- **Data Analysis with Python**
- **Week 1**
-
- **Importing Datasets**

- **The Problem**
- **Video 1 min**
- ❌

- (
- we'll be looking at the dataset on used car prices
- **Practice Quiz: Understanding the Data**
-
- **Python Packages for Data Science**
- **2 min video** ❌

- **(**
- In this video, we'll be talking about data analysis and the scenario in which we'll be playing the data analyst or data scientist.
- we'll be going into how to understand the data, how to import it into Python, and how to begin looking into some basic insights from the data

- **Understanding the Data**
- Video ❌
- 2 min

- In order to do data analysis in Python, we should first tell you a little bit about the main packages relevant to analysis in Python
- **Practice Quiz: Python Packages for Data Science**

- **Importing and Exporting Data in Python**
- **4 min video** ❌

- In this video, we'll look at how to read any data using python's pandas package.
- **Practice Quiz: Importing and Exporting Data in Python**

- **Getting Started Analyzing Data in Python**
- **Video** ❌

- In this video, we introduce some simple Pandas methods that all data scientists and analysts should know when using Python, Pandas and data
- **Practice Quiz: Getting Started Analyzing Data in Python**

- **Accessing Databases with Python**
- **Video** ❌

- Hello, in this video you will learn how to access databases using Python
- Databases are powerful tools for data scientists. After completing this module, you'll be able to explain the basic concepts related to using Python to connect to databases.
- **Lesson Summary**
- In this lesson, you have learned how to:
- **Define the Business Problem**: Look at the data and make some high-level decision on what kind of analysis should be done
- **Import and Export Data in Python**: How to import data from multiple data sources using the Pandas library and how to export files into different formats.
- **Analyze Data in Python**: How to do some introductory analysis in Python using functions like
  - **dataframe.head()** to view the first few lines of the dataset,
  - **dataframe.info()** to view the column names and data types.
- **Lab 1:Importing Datasets**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/Module%201/DA0101EN-Review-Introduction.ipynb
- **Graded Quiz: Importing Datasets**
- **Week 1 completed**

- **Week 2**
- **<u>Data Wrangling</u>**
- **Pre Processing Data in Python**
- Video 2 min

- (In this video, we'll be going through some data preprocessing techniques.
- **Dealing with Missing values in Python**
- 6 min Video ❌

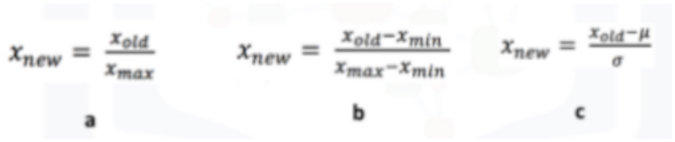
- (
- In this video, we will introduce the pervasive problem of missing values as well as strategies on what to do when you encounter missing values in your data
- we've gone through two ways in Python to deal with missing data. We learnt to drop problematic rows or columns containing missing values and
- then we learnt how to replace missing values with other values. But don't forget the other ways to deal with missing data. You can always check for a higher quality data set or source or in some cases you may want to leave the missing data as missing data.
- **Practice Quiz : Dealing with Missing Values in Python**
- **Data Formatting in Python**
- Video 3 min ❌

- (
- In this video, we'll look at the problem of data with different formats, units and conventions and the pandas methods that help us deal with these issues
- **Practice Quiz: Data Formatting in Python**
- Done
- **Data Normalisation in Python**
- Video 3 min ❌

- **Turning categorical variables into quantitative variables in Python**
- **2 min video**

- we'll be talking about data normalization.
- **Practice Quiz: Data Normalization in Python**

1. Which of the following is the correct formula for z -score or data standardization?



c

Correct
You simply subtract the mean and divide by the standard deviation.

•In this video, we'll discuss how to turn categorical variables into quantitative variables in Python

- **Binning in Python**
- **Video 1 min**

- **Practice Quiz: Turning categorical variables into quantitative variables in Python**
- Done
- **Lesson Summary**
- In this lesson, you have learned how to:
- Identify and Handle Missing Values: Drop rows with incomplete information and impute missing data using the mean values.
- Understand Data Formatting: Wrangle features in a dataset and make them meaningful for data analysis.

- **(**
- In this video, we'll talk about binning as a method of data pre-processing.

- Apply normalization to a data set: By understanding the relevance of using feature scaling on your data and how normalization and standardization have varying effects on your data analysis.
- **Lab 2 : Data Wrangling**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/Module%202/DA0101EN-2-Review-Data-Wrangling.ipynb
- **Graded Quiz : DataWrangling**
- Done
- **Exploratory Data Analysis**
- Video 1 min

- In this module we're going to cover the basics of exploratory data analysis using python.
- **Descriptive statistics**

- Video 4 min
- we'll be talking about Descriptive Statistics

- **Group by in Python**
- 3 min video

- (
- we'll cover the basics of grouping and how this can help to transform our dataset.

1. Which of the following tables representing number of drive wheels, body style and price is a Pivot Table? **0 / 1 point**

	price				
body-style	convertible	hardtop	hatchback	sedan	wagon
drive-wheels					
4wd	20239.229524	20239.229524	7603.000000	12647.333333	9095.750000
fwd	11595.000000	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.600000	24202.714286	14337.777778	21711.833333	16994.222222

a)

	drive-wheels	body-style	price
0	4wd	hatchback	7603.000000
1	4wd	sedan	12647.333333
2	4wd	wagon	9095.750000
3	fwd	convertible	11595.000000
4	fwd	hardtop	8249.000000
5	fwd	hatchback	8396.387755
6	fwd	sedan	9811.800000
7	fwd	wagon	9997.333333
8	rwd	convertible	23949.600000
9	rwd	hardtop	24202.714286
10	rwd	hatchback	14337.777778
11	rwd	sedan	21711.833333
12	rwd	wagon	16994.222222

b)

Incorrect
incorrect, a pivot table has one variable displayed along the columns and the other variable displayed along the rows.

- **Correlation**
- Video 2 min

- (
- In this video, we'll talk about the correlation between different variables.
- **Practice Quiz : Correlation**

- **Correlation statistics**
- Video 2 min

- (
- In this video, we'll introduce you to various correlations statistical methods

- **Practice Quiz : correlation statistics**

-
- **Association between two categorical variables : Chi-Square**
- **Video 4 min**

- In this video, we will learn how to find out if there is a relationship between two categorical variables

- **Lesson summary**
- **Describe Exploratory Data Analysis:** By summarizing the main characteristics of the data and extracting valuable insights.
- **Compute basic descriptive statistics**: Calculate the mean, median, and mode using python and use it as a basis in understanding the distribution of the data.
- **Create data groups**: How and why you put continuous data in groups and how to visualize them.

- **Define correlation as the linear association between two numerical variables**: Use Pearson correlation as a measure of the correlation between two continuous variables
- **Define the association between two categorical variables:** Understand how to find the association of two variables using the Chi-square test for association and how to interpret them.
- **Lab 3 : Exploratory Data Analysis**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/Module%203/DA0101EN-3-Review-Exploratory-Data-Analysis.ipynb
- **Graded Quiz: Exploratory Data Analysis**
- Week 3 done ✅
- **Week 4**
- **MODEL DEVELOPMENT**
- **Model Development**
- Video 1 min ❌

- In this video we will examine model development by trying to predict the price of a car using our dataset. In this module, you'll learn about simple and multiple linear regression, model evaluation using visualization, polynomial regression and pipelines, R-squared and MSE for in-sample evaluation, prediction and decision making, and how you can determine a fair value for a used car.
- In this course, you will learn about simple linear regression, multiple linear regression and polynomial regression.
- **Linear Regression and Multiple Linear Regression**
- 6 min Video

- In this video, we'll be talking about simple linear regression and multiple linear regression
- **Practice Quiz: Linear Regression and Multiple Linear Regression**

2. consider the following equation:

$$y = b_0 + b_1 x$$

the variable y is?

- ○ the intercept
- ○ the predictor or independent variable
- ◉ the target or dependent variable

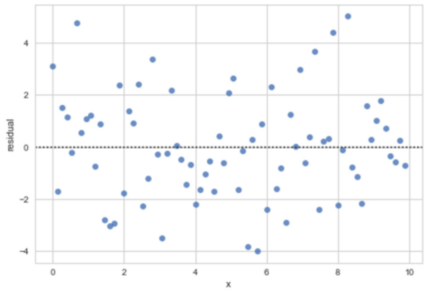
✓ **Correct**
correct

- **Model Evaluation using Visualization**
- 4 min video

- In this video, we'll look at Model Evaluation using Visualization

- **Practice Quiz: Model Evaluation using Visualization**

1. Consider the following **Residual Plot**, is our linear model correct :



- ○ yes
- ◉ incorrect

✗ **Incorrect**
incorrect, as we expect to see the results to have zero mean, the error is distributed evenly around the x-axis with similar variance and there is no curvature

- **Polynomial Regression and Pipelines**
- **4 min video**

- In this video, we will cover polynomial regression and pipelines

- **Practice Quiz: Polynomial Regression and Pipelines**
- Done

- **Measures for In-Sample Evaluation**
- 3 min video

- (
- Now that we've seen how we can evaluate a model by using visualization, we want to numerically evaluate our models. Let's look at some of the measures that we use for in-sample evaluation.
- **Practice Quiz: Measures for In-Sample Evaluation**
-

1. Consider the following lines of code; what value does the variable **out** contain?

```
lm = LinearRegression()
lm.score(X,y)
X = df[['highway-mpg']]
Y = df['price']
lm.fit(X, Y)
out=lm.score(X,y)
```

- ○ Mean Squared Error
- ◉ The Coefficient of Determination or R^2

Correct
correct

- **Prediction and Decision Making**
- 5 min video

- In this video our final topic will be on prediction and decision making. How can we determine if our model is correct?

- in the next section we will look at more accurate ways to evaluate the model
- **Lesson Summary**

Define the explanatory variable and the response variable: Define the response variable (y) as the focus of the experiment and the explanatory variable (x) as a variable used to explain the change of the response variable. Understand the differences between Simple Linear Regression because it concerns the study of only one explanatory variable and Multiple Linear Regression because it concerns the study of two or more explanatory variables.

Evaluate the model using Visualization: By visually representing the errors of a variable using scatterplots and interpreting the results of the model.

Identify alternative regression approaches: Use a Polynomial Regression when the Linear regression does not capture the curvilinear relationship between variables and how to pick the optimal order to use in a model.

Interpret the R-square and the Mean Square Error: Interpret R-square (x 100) as the percentage of the variation in the response variable *y* that is explained by the variation in explanatory variable(s) *x*. The Mean Squared Error tells you how close a regression line is to a set of points. It does this by taking the average distances from the actual points to the predicted points and squaring them.

- **Lab: Model Development**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/Module%204/DA0101EN-4-Review-Model-Development.ipynb
- **Graded Quiz: Model Development**

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

- -100
- -1
- 0
- **1**

Question 3: What statement is true about Polynomial linear regression

- Polynomial linear regression is not linear in any way
- **Although the predictor variables of Polynomial linear regression are not linear the relationship between the parameters or coefficients is linear.**
- Polynomial linear regression uses wavelets

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

- **False**
- True

Question 5: Assume all the libraries are imported, y is the target and X is the features or dependent variables, consider the following lines of code:

```
Input = [('scale', StandardScaler()), ('model', LinearRegression())]

pipe = Pipeline(Input)

pipe.fit(X,y)

ypipe = pipe.predict(X)
```

What have we just done in the above code?

- Polynomial transform, Standardize the data, then perform a prediction using a linear regression model
- **Standardize the data, then perform prediction using a linear regression model**
- Polynomial transform then Standardize the data

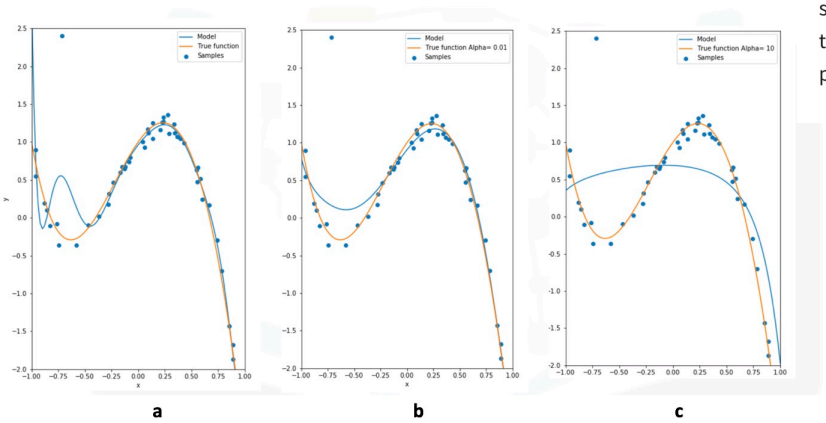
- **Week 5**
- **MODEL EVALUATION AND REFINEMENT**
- **Model Evaluation and Refinement**
- **7 min Video**

- Model evaluation tells us how our model performs in the real world. In the previous module, we talked about in-sample evaluation
- Let's illustrate the process. First, we split the data into three folds. We use two folds for training, the remaining fold for testing. The model will produce an output, and we will store it in an array. We will repeat the process using two folds for training, one for testing. The model produces an output again.
- Finally, we use the last two folds for training. Then we use the testing data. This final testing fold produces an output. These predictions are stored in an array
- **Practice Quiz: Model Evaluation**
- Done
- **Overfitting , Under-fitting and Model Selection**
- **Video 4 min**

- **(**
- If you recall, in the last module, we discussed polynomial regression. In this section, we will discuss how to pick the best polynomial order and problems that arise when selecting the wrong order polynomial.
- **Practice Quiz: Overfitting, Underfitting and Model Selection**
- Done
- **Ridge Regression Introduction**
  - Ridge regression is a regression that is employed in a Multiple regression model when Multicollinearity occurs. Multicollinearity is when there is a strong relationship among the independent variables. Ridge regression is very common with polynomial regression. The next video shows how Ridge regression is used to regularize and reduce the standard errors to avoid over-fitting a regression model
- **Ridge Regression**
- Video 4 min

- In this video, we'll discuss ridge regression. Ridge regression prevents overfitting. In this video, we will focus on polynomial regression for visualization, but overfitting is also a big problem when you have multiple independent variables, or features

- **Practice Quiz : Ridge Regression**
- the following models were all trained on the same data, select the model with the lowest value for alpha



a b c

- A is answer
- **Grid Search**
- Video 4 min

- Grid Search allows us to scan through multiple free parameters with few lines of code.
- Parameters like the alpha term discussed in the previous video are not part of the fitting or training process. These values are called hyperparameters
- **Lesson Summary**

Identify over-fitting and under-fitting in a predictive model: Overfitting occurs when a function is too closely fit to the training data points and captures the noise of the data. Underfitting refers to a model that can't model the training data or capture the trend of the data.

Apply Ridge Regression to linear regression models: Ridge regression is a regression that is employed in a Multiple regression model when Multicollinearity occurs.

Tune hyper-parameters of an estimator using Grid search: Grid search is a time-efficient tuning technique that exhaustively computes the optimum values of hyperparameters performed on specific parameter values of estimators.

- **Lab: Model Evaluation and Refinement**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/Module%205/DA0101EN-5-Review-Model-Evaluation-and-Refinement.ipynb
- **Graded Quiz : Model Refinement**

Question 1: In the following plot, the vertical access shows the mean square error andthe horizontal axis represents the order of the polynomial. The red line represents the training error the blue line is the test error. What is the best order of the polynomial given the possible choices in the horizontal axis?

- 2
- **8**
- 16

Question 2: What is the use of the "train_test_split" function such that 40% of the data samples will be utilized for testing, the parameter "random_state" is set to zero, and the input variables for the features and targets are_data, y_data respectively.

- train_test_split(x_data, y_data, test_size=0, random_state=0.4)
- **train_test_split(x_data, y_data, test_size=0.4, random_state=0)**
- train_test_split(x_data, y_data)

Question 3: What is the output of cross_val_score(lre, x_data, y_data, cv=2)?

- The predicted values of the test data using cross validation.
- **The average R^2 on the test data for each of the two folds**
- This function finds the free parameter alpha

Question 4: What is the code to create a ridge regression object "RR" with an alpha term equal 10

- RR=LinearRegression(alpha=10)
- **RR=Ridge(alpha=10)**
- RR=Ridge(alpha=1)

Question 5: What dictionary value would we use to perform a grid search for the following values of alpha: 1,10, 100. No other parameter values should be tested

- alpha=[1,10,100]
- **[{'alpha': [1,10,100]}]**
- [{'alpha': [0.001,0.1,1, 10, 100, 1000,10000,100000,100000],'normalize':[True,False]}]

- **Week 6**
- **Final Assignment**
- **House Sales in King County, USA**
- **Project Case Scenario**
- In this assignment, you are a Data Analyst working at a Real Estate Investment Trust. The Trust would like to start investing in Residential real estate. You are tasked with determining the market price of a house given a set of features. You will analyze and predict housing prices using attributes or features such as square footage, number of bedrooms, number of floors, and so on. A template notebook is provided in the lab; your job is to complete the ten questions. Some hints to the questions are given in the template notebook.
- You will use Watson Studio to perform the analysis, and share an image of your finished Jupyter notebook with a URL. If you are not familiar with how to set up Watson Studio, the next section will work you through creating an instance in Watson Studio otherwise, use the following information to get the final Notebook and get started:
- Notebook URL:
- Create a Notebook in Watson Studio and use the option 'From URL' to import the final notebook. Copy the link given below: Notebook link House Sales

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/

FinalModule_Coursera/House_Sales_in_King_Count_USA.ipynb

- **Project Overview**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/FinalModule_Coursera/instructions.md.html?origin=www.coursera.org
- **Share your Jupyter Notebook**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/FinalModule_Coursera/UploadnotebookDA.md.html?origin=www.coursera.org
- **Peer-graded Assignment: House Sales in King County, USA**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/FinalModule_Coursera/House_Sales_in_King_Count_USA.ipynb
- **Final Exam Answers**

Data Analysis with Python Final Exam Answers

Question 1: Question 1: What does the following command do:

df.dropna(subset=["price"], axis=0)

- **Drop the "not a number" from the column price**
- Drop the row price
- Rename the data frame price

Question 2: How would you provide many of the summery statistics for all the columns in the dataframe "df":

- **df.describe(include = "all")**
- df.head()
- type(df)
- df.shape

Question 7: How do you "one hot encode" the column 'fuel-type' in the dataframe df

- **pd.get_dummies(df["fuel-type"])**
- df.mean(["fuel-type"])
- df[df["fuel-type"])==1]=1

Question 8: What does the vertical axis in a scatter plot represent

- independent variable
- **dependent variable**

Question 9: What does the horizontal axis in a scatter plot represent

- **independent variable**
- dependent variable

Question 10: If we have 10 columns and 100 samples how large is the output of df.corr()

- 10 x 100
- **10 x 10**
- 100×100
- 100×100

Question 11: what is the largest possible element resulting in the following operation "df.corr()"

- 100
- 1000
- **1**

Question 12: if the Pearson Correlation of two variables is zero:

- the two variable have zero mean
- **the two variables are not correlated**

Question 13: if the p value of the Pearson Correlation is 1:

- the variables are correlated
- the variables are not correlated
- **none of the above**

Question 14: What does the following line of code do: lm = LinearRegression()

- fit a regression object lm
- **create a linear regression object**
- predict a value

Question 16: 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)

- Standardize the data, then perform a polynomial transform on the features Z
- find the correlation between Z and y
- **Standardize the data, then perform a prediction using a linear regression model using the features Z and targets y**

Question 17: What is the maximum value of R^2 that can be obtained

- 10
- **1**
- 0

Question 18: We create a polynomial feature as follows "PolynomialFeatures(degree=2)", what is the order of the polynomial

- 0
- 1
- **2**

Question 19: You have a linear model the average R^2 value on your training data is 0.5, you perform a 100th order polynomial transform on your data then use these values to train another model, your average R^2 is 0.99 which comment is correct

- 100-th order polynomial will work better on unseen data
- You should always use the simplest model
- **the results on your training data is not the best indicator of how your model performs, you should use your test data to get a beter idea**

Question 20:You train a ridge regression model, you get a R^2 of 1 on your training data and you get a R^2 of 0 on your validation data, what should you do:

- **Nothing your model performs flawlessly on your test data**
- your model is under fitting perform a polynomial transform
- your model is overfitting, increase the parameter alpha