**Notes for Interview**

* **Pandas** is great because it uses multi-dimensional arrays and fast operations internally to provide higher level methods for manipulation and analysis.
* Do data analysts use **Python**?That makes Python a must-have tool not only for data analysis but for all data science. You can make the data more accessible and easier-to-use by means of creating various charts and graphics, as well as web-ready interactive plots. Yes, Python provides you with the capability to get a good sense of data.
* Embed **Tableau** scripts on to our website
  + Tableau is a BI tool that is designed to help the user to create visuals and graphics. One can do this without any technical help or prior knowledge of programming.
  + There are a lot of advantages of Tableau such as:
    - Speed
    - Ease of use
    - Beautiful and interactive dashboard
    - Direct connection
    - Easy publishing and sharing
    - Growing market share and popularity
* **Machine Learning** is the core subarea of artificial intelligence. It makes computers get into a self-learning mode without explicit programming. When fed new data, these computers learn, grow, change, and develop by themselves.
  + Machine learning can be the key to unlocking the value of corporate and customer data and enacting decisions that keep a company ahead of the competition.
  + **Regression analysis** consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x). Briefly, the goal of regression model is to build a mathematical equation that defines y as a function of the x variables.
    - Regression is a supervised machine learning technique which is used to predict continuous values. The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data. The three main metrics that are used for evaluating the trained regression model are variance, bias and error.
    - **Linear regression** is used to predict the continuous dependent variable using a given set of independent variables. In Linear regression, we predict the value of continuous variables.
    - **Logistic Regression** is used to predict the categorical dependent variable using a given set of independent variables. (Binary variables) In logistic Regression, we predict the values of categorical variables.
* **HTML** is the standard markup language for creating websites and CSS is the language that describes the style of an HTML document. We combined HTML and **CSS** to create a basic web page
* **Matplotlib** is an incredibly popular and useful visualization tool.
* **California Wildfires Dataset & Our Project**
  + <https://www.kaggle.com/ananthu017/california-wildfire-incidents-20132020/code>
  + List of WildFires in California between 2013 and 2020
  + About this file - The file contains the data of over 1600 wildfires that have occurred in California between 2013 and 2020.
  + Description: California is one of the places having the most deadliest and destructive wildfire seasons. The dataset contains the list of Wildfires that has occurred in California between 2013 and 2020. The dataset contains the location where wildfires have occurred including the County name, latitude and longitude values and also details on when the wildfire has started. 40 columns.
  + This data helps to generate insights on what locations in California are under fire threat, what time do Wildfires usually occur and how frequent and devastating they are!!
  + **WILDFIRE MAP:** Each circle is plotted by latitude and longitude coordinates from the data set. Each circle's radius is determined by the amount of acres burned in that area. The color intensity is plotted to show the length of time the fire burned in hours.
  + **FIRE STATS:** Converted data into visual analytics creating powerful visual information that communicates Year/County comparison for wildfire data. Thus, making user friendly, interactive graphic to use on our website.
  + **MACHINE LEARNING:** Regression is a supervised machine learning technique which is used to predict continuous values. We used a heat a
    - Used heat map to see that variables of “Started Year” and “Acres Burned(h)” that had the largest correlation coefficient at 0.47 = moderate positive linear relationship (measure of strength of the straight-line or linear relationship between two variables)
    - County fire prediction score: 0.05.

Looking to see if there was a relationship between if when a fire started (month) would predict where the fire would happen

Used logistic regression.

You have a better chance of flipping a coin that using this model to predict that.

* **Seaborn** provides an API on top of Matplotlib that offers choices for plot style and color defaults, defines simple high-level functions for common statistical plot types.
  + *Exploratory Data Analysis Process: Heat Map*

This visual displays the correlation matrices for our data. This allows us to decide which feature affects the target variable the most and should, in turn, be used in predicting this target variable (used for feature selection in machine learning).

* **WHY SHOULD WE HIRE YOU?** Knowledge, experience, & excellent communication ability to be an asset to your company. Variety of skills learned through this course, multi-diverse knowledge set of different programs that can be useful in a various amount of projects
* **Hardest part of our project and how did we handle/overcome this?** Data cleaning to find accurate results with so much data. (don’t see it until error; lots of n/a’s, delving into formatting time converting to all hours), working backward to figure out resolutions

Testing dependent variables for correlation –which variables to use to develop machine learning model

* **Why is predictive modeling important?** In order to get an in-depth insight inside data and make decisions that will drive the businesses, we need predictive modeling. Predictive modeling makes use of statistics to forecast the outcomes.
  + Benefits of predictive analytics
    - Gain a competitive advantage.
    - Find new product/service opportunities.
    - Optimize product and performance.
    - Gain a deeper understanding of customers.
    - Reduce cost and risk.
    - Address problems before they occur.
    - Meet consumer expectations.
    - Improved collaboration.
* **Why is data science important/relevant?** Data Science enables enterprises to measure, track, and record performance metrics for facilitating enterprise-wide enhanced decision making. Companies can analyze trends to make critical decisions to engage customers better, enhance company performance, and increase profitability.
* **What would we do if we had more time on this assignment?** Look at how weather impacts fire stats; would this be a better predictor/develop a stronger machine learning model

**What is the difference between unsupervised learning and supervised learning?** Unsupervised learning is a machine learning technique, where you do not need to supervise the model. Unsupervised machine learning helps you to finds all kind of unknown patterns in data.

Supervised learning allows you to collect data or produce a data output from the previous experience. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.

* **What are various steps involved in an analytics project?**

1. Understand the business problem
2. Explore the data and become familiar with it.
3. Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.
4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step till the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over the period of time.