**Independent Learning Reports:**

For my Independent Learning, I chose **Master Python for Data Science**. This online course is 29 hours learning pathway offered by LinkedIn Learning. It include the following programs: Python Essential Training, Data Ingestion with Python, Python for Data Science I & II, Python for Data Visualization & Analysis, and Natural Language Processing. So, below are the brief summary of what I have learning through these training and a screenshot of each certification of completion.

1. **NLP with Python For Machine Learning**

Though the 4h 14 min course, I gained some basic knowledge I need to tackle complex problem using machine learning. It introduced me to a lot of Natural Language Processing(NLP) concepts such as advanced data cleaning, vectoring techniques, and building machine learning classifiers. My takeaways from NLP definition is that it is a field concerned with the ability of a computer to understand, analyze, manipulate, and potentially generate human language(English, French, and Spanish). NLP can be to solve various problems such as Auto-Complete search, spam filter, and Auto-correct editing just to cite few. It encompasses many topics such as Sentiment Analysis, Topic Modeling, Text Classification, Sentence segmentation or speech tagging. The NLP toolkit that I used for exercises is called NLTK. In addition, in this course I learned how to use regular expression(re) to identify whitespace between words(tokens), identify or create delimiters, remove punctuation, identify textual patterns, and clean HTML tags just to cite these. I learned useful methods for tokenizing including findall() and split(); and useful regexes for tokenizing words(‘\w’ & ‘\W’) and whitespace (‘\S’& ‘\s’). In addition, I learning the concepts of stemming(reducing word to its root), lemmatizing (grouping together words), and vectorizing(converting words in numeric form as shown in my code snippets below. Finally, my learning takeaways can be summarized with this machine learning pipeline that include: reading raw text, tokenizing, cleaning by removing stops words, punctuation, stemming, machine learning algorithm(fit/train), and etc., as shown below.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

Text

Description automatically generated

Finally, I have learned several skillsets that are actually helping me to professionally. The corresponding certification of completion is in the appendix.

1. **Python Data Ingestion**

Through this course, I learned how to extract, transform, and load data with python. Also, I gained useful skillsets such as working with python, calling APIs, web scraping, and processing even-base data. So very briefly, some of my takeaways include :

**Big Data format**: Prior to taking this course, I was not familiar with other big data formats such as Parquet, Avro, and ORC. Because of performance bottleneck from processing large text files, these news big data formats are more efficient. We use each on those file format, we just need to find its library and load it into Pandas. The screenshot below shows how I load parquet format into Pandas to create a data frame

Text

Description automatically generated with medium confidence

Furthermore, concerning web scraping, I was introduced to Beautiful Soup library as best approach to parse HTML data. I will be instantiating the use of Beautiful Soup below in my Data Science II pipeline. In addition, this course give me an opportunity to review how to fill missing values and find outliers. The certificate of completion is shown in the appendix.

1. **Python Essential Training I & II**

These two courses were most a review for me. I am proficient with python language. I use its various environment daily to perform complex tasks related to large scale software development such as: using ArcGIS Pro python Conda environment to write robust test suits for ArcGIS Pro and others. So, many concepts such data structures, conditionals and operators, files input/output, creating classes, creating and consuming modules, amply explained in these courses are not new to me. The certificates of completion are in the appendix.

1. **Python for Data Science Essential Training I**

This course took me steps by through a practical data science project: Some of the exercises include a web scraper that downloads and analyzes data from the web; techniques to clean, reformat, transform, and describe raw data; generate visualizations; remove outliers; and perform simple data analysis just to cite a few. I learned a great deal going through this training. Some of my takeaways include:

**Times Series**: I have learned built-in python datetime module that supplies classes for manipulating times and dates in various ways (simple and complex). For instance, the code snippet shows how I one can read data from a CSV file and process and display time series.

Graphical user interface, text, application, email

Description automatically generated

In addition, I learned how to use scatter\_matrix from pandas.plotting, matplotlib.pyplot, pylab, and seaborn to perform statistics and display data in various ways. Then, how to perfom parametric and non-parmetric correlation analysis. An example of non-parametric correlation analysis is chi-square hypothesis test of independence. Scipy.stats in addition to the library above to progress data and perform a chi-square test as shown in this code snippet:

Graphical user interface, text, application

Description automatically generated

**Beautiful Soup Object**: The important tool help with web scraping and tag manipulation. To consume this library, we do the following “ from bs4 import BeautifulSoup”, as shown in the code snippet below. Finally, this course also introduced me to NLP concepts. I had an opportunity to work on a sample machine learning pipeline such as the one I shared earlier.

**Beautiful Soup Code Snippet**:

Graphical user interface, text, application, email

Description automatically generated

1. **Python for Data Science Essential Training II**

The course is more hands-on experience with python. I really learned a lot from this and am still going through documentation to better understand some of the concepts briefly discussed. Through the course, I had an opportunity to go step by step through a practical data science project that involves: building machine learning models that can generate predictions and recommendations and automate routine tasks. I learned a couple regression and clustering models. Also, I was introduced to dimensional reduction methods and some popular machine learning methods such as ensemble, neural networks and Bayesian models. Below are some of my takeaways from the training:

**Logistic Regression**: I learned that for a successful logic regression model, data should be free of missing values; the predicted variable should be binary or ordinal, all predictors should be independent from each other, and there should be at least 50 observations per predictor variable to ensure reliable results. Below is a code snippet for logistic regression model

Graphical user interface, text, application, email

Description automatically generated

Furthermore, I was exposed to a lot of features engineering approach related to logistic regression, model evaluation and test prediction as shown below.

Graphical user interface, text, application, email

Description automatically generated

**K-Means Clustering**: It is said that “this method of predictions are based on the number of centroids present (k) and nearest mean values, given a Euclidean distance measurement between observations”. I learned that to correctly use K-means, you should scale your variables and look at a scatterplot or data table to estimate the appropriate number of centroids to use for the K parameter values. Below is a code snippet for K-means method:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Another unsupervised method for cluster analysis is DBSCAN for outlier detection. I won’t expand on that here.

**Dimension Reduction Methods**: I have learned both explanatory factors analysis and principal component analysis (PCA). **Factor Analysis** is for variables that are quite meaningful but that are inferred and not directly observable. Below is a code snippet show libraries to perform basic for factor analysis

Graphical user interface, text

Description automatically generated

**Principal Component Analysis**: It helps to reduce information redundancy and noise. Below are library that I used to perform analysis.

Graphical user interface, text, application, chat or text message

Description automatically generated

Finally, I was introduced to few popular machine learning methods including Association rules models with Apriori, Neural networks with a perceptron, Instance-based learning with KNN, Decision three models with CART, Bayesian models with Naïve Bayes, and Ensemble models with random forests just to cite few.

**Neural Network with a perception**: I learned that a perceptron is a neural network with just one layer. The reason it is called “linear binary classifier” is that it outputs a binary response variable. Data must be linearly separable for a perceptron to operate properly. They are four essential elements of a perceptron including (Input layer, Weights and bias, a weighted sum and activation function) as shown on the training material below.

Diagram

Description automatically generated

Below is a sample pipeline for Neural network:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated with medium confidence

All in all, like the Neural network, I have code snippets for other machine learning methods I cited above. Please, see below my certificates of completion and total number of hours. It was indeed a good independent learning experience. I was exposed to a lot of useful learning materials that I will continue to refer to.

**Appendix:**

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated