz Mission - Introduction to Machine Learning

Write this to student (Written Only One time)

Quiz is used to test learners’ knowledge of the lesson which they just completed. The learners are asked an average of 4 questions. For each question, they are presented with several options and are expected to select the correct answer. Before selecting the correct answer, they need to specify their confidence level about the answer they selected.

**Mission 2 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Understanding Machine Learning types and Applications

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 2 Instructions:**











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A machine learning model is aimed at identifying the presence of a child in an image based on previously learned images of children. What type of learning problems is that example? | | | | | | |
|  |  | a. | Unsupervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Semi-supervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Supervised learning. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Reinforcement learning. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The initial images are already labeled which means this is an example of supervised learning problems. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Notice that the initial images’ labels are known in advance. | |  |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A machine learning model is aimed at using the similarities in 10000 articles to group articles of the same topic together. What type of learning problems is that example? | | | | | | |
|  |  | a. | Supervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Semi-supervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Unsupervised learning. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Reinforcement learning. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The articles have no labels which leads the algorithm to recognize similarities in the data and cluster similar data together. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Notice that the articles have no labels. | |  |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which of these are machine learning applications?  Select all that apply. | | | | | | |
|  |  | a. | Camera that tracks the pattern of movement of cars in a certain area. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Predicting stock market price movements. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Creating database to retrieve data of students. | | |  |
|  |  | | |  |  |  |
|  |  | d. | A Matrix Calculator which calculates the determinant of a given matrix. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

MCQ Responses Table

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Correct | Great! This is an example of a machine learning application. |
| b. | Correct | Great! This is an example of a machine learning application. |
| c. | Incorrect | Unfortunately, this is incorrect. This is not a machine learning application. |
| d. | Incorrect | Unfortunately, this is incorrect. This is not a machine learning application. |

Mission 3 – Quiz Mission Score Booster : Introduction to Machine Learning

**Mission 3 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Understanding Machine Learning Applications and History  
For more information regarding the grading of the Booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 3 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| The idea of a neuron was introduced in ….. | | | | | | |
|  |  | a. | 21st century. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 1960s. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | 1940s | | |  |
|  |  | | |  |  |  |
|  |  | d. | None of the above. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. The neuron was not introduced in the 21st century. |
| b. | Incorrect | Unfortunately, this is incorrect. The neuron was not introduced in the 1960s. |
| c. | Correct | Great! In the early 1940s, a mathematician and a neurophysiologist by the names of Walter Pitts and Warren McCulloch have introduced the idea of creating a mathematical model for what they called a neuron. |
| d. | Incorrect | Unfortunately, this is incorrect. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A machine learning model is aimed at making a videogame play itself. What type of learning problems is that example? | | | | | | |
|  |  | a. | Supervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Semi-supervised learning. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Reinforcement learning. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Unsupervised learning. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Reinforcement learning is the type of learning which is mainly accompanied by making decisions based on a cost and reward system. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Please read “Section 03: Machine Learning Problem Types” again for different type of learning. | |  |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Detecting handwritten digits is an example of what? | | | | | | |
|  |  | a. | Personal assistants. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Machine learning in medicine. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Image recognition. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Sequence recognition. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Hand writing recognition is an application of image recognition where the user could use machine learning to let the program understand different hand-written digits and letters and eventually read human handwriting. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Remember that the digits are given to us as an image. | |  |
|  |  |  |  |

# Submodule 2. Introduction to Data

## **Submodule 2 Activities Summary:**

List the activities learners will be completing during the module, and in the sequence, they will be completed. Please indicate which activities are graded, and which ones are not. All activities must match the stated module objectives listed above. The following list displays examples of activities that may be included. If you have questions about the activities you would like to include in the course, please speak to Learning Services.

1. Interactive lesson mission (Complete reading of the lesson is graded, and this is confirmed by finishing the informal assessment inside the lesson)
2. Quiz mission (Graded)
3. Editor mission (Graded)
4. Emulator mission (Graded)

**Submodule 2 Essential Question (Purpose):**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

How do you use Pandas library and Scikit-learn library to interpret and preprocess data? These Libraries have some built-in dataset and function that you can use to manipulate data. You can use any dataset not only built-in ones.

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation and Anaconda installation guides (See The Installation Manual Here). **Develop Submodule 2 Content**

Repeat this section for every mission.

Start explaining, based on the course structure, the objectives to be covered. You may do this in a table if you find it beneficial. In this section, you will develop content. This section is where you use your knowledge to build connection pieces between theory, resources, and assessments. Use the module objectives as the base for the structure of your content. This section is not focused on curation of resources; it is focused on you, as the SME, providing your expertise to create content. However, if you are referencing any type of source, please cite it in APA format (including open source materials and creative commons material).

Write Here:

Mission 1 - Interactive Lesson Mission: Introduction to Data

Section 0: introduction

|  |
| --- |
| What will you learn in this lesson? |
| In this lesson, You will be introduced to Datasets. By the end of this lesson, you will know about:   1. Different dataset structures. 2. Scikit-learn built-in datasets. 3. Formal introduction to Pandas package. 4. Various Dataset Examples. |

Section 01: Overview

In this section, you will learn what data is and how to use it.

|  |  |
| --- | --- |
| Do you recognize data? | |
| Everything around is distinguished by data. Data takes information we gain and puts it into proper context or in analysis.  The words data and information are used interchangeably. Data is what makes information. | Citation: RoboGarden Course Resources Copyright 2019 |

Next are some important definitions about data and attributes. Remember the blue boxes indicate definitions. To know more about the meaning of each boxes’ color please check the Orientation of the course.

|  |  |  |
| --- | --- | --- |
|  |  | Data  Data is a group of attributes that characterizes different objects. It refers to different sets of attribute values that give meaning to different quantitative or qualitative variables. |

|  |  |  |
| --- | --- | --- |
|  |  | Attribute  An attribute is a certain property of an object. A set of attributes defines any data object. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute types  There are multiple types of attributes.  Click through the carousel to learn more about different attribute types. | | | | |
|  |  | Nominal Attributes  Nominal attributes are either discrete or categorical attributes. These define an object and cannot be ordered.  **Example:**   * Eye Colour * Zip Codes |  |  |
|  |  |     |  |  |

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Nominal Attributes  Nominal attributes are either discrete or categorical attributes that define an object and cannot be ordered.  **Example:**   * Eye Colour * Zip Codes |
|  | Ordinal Attributes  Ordinal attributes are types of categorical data that have natural order or rankings.  **Example:**   * Movie ratings (1 – 5) |
|  | Interval Attributes  Interval attributes are the type of attributes which differentiate between values.  **Example:**   * Temperature * Calendar Dates |
|  | Ratio Attributes  Ratio attributes have all the properties of the interval attribute in addition to the accurate definition of the 0.0 value.  **Example:**   * Weight * Height |

|  |  |  |
| --- | --- | --- |
|  |  | Difference Between Discrete and Continuous attributes;  **Discrete**: Attributes with only finite sets of values  **Continuous:** Attributes with real numbers as real values |



Section 2: Introduction to Datasets

Now it is the time to know more about Datasets.

|  |  |  |
| --- | --- | --- |
|  |  | Dataset  Datasets are large collections of data points that are commonly used for analysis, study, or visualizing data. |

|  |  |  |
| --- | --- | --- |
|  |  | Public Datasets  Several organizations have disclosed their collected data to the public for access, modification or reuse.  The aim is to help the community in contributing to the technology of today and use the data to improve different machine learning techniques, data mining techniques or data visualization techniques. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Types of Datasets  There are different types of collections for data points; click through the carousel to learn more about them: | | | | |
|  |  |  |  |  |
| Citation: RoboGarden Course Resources Copyright 2019  Ordered Data  Data collected in a sequential manner over an interval of time.  **Example:**  Spatial-temporal data: shipping movements across a geographic area over time.  Click through the carrousel below to learn more about ordered data.  Citation: RoboGarden Course Resources Copyright 2019. | | | | |
|  |  |    |  |  |

|  |  |  |
| --- | --- | --- |
| **#** | **Image Description** | **Image** |
|  | Tabular Data  Collection of entries where each entry consists of several attributes to distinguish it.  Example:   * Banking data. | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Graph Data  Data distributed in nodes and edges.  Example:  Each person in Facebook is a node and is connected with edges to his/her friends. | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Ordered Data  Data collected in a sequential manner over an interval of time.  **Example:**  Spatial-temporal data: shipping movements across a geographic area over time. | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Which type of dataset does the data in a Bank Database belong to? | | | | | | |
|  |  | a. | Graph Data | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Ordered Data | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Tabular Data | | |  |
|  |  | | |  |  |  |
|  |  | d. | Matrix Data | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! A Bank structured database belongs to tabular datasets. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect. A Bank structured database is represented mainly by different attributes or columns and multiple rows of different clients or transactions. | |  |
|  |  |  |  |

Section 3: Libraries

In this section, you will be introduced to the concept of libraries and further you will learn how to use some of the famous python libraries such as Scikit-learn and Pandas. These libraries contain some built-in datasets that let you start exploring the commands without needing to download a dataset from an outside source. You can also download any interesting dataset from an outside source and start manipulating and processing the dataset using Scikit-learn and Pandas Library. Using the Libraries function let you use very complex codes without needing to write them from scratch.

|  |  |
| --- | --- |
|  | Python Libraries  Python library is a collection of functions and methods that allows you to perform many actions on a given dataset without writing your code. |

|  |
| --- |
| Scikit-Learn Library  Scikit-learn is a free software machine learning library for the Python programming language. It features various implementation to machine learning algorithms and data preprocessing techniques.  Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.  The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as ease of use, code quality, collaboration, documentation and performance. |
| The Scikit-learn library also provides some built-in datasets that can be used for machine learning experimentation. |
| Citation: Scikit-Learn [Digital Image]. Retrieved from https://scikit-learn.org/ |

|  |  |  |
| --- | --- | --- |
|  |  | **Notes about ScikitLearn**  To learn more about available datasets in Scikit-learn and the How to of working with datasets using Scikit-learn, see https://scikit-learn.org/stable/datasets/index.html |

Example Dataset

|  |  |
| --- | --- |
| The Digits Dataset | |
| The digits dataset consists of images of handwritten digits from 0 to 9. This dataset is used for training classification models to recognize handwritten digits. | Citation: RoboGarden Course Resouces Copyright 2019 |

Using Scikit-learn datasets to read a specific dataset:

In this section, you will learn how to load digit datasets from the Scikit-learn dataset and how to set the output of your load function and then read information regarding the size of the dataset. Then you will learn how to visualize the data in the “digit dataset” using the Matplotlib Library from Python. All these instructions happen in five steps that you need to enter into an editor such as Jupyter notebook or Spider to be able to see the results as it is shown here. To be able to develop any codes please make sure you have Anaconda installed on your computer (See The installation Manual Here).

|  |  |  |
| --- | --- | --- |
| 1. Loading the Dataset | | |
| To load the digits dataset, use the load\_digits function from the sklearn.datasets library. | | |
|  |  |  |
|  | from sklearn.datasets import load\_digits dataset = load\_digits() |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Input and Target Sets | | |
| You can use the return\_X\_y=true parameter to tell the function to return the dataset as x and y instead of a dictionary-like dataset. | | |
|  |  |  |
|  | from sklearn.datasets import load\_digits (x,y) = load\_digits(return\_X\_y=True) |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Show Dataset Size | | |
| The output of this line of code is (1797, 64), which means that the dataset has 1797 samples, and each has a dimension of 64 (each image is 8 by 8 pixels). | | |
|  |  |  |
|  | print(dataset.data.shape) |  |
|  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Showing a sample image. | | | |
| The following line of code simply shows you an example of the images in this dataset.  Note:  You will be introduced to visualization techniques and Matplotlib library later in this module. | | | |
|  |  | |  |
|  | import matplotlib.pyplot as plt plt.imshow(dataset.images[400]) | |  |
|  |  | |  |
| 1. Showing output | | | |
| The output is equal to 4   |  |  |  | | --- | --- | --- | |  |  |  | |  | print(dataset.target[400]) |  | |  |  |  | | |  | |
| Citation: RoboGarden Course Resources Copyright 2019 | |  | |











Section 04: Introduction to Pandas Library

In this section, you will read about Pandas library and you will learn how to use it to manipulate data.

Before start with Pandas Library, Let us recall what the definition of a Library is:

|  |  |
| --- | --- |
|  | Python Libraries  Python library is a collection of functions and methods that allows you to perform many actions on a given dataset without writing your code. |

|  |
| --- |
| What is Pandas?  Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.  The name pandas is derived from the term “Panel data” defined as the observation of the same data samples over a period of time. |
| The Pandas library is considered the number one tool in Data Science and Data Analysis. Pandas and dealing with datasets Pandas library is one of the main tools to manipulate datasets as a prior step to training a machine learning model.  Pandas helps the user in:   * Exploring datasets. * Describing meta-data of different attributes in datasets. * Operations on the dataset. * Dataset Visualization. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pandas Core Components  Pandas has two main components:  Click through the carousel to learn more about Pandas Core components. | | | | |
|  |  | Series  The Pandas series is simply a column. It is a one-dimensional array with axis labels.  A Pandas series has a name. |  |  |
|  |  |     |  |  |

Text Carousel Content Table

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Series  The Pandas series is simply a column. It is a one-dimensional array with axis labels.  A Pandas series has a name. |
|  | Data Frame  The Pandas data frame is a multi-dimensional table made up of a collection of Series.    Citation: DataFrames [Digital Image]. Retrieved from https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/ |

|  |  |  |
| --- | --- | --- |
|  |  | Notes about Pandas and Numpy  Only Pandas and Numpy Library are similar regarding having components. Other Libraries such as Scikit Learn has no basic components. |

For going through the instructions, you will need to have Anaconda installed on your system and then install libraries such as Pandas using the Anaconda environment. You will find the instructions for installing Anaconda and Libraries Here. When the installation is complete you can use any editor such as Spider and Jupyter notebook to go through the instructions of this section. The instructions will help you to read a dataset in CSV format and manipulate its information.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Using Pandas  Pandas is considered an easy tool for data scientists to manipulate, explore, visualize and process data. In step 1, you will see how to install a Library in the Anaconda environment. For more information, you can see the Anaconda installation guide Here. | | | | | |
|  | Step 1 | Installing Pandas  To install the Pandas library on your machine, type the following command on your terminal.  pip install pandas | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Importing Pandas  After installing the Pandas library, you can now import it in a python script to use its functionalities.  In this lesson, Pandas will be explicitly imported with the abbreviated name of pd as in the following command  Import pandas as pd | | |  |
|  |  | |  |  |  |
|  | Step 3 | Reading CSV Files  CSV is short for Comma Separated Values. A CSV file is a text file that contains data in a comma-separated format.  Each line represents a row, and columns of each row are separated by commas.  In Pandas reading a CSV file can be easily done using the function pd.read\_csv()  Assume we have the shown dataset stored in a CSV file. The data represents high school GPA versus university GPA for six students. The data could be read and stored in a Pandas data frame called df using the following line of code:    df = pd.read\_csv(‘gpa.csv’) | | |  |
|  |  | |  |  |  |
|  | Step 4 | Viewing the data  View the dataframe to view the data. | | |  |
|  |  |  |  |  |  |
|  | Step 5 | DataFrame.head()  Pandas head() function fetches the data samples at the top of the dataframe. | | |  |
|  |  | |  |  |  |
|  | Step 6 | Inputs and Target  Assuming you’d like to find a relationship between a student’s GPA in high school and a student’s GPA in the university. High school GPA is the independent variable that you’d want to represent using the variable X and University GPA is the dependent variable that you’d like to represent using the variable Y.  In general, x can be more than one column. y is typically a single column (usually the last one).  In Pandas you can access only a subset of the columns using the lines below: | | |  |
|  |  | |  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  | Note about creating CSV Files  Any text editor (notepad, notepad++, vi …) can create, open, and edit CSV files.  Also, Microsoft Excel and OpenOffice can Save spreadsheet files as CSV files. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Extracting Data from Pandas Data Frames  Needless to say, you will need to extract data from columns or from rows.  In the following section, you will learn about accessing elements of a pandas data frame and extracting data from specific cells. | | | | | |
|  | 1 | Columns  As mentioned before, extracting specific columns could be easily done in Pandas using the column names.  **Example:**  Assume we have the shown dataset stored in a CSV file.    We want to extract only two columns, the High school GPA and the University GPA. Just pass the names inside of the square brackets:  df2 = df[[‘HighSchool\_GPA’,’University\_GPA’]] | | |  |
|  |  |  |  |  |  |
|  | 2 | Rows  For rows, we have two options:   * DataFrame.loc. * DataFrame.iloc.   Both .loc and .iloc are quite similar in many aspects.  loc and iloc can be thought of as similar to Python list slicing. To show this even further, Let us select multiple rows.  Example:  For the dataset shown above.    Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  | Important Note about .loc and .iloc  If we changed the .loc with iloc in the previous example, the result slightly changes.  The reason behind this is that iloc acts exactly like the Python slicing functions where the end index is not included in the returned slice while loc returns the row at the end index as well.    Citation: RoboGarden Course Resources Copyright 2019 |

Section 05: Understand the datasets used in the missions

In this section, we are providing you with information about Mushroom dataset and Iris dataset. You will read about all the details about these datasets and their attributes. It is an assistant section that you will use its information to complete upcoming missions.

Mushrooms Dataset

|  |
| --- |
| The Mushrooms dataset is one of the most popular educational datasets available with more than 8 thousand hypothetical samples of gilled mushrooms corresponding to approximately 20 different species of mushrooms.  You are required to Explore the mushrooms dataset, Visualize the data and Manipulate the features to fit your needs.  *Labels: The dataset defines the data as definitely edible samples of mushrooms or definitely poisonous samples.* |

Attributes information

|  |  |
| --- | --- |
| These are the most important attributes in the dataset. | |
|  | cap-shape:   * bell= b * conical= c * convex= x * flat= f * knobbed= k * sunken=s |
|  | cap-surface:   * fibrous= f * grooves= g * scaly= y * smooth= s |
|  | Cap-color defines the color of the mushroom cap |
|  | Bruises is a binary attribute defining whether the mushroom sample has bruises or not. |
|  | Odor defines the type of odor coming from the mushroom cap. |
| Gill Properties | |
|  | Gill attachment |
|  | Gill spacing |
|  | Gill size:   * Broad = b * Narrow = n |
|  | Gill color |
| Stalk (stem) properties | |
|  | Stalk shape |
|  | Stalk root |
|  | Stalk surface above ring |
|  | Stalk surface below ring |
|  | Stalk color above ring |
|  | Stalk color below ring |
| General | |
|  | veil-type:   * partial= p * universal= u |
|  | ring-number:   * none=n * one=o * two=t |
|  | spore-print-color |
|  | population:   * abundant= a * clustered= c * numerous= n * scattered= s * several= v * solitary=y |
|  | habitat:   * grasses= g * leaves= l * meadows= m * paths= p * urban= u * waste= w * woods= d |

Iris Dataset

|  |
| --- |
| The data set consists of 50 samples from each of three species of Iris flower:   1. Iris setosa. 2. Iris virginica 3. Iris versicolor   Four features were measured from each sample:   * The sepal length * The sepal width * The petal length * The petal width   Based on the combination of these four features, Fisher (the scientist who developed this dataset) developed a linear discriminant model to distinguish the species from each other.  Citation: Iris [Digital Image]. Retrieved from http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html |

Section 6: ~~Conclusion Summary~~

|  |
| --- |
| What did we cover in this lesson?  Now you have learned what a dataset is and you have learned to manipulate several of them. Some libraries such as Pandas Library and Scikit learn Library were introduced to assist you in visualization and working with data. |
| You were introduced to data attributes and datasets. You were introduced to the Pandas Library basics in Python.  You learned that data is what constitutes the information we gain every day. Datasets are large collections of data points that are commonly used for analysis, study, or visualizing data. You were introduced to Scikit-learn library which is one of the most essential libraries to know to implement classical machine learning models in Python. |



## **Submodule 2 Application Mission(s) Details**

Repeat this section for every mission. The activities listed below are examples. See page 7 for the activities summary.

Below you will see the details of all the application missions within the submodule 2.

Mission 2 - Editor Mission: Read and Modify Datasets

**Mission 2 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming in Pandas: Pandas Data access

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Mission 2 ~~Content~~ instructions:**

In this mission, you will work with a large dataset and will practice how to extract and query information in the Mushrooms dataset using Pandas.

Purpose: Programming using the Pandas library to access in a dataset.

Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Now, let us move on to a bigger (but not huge) datasets with more attributes! Do not worry, big datasets load and behave exactly like small datasets!  This dataset is all about mushrooms! Mushrooms are rather remarkable when viewed as food. While a lot of people like them a part of their meals, some mushroom species are poisonous! This dataset lists different species' properties along with their edible/poisonous classifications.  You can revisit the mushroom dataset Here  *License* This is public **CC0: Public Domain** dataset available [here](https://www.kaggle.com/uciml/mushroom-classification).  Your code should perform the following tasks to complete this mission:   1. Load the mushrooms dataset from mushrooms.csv. 2. Extract data tuple with index 7 and its target (including the class as its first element). 3. Find out whether or not the mushroom with index 7 is edible or not.   Your code should return one boolean value edible.  HINT: you can index pandas.DataFrame with column names and the edible/poisonous classification is in a column called class. | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  # Write your code here. Do not change any other parts of the code  print(edible)  return edible |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  edible = dataset['class'][7] == 'e'  print(edible)  return edible |  |
|  |  |  |

Mission 3 - Editor Mission: Extract information from a dataset

**Mission 3 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming in Pandas: Pandas Data access

For more information regarding the grading of the Editor missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 3 Instructions:**

## In this mission, you will work with the Iris dataset and will practice how to extract and use information from a dataset.

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Another rather popular built-in dataset is the [Iris dataset](https://scikit-learn.org/stable/datasets/index.html" \l "iris-dataset). Let us load it! The Iris dataset contains information about 3 types of the irisIn plant. The dataset contains 150 data tuples and their corresponding target labels. Each tuple consists of 4 attributes about a certain plant; sepal length, sepal width, petal length and petal width (all in cm units) in that order. The target label is an integer 0, 1 or 2 representing the tuple's classification as an Iris-Setosa, an Iris-Versicolor or an Iris-Virginica, respectively.  Your code should perform the following tasks to complete this mission:   1. Load the Iris dataset from scikit-learn. 2. Extract data tuple with index 7 and its target. 3. Extract the sepal width (in cm) of the plant with index 7 and its target label as in integer (0, 1 or 2).   Your code should return the image of the 7th index’s sepal\_width, and its label  HINT: the sepal width (in cm) is the element with index 1 in the data tuple.  Revisit Pandas basic Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  # Write your code here. Do not change any other parts of the code  print(label)  return sepal\_width, label |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  from sklearn.datasets import load\_iris  dataset = load\_iris()  sepal\_width = dataset.data[7][1]  label = dataset.target[7]  print(label)  return sepal\_width, label |  |
|  |  |  |

Mission 4 - Editor Mission Score Booster : Read and Modify a dataset

**Mission 4 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming in Pandas: Pandas Data access

For more information regarding the grading of the Editor missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 4 Instructions:**

## In this mission, you will read the Mushroom dataset and check whether the selected mushroom is edible or not.

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task in the editor: | | | | | | |
|  | Step 1 | Load the Mushroom dataset.  Your code should perform the following tasks to complete this mission:   1. Load the Mushroom dataset from mushrooms.csv. 2. Extract data tuple with index 9 and its target (including the class as its first element). 3. Find out whether or not the mushroom with index 9 is edible or not.   Your code should return one boolean value edible.  Re-visit the Mushroom dataset Here  HINT: you can index pandas.DataFrame with column names and the edible/poisonous classification is in a column called class. | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  # Write your code here. Do not change any other parts of the code  print(edible)  return edible |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  edible = dataset['class'][9] == 'e'  print(edible)  return edible |  |
|  |  |  |

Mission 5 - Editor Mission Score Booster : Read, Modify and extract data

**Mission 5 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming in Pandas: Pandas Data access

For more information regarding the grading of the Booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 5 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task in an editor: | | | | | | |
|  | Step 1 | Let us retry loading the [Iris dataset](https://scikit-learn.org/stable/datasets/index.html" \l "iris-dataset).  Your code should perform the following tasks to complete this mission:   1. Load the Iris dataset from scikit-learn. 2. Extract data tuple with index 9 and its target. 3. Extract the sepal width (in cm) of the plant with index 9 and its target label as in integer (0, 1 or 2).   Your code should return the image of the 9th index’s sepal\_width, and its label.  HINT: the sepal width (in cm) is the element with index 1 in the data tuple. | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  # Write your code here. Do not change any other parts of the code  print(label)  return sepal\_width, label |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  from sklearn.datasets import load\_iris  dataset = load\_iris()  sepal\_width = dataset.data[9][1]  label = dataset.target[9]  print(label)  return sepal\_width, label |  |
|  |  |  |

Mission 6 - Quiz Mission: Introduction to Data

**Mission 6 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Understanding Data, Datasets and Pandas Library.

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 6 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Consider readings from a sensor at a constant rate, which type of dataset does the data belong to? | | | | | | |
|  |  | a. | Graph Data. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Ordered Data. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Tabular Data. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Matrix Data. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Readings from a sensor are data collected sequentially over an interval of time, which is the definition of ordered data. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is Incorrect. Readings from a sensor are collected in multiple time periods. Think of the type of dataset that supports sequential data collection. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which kind of attribute is Manufacturer Brand Name? | | | | | | |
|  |  | a. | Ordinal. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Interval. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Nominal. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Ratio. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Nominal attributes are either discrete or categorical attributes that define an object and cannot be ordered, which makes the brand manufacturer a nominal attribute. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Manufacturer Brand Name is a discrete value that takes unordered definite values. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which kind of attribute is Manufacturer Meal Price? | | | | | | |
|  |  | a. | Nominal. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Ordinal. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Ratio. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Interval. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! A meal price is a numeric value where the range or the difference between values does not necessarily give meaning. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. The meal price is a continuous value attribute in which differences in values are distinctive and do not give an extra meaning. | |  |
|  |  |  |  |











Mission 7 - Quiz Mission Score Booster : Introduction to Data

**Mission 7 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Understanding Data, Datasets and Pandas

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 7 Instructions:**

Lesson 2: Quiz – Score Booster

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Consider a Bank Database, which type of dataset does the data belong to? | | | | | | |
|  |  | a. | Graph Data. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Ordered Data. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Tabular Data. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Matrix Data. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! A Bank structured database belongs to tabular datasets. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect A Bank structured database is represented mainly by different attributes or columns and multiple rows of different clients or transactions. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Consider readings from a sensor at a constant rate, which type of dataset does the data belong to? | | | | | | |
|  |  | a. | Graph Data. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Ordered Data. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Tabular Data. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Matrix Data. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Readings from a sensor are data collected sequentially over an interval of time, which is the definition of ordered data. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is Incorrect. Readings from a sensor are collected in multiple time periods. Think of the type of dataset that supports sequential data collection. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Consider the Social media connections on Facebook as an example, which type of dataset does the data belong to? | | | | | | |
|  |  | a. | Graph Data. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Ordered Data. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Tabular Data. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Matrix Data. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Social media connections are best represented as graphs, with nodes (people, organizations, etc.) and edges (friendships, follows, subscriptions, etc.). | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is Incorrect. Facebook connections’ data can be described by the members including the people and the organizations and the relationships between these members (follows, likes, etc.). | |  |
|  |  |  |  |











# **Data Exploration**

## **Submodule 3 Activities Summary:**

List the activities learners will be completing during the module, and in the sequence, they will be completed. Please indicate which activities are graded, and which ones are not. All activities must match the stated module objectives listed above. The following list displays examples of activities that may be included. If you have questions about the activities you would like to include in the course, please speak to Learning Services.

1. Interactive lesson mission (Complete reading of the lesson is graded, and this is confirmed by finishing the informal assessment inside the lesson)
2. Quiz mission (Graded)
3. Editor mission (Graded)
4. Emulator mission (Graded)

**Submodule 3 Essential Question (Purpose):**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

How to properly interpret your collected data? And How to use Numpy library for data interpretation?   
  
For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Develop Submodule 3 Content**

Repeat this section for every mission.

Start explaining, based on the course structure, the objectives to be covered. You may do this in a table if you find it beneficial. In this section, you will develop content. This section is where you use your knowledge to build connection pieces between theory, resources, and assessments. Use the module objectives as the base for the structure of your content. This section is not focused on curation of resources; it is focused on you, as the SME, providing your expertise to create content. However, if you are referencing any type of source, please cite it in APA format (including open source materials and creative commons material).

Write Here:

Mission 1 - Interactive Lesson Mission: Data Exploration

Section 0: introduction

|  |
| --- |
| What will you learn in this lesson? |
| In this lesson, You will be introduced to Summary Statistics. By the end of this lesson, you will know about:   1. The Mean, Mode, Median, Standard Deviation and Variance. 2. Correlation between different variables. 3. The correlation matrix and how to compute it using Pandas. 4. Dealing with missing data, noise and outliers. 5. Distance measures and what is the meaning of dis-similarity or similarity between data points. |
|  |

Section 1: statistic

In this section, you will learn about some simple but necessary statistic knowledge.

|  |
| --- |
| Exploring the data is one of the main steps of any machine learning problem. |

|  |  |  |
| --- | --- | --- |
|  |  | Summary Statistics  The key to understanding different datasets is in defining the structure of the dataset using summary statistics.  **Examples:** frequency, mode, median, or mean. |

|  |  |  |
| --- | --- | --- |
|  |  | Note about summary statistics  Almost all summary statistics can be calculated after analyzing the whole dataset once. |

Summary Statistics

|  |
| --- |
| The Frequency |
| The frequency defines the amount of time a certain value occurs in a dataset.  **Example: {male, male, female, female, male, female}**  The frequency of the value "female" is **50%.** |

|  |
| --- |
| The Mode |
| The most frequent value that occurs in a given dataset.  **Example: {male, male, male, female, male, female}**  The mode is the value **"male".** |

|  |  |
| --- | --- |
| The Median | |
| The Median is the middle data point of all the data points.  $$\text {Median}=\left\{\begin{array}{ll}{X(r+1)} & {\text { if } \mathrm{m} \text { is odd }, m=2 r+1} \\ {0.5(X r+X(r+1))} & {\text { if } \mathrm{m} \text { is even }, m=2 r}\end{array}\right.$$   * $$X(r+1)$$ is the right middle input in a set * Xr is the left middle point in a set * m is the number of data points |  |
|  | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |
| --- | --- |
| The Mean | |
| The Mean is known as the average of data points in a dataset.   * $$\mu=\frac{1}{m} \sum\_{1}^{m} X i$$ is the Mean * m is the number of data points * Xi is a single data point | Citation: RoboGarden Course resources Copyright 2019 |

|  |  |
| --- | --- |
| The Standard Deviation | |
| Standard deviation () is a measure of the amount of variation or dispersion of a set of values.  $$  s\_{N}=\sqrt{\frac{1}{N} \sum\_{i=1}^{N}\left(x\_{i}-\bar{x}\right)^{2}}  $$   * N is the number of data points * Xi is a single data point * is the mean of the dataset? |  |
| Citation: Standard Deviation [Digital Image]. Retrieved from https://www.students4bestevidence.net/blog/2018/09/26/a-beginners-guide-to-standard-deviation-and-standard-error/ |  |

|  |  |
| --- | --- |
| The Variance | |
| Variance is the Expected value of the squared standard deviation of the data from the mean.  It measures the spread of the data in the dataset from their average value.   * $$\operatorname{Var}(X)=\frac{1}{m} \sum\_{i=1}^{m}\left(x\_{k}-\mu\_{x}\right)^{2}$$ is Kth feature * is the Mean of X * m is the number of data points | Citation: Variance [Digital Image]. Retrieved from https://en.wikibooks.org/wiki/Statistics/Summary/Variance |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Dataset X = {2, 2, 3, 4, 2, 5, 9, 3} | | | | | | |
| The mode of the dataset is | | Dropdown | | |  | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Correct Answer!  Since ‘2’ is the most repeated number in the dataset, so ‘2’ is the mode for this data set. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect answer! The number you choose is not the most repeated in the dataset. | |  |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Dataset X = {2, 2, 3, 4, 2, 5, 9, 3}   1. The mean of the dataset is | | ……… | | | . | |
| 2.5 | | |
| 3.75 | | |
| 4 | | |
| 10 | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Correct Answer!  The Mean is calculated through this formula: $$\mu=\frac{1}{m} \sum\_{1}^{m} X i$$Which is ‘3.75’. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect Answer!  The Mean is calculated through this formula: $$\mu=\frac{1}{m} \sum\_{1}^{m} X i$$. For example: for a simple dataset of {2,3,10} the mean would be the sum of the three numbers divided by their number.  Mean of {2,3,10} = (2+3+10)/3 = 5 | |  |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Dataset X = {2.5, 3.1, 3.4, 3.5, 3.5, 4, 4.1 } | | | | | | |
| The median of the dataset is | | Dropdown | | |  | |
| 3.5 | | |
| 3.4 | | |
| 3.45 | | |
| 3.55 | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Correct Answer!  The median is calculated through this formula: $$\text {Median}=\left\{\begin{array}{ll}{X(r+1)} & {\text { if } \mathrm{m} \text { is odd }, m=2 r+1} \\ {0.5(X r+X(r+1))} & {\text { if } \mathrm{m} \text { is even }, m=2 r}\end{array}\right.$$Which is ‘3.5’ | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect Answer!  The median is calculated through this formula: $$\text {Median}=\left\{\begin{array}{ll}{X(r+1)} & {\text { if } \mathrm{m} \text { is odd }, m=2 r+1} \\ {0.5(X r+X(r+1))} & {\text { if } \mathrm{m} \text { is even }, m=2 r}\end{array}\right.$$.  For example: for dataset {2,3,10} since the number of elements is odd the median is simply the number in the middle. In this case, 3 is the median. | |  |
|  |  |  |  |

|  |  |
| --- | --- |
| Correlation | |
| Correlation is an essential statistical technique in Exploring data which aims at finding how strongly different pairs of variables relate to each other.  **Example:**  Consider the relationship between height and weight. You can logically assume that taller people generally weigh more than shorter people. However, this relationship is not ideal as people in the same height might actually vary in weight. Correlation between the two variables height and weight can tell you exactly how much people’s weights vary with respect to their heights.  Although this correlation is fairly obvious, your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.  **Note:**  The correlation coefficient is a value that indicates the strength of the relationship. Correlation coefficient varies in the range [-1,1] where:  **-1:** Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).  **0:** No correlation. The variables do not have a relationship with each other.  **1:** Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).    Image Citation: RoboGarden Course Resources Copyright 2019   * rxy – the correlation coefficient of the linear relationship between the variables x and y * xi – the values of the x-variable in a sample * x̅ – the mean of the values of the x-variable * yi – the values of the y-variable in a sample * ȳ – the mean of the values of the y-variable     Citation: Correlation [Digital Image]. Retrieved from https://medium.com/@dipti.rohan.pawar/correlation-statistical-analysis-9471411f0431 |  |

|  |
| --- |
| Correlation Matrix |
| A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.  The function pd.DataFrame.corr() computes the correlation matrix between the columns of the given data frame.  **Example:**  Consider the GPA dataset discussed before.    Citation: RoboGarden Course Resources Copyright 2019  **Using Pandas**    Citation: RoboGarden Course Resources Copyright 2019 |

Section 02: Introduction to Numpy

In this section, you will be introduced to the Numpy library. The library is useful for handling N-Dimensional array of data.

|  |
| --- |
| What is Numpy?  NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array. Numpy provides ways to process multidimensional data making it one of the essential libraries needed by a machine learning engineer. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NumPy Ndarray Initialization  Ndarray is the main component in NumPy. It represents an N-dimensional array of objects of the same type.  Firstly: you should write these lines of code into your notebook to import the Numpy Library. Notebook refers to the Jupyter notebook that comes with the installation of Anaconda. Notebooks let you combine coding and writing notes at the same time. For more information about Anaconda installation and Jupyter notebook please click on this link.  Import the Numpy Library using the alias np  import Numpy as np  The following are ways of initializing a NumPy array.  Click through the carousel to learn more about different initializations | | | | |
|  |  | Array of zeros  numpy.zeros(shape, dtype = float, order = 'C') |  |  |
|  |  |     |  |  |

Text Carousel Content Table

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Array of zeros  Initializes a multidimensional array filled with zeros with the given shape.  numpy.zeros(shape, dtype = float, order = 'C')  Citation: RoboGarden Course Resources Copyright 2019 |
|  | Array of ones  Initializes a multidimensional array filled with ones with the given shape.  numpy.ones(shape, dtype = float, order = 'C')  Citation: RoboGarden Course Resources Copyright 2019 |
|  | Array of Existing Data  Initializes a multidimensional array from a Python sequence.  numpy.asarray(a)  numpy.array(a)  Citation: RoboGarden Course Resources Copyright 2019 |
|  | Array from definite range of numbers  Initializes a multidimensional array within the defined range between the start and the stop. The number of items is defined based on the step provided.  -numpy.arange(start, stop, step, dtype)  Citation: RoboGarden Course Resources Copyright 2019  - numpy.linspace(start, stop, num, endpoint, retstep, dtype)  Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Using Numpy  Consider using Numpy for manipulating and applying computation on array-like data. Data such as images are considered array. | | | | | |
|  | Step 1 | Installing and using NumPy  To install the NumPy library on your machine, type the following command on your computer terminal.  pip install NumPy  For package installation Guide. See Anaconda installation guide Here | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Importing Numpy  After installing the Numpy library, you can now import it in a Python script to use its functionalities.  In this lesson, Numpy will be explicitly imported with the abbreviated name of np as in the following command  Import Numpy as np | | |  |
|  |  | |  |  |  |
|  | Step 3 | Slicing and Indexing   * Indicies in a Numpy array start at 0 just like the Python list.   Citation: RoboGarden Course Resources Copyright 2019     * Using a colon, you can indicate the start and the end of your slice.   **Hint:**  The end of your slice is the index of the number that the slice was taken before. The number with the index of the end does not appear in the sliced Ndarray.  Citation: RoboGarden Course Resources Copyright 2019     * You can use negative indices for accessing the elements of the array in reverse.   Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |
|  | Step 4 | Statistical Analysis  Numpy helps greatly in quickly calculating statistical measures for the Numpy arrays.  **Examples:**   * Mean()   Citation: RoboGarden Course Resources Copyright 2019     * Median()   Citation: RoboGarden Course Resources Copyright 2019     * Standard Deviation()   Citation: RoboGarden Course Resources Copyright 2019     * Variance()   Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  |  |  |  |  |
|  |  | |  |  |  |
|  |  | |  |  |  |

Section 3: Important Terms

In this section, you will learn some important terms that you will need to know them when working with datasets. The terms are noise, outliers and missing data which refer to the existence of problems in your dataset. You need to fix them before using the dataset.

Noise

|  |
| --- |
| What is Noise in Data? |
| In analyzing data, we define noise as any distortion to the actual data points. Noise in data has two main types:  **Noise in Objects**: An extraneous object.  **Noise in Attributes:** Distortion in original values of Data.  Example:  Distortion of voice as a result of a bad phone signal. |
| Citation: RoboGarden Course resources Copyright 2019 |

|  |  |  |
| --- | --- | --- |
|  |  | Data Outliers  Data outliers are data points whose attribute values differ noticeably from other data points in the dataset. |

|  |  |
| --- | --- |
| Should you always eliminate outliers? | |
| No, sometimes outliers are the goal of our search. Outliers are different than noise.  Anomaly detection of one of the most important fields of study in machine learning which aims at accurately detecting outliers in Data.  **Examples:**   * *Fraud detection* * *Spam mail classification* * *Intrusion detection* | RoboGarden Course Resources Copyright 2019 |

|  |  |  |
| --- | --- | --- |
|  |  | The Abnormality of the Outliers  Outliers are defined by the large distance between them and most of the data points in the dataset.  Distance resembles the differences between the data points. |

|  |  |  |
| --- | --- | --- |
|  |  | Missing Data  Missing data happens if no value is stored for a certain attribute in observation, and is considered a major factor that contributes to any conclusions drawn from the same data. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods of Dealing with Missing Data | | | | | |
|  | 1 | Eliminate Data Objects completely or only eliminate the missing values. | | |  |
|  |  |  |  |  |  |
|  | 2 | Fill in the missing data putting the mode of the attribute or using an unbiased average of data. | | |  |
|  |  | |  |  |  |
|  | 3 | Ignore missing values during analysis and assume value during testing. | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  From a list of movie reviews, a system needs to detect fake reviews made by programmed bots. Which of these choices are you looking for? | | | | | | | | |
|  |  | a. | Noise | | | |  | |
|  |  |  | | |  |  |  | |
|  |  | b. | Duplicate Data | | | |  | |
|  |  |  | | |  |  |  | |
|  |  | c. | Outliers | | | |  | |
|  |  | | | |  |  |  | |
|  |  | d. | Missing values | | | |  | |
|  |  | | | |  |  |  | |
|  |  | | | |  |  |  | |
|  |  | | |  | | | |  | |

MCQ Responses Table

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, that is incorrect. Noise is the data points resulting from an extraneous source. Noise is never desired. One of the main tasks in data preprocessing is to eliminate the noise in the dataset. |
| b. | Incorrect | Unfortunately, that is incorrect. The fake reviews are not duplicate data. |
| c. | Correct | Awesome! In this scenario, the system seeks a few reviews with characteristics that are noticeably different from human reviews. In other words, the system seeks to detect data points whose attribute values are significantly different from those of the rest of the data. The system seeks outliers. |
| d. | Incorrect | Unfortunately, that is incorrect. The fake reviews are reviews with characteristics different from the majority of real reviews. They do not represent missing data. |

Section 3: Dis-similarity measures

|  |
| --- |
| Similarity and Dis-similarity measures were introduced to offer an actual measure to the difference between different data points. |

|  |
| --- |
| What are the properties of Dis-similarity measures? |
| * A numerical measure of how two data points are different from each other. * Zero for two duplicate data points. * No fixed upper limit (depending on the type of dis-similarity). |

|  |  |
| --- | --- |
|  | Proximity  The proximity term refers to the function that measures features proximity. Proximity refers to both similarity and dissimilarity. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Distance Examples  There are many types of dissimilarity measures. All of these distance metrics are used with numeric data.  Click through the carousel to learn more about different distance measures. | | | | |
|  |  | Content Text |  |  |
|  |  |     |  |  |

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Euclidean Distance  Defined as the straight-line distance between two points in multidimensional spaces.   * $$d(x, y)=\sqrt{\sum\_{i=1}^{m}\left(X\_{k}-Y\_{x}\right)^{2}}$$m is the number of features per data point * X and Y are data objects |
|  | Manhattan Distance  The sum of the absolute differences of their Cartesian coordinates of two points.  $$d(x, y)=\sum\_{i=1}^{m}\left|X\_{k}-Y\_{x}\right|$$   * m is the number of features per data point * X and Y are data objects |
|  | Minkowski Distance  It is a Generalization of the Euclidean Distance.   * $$d(x, y)=\sqrt[r]{\sum\_{i=1}^{m}\left|X\_{k}-Y\_{x}\right|^{r}}$$m is the number of features per data point * r is a parameter * X and Y are data objects |

Section 06: Lesson Summary

|  |
| --- |
| What did we cover in this lesson? |
| Great work! You were introduced to different statistical measures that are important to interpret datasets including mean, median, standard deviations and correlation. You were introduced to the Numpy library and were able to interpret the difference between the noise and outliers in different datasets. |



## **Submodule 3 Application Mission(s) Details**

Repeat this section for every mission. The activities listed below are examples. See page 7 for the activities summary.

Below you will see the details of all the application missions within the submodule 3.

Mission 2 - Editor Mission: Exploring data

**Mission 2 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

The student will learn how to use functions in the Pandas library to explore and manipulate datasets.

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Mission 2 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Now, Let us explore the Mushrooms dataset!  Your code should perform the following tasks to complete this mission:   1. Count the edible mushrooms species and calculate their percentage in the dataset. 2. Extract all the *unique* cap colors in the dataset (n, y, w, etc).   Your code should return edible\_count, edible\_percentage and cap\_colors.  Revisit the Mushroom dataset Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  # Write your code here. Do not change any other parts of the code  print(edible\_percentage)  return edible\_count, edible\_percentage, cap\_colors |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  import numpy as np  edible\_count = np.count\_nonzero(dataset.iloc[:, 0].values == 'e')  cap\_colors = set(np.unique(dataset.iloc[:, 3].values))    total\_count = dataset.shape[0]  edible\_percentage = edible\_count / total\_count  print(edible\_percentage)  return edible\_count, edible\_percentage, cap\_colors |  |
|  |  |  |

Mission 3 - Editor Mission: Apply Statistics to the Iris dataset

**Mission 3 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming in Numpy

For more information regarding the grading of the Editor missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 3 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Now Let us get a better look at the Iris dataset!  Your code should perform the following tasks to complete this mission:   1. Calculate the mean of all the attributes; sepal length, sepal width, petal length and petal width. 2. Calculate the variance of all the attributes; sepal length, sepal width, petal length and petal width.   Your code should return means and variances; each is a 1D array with 4 values for the sepal length, sepal width, petal length and petal width, respectively.  Hint: there are a lot of ways to do this, but you may consider taking a look at the built-in functions mean and var in *Numpy* or *Pandas*.  Revisit the Iris dataset Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | import numpy as np  def main():    from sklearn.datasets import load\_iris  dataset = load\_iris()  # Write your code here. Do not change any other parts of the code  print(np.max(means))  return means, variances |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():    from sklearn.datasets import load\_iris  dataset = load\_iris()  import numpy as np  means = np.mean(dataset.data, axis=0)  variances = np.var(dataset.data, axis=0)  print(np.max(means))  return means, variances |  |
|  |  |  |

Mission 4 - Editor Mission Score Booster : Count and extract information

**Mission 4 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Exploring Pandas Functions

For more information regarding the grading of the Booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 4 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Let us try exploring the Mushrooms dataset again!  Your code should perform the following tasks to complete this mission:   1. Count the poisonous mushrooms species and calculate their percentage in the dataset. 2. Extract all the *unique* cap shapes in the dataset (x, b, f, etc).   Your code should return poisonous\_count, poisonous\_percentage and cap\_shapes. | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  # Write your code here. Do not change any other parts of the code  print(poisonous\_percentage)  return edible\_count, edible\_percentage, cap\_colors |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  import numpy as np  poisonous\_count = np.count\_nonzero(dataset.iloc[:, 0].values == 'p')  cap\_shapes = set(np.unique(dataset.iloc[:, 3].values))    total\_count = dataset.shape[0]  poisonous\_percentage = poisonous\_count / total\_count  print(poisonous\_percentage)  return poisonous\_count, poisonous\_percentage, cap\_shapes |  |
|  |  |  |

Mission 5 – Editor Mission Score Booster: Apply more statistics on the Iris dataset!

**Mission 5 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Learning how to use Numpy in their codes

For more information regarding the grading of the Booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 5 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Let us get a better look at the Iris dataset!  Your code should perform the following tasks to complete this mission:   1. Calculate the median of all the attributes; sepal length, sepal width, petal length and petal width. 2. Calculate the standard deviation of all the attributes; sepal length, sepal width, petal length and petal width.   Your code should return medians and stds; each is a 1D array with 4 values for the sepal length, sepal width, petal length and petal width, respectively.  Hint: there are a lot of ways to do this, but you may consider taking a look at the built-in functions median and std in *Numpy* or *Pandas*. | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():    from sklearn.datasets import load\_iris  dataset = load\_iris()  # Write your code here. Do not change any other parts of the code  print(np.max(medians))  return medians, stds |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():    from sklearn.datasets import load\_iris  dataset = load\_iris()  import numpy as np  medians = np.median(dataset.data, axis=0)  stds = np.std(dataset.data, axis=0)  print(np.max(medians))  return medians, stds |  |
|  |  |  |

Mission 6 - Quiz Mission: Data Exploration

**Mission 6 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Understanding Data Statistical Measures. E.g. Mean, Median, etc.

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 6 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| The most frequent value that occurs in a given dataset is: | | | | | | |
|  |  | a. | Mode. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Median. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Mean. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | * A) Mode: The most frequent value that occurs in a given dataset * B) Median: The middle data point of all the data points. * C) Mean: The average of data points in a dataset. | |  |
|  |  |  |  |







|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset X = {2, 2, 2, 3, 4, 8, 3, 4, 2, 5, 9, 3}  The mode of the dataset is: | | | | | | |
|  |  | a. | 2 | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 3 | | |  |
|  |  |  | |  |  |  |
|  |  | c. | 4 | | |  |
|  |  | | |  |  |  |
|  |  | d. | 5 | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The mode of the given dataset is 2. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is not the correct answer. Remember the mode is the most occurring value in the dataset.  Example:  In the dataset {1,2,2,3}, the mode is equal to 2 | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset X = {2, 2, 2, 3, 4, 8, 3, 4, 2, 5, 9, 3}  The mean of the dataset X is approximately: | | | | | | |
|  |  | a. | 2.5 | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 3.75 | | |  |
|  |  |  | |  |  |  |
|  |  | c. | 4 | | |  |
|  |  | | |  |  |  |
|  |  | d. | 3.92 | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The mean of the given dataset is 3.92. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is not the correct answer.  The Mean is calculated through this formula: $$\mu=\frac{1}{m} \sum\_{1}^{m} X i$$. For example, for a simple dataset {2,3,10} the mean would be the sum of the three numbers divided by their number.  Mean of {2,3,10} = (2+3+10)/3 = 5 | |  |











Mission 7 - Quiz Mission Score Booster : Data Exploration

**Mission 7 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Understanding Data Statistical Measures.

For more information regarding the grading of the Quiz mission please refer to Module 1 Introduction section as well as the Orientation.

**Mission 7 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset X = {2, 2, 2, 3, 4, 8, 3, 4, 2, 5, 9, 3}  The median of the dataset is: | | | | | | |
|  |  | a. | 2 | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 4.5 | | |  |
|  |  |  | |  |  |  |
|  |  | c. | 3 | | |  |
|  |  | | |  |  |  |
|  |  | d. | 9 | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The median of the given dataset is 3. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is not the correct answer.  The median is calculated through this formula: $$\text {Median}=\left\{\begin{array}{ll}{X(r+1)} & {\text { if } \mathrm{m} \text { is odd }, m=2 r+1} \\ {0.5(X r+X(r+1))} & {\text { if } \mathrm{m} \text { is even }, m=2 r}\end{array}\right.$$.  For example: for dataset {2,3,10} since the number of elements is odd the median is simply the number in the middle. In this case, 3 is the median. | |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which of these terms get most affected by outliers? | | | | | | |
|  |  | a. | Mean | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Median | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Max/Min | | |  |
|  |  | | |  |  |  |
|  |  | d. | Variance | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! All the statistical descriptors, mean, median, variance, standard deviation and mode are quite immune to outlier as they form a tiny portion of the data, which makes their contribution to the calculations of such descriptors quite negligible. The minimum and maximum, prone to outliers. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Think about what formulas could be immune to outliers. For example: for dataset {2,3,10} since the number of elements is odd the median is simply the number in the middle. In this case, 3 is the median. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| What is the maximum number of dimensions where someone could calculate the distance between data points? | | | | | | |
|  |  | a. | 1 Dimension. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 2 Dimensions. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Any countable number of dimensions. | | |  |
|  |  | | |  |  |  |
|  |  | d. | 4 Dimensions. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! While it is easy to visualize the distance between points in 1D, 2D and 3D spaces, the notion of distance applies to data of any dimensions. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Remember the Euclidean Distance! Is it bounded to a finite number of dimensions? | |  |
|  |  |  |  |











# **Data Pre-Processing**

## **Submodule 4 Activities Summary:**

List the activities learners will be completing during the module, and in the sequence, they will be completed. Please indicate which activities are graded, and which ones are not. All activities must match the stated module objectives listed above. The following list displays examples of activities that may be included. If you have questions about the activities you would like to include in the course, please speak to Learning Services.

1. Interactive lesson mission (Complete reading of the lesson is graded, and this is confirmed by finishing the informal assessment inside the lesson)
2. Quiz mission (Graded)
3. Editor mission (Graded)
4. Emulator mission (Graded)

**Submodule 4 Essential Question (Purpose):**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

What is preprocessing? And How to use Sklearn to do it?

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Develop Submodule 4 Content**

Repeat this section for every mission.

Start explaining, based on the course structure, the objectives to be covered. You may do this in a table if you find it beneficial. In this section, you will develop content. This section is where you use your knowledge to build connection pieces between theory, resources, and assessments. Use the module objectives as the base for the structure of your content. This section is not focused on curation of resources; it is focused on you, as the SME, providing your expertise to create content. However, if you are referencing any type of source, please cite it in APA format (including open source materials and creative commons material).

Write Here:

Mission 1 – Interactive Lesson Mission: Data Pre-Processing

Section 0: introduction:

|  |
| --- |
| What will you learn in this lesson? |
| In this lesson, You will be introduced to Machine Learning. By the end of this lesson, you will know the following:   1. What is preprocessing? 2. Encoding and different types of encoders. 3. The idea behind scaling features. |

Section 01: Data Preprocessing Overview

In this section, you will learn about how to preprocess data or clean it.

|  |  |  |
| --- | --- | --- |
|  |  | Data Preprocessing  Data Preprocessing is an important step in data mining and machine learning. As Data Acquisition techniques almost always lead to some noisy data samples with out-of-range values, data outliers, missing data points or data attributes, etc. Processing unclean data will lead to misleading results.  Usually data preprocessing is the most important step in any machine learning project. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Steps of data preprocessing  Preprocessing mainly concerns with four main tasks:   1. Cleaning. 2. Encoding. 3. Scaling. 4. Feature reduction.   Click through the carousel to learn more about the four main preprocessing tasks. | | | | |
|  |  | Cleaning.  The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.  After cleansing, a data set should be consistent with other similar data sets in the system. |  |  |
|  |  |     |  |  |

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Cleaning.  The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.  After cleansing, a data set should be consistent with other similar data sets in the system. |
|  | Encoding  The process involving the review and adjustment of collected categorical survey data in order to fit the following processes in the machine learning project cycle. |
|  | Scaling.  The process of normalizing different values of the attributes to have the same range. This assures different attributes measured at different units to be treated similarly by any mathematical model used for machine learning. |
|  | Feature Reduction.  In machine learning, feature reduction is the process of reducing the number of the variables in the dataset to obtain only a subset of principal values representing the most percentage of the variation in the original dataset. |

Section 01: Encoding

In this section, you will learn about the definition of encoding and you will know to what kind of data types we need to apply encoding. Then you will try an example about encoding data.

|  |
| --- |
| Categorical data and Machine learning |
| Categorical data (as mentioned before) must be represented in a numerical form to aid different machine learning models to train on them. This is called Encoding. |

|  |  |  |
| --- | --- | --- |
|  |  | Categorical Data  Categorical data is the type of data that can only have one of a limited set of values assigning each data object to a specific group or a nominal category. |

|  |
| --- |
| Examples of Categorical Data:   1. Male/Female 2. Yes/No 3. Country Names |

|  |  |  |
| --- | --- | --- |
|  |  | Important note about categorical data  To train a machine learning model using categorical data, encode it into numerical values first. |

Introduction to Encoders

Encoders are a module of machine learning models whose task is converting given information to a particular form so that system can work with it. For example, if the input of encoder is categorical data, the encoder will convert it to an integer array. Below you will read more about different types of encoders.

|  |
| --- |
| Types of Encoders  The Scikit-Learn library supports the following types of encoders:  • **Ordinal Encoder**: encodes categorical features as an integer array.  • **One-hot Encoder**: encodes only one category as 1 and all other categories as 0.  • **Label Encoder**: encodes target labels with values between 0 and n\_categories-1. |
|  |

|  |  |  |
| --- | --- | --- |
|  |  | Important note about ordinal encoders and label encoders  Both label encoders and ordinal encoders have the same functionality. That is, they both generate a column of integers with values in the range from 0 to n\_categories - 1.  The only difference is that an ordinal encoder is typically used for encoding input features, while a label encoder is used for encoding the target (output) variable. Therefore, an ordinal encoder can fit data that has the shape of (n\_samples, n\_features), while a label encoder can only fit data that has the shape of (n\_samples,).  For only one categorical feature, either one can be used. |

|  |  |
| --- | --- |
| Label Encoder | |
| A label encoder encodes categories with value between 0 and n\_categories-1. | Citation: RoboGarden Course Resources Copyright 2019 |

|  |
| --- |
| One Hot Encoder |
| One-hot encoding is typically used when we consider only one of the categories, i.e. one of the categories takes the value of 1 while all other categories take the value of 0. Each category value will be encoded in a column.  One-hot Encoder works on numerical values encoded by Label Encoder (not on direct categorical data). |

Example Dataset

|  |
| --- |
|  |
| In this example, you will learn how to encode the target data. The encoded data are used in machine learning methods such as classification. |

Encoding Targets

|  |  |  |
| --- | --- | --- |
| 1. Initialize Encoder | | |
| Use the Label Encoder from the sklearn.preprocessing library: | | |
|  |  |  |
|  | from sklearn.preprocessing import LabelEncoder  encoder = LabelEncoder() |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Fit Encoder to Data | | |
| Use the fit() function for the encoder to learn the available categories: | | |
|  |  |  |
|  | encoder.fit(df[‘Country’]) |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Encode Categorical Features | | |
| Now, use the transform function to encode the categorical features: | | |
|  |  |  |
|  | df[‘Country’]= encoder.transform(df[‘Country’]) |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Combine Fitting and Encoding | | |
| Alternatively, we can use the fit transform() function to perform both fitting and encoding in one operation: | | |
|  |  |  |
|  | df[‘Country’]= encoder.fit\_transform(df[‘Country’]) |  |
|  |  |  |

|  |
| --- |
| 1. View Encoded Labels |
| The country names should now be encoded to three classes 0, 1 and 2. |
| Citation: RoboGarden Course Resources Copyright 2019 |

Section 02: Scaling Features

In this part, you will learn about what scaling of data is and how you can scale your data using the Scikit-learn library.

|  |
| --- |
|  |
| The range of values of raw data varies widely. In some machine learning algorithms, such as SVM and ANN, objective functions will not work correctly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature.  Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. |

|  |  |  |
| --- | --- | --- |
|  |  | Feature Scaling  Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as *data normalization* and is generally performed during the data preprocessing step. |

|  |
| --- |
| Feature Standardization |
| Some machine learning models require input features to have zero mean and unit variance for better accuracy or to be able to learn from available data. To achieve that, we calculate the mean and variance of the data, subtract the mean from all data samples, then divide all data samples by the variance. That does not affect how the data looks, as we will see shortly, but it helps the machine learning model learn from the data.  This is called feature standardization. |
|  |
| Citation: RoboGarden Course Resources Copyright 2019 |

Scaling In Scikit-learn

In this example, you will learn how to scale and fit data so that you can remove the mean and scale the data to have a variance of one.

|  |
| --- |
| In this example, you will learn how to scale and fit data so that you can remove the mean and scale the data to have a variance of one. |

|  |  |  |
| --- | --- | --- |
| 1. Initialize Scaler | | |
| To scale features, we use the StandardScaler from the sklearn.preprocessing library: | | |
|  |  |  |
|  | from sklearn.preprocessing import StandardScaler  scaler = StandardScaler() |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Fit Scaler to Training Data | | |
| Use the fit() function for the Scaler to scale x\_train: | | |
|  |  |  |
|  | scaler = scaler.fit(x\_train) |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Scale Training Data | | |
| The scaler can now be used to scale training data using the transform function: | | |
|  |  |  |
|  | x\_train = scaler.transform(x\_train) |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 1. Combine Fitting and Scaling | | |
| Alternatively, we can use the fit\_transform() function to perform both fitting and scaling in one operation: | | |
|  |  |  |
|  | x\_train = scaler.fit\_transform(x\_train) |  |
|  |  |  |

|  |
| --- |
| 1. Compare Scaled Data to Original Data |
| By plotting the training data before and after scaling, we can see that both figures are the same except for the values of the x-axis. |
| Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |  |
| --- | --- | --- |
| 1. Scale Test Features | | |
| We also need to scale test features using the same scaler.  Since the scaler is already fitted, we only use transform() for test data: | | |
|  |  |  |
|  | x\_test = scaler.transform(x\_test) |  |
|  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  What does the piece of code below do? | | | | | | |
|  |  | | | | |  |
|  | x\_train = scaler.fit\_transform(x\_train) | | | | |  |
|  |  | | | | |  |
|  |  | a. | Fits and scales the training data. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Fits the training data. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Transforms the shape of the training data. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Extract the mean of the training data. | | |  |
|  |  | | |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Correct | Great! The function fit\_transform() combines the two operations of fitting the training data and then scales the result. |
| b. | Incorrect | Incorrect. The function fit() is the function that aims at only fitting the scaler to the input data without transforming the data. |
| c. | Incorrect | Incorrect. The function transform() is the function that aims at transforming the data to a previously fitted scaler instance. |
| d. | Incorrect | Incorrect. The function fit\_transform() is not responsible for extracting any statistical data from the input data. |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Section 06: Lesson Summary

|  |
| --- |
| What did we cover in this lesson? |
| Great work! You were introduced to preprocessing in Scikit-learn, encoding the categorical data and feature scaling. Encoding categorical attributes is essential for use in machine learning models’ training phases. In addition to that, you were introduced to the importance of scaling and how to implement scaling in Python with the help of Scikit-learn library. |



## **Submodule 4 Application Mission(s) Details**

Repeat this section for every mission. The activities listed below are examples. See page 7 for the activities summary.

Below you will see the details of all the application missions within submodule 4.

Mission 2 - Editor Mission: Data Encoding

**Mission 2 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Sklearn Library: Using Scikit learn Encoders to encode categorical data

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

Mission 2 Instructions:

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | The Mushrooms dataset has a plethora of categorical attributes, so Let us use it for practicing!  Your code should perform the following tasks to complete this mission:   1. Encode y (the class attribute) using LabelEncoder. 2. Encode selected\_X (a subset of attributes) using OneHotEncoder.   Your code should return encoded\_y and encoded\_X; arrays with the transformed result of applying the encoders.  Hint: encoded\_y should be a 1D array with only 0s and 1s. Hint: encoded\_X should be a 2D array with only 0s and 1s (the fit\_transform function for a OneHotEncoder returns a sparse matrix by default, hence, to get an array, you can either disable the sparse option in the constructor (OneHotEncoder(sparse=False)) or call toarray() on the sparse matrix returned). HINT: knowing that selected\_X has 2 columns with 6 unique values in the first and 4 unique values in the second, how many columns should encoded\_X have?  Revisit the Mushrooms dataset Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  y = dataset.iloc[:, 0].values  selected\_X = dataset.iloc[:, 1:3].values  # Write your code here. Do not change any other parts of the code  print(encoded\_X[0,:])  return encoded\_y, encoded\_X |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  y = dataset.iloc[:, 0].values  selected\_X = dataset.iloc[:, 1:3].values  from sklearn.preprocessing import LabelEncoder, OneHotEncoder  encoded\_y = LabelEncoder().fit\_transform(y)  encoded\_X = OneHotEncoder().fit\_transform(selected\_X).toarray()  print(encoded\_X[0,:])  return encoded\_y, encoded\_X |  |
|  |  |  |

Mission 3 - Editor Mission: Data Scaling

**Mission 3 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Sklearn Library: Using Scikit Learn Scalers to scale numerical data

For more information regarding the grading of the Editor missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 3 Insructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Time to practice scaling features! Let us go to our friendly Iris dataset!  Your code should perform the following tasks to complete this mission:   1. Load the Iris dataset from scikit-learn as a data vector X and a label vector y. 2. Scale the data vector X with a standard scaler (normalize X).   Your code should return the scaled data vector scaled\_X.  SELF-CHECK: notice the result of calling mean(axis=0) and var(axis=0) (or var(axis=0)) for scaled\_X!  Revisit the Iris dataset Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():      from sklearn.datasets import load\_iris      X, y = load\_iris(return\_X\_y=True)      # Write your code here. Do not change any other parts of the code      import numpy as np      #uncomment the two lines after asigning the values to "scaled\_X" inorder to complete this mission      #print(np.mean(scaled\_X))      #return scaled\_X |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():      from sklearn.datasets import load\_iris      X, y= load\_iris(return\_X\_y=True)      from sklearn.preprocessing import StandardScaler      scaler = StandardScaler()      scaled\_X = scaler.fit\_transform(X)      import numpy as np      print(np.mean(scaled\_X))      return scaled\_X |  |
|  |  |  |

Mission 4 – Editor Mission Score Booster: Data Encoding

**Mission 4 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Sklearn Library: You will do several missions with Sklearn. For example, you will Use Scikit-Learn Encoders to encode categorical data.

For more information regarding the grading of the Booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 4 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Let us retry encoding some Mushrooms dataset categorical attributes!  Your code should perform the following tasks to complete this mission:   1. Encode y (the class attribute) using LabelEncoder. 2. Encode selected\_X (a subset of attributes) using OneHotEncoder.   Your code should return encoded\_y and encoded\_X; arrays with the transformed result of applying the encoders.  Hint:  encoded\_y should be a 1D array with only 0s and 1s. Hint: encoded\_X should be a 2D array with only 0s and 1s (the fit\_transform function for a OneHotEncoder returns a sparse matrix by default, hence, to get an array, you can either disable the sparse option in the constructor (OneHotEncoder(sparse=False)) or call toarray() on the sparse matrix returned). HINT: knowing that selected\_X has 2 columns with 10 unique values in the first and 2 unique values in the second, how many columns should encoded\_X have?  Revisit the Mushrooms dataset Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  y = dataset.iloc[:, 0].values  selected\_X = dataset.iloc[:, 3:5].values  # Write your code here. Do not change any other parts of the code  print(encoded\_X[0,:])  return encoded\_y, encoded\_X |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():  import pandas as pd  dataset = pd.read\_csv('mushrooms.csv')  y = dataset.iloc[:, 0].values  selected\_X = dataset.iloc[:, 3:5].values  from sklearn.preprocessing import LabelEncoder, OneHotEncoder  encoded\_y = LabelEncoder().fit\_transform(y)  encoded\_X = OneHotEncoder().fit\_transform(selected\_X).toarray()  print(encoded\_X[0,:])  return encoded\_y, encoded\_X |  |
|  |  |  |

Mission 5 – Editor Mission Score Booster: Data Scaling

**Mission 5 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Sklearn Library: Using Scikit learn Scalers to scale numerical data

For more information regarding the grading of the booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 5 Instructions:**

## Instructions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| To complete this mission, perform the following task: | | | | | | |
|  | Step 1 | Let us retry scaling our friendly Iris dataset!  Your code should perform the following tasks to complete this mission:   1. Load the Iris dataset from scikit-learn as a data vector X and a label vector y. 2. Scale *the first two columns of* the data vector X with a standard scaler (normalize them).   Your code should return the scaled data vector scaled\_columns (a 2D array with *only* the 2 scaled columns).  SELF-CHECK: notice the result of calling mean(axis=0) and var(axis=0) (or var(axis=0)) for scaled\_columns!  Revisit the Iris dataset Here | | |  | |
|  |  | |  |  |  | |
|  |  | | | | |  |
|  |  | | | | |  |

|  |  |  |
| --- | --- | --- |
| Initial Code | | |
| Python | | |
|  |  |  |
|  | def main():  # Write your code here. Do not change any other parts of the code  Import numpy as np  print(np.mean(scaled\_ columns))  return scaled\_columns |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Model Code | | |
| Python | | |
|  |  |  |
|  | def main():      from sklearn.datasets import load\_iris      X, y = load\_iris(return\_X\_y=True)      from sklearn.preprocessing import StandardScaler      scaler = StandardScaler()      scaled\_columns = scaler.fit\_transform(X[:, :2])      import numpy as np      print(np.mean(scaled\_columns))      return scaled\_columns |  |
|  |  |  |

Mission 6 - Quiz Mission: Data Pre-Processing

**Mission 6 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Preprocessing Techniques: Scaling and Encoding

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 6 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Feature scaling is particularly important when mixing what types of input features? | | | | | | |
|  |  | a. | Categorical and numerical features. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Features in different ranges. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Ordinal features. | | |  |
|  |  | | |  |  |  |
|  |  | d. | None of the above. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Data scaling is particularly important when mixing features with different ranges so that the feature with larger numbers in its range does not dominate the calculations. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect. Scaling aims at standardizing the range of values of the data. The range of values of raw data varies widely. In some machine learning algorithms, such as SVM and ANN, objective functions will not work correctly without normalization. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Given the following code snippet, what does this piece of code do?   |  |  |  | | --- | --- | --- | |  |  |  | |  | x\_train = scaler.fit\_transform(x\_train) |  | |  |  |  | | | | | | | |
|  |  | a. | Fits and scales the training data. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Fits the training data. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Transform the shape of the training data. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Scales the training data. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Correct | Great! The function fit\_transform() combines the two operations of fitting the training data and then scales the result. |
| b. | Incorrect | Incorrect. The function fit() is the function that aims at only fitting the scaler to the input data without transforming the data. |
| c. | Incorrect | Incorrect. The function transform() is the function that aims at transforming the data to a previously fitted scaler instance. |
| d. | Incorrect | Incorrect. The function fit\_transform() is not responsible for extracting any statistical data from the input data. |







|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| The ordinal encoder is typically used to encode what? | | | | | | |
|  |  | a. | Numerical values. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Target labels only. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Multi-option input features. | | |  |
|  |  | | |  |  |  |
|  |  | d. | None of the above. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! While the ordinal encoder can be used to encode target labels, it is typically used for multiple input features. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Remember the difference between ordinal and label encoders. | |  |
|  |  |  |  |











Mission 7 - Quiz Mission Score Booster: Data Pre-Processing

**Mission 7 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Scaling and Encoding

For more information regarding the grading of the Booster missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 7 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| The label encoder is typically used to encode what? | | | | | | |
|  |  | a. | Numerical values. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Target labels. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Multi-option input features. | | |  |
|  |  | | |  |  |  |
|  |  | d. | None of the above. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The target encoder can encode one input feature, but it is typically used for target labels. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. Remember the difference between label encoders and ordinal encoders. | |  |
|  |  |  |  |











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| When encoding 3 columns with 6 unique values in the first, 2 unique values in the second and 4 unique values in the third using a one-hot encoder, how many columns will be in the output? | | | | | | |
|  |  | a. | 144. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 12. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | 48. | | |  |
|  |  | | |  |  |  |
|  |  | d. | 6. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! The one-hot encoder makes a column for each unique value with 1 in its place and 0s elsewhere. This makes the output of 6+2+4 = 12 columns. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. The one-hot encoder makes a column for each unique value with 1 in its place and 0s elsewhere. | |  |
|  |  |  |  |











# **Data Visualization**

## **Submodule 5 Activities Summary:**

List the activities learners will be completing during the module, and in the sequence, they will be completed. Please indicate which activities are graded, and which ones are not. All activities must match the stated module objectives listed above. The following list displays examples of activities that may be included. If you have questions about the activities you would like to include in the course, please speak to Learning Services.

1. Interactive lesson mission (Complete reading of the lesson is graded, and this is confirmed by finishing the informal assessment inside the lesson)
2. Quiz mission (Graded)
3. Editor mission (Graded)
4. Emulator mission (Graded)

**Submodule 5 Essential Question (Purpose):**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

What is the importance of data visualization? How do you visualize your data using Matplotlib and Seaborn Libraries?

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Develop Submodule 5 Content**

Repeat this section for every mission.

Start explaining, based on the course structure, the objectives to be covered. You may do this in a table if you find it beneficial. In this section, you will develop content. This section is where you use your knowledge to build connection pieces between theory, resources, and assessments. Use the module objectives as the base for the structure of your content. This section is not focused on curation of resources; it is focused on you, as the SME, providing your expertise to create content. However, if you are referencing any type of source, please cite it in APA format (including open source materials and creative commons material).

Write Here:

Mission 1 - Interactive Lesson Mission: Introduction to Visualization

Section 0: Introduction

|  |
| --- |
| What will you learn in this lesson? |
| In this lesson, you will be introduced to Data Visualization. By the end of this lesson, you will know about:   1. Data visualization importance. 2. Using Matplotlib. 3. Advanced visualizations. 4. Using Seaborn. |

Section 01: Introduction to Visualization

In this section, you will learn how to visualize your data using a Python library called Matplotlib.

|  |
| --- |
| Why Visualization? |
| Representing the data with tools like graphs, charts, maps, and other data visualization tools makes it easier to observe and understand trends and patterns present within the information.  Thus, data analysts rely mainly on the graphical representation of massive amounts of information to make it possible to provide data-driven decisions.  We typically need to visualize data to select the most suitable machine learning model. We may also need to visualize the model to see how well it fits our data. We will use similar Python commands for that. |
| Citation: Visualization[Digital Image]. Retreived from https://www.tapclicks.com/resources/blog/the-ultimate-guide-to-data-visualization/ |



|  |  |  |
| --- | --- | --- |
|  |  | Visualization  Data visualization is the act of organizing the data and information in graphical representation. |

Section 2: The Matplotlib Library

In this section, you will be introduced to the Matplotlib library that is useful for the purpose of visualization.

|  |  |  |
| --- | --- | --- |
|  |  | What is Matplotlib?  A massive data Python plotting library that is used to plot and display several types of graphs, each useful for various representations. |

|  |  |  |
| --- | --- | --- |
|  |  | Important note about Matplotlib:  Matplotlib contains two different modules, Pyplot, and Pylab. We mainly work with Pyplot for representation. |

Installing the Library

|  |
| --- |
| How do you install Matplotlib on Python? |
| To install the Matplotlib library on your machine, type the following command on your terminal.  pip install matplotlib |
| For package installation Guide. See Anaconda installation guide Here  Citation: RoboGarden Course Resources Copyright 2019 |
|  |

Importing the Library

|  |  |  |
| --- | --- | --- |
| How do you import the library in the code? | | |
| After installing the matplotlib library, you can now import it in a Python script to use its functionalities.  When using the pyplot module, it should be explicitly imported with the abbreviated name of plt as in the following command: | | |
|  |  |  |
|  | import matplotlib.pylot as plt |  |
|  |  |  |

Section 3: Basic Plot Components

In this section, you will learn what component a plot can have to look complete and convey information correctly. Each of these components are not a complete visualization tool when considering individually but their combination will make the visualization comprehensive.

|  |
| --- |
| Overview  Before discussing some of the main visualizations supported by the Matplotlib library, let us discuss the main components of matplotlib pyplot.   * **Figure**: the figure is the entire window where the plot is drawn, and it is the top-level component. It may also contain several subcomponents within to better describe the plotted graph. * **Axes:** they represent the area which the plot functions on Matplotlib would plot the data on. * **Axes labels**: a description of the values of each axis. * **Plot title**: a descriptive title of the plot. * **Legend**: a box showing description of each graph of a multi-graph plot. |
| Hint: plt is the alias used for the matplotlib.pyplot package throughout this tutorial.  When displaying a plot using the plt.show () method, a figure with the required plot is displayed.  There are several features you could apply to your plot graph to further display certain properties or to personalize it. Among the customizable features in the pyplot are the following: |

|  |  |  |
| --- | --- | --- |
|  |  | Notes about matplotlib extra information:  You might hear of some Matplotlib components under the name of artist objects. Don’t worry, artist objects are virtually any object that the Matplotlib package offers any developer. Axes, figures and various other objects in Matplotlib are considered artist objects. |

Adjusting Basic Plot Components

In this section, you will see three modifications to plot and you will learn how to plot information.

|  |
| --- |
| Overview  You will need to modify these main components of plots in order to properly visualize your data.  In this section, you will be introduced to some of the basic modifications you will most probably need while working with Matplotlib. |

|  |  |
| --- | --- |
| 1. Setting Plot Title | |
| Plot title is set using the method plt.title (“title of figure”), which takes a string as input.  plt.title('Cutom Figure') | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |
| --- | --- |
| 1. Customizing Axes Labels | |
| The Pyplot module allows you to set labels to the figure axes to describe the kind or type of values they represent. This is done using the methods:  • plt.xlabel() to label the horizontal (x) axis.  • plt.ylabel() to label the vertical (y) axis.  plt.xlabel("Horizontal label")  plt.ylabel("Vertical label") | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |
| --- | --- |
| 1. Adding Legends | |
| When plotting several graphs on the same axes, you may want to clarify the graphs from each other and what they represent using a legend. This could be done using the plt.legend(loc='') method.  The loc parameter string consists of two parts; the vertical location of the legend (upper or lower), followed by the horizontal location of the legend (left or right).  plt.plot(x,y1, color='red', label='first y values')  plt.plot(x,y2, color='blue', label='second y values')  plt.legend(loc='upper left') | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  To change the title of the plot to "Employee status" and change the axes labels to age on the y-axis and salary on the x-axis, what lines of code do you write? | | | | | | |
|  |  | a. | plt.title('Employee status')  plt.xlabel('Age')  plt.ylabel('Salary') | | |  |
|  |  |  | |  |  |  |
|  |  | b. | plt.title\_labels(['Employee status, Age, Salary']) | | |  |
|  |  |  | |  |  |  |
|  |  | c. | plt.title('Employee status')  plt.ylabel('Age')  plt.xlabel('Salary'). | | |  |
|  |  | | |  |  |  |
|  |  | d. | None of the above. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

MCQ Responses Table

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. Notice the reversed axes labels. |
| b. | Incorrect | Unfortunately, this is the wrong syntax. |
| c. | Correct | Awesome! |
| d. | Incorrect | Unfortunately, this is incorrect. One of the answers is correct. |

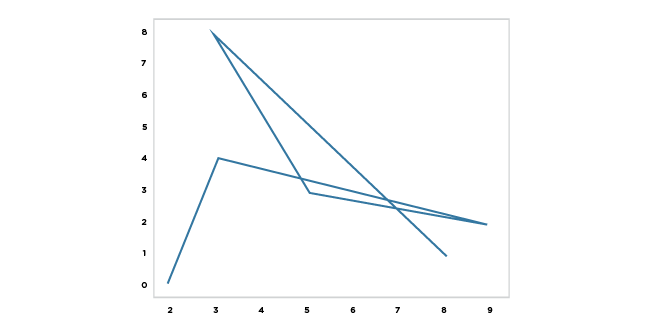
|  |
| --- |
| What is a Matplotlib Subplot? |
| The Subplot is exactly what it sounds like. It is a way of adding multiple subplots to the same figure.  Is adding subplots synonymous to adding new axes? Yes. In most cases, axes and subplots in Matplotlib resemble the same thing. |
| Citation: RoboGarden Course Resources Copyright 2019 |

Section 4: Basic Visualizations

In this section, you will learn some basic type of plots to visualize your data.

|  |
| --- |
| What are the four basic plots for data visualization? |
| Example plots that are supported by the Pyplot module of the Matplotlib library are:   * Line plots * Scatter plots * Bar plots * Histograms   These visualizations are available in most of the visualization libraries. Together they visualize the direct relationships within the attributes and between these attributes.  The Pyplot plots can be used to plot lists, Numpy arrays and matrices, and Pandas data frames by providing them as arguments to the different methods (plot types) of the module. |

|  |  |  |
| --- | --- | --- |
| Visualization Basic Plots  Click through the carousel to learn more about Visualization of these plots. | | |
|  |  |

Citation: RoboGarden Course Resources Copyright 2019

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Line Plots  The method plt.plot(x,y) is used to plot a line graph with the given data in arrays x and y as the axes values.  To display the figure, call the method plt.show() at the end.  import matplotlib.pyplot as plt  x=[1,2,3,4,5,6]  y=[7,4,2,3,8,1]  plt.plot(x, y)  plt.show() | | | | |
|  |  |    |  |  |

|  |  |  |
| --- | --- | --- |
| **#** | **Image Description** | **Image** |
|  | Line Plots  The method plt.plot(x,y) is used to plot a line graph with the given data in arrays x and y as the axes values.  To display the figure, call the method plt.show() at the end.  import matplotlib.pyplot as plt  x=[1,2,3,4,5,6]  y=[7,4,2,3,8,1]  plt.plot(x, y)  plt.show() | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Bar Plots  The method plt.bar(x,y) is used to plot a bar graph with the given data in arrays x and y as the axes values.  import matplotlib.pyplot as plt  x=[1,2,3,4,5,6]  y=[7,4,2,3,8,1]  plt.bar(x, y)  plt.show()  **Note:**  Matplotlib also supports the horizontal bar plots using the function plt.barh()  import matplotlib.pyplot as plt  x=[1,2,3,4,5,6]  y=[7,4,2,3,8,1]  plt.barh(x, y)  plt.show() | Citation: RoboGarden Course Resources Copyright 2019    Citation: RoboGarden Course Resources Copyright 2019 |
|  | Scatter Plots  The Pyplot module also gives the ability to represent the data as a scatter plot using the method plt.scatter(x,y).  import matplotlib.pyplot as plt  x=[1,2,3,4,5,6]  y=[7,4,2,3,8,1]  plt.scatter(x, y)  plt.show() | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Histograms  The Pyplot module allows the representation of 1D and 2D histograms using the method plt.hist(x) for plotting one-dimensional histograms and the method plt.hist2d(x,y) for plotting two-dimensional histograms.  import matplotlib.pyplot as plt  x=[1,2,3,4,5,6]  y=[7,4,2,3,8,1]  plt.hist2d(x, y)  plt.show() | Citation: RoboGarden Course Resources Copyright 2019 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Fill in the Missing Blank | | | | | | |
| To create a bar plot in Matplotlib, | | Dropdown  | | |  | |
| plt.show() | | |
| plt.plot() | | |
| plt.bar() | | |
| plt.scatter() | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. plt.show() opens a window to show the resulting plot by the Matplotlib plotting module. |
| b. | Incorrect | Unfortunately, this is incorrect. Matplotlib’s plot() function Plots y versus x as lines and/or markers. |
| c. | Correct | Great! plt.bar() is used to create bar plots. |
| d. | Incorrect | Unfortunately, this is incorrect. Matplotlib’s scatter() plots a scatter plot of x and y with varying marker size or color. |

## Test your knowledge

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Practice Question | | | | | | |
| To represent data of different employees in a company based on salary and age, | | ……..  | | | can be used for visualization. | |
| Histogram | | |
| Scatter plot | | |
| Bar plot | | |
| Line plot | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Correct Answer!  Scatter plot is the best graph to represent this type of discrete data. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Incorrect Answer! This data should be represented as discrete data having two axes, so the option you chose is not the best option. | |  |
|  |  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Section 5: Visualization for EDA

In this section, you will learn how to explore data using statistic concepts.

|  |  |  |
| --- | --- | --- |
|  |  | Exploratory Data Analysis (EDA)  EDA refers to the process of thoroughly investigating the different patterns and signs in data in order to spot outliers, test initial hypotheses or understand your data based on different statistical approaches and graphical representations. |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Meaningful Statistical visualizations  Some visualizations are important for Exploratory data analysis.  In this section, you will be introduced to three of the most important visualizations:   * Boxplots * Violin plots * Heat maps   Now you will learn about some of the most important visualizations in the field of EDA.  Click through the carousel to learn more about these visualizations. | | | | |
|  |  | Citation: Boxplots [Digital Image]. Retrieved from https://www.simplypsychology.org/boxplots.html |  |  |
| 1. Box plots   Boxplots are visualizations that aim at displaying mainly the spread and the dispersion of the data in your dataset. It displays the distribution based on a **five-number summary** including the maximum, the minimum, the absolute median (the median of the whole dataset), the lower median (the median of the elements whose values are less than the absolute median) and the upper median (the median of the elements whose values are more than the absolute median).  Boxplots can tell you exactly if your dataset is symmetrical or skewed and how skewed is it.  Boxplots (as shown in the figure) consist of:   * The box: it represents the data between the lower median and the upper median. The line in the middle of the box represents the absolute median. The location of the median with respect to your data can show if the data is skewed and how skewed the data is. * The whiskers: They represent the regions between the minimum and maximum with both the lower and the upper medians respectively. | | | | |
|  |  |    |  |  |

Image Carousel Content Table

|  |  |  |
| --- | --- | --- |
| **#** | **Image Description** | **Image** |
|  |  | 1. Box plots   Boxplots are visualizations that aim at displaying mainly the spread and the dispersion of the data in your dataset. It displays the distribution based on a **five-number summary** including the maximum, the minimum, the absolute median (the median of the whole dataset), the lower median (the median of the elements whose values are less than the absolute median) and the upper median (the median of the elements whose values are more than the absolute median).  Boxplots can tell you exactly if your dataset is symmetrical or skewed and how skewed is it.   * Boxplots (as shown in the figure) consist of: The box: it represents the data between the lower median and the upper median. The line in the middle of the box represents the absolute median. The location of the median with respect to your data can show if the data is skewed and how skewed the data is.   The whiskers: They represent the regions between the minimum and maximum with both the lower and the upper medians respectively. |
|  | Citation: Violin plots [Digital Image]. Retreived from https://medium.com/@shobhitsrivastava18th/boxplot-v-s-violinplot-93a1d9f3a831 | 1. Violin Plots   Similar to boxplots, violin plots are visualizations that aim at displaying mainly the spread and the dispersion of the data in your dataset. Based on the **five-number summary** including the maximum, the minimum, the absolute median, the lower median and the upper median. However, the major advantage of violin plots to boxplots is that violin plots have the ability of showing the exact entire distribution of the data. |
|  | Citation: Heat map [Digital Imag]. Retreived from https://seaborn.pydata.org/generated/seaborn.heatmap.html  Image result for heatmap | Heat maps  Heat maps are not strictly statistical. Yet they are one of the most important visualization approaches when dealing with different datasets.  A heat map (or heat map) is a graphical representation of data where values are depicted by color. Heat maps make it easy to visualize complex data and understand it at a glance. The practice we now call heat maps is thought to have originated in the 19th century, where manual gray-scale shading was used to depict data patterns in matrices and tables. |

|  |
| --- |
| Quartiles  An important aspect to properly grasp the concepts of box plots (one of the visualization types) are the quartiles. |
| median (Q2/50th Percentile): the middle value of the dataset.  first quartile (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the dataset.  third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the dataset.  interquartile range (IQR): 25th to the 75th percentile. |
| Citation: Robogarden Course resources Copyright 2019 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Knowledge Check  Consider the dataset {2, 3, 4, 3, 1, 12, 5}, which of these numbers fall within the IQR? | | | | | | |
|  |  | a. | 12. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | 1. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | 4. | | |  |
|  |  | | |  |  |  |
|  |  | d. | None of the above. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

MCQ Responses Table

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. 12 is the maximum in the given sample dataset which means it lies outside of the inter-quartile range. |
| b. | Incorrect | Unfortunately, this is incorrect. 1 is the minimum in the given sample dataset which means it lies outside of the inter-quartile range. |
| c. | Correct | Great! 4 lies between the absolute median and the upper median. |
| d. | Incorrect | Unfortunately, this is incorrect. There exists a correct answer. |

Section 6: Seaborn

In this section, you will be introduced to the Seaborn library that is a useful library for visualization of statistical notions.

|  |  |  |
| --- | --- | --- |
|  |  | What is Seaborn?  Seaborn is a visualization library in Python. It is built on top of Matplotlib. It mainly provides a high-level interface to draw statistical graphics. |

|  |  |  |
| --- | --- | --- |
| Importing Seaborn | | |
| After installing the Seaborn library, you can now import it in a Python script to use its functionalities.  In this example, Seaborn will be explicitly imported with the abbreviated name of sns as in the following command: | | |
|  |  |  |
|  | Import seaborn as sns |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  | Important note about Seaborn library:  In the following section, You will use the built-in datasets in the Seaborn library. However, Seaborn mainly deals with data in the form of a Pandas data frame. You can easily use your own data just by understanding the simple examples on the built-in datasets. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample Code  In the following section, you will use Seaborn to visualize the Iris dataset that you previously encountered in the missions. You can follow instructions by entering the following codes in an editor such as Spider in your own computer. | | | | | |
|  | Step 1 | Using a Built-in Dataset  You can load any of the built-in datasets in Seaborn using the load\_dataset()  Hint:  The load\_dataset() returns a pandas data frame. You can easily replace that with your own data. You may want to try this code by entering the code in an editor in your computer.  df = sns.load\_dataset("iris") | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Boxplots  Using the sns.boxplot(), you can plot the boxplot of any specific feature in your dataset. You may want to try this code by entering the code in an editor in your computer. Please import the Seaborn dataset first (import seaborn as sns) and load your dataset using (df = sns.load\_dataset("iris")) and then add the following commands.  sns.boxplot(x="petal\_width", data= df)    Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |
|  | Step 3 | Violin Plots  Using the sns.violinplot(), you can plot the violin plot of any specific feature in your dataset. You may want to try this code by entering the code in an editor in your computer. Please import the Seaborn dataset first (import seaborn as sns) and load your dataset using (df = sns.load\_dataset("iris")) and then add the following commands.  sns.violinplot(x="petal\_width", data= df)    Citation: RoboGarden Course resources Copyright 2019 | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Step 4 | Heatmaps  Using the sns.heatmap(), you can plot heatmaps easily. In case you want to try this command. Please import the Seaborn dataset first (import seaborn as sns) and load your dataset using (df = sns.load\_dataset("iris")) and then add the following commands.  Notes:   * In the following example df.corr() represents the correlation matrix of the iris dataset using the corr() function of the pandas dataframe. * Check the arguments of the heatmap() in the seaborn documentation!   sns.heatmap(data= df.corr())    Citation: RoboGarden Course resources Copyright 2019 | | |  |
|  |  | |  |  |  |

Section 06: Lesson Summary

|  |
| --- |
| What did we cover in this lesson? |
| Great work! You were introduced to data visualization, EDA and you used Matplollib and Seaborn for plotting data.  You were introduced to Boxplots, Violin plots and heatmaps. These plots are important plots that are used in Exploratory data analysis based on the information of the median and the interquartile range. |



## **Submodule 5 Application Mission(s) Details**

Repeat this section for every mission. The activities listed below are examples. See page 7 for the activities summary.

Below you will see the details of all the application missions within submodule 5.

Mission 2 - Emulator Mission: Plot a parabola

Emulators are similar to quizzes. The emulator shows the student a code output then asks the student to choose the correct code snippet to produce the same output.

**Mission 2 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming using Matplotlib

For more information regarding the grading of the Emulator missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 2 Instructions:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Which of the following lines of code produces a continuous plot of the function x\*\*2?   |  |  |  | | --- | --- | --- | |  |  |  | |  | from matplotlib import pyplot as plt  import numpy as np  x = np.linspace(-100, 100) |  | |  |  |  | | | | | | | | |
|  |  | | | | | |  |
|  | Citation: RoboGarden Course resources Copyright 2019 | | | | | |  |
|  |  | | | | | |  |
|  |  | a. | plt.scatter(x, x\*\*2) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | b. | plt.plot(x, x\*\*2) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | c. | plt.bar(x, x\*\*2) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | d. | plt.hist(x, x\*\*2) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | | |  | |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. plt.scatter will create a scatter plot of the points, as shown in the figure. |
| b. | Incorrect | Yes, this code will achieve the desired result. |
| c. | Correct | Unfortunately, this is incorrect. plt.bar will create a plot of vertical bars at the equally distributed points, as shown in the figure. |
| d. | Incorrect | Unfortunately, this is incorrect. This will result in an error. |







Mission 3 - Emulator Quiz Mission: Plot independent points

**Mission 3 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming using Matplotlib

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 3 Instructions:**

In this section, you will see a quiz based on the code that they select to input in the emulator. Your code should simulate the given image as its output.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Which of the following commands is used to visualize an ordered dataset with one independent and one dependent variable?    Citation: RoboGarden Course resources | | | | | | | |
|  |  | | | | | |  |
|  |  | | | | | |  |
|  |  | a. | plt.scatter(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | b. | plt.plot(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | c. | plt.bar(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | d. | plt.hist(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | | |  | |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Correct. | Yes, this code will achieve the desired result. |
| b. | Incorrect | This code will not achieve the desired result. |
| c. | Incorrect | This code will not achieve the desired result. |
| d. | Incorrect | This code will not achieve the desired result. |







Mission 4 – Emulator Mission Score Booster: Plot a line

**Mission 4 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming using Matplotlib

For more information regarding the grading of the Emulator missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 4 Instructions:**

|  |  |
| --- | --- |
|  | Lesson 5: Score Booster Emulator – Plot a line |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Which of the following lines of code produces a continuous plot of the function x?   |  |  |  | | --- | --- | --- | |  |  |  | |  | from matplotlib import pyplot as plt  import numpy as np  x = np.linspace(-100, 100) |  | |  |  |  | | | | | | | | |
|  |  | | | | | |  |
|  | Citation: RoboGarden Course resources Copyright 2019 | | | | | |  |
|  |  | | | | | |  |
|  |  | a. | plt.scatter(x, x) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | b. | plt.plot(x, x) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | c. | plt.bar(x, x) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | d. | plt.hist(x, x) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | | |  | |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. plt.scatter will create a scatter plot of the points, as shown in the figure. |
| b. | Incorrect | Yes, this code will achieve the desired result. |
| c. | Correct | Unfortunately, this is incorrect. plt.bar will create a plot of vertical bars at the equally distributed points, as shown in the figure. |
| d. | Incorrect | Unfortunately, this is incorrect. This will result in an error. |







Mission 5 – Emulator Mission Score Booster: Scattered points

**Mission 5 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Programming using Matplotlib

For more information regarding the grading of the Emulator missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 5 Instructions:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Which of the following commands is used to visualize an ordered dataset with one independent and one dependent variable?    Citation: RoboGarden Course resources | | | | | | | |
|  |  | | | | | |  |
|  |  | | | | | |  |
|  |  | a. | plt.scatter(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | b. | plt.plot(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  |  | |  | |  |  |
|  |  | c. | plt.bar(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | d. | plt.hist(x, y) | | Citation: RoboGarden Course resources Copyright 2019 | |  |
|  |  | | |  | |  |  |
|  |  | | |  | |  |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Correct. | Yes, this code will achieve the desired result. |
| b. | Incorrect | This code will not achieve the desired result. |
| c. | Incorrect | This code will not achieve the desired result. |
| d. | Incorrect | This code will not achieve the desired result. |







Mission 6 - Quiz Mission: Data Visualization

**Mission 6 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Matplotlib

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 6 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fill in the missing blank | | | | | | |
| * To create a bar plot in Matplotlib, | | …….. | | | Is used. | |
| plt.show() | | |
| plt.plot() | | |
| plt.bar() | | |
| plt.barplot() | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |
|  | | | | | | |

Responses Table

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. plt.show() displays the required plot. |
| b. | Incorrect | Unfortunately, this is incorrect. plt.plot() plots the line plot between different variables. |
| c. | Correct | Great! The method plt.bar(x,y) is used to plot a bar graph with the given data in arrays x and y as the axes values. |
| d. | Incorrect | Unfortunately, this is incorrect. There is no function with the name plt.barplot() |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| * To represent data of different competitors for weight and height features | | …….. | | | Can be used for visualization. | |
| histogram | | |
| Scatter plot | | |
| Bar plot | | |
| box plot | | |
|  |  | |  |  | |  |
|  |  | |  |  | |  |

|  |  |  |
| --- | --- | --- |
| **Item** | **Status** | **Explanation Text/Images** |
| a. | Incorrect | Unfortunately, this is incorrect. A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. |
| b. | Correct | Great! Scatter plots can be used to represent scattered data points along different axes representing different features. |
| c. | Incorrect | Unfortunately, this is incorrect. a bar plot is not the best representation of the relationship between two variables. |
| d. | Incorrect | Unfortunately, this is incorrect. Boxplots are visualizations that aim at displaying mainly the spread and the dispersion of the data in your dataset. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which of the following Python libraries is used for visualizations? | | | | | | |
|  |  | a. | Pandas. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Numpy. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Matplotlib. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Pickle. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Matplotlib is a very popular visualization library for Python. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. The library you chose is not a visualization library. | |  |
|  |  |  |  |



























Mission 7 – Quiz Mission Score Booster: Data Visualization

**Mission 7 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Knowledge: Matplotlib

For more information regarding the grading of the Quiz missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 7 Instructions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which kind of plot should be used to visualize unordered 2D data, e.g., prices of houses against their sizes? | | | | | | |
|  |  | a. | Bar. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Box. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Scatter. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Violin. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Scatter plots are ideal for visualizing unordered 2D data as points without much clutter. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. The visualization technique you chose is not the best for visualizing unordered data. | |  |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Which kind of plot should be used to visualize dense ordered data, e.g., continuous readings from a sensor against time? | | | | | | |
|  |  | a. | Scatter. | | |  |
|  |  |  | |  |  |  |
|  |  | b. | Histogram. | | |  |
|  |  |  | |  |  |  |
|  |  | c. | Line plots. | | |  |
|  |  | | |  |  |  |
|  |  | d. | Box. | | |  |
|  |  | | |  |  |  |
|  |  | | |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Correct |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Great! Line plots are ideal for visualizing ordered data, especially when they are dense as they show the ups and downs without gaps. | |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | That is Incorrect |  |
|  | Explanation |  |  |
|  |  |  |  |
|  | Unfortunately, this is incorrect. For dense ordered data, you would like to notice the different increases and decreases in the slope and the direction of any change and a general representation of the point. What kind of plot would you think best satisfies these needs? | |  |
|  |  |  |  |























# Submodule 6

## **Submodule 6 Activities Summary:**

List the activities learners will be completing during the module, and in the sequence they will be completed. Please indicate which activities are graded, and which ones are not. All activities must match the stated module objectives listed above. The following list displays examples of activities that may be included. If you have questions about the activities you would like to include in the course, please speak to Learning Services.

1. Interactive lesson mission (Complete reading of the lesson is graded, and this is confirmed by finishing the informal assessment inside the lesson)

**Submodule 6 Essential Question (Purpose):**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Implement what you’ve learnt!

Learn Git and GitHub.

For more information regarding the grading of the missions (quiz, emulators and editors) of this submodule please refer to Module 1 Introduction section as well as the Orientation.

**Develop Submodule 6 Content**

Repeat this section for every mission.

Start explaining, based on the course structure, the objectives to be covered. You may do this in a table if you find it beneficial. In this section, you will develop content. This section is where you use your knowledge to build connection pieces between theory, resources, and assessments. Use the module objectives as the base for the structure of your content. This section is not focused on curation of resources; it is focused on you, as the SME, providing your expertise to create content. However, if you are referencing any type of source, please cite it in APA format (including open source materials and creative commons material).

Write Here:

Mission 1 - Interactive Lesson Mission: Git and GitHub

Section 00: introduction

|  |
| --- |
| What will you learn in this lesson? |
| In this lesson, You will be introduced to Git. By the end of this lesson, you will understand:   1. How to use control-version systems to keep track of changes in your code. 2. How to coordinate work between programmers. |

Section 01: What is Git?

|  |  |  |
| --- | --- | --- |
|  |  | Note about Git  A system that keep tracks or records changes to a file or set of files overtime so that you can recall a specific version later |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| But why do you need Git?  Click through the carousel to learn about what Git could enable the user to do | | | | |
|  |  | revert files to a previous state |  |  |
|  |  |     |  |  |

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Revert files to a previous state. |
|  | Revert the entire project back to a previous state. |
|  | Compare changes overtime. |
|  | Know who changed what and when. |
|  | Keep the entire code and history locally on your machine. |

Section 02: Git Terminology

|  |  |  |
| --- | --- | --- |
|  |  | Snapshot  What all your files look like at a given point in time. |

|  |  |  |
| --- | --- | --- |
|  |  | Commit  It is the act of creating a snapshot.  commit contains three pieces of information:   * Information about how the files changed previously. * A reference to the commit that came before it. * A commit ID or hash |

|  |  |  |
| --- | --- | --- |
|  |  | Repository  A collection of all the files, and the history of those files; it consists of all your commits.  Repository can be on a local machine or a remote server like GitHub.  Note:  Repository sometimes called repo for short. |

|  |  |  |
| --- | --- | --- |
|  |  | Pull  Downloading commits that don not exist on your machine from a remote server. |

|  |  |  |
| --- | --- | --- |
|  |  | Branch  It is a place where commits live in.  A project can have many branches. The main branch in a repository is called the master branch.  A project is a bunch of commits linked together that live on some branch contained in a repository.  Hint:  If you want to make a change to your project or add a new feature without effecting it, add a branch. Your project won’t be effected; you can restore the snapshot of your project before adding this feature.  Hint:  You can merge the new branch to the master branch if the feature has been added successfully. |

Section 03: How to use Git?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| For a simple project:  The following is some of the main steps that are most probably used to use Git:  Click through the carousel to view the steps used or Git. | | | | |
|  |  | Citation: RoboGarden Course Resources Copyright 2019 |  |  |
| You have a master branch | | | | |
|  |  |    |  |  |

Image Carousel Content Table

|  |  |  |
| --- | --- | --- |
| **#** | **Image Description** | **Image** |
|  | You have a master branch | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Adding a branch | Citation: RoboGarden Course Resources Copyright 2019 |
|  | Merging a branch | Citation: RoboGarden Course Resources |

Section 04: what is GitHub

|  |  |  |
| --- | --- | --- |
|  |  | What is GitHub?  GitHub is a Git repository hosting service. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| But why do you need GitHub?  Using GitHub enables the user to:  Click through the carousel to learn more about what GitHub enables the user. | | | | |
|  |  | Share code with other developers |  |  |
|  |  |     |  |  |

|  |  |
| --- | --- |
| **Item** | **Text** |
|  | Share code with other developers. |
|  | Download projects and work on them. |
|  | Upload projects, edit and merge them with the main codebase. |

Section 05: Guidelines to using GitHub

In this section, you will be introduced to Git that is a version control system and is used for tracking change of your code and coordinating work among programmers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Installing Git | | | | | |
|  | Step 1 | Visit the following link to download Git version which is compatible with your platform:  <https://git-scm.com/book/en/v2/Getting-Started-Installing-Git> | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Type the following command to check Git version:  git --version | | |  |
|  |  | |  |  |  |
|  | Step 3 | Enter your username and email: git config --global user.name your\_name git config --global user.email your\_email | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Create a Repository | | | | | |
|  | Step 1 | Create a new folder. | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Open the terminal inside the new folder and type the following command:  git init  You will get the following message:  Initialized empty Git repository in path/new folder/.git/ | | |  |
|  |  | |  |  |  |
|  | Step 3 | Type the following command to check the status of the repository  git status  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modify a File in Git | | | | | |
|  | Step 1 | Create a new python file, and call it index.py | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Add the following lines inside this python file x = 2 | | |  |
|  |  | |  |  |  |
|  | Step 3 | Save the file and check the repository status again:  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Make a commit | | | | | |
|  | Step 1 | Type the following commands: git add . git commit -m “create python file” | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |
|  | Step 3 | Check the status again:  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Create new branch | | | | | |
|  | Step 1 | To show the current branches type the following command: git branch  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Create an additional file.  Create a branch call it “adding\_file” and log to it.  Show the current branches again. | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Merge to the master branch | | | | | |
|  | Step 1 | Switch to the master branch:  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Type the following command:  $git merge adding\_file | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Create a GitHub repository | | | | | |
|  | Step 1 | Create an account at Github:  Link: www.github.com | | |  |
|  |  |  |  |  |  |
|  | Step 2 | Click on the’+’ sign on the top right corner of the GitHub home page and select New Repository:  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |
|  | Step 3 | Type the repository name:  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Step 4 | Choose the second option, because you already have a repository. Then copy these commands and run it on the terminal in the root folder of the project. | | |  |
|  |  |  |  |  |  |
|  | Step 5 | You will be asked to enter your username and password:  Citation: RoboGarden Course Resources Copyright 2019 | | |  |
|  |  | |  |  |  |
|  | Step 6 | * Note that you can see the 3 commits we had and the one branch. * The URL of the current page is the link for your repository. | | |  |
|  |  | |  |  |  |

**Module 1 Summary**

Please write a summary for students to capture what they learned in this module. Make sure the essential question is answered here.

Write here:

Section 1: Summary of the Module

|  |
| --- |
| In this module, you learned how to:   * Discuss the concept of data and the types of data. * Load different online available data-sets. * Process data for machine learning. * Visualize data using Matplotlib. |
|  |

|  |
| --- |
| What will you learn next? |
| **You have been introduced to the concept of data processing and visualization without really using them in applying machine learning. In the next module, you will become familiar with machine learning, especially classification. In the subsequent module, you will learn how to classify data using machine learning algorithms.** |
| Citation: RoboGarden Course Resources Copyright 2019 |

**Module 1 Assessment(s) Define and map this – quizzes, mini projects, etc.**

Define the graded module assessments that are being used.

Write summary here:

Mini-project: Chocolate!

**Mission 1 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Skills Tested:

* Using Pandas Programming to read the dataset.
* Using Scikit-learn to preprocess your data.
* Data Visualization Using Matplotlib

For more information regarding the grading of the Mini-Project missions please refer to Module 1 Introduction section as well as the Orientation.

**Mission 1 Instructions:**

Description

|  |
| --- |
| Chocolate's deliciousness is one of the few things that almost all people agree on. However, which chocolate bars are worth more than others? In this project, we shall try to find out!  You are required to Explore the Chocolates dataset, visualize the data and manipulate the features to fit your needs.  *License*: These ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society and it is publicly available online.  You will need to use GitHub to complete this mini-project. Find Guidelines of Using GitHub Here |

|  |
| --- |
| Expected Output |
| By the end of this mini-project, you are supposed to deliver within your code:   * The count of the tuples in the given dataset. * The count of the names of unique company names from the attributes. * The count of reviews in 2013 from the attributes. * The count of missing values in a specifically given * An output plot of the Histogram of the values in the column named Ratings. * An output plot of the scatter plot between the Cocoa Percent values against the Rating values. * The Normalized ratings column values. |

Instructions

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Follow these instructions to complete your mini project. | | | | | | | | |
|  | Step 1 | Download the Dataset  Download the Dataset from the following link:  <https://www.kaggle.com/rtatman/chocolate-bar-ratings> | | | | | |  |
|  |  |  | | |  |  | |  |
|  | Step 2 | **Reading the Dataset** Read the dataset into a Pandas DataFrame!  Does the dataset include any missing values? If so, drop them!  **Hint:**  Pandas can do that with one line of code! | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 3 | **Exploring the Dataset**  Answer the following questions about the dataset using Python commands   * How many tuples are there in the dataset? * How many unique company names are there in the dataset? * How many reviews are made in 2013 in the dataset? * In the BeanType Column, how many missing values are there?   **Hint:**  Each question should require few lines of code! | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 4 | **Visualization**  Visualize the Ratings column with a histogram!  Comment on the resulting figure! | | | | | |  |
|  | Step 5 | **Convert and Visualize**   * **Convert the column Percent**   1. Change the type of values in the column percent from string values to numerical values. * **Visualize**   Plot the converted numerical Cocoa Percent values against the Rating values!  From what you see, does more cocoa in a bar correspond to higher rating?  Hint:  try a scatter plot with small alpha, e.g., 0.1, to flush out the density of each point. | | | | | |  |
|  | Step 6 | **Normalization**  Normalize the Ratings Column and print the Results. | | | | | |  |
|  | Step 7 | **Challenge yourself (Optional)**  Can you make a list of the companies ordered by their average score (averaged over each company's review). | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 8 | **Encoding**  Suppose we are interested in the companies names and locations for some collective analysis, encode the two categorical columns with the encoder you think is best for the job! | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 9 | Submit your Jupyter Notebook to Github | | | | | |  |
|  |  | Copy and paste your GitHub link in the field below, then click *Submit*. | | | | | |  |
|  |  |  |  | | | |  |  |
|  |  | https://www.github.com | | |  |
|  |  | | | |  |
|  |  | | | |  |  | |  |

|  |
| --- |
| Submitted Successfully |

Mini-project: Football!

**Mission 2 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Skills Tested:

* Using Pandas Programming to read the dataset.
* Using Scikit-learn to preprocess your data.
* Data Visualization Using Matplotlib

For more information regarding the grading of the Mini-Project missions of please refer to Module 1 Introduction section as well as the Orientation.

**Mission 2 Instructions:**

Description

|  |
| --- |
| Football is one of the most popular sports in the world. Currently, the hype for the matches cannot be understated and skilled players have gone beyond celebrity status! In this mini-project, we shall look at data collected about matches from 1872 to 2019!  *License*: The dataset is **CC0: Public Domain** and it is publicly available online.  You will need to use GitHub to complete this mini-project. Find Guidelines of Using GitHub Here |

|  |
| --- |
| Expected Output |
| By the end of this mini-project, you are supposed to deliver within your code:   * The count of the tuples in the given dataset. * The count of the names of unique tournament names from the attributes. * The count of matches in 2018 from the attributes. * The count of wins, losses or draws given in the data columns. * An output plot of the Pie chart of the wins, losses or draws. * An output plot of the Pie chart of the neutral column in the input dataset. * The count of the names of unique team names from the attributes. |

Instructions

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Follow these instructions to complete your mini project. | | | | | | | | |
|  | Step 1 | Download the Dataset  Download the Dataset from the following link:  <https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017> | | | | | |  |
|  |  |  | | |  |  | |  |
|  | Step 2 | **Reading the Dataset** Read the dataset into a Pandas DataFrame!  Does the dataset include any missing values? If so, drop them!  **Hint:**  Pandas can do that with one line of code! | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 3 | **Exploring the Dataset**  Answer the following questions about the dataset using Python commands   1. How many tuples are there in the dataset? 2. How many tournaments are there in the dataset?   **Hint:**  Each question should require few lines of code! | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 4 | **Convert and Deduce**  Convert the column date to timestamps!  Find out how many matches in the dataset were played in 2018.  **Hint:**  Use the date column. | | | | | |  |
|  | Step 5 | **Team Statistics**  Calculate How many times the home team won, lost or had a draw. | | | | | |  |
|  | Step 6 | **Visualization**   1. Plot the numbers extracted from step 5 in a Pie chart. 2. Plot the neutral column as a pie chart.   **Hint:**  Try to Visualize the neutral column Using Pandas (Only one line of code). | | | | | |  |
|  | Step 7 | **Unique teams**  How many unique teams are there in the dataset? Find out! | | | | | |  |
|  |  | | | |  |  | |  |
|  |  | | | |  |  | |  |
|  | Step 9 | Submit your Jupyter Notebook to Github | | | | | |  |
|  |  | Copy and paste your GitHub link in the field below, then click *Submit*. | | | | | |  |
|  |  |  |  | | | |  |  |
|  |  | https://www.github.com | | |  |
|  |  | | | |  |
|  |  | | | |  |  | |  |

|  |
| --- |
| Submitted Successfully |

Mission 3 – Module 1 overview project (Accumulative project): Breast Cancer Campaign

**Mission 3 Essential Purpose:**

Repeat this section for every mission.

Write the essential question learners should explore during this lesson. Write the purpose as a question that will be answered throughout the mission. Ie. Why is Python important for Machine Learning?

Write Here:

Skills Tested:

* Pandas Programming.
* Preprocessing in Scikit-learn.
* Visualization in Seaborn.

**Mission 3 Instructions:**

Description

|  |
| --- |
| An insurance company decided to start a breast cancer awareness campaign.  The first step for raising awareness is spreading the word about the facts and the statistics of the problem to the public. Requirements In this part of your project, you are required to:   * Perform exploratory data analysis on the data from the Wisconsin breast cancer dataset. * Visualize the data and show your visualizations. * Preprocess the dataset to be ready for future work.   Note:  You will need to use GitHub to complete this project. Find Guidelines of Using GitHub Here  Citation: RoboGarden Course Resources Copyright 2019 |

Instructions

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Follow these instructions to complete your project. | | | | | | | | |
|  | Step 1 | Download the Dataset  Download the Dataset from the following link:  https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/download | | | | | |  |
|  |  |  | | |  |  | |  |
|  | Step 2 | **Reading the Dataset** Read the dataset into a Pandas DataFrame!  Does the dataset include any missing values? If so, drop them!  **Hint:**  Pandas can do that with one line of code! | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 3 | **Dropping Un-necessary columns**  Choose the features you think are relevant to our analysis! There are A LOT of features in this dataset but we have to make our models training time reasonable.  **Hint:**  Notice the fact that some of the data in the Breast Cancer dataset is irrelevant to the research such as id attribute. | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 4 | **Preprocessing**  Perform any needed pre-processing on the chosen features including:   * Scaling. * Encoding. * Dealing with Nan values. * Normalization.   **Note:**  You need to output the result of your pre-processing to an output CSV called “data\_refined.csv” | | | | | |  |
|  | Step 5 | Visualization  You are needed to deliver a number of Visualization for your dataset including:   * Pair Plots for the features. * Correlation Matrix heat map. * Box plots for the features.   **Note:**  Feel Free to add the Visualizations that you like! | | | | | |  |
|  | Step 6 | **Challenge Yourself**   * Visualize your data in Violin plots. * Describe what a Violin plot is. * Determine whether or not some of the features have outliers based on your Violin plots. | | | | | |  |
|  |  | | | |  |  | |  |
|  | Step 7 | Submit your Jupyter Notebook to Github | | | | | |  |
|  |  | Copy and paste your GitHub link in the field below, then click *Submit*. | | | | | |  |
|  |  |  |  | | | |  |  |
|  |  | https://www.github.com | | |  |
|  |  | | | |  |
|  |  | | | |  |  | |  |

|  |
| --- |
| Submitted Successfully |

Assessment Rubric(s) for Module 1

## The following is an example of an assignment rubric, feel free to edit the criteria and points as needed. Please provide a rubric for each graded assessment.

*This assignment is graded out of 20 points using the following rubric and is worth XX% of the final grade.*

*Learners may receive partial scores or a zero for unacceptable work.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Criteria | 0 | 1 | 2 | Score |
| Recognize Requirement | Little or no evidence that student clarified and understood the key concepts of the module by recognizing the module requirement. | Some evidence that student clarified and understood the key concepts of the module by recognizing the module requirement. | Multiple demonstrations of evidence that student clarified and understood the key concepts of the module by recognizing the module requirement. |  |
| Steps in the mini-projects | Little or no evidence that steps were completed | Some evidence that steps were completed | Multiple demonstrations of evidence that steps were completed |  |
| Using Tools | Little or no evidence that the tools are were used effectively | Some evidence that the tools are were used effectively | Multiple demonstrations of evidence that the tools are were used effectively |  |
| Filling the Required Code | Little or no evidence that the required code is completed | Some evidence that the required code is completed | Multiple demonstrations of evidence that the required code is completed |  |
| Total Score (10) | | | |  |