: Getting Started Analyzing Data in Python :

1. What attribute or function returns the data types of each column?

Ans: dtypes

2. To enable a summary of all the columns, what must the parameter include be set to for the method describe?

Ans: df.describe(include="all")

: Importing Datasets :

1. What is the name of what we want to predict?

Ans: target

2. What does csv stand for?

Ans: Comma Separated Values

3. What library is primarily used for data analysis?

Ans: pandas

4. What library is primarily used for machine learning?

Ans: scikit-learn

5. How would you check the bottom 10 rows of dataframe df.

Ans: df.tail(10)

6. What task does the following command perform:

```
import pandas as pd
df = pd.read_csv("data.csv", header=None)
df.to_csv("A.csv")
```

Ans: Save the dataframe df to a csv file called "A.csv"

7. If you use the method describe() without changing any of the arguments you will get a statistical summary of all the columns of type object?

Ans: false

: Importing and Exporting Data in Python :

1. Some common encodings are?

Ans: csv, xlsx

2. What does the following method do to the dataframedf : df.head(12)?

Ans: show the first 12 rows of dataframe

: Python Packages for Data Science :

1. What is a Python library?

Ans: a collection of functions and methods that allow you to perform lots of actions without writing your code

2. What is the primary instrument or data-structure of Pandas?

Ans: dataframe

: Understanding the Data :

1. What does csv file stand for?

Ans: comma separated values

: Data Formatting in Python :

1. How would you multiply each element in the column df["a"] by 2 and assign it back to the column df["a"]?

Ans: df["a"]=2*df["a"]

2. How would you cast each element in the column "price" to an integer?

Ans: df["price"] = df["price"].astype("int")

: Data Normalization in Python :

1.Consider the column "length", select the correct code for z-score or standard score.

Ans: df["length"] = (df["length"]-df["length"].mean())/df["length"].std()

: Data Wrangling :

1. What task does the following line of code perform:

df['peak-rpm'].replace(np.nan, 5,inplace=True)

Ans: replace the not a number values with 5 in the column 'peak-rpm'

2. What task does the following lines of code perform:

```
avg=df['horsepower'].mean(axis=0)
df['horsepower'].replace(np.nan, avg)
```

Ans: nothing the parameter inplace is not set to true

3. How would you rename column name from "highway-mpg" to "highway-L/100km"

Ans: df.rename(columns={'"highway-mpg"':'highway-L/100km'}, inplace=True)

: Categorical Values :

1. Why do we use Categorical Variables

Ans: Most statistical models cannot take in objects or strings as inputs

2. What is the correct line of code to perform One-hot encoding on the column 'fuel'

Ans: pd.get_dummies(df['fuel'])

: Exploratory Data Analysis :

1. What task does the method value_counts perform

Ans: Returns counts of unique values

2. What does the horizontal axis in a scatter plot represent

Ans: independent variable

3. what is the largest possible element resulting in the following operation df.corr()

Ans: 1

4. if the Pearson Correlation between two variables is zero

Ans: the two variables are not correlated

5. Consider the dataframe df what method provides the summary statistics?

Ans: df.describe()

6. What is the minimum possible value of Pearson's Correlation :

Ans: -1

7. What is the Pearson correlation between variables X and Y, if X=Y:

Ans: 1

: Model Development :

1. What steps do the following lines of code perform:

Input=[('scale',StandardScaler()),('model',LinearRegression())] pipe=Pipeline(Input) pipe.fit(Z,y) ypipe=pipe.predict(Z)

Ans: Standardize the data, then perform a prediction using a linear regression model using the features Z and targets y

2. We create a polynomial feature as follows "PolynomialFeatures(degree=2)", what is the order of the polynomial

Ans: 0

3. What value of R^2 (coefficient of determination) indicates your model performs best ?

Ans: 1

4. The larger the mean square error, the better your model has performed

Ans: False

: Model Evaluation :

1. What is the correct use of the "train_test_split" function such that 90% of the data samples will be utilized for training, the parameter "random_state" is set to zero, and the input variables for the features and targets are x_data, y_data respectively.

Ans:

train_test_split(x_data, y_data, test_size=0.1, random_state=0)