Data Scientist

Core Question:

Is the data of sufficient quality for your use case.

**1. Programming Languages:**

Python:

Basics:

* *for* loop
* *in*
* *while* loop
* Data Structures
  + Base Types
  + Lists
  + Dictionaries
* Control Flow and Operators
* Functions and Classes

Libraries:

* Pandas [sort data]
* Numpy [array and mathematic functions]
* Scipy [viewing and transforming data]
* Matplotlib [visualization]

Dev Software:

* Jupyter Notebook
* Jupyter Lab

Publishing Software:

* Flask
* Django

**2. Be able to identify Data Science tasks and Algorithms used for each**

Tasks => Algorithms:

* Churn prediction => Binary Classification
  + (structured) (supervised)
* Sales forcasting => Regression
  + (structured) (supervised)
* Sentiment/Market Based Analysis => Association Rules
  + (structured) (supervised)
* Sign Language Recognition => Object Detection
  + (unstructured)
* Image Classification
  + (unstructured)
* Defect Analysis => Semantic Segmentation
* Human Pose Modelling => Posenet

**3. Understanding the Different Types of Data**

Types of Data:

* + Structure (csv, txt, xlsx, sql views) (pandas)
  + Unstructured (image, video, audio, text-based data)

**4. Analyzing and Visualizing Data**

Structured Data:

* + Viewing and Transforming (Pandas)
  + Visualization (MatPlotLib, Seaborn)

Images and Videos:

* + Viewing and Transforming (OpenCV)
  + Visualization (MatPlotLib)

Text and Natural Language:

* + Viewing and Transforming (NLTK, Textblob, Spacy)
  + Visualization (MatPlotLib [plt.imshow()])

Audio:

* + Viewing and Transforming (Scipy)
  + Visualization (Numpy)

**5. Preprocessing and Feature Engineering**

*Get data ready for modeling (this is preprocessing)*

Structured:

* + Fill in missing values
  + Normalizing and Scaling data
  + Setup independent and dependent variables
  + Split out into a training and testing datasets

Text files:

* + Learn how to remove punctuation
  + Limitized data (take words and return them to a base format)
  + Perform tokenization

Images:

*Tensorflow has preprocessing scripts to get data in the right format*

* + Check if images are valid
  + Labelling images using labelme and labelimg
  + Performing image augmentation using OpenCV

Audio files:

* + Convert .wav to spectrograms

**6. Models, Algorithms, and Evaluation**

*Use preprocessed data and apply models and algorithms to solve problems*

Supervised:

*Have a defined outcome, use labeled data, use input features to predict an output feature.*

* + Structured Regression
    - Random Forest Regressor
    - Gradient Boosting Regressor
  + Structured Classification
    - Random Forest Classifier
    - Gradient Boosting Classifier
  + Image Classification
    - Keras Sequential Neural Network
  + Object Detection
    - Tensorflow Single Shot Detector
  + Semantic Segmentation
    - Tensorflow Mark R-CNN
  + Pose Estimation
    - PoseNet
  + Reinforcement Learning
    - Stable Baselines
  + Sentiment Analysis
    - Text Blob Sentiment

Unsupervised:

* + Clustering
    - K-Means
  + Anomaly Detection
    - One Class SVM
  + Dimesionality Reduction
    - Principal Component Analysis

Know which algorithms to use and when

Evaluate their performance

**7. Deployment and Integration**

*Be able to make end outcomes (this is a full stack data science skill)*

Learn how to deploy using Cloud service providers:

* + IBM Watson Machine Learning
  + Microsoft Azure ML
  + Amazon Web Services SageMaker

Open-source Deployment:

* + ML app using Django
  + API using Flask
  + FastAPI

**8. Domain Expertise and Presentation**

Look at an industry and what their current issues are to be able to come up with a data science solution.

Go to Github and look at practical examples.

Have GREAT presentation skills.