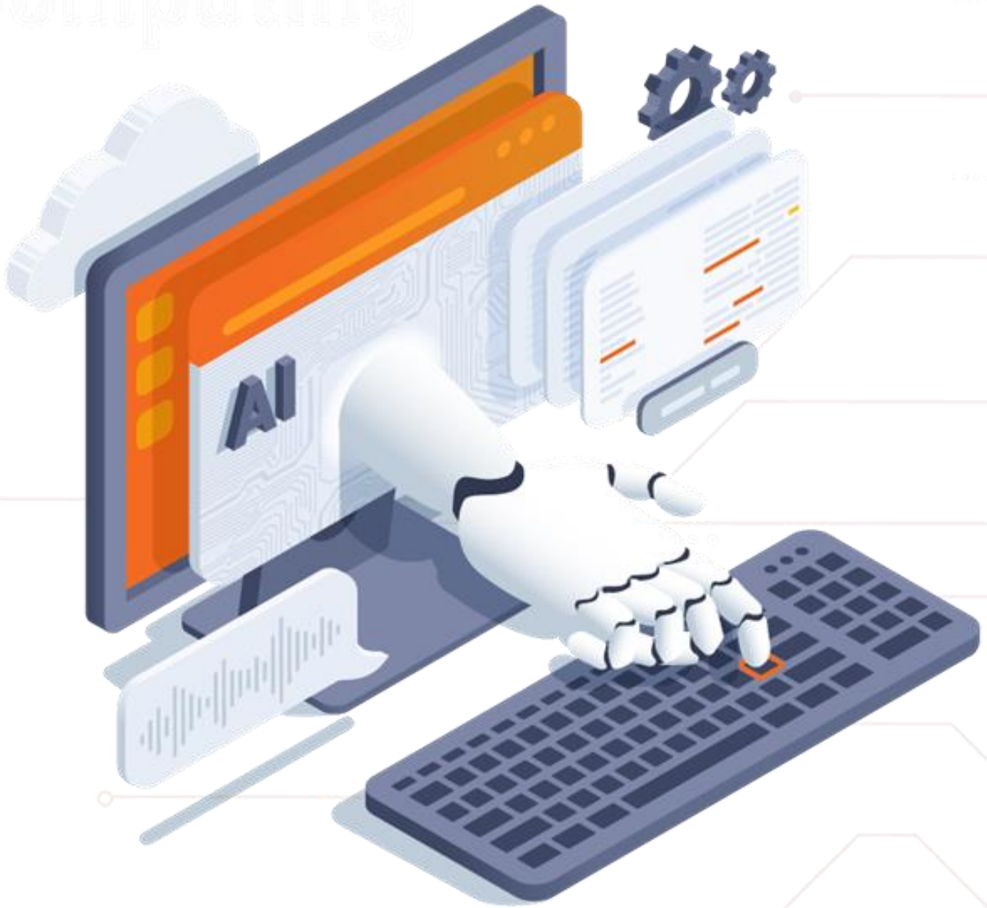
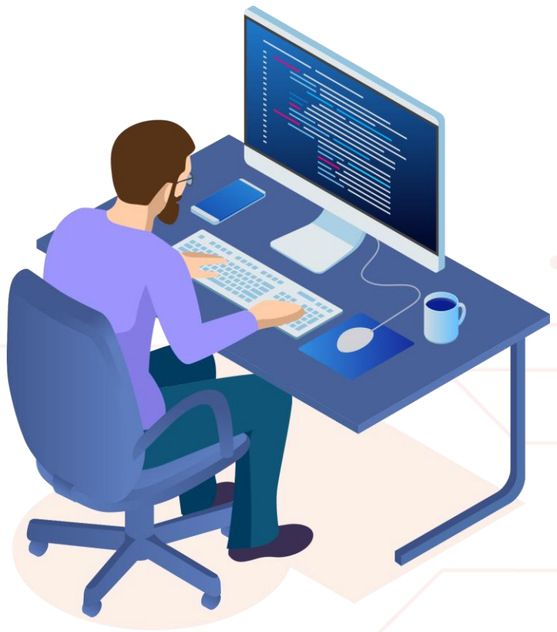


Caltech | Center for Technology & Management Education

Applied Data Science with Python

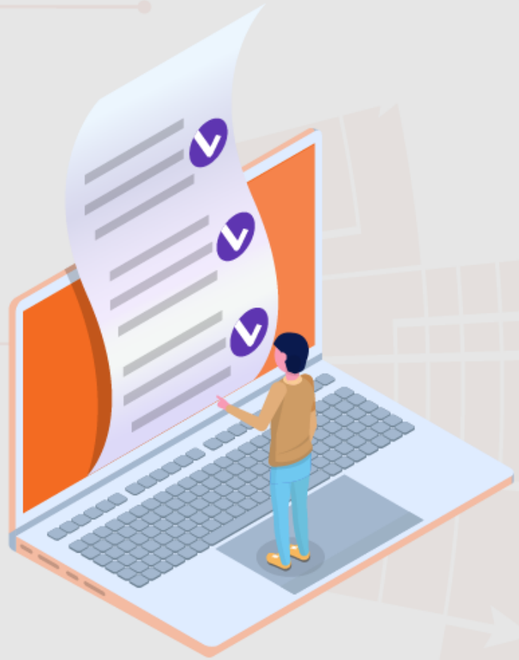
Course-end Project



Feature Engineering Real Estate Analytics

Objectives

- To assess the data and prepare a fresh dataset for training and prediction
- To create a box plot to identify the variables with outliers



Prerequisites

- Basics of Python
- Application of Python libraries in data science
- Skewness
- Perform analysis on a dataset
- Knowledge of DataFrame
- Train and perform prediction on a dataset

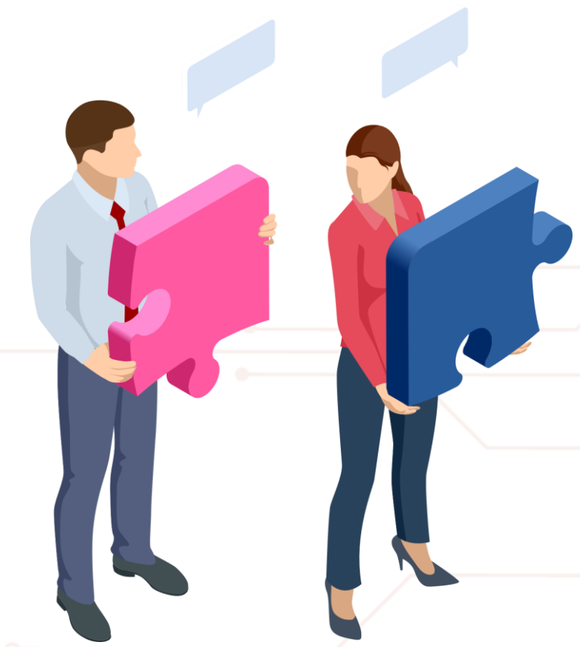
Industry Relevance

- **Basics of Python:** It is used for web development, data science and data analysis, machine learning, startups, and the finance industry.
- **Application of Python libraries in data science**: Python's large library ecosystem makes it possible to perform a wide range of functions, particularly in data science and machine learning.
- **Skewness**: Skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution.
- **Perform analysis on a dataset:** Data analysis refers to the process of manipulating raw data to uncover useful insights and draw conclusions.

Industry Relevance

- **Knowledge of DataFrame:** DataFrames are one of the most common data structures used in modern data analytics because they are a flexible and intuitive way of storing and working with data.
- **Train and perform prediction on a dataset**: The initial dataset used to train machine learning algorithms is known as training data. Models use these data to develop and improve their rules.

Problem Statement



While searching for the dream house, the buyer looks at various factors, not just the height of the basement ceiling or the proximity to an east-west railroad.

Using the dataset, find the factors that influence price negotiations while buying a house.

There are 79 explanatory variables describing every aspect of residential homes in Ames, Iowa.

Dataset Description

Variable	Description
SalePrice	The property's sale price is in dollars. This is the target variable that you're trying to predict
MSSubClass	The building class
MSZoning	The general zoning classification
LotFrontage	Linear feet of street connected to property
LotArea	Lot size in square feet
Street	Type of road access
Alley	Type of alley access
LotShape	General shape of property
LandContour	Flatness of the property
Utilities	Type of utilities available
LotConfig	Lot configuration

Dataset Description

Variable	Description
LandSlope	Slope of property
Neighborhood	Physical locations within Ames city limits
Condition1	Proximity to main road or railroad
Condition2	Proximity to main road or railroad(if a second is present)
BldgType	Type of dwelling
HouseStyle	Style of dwelling
OverallQual	Overall material and finish quality
OverallCond	Overall condition rating
YearBuilt	Original construction date
YearRemodAdd	Remodel date
RoofStyle	Type of roof

Dataset Description

Variable	Description
RoofMatl	Roof material
Exterior1st	Exterior covering on house
Exterior2nd	Exterior covering on house (more than one material)
MasVnrType	Masonry veneer type
MasVnrArea	Masonry veneer area and square feet
ExterQual	Exterior material quality
ExterCond	The present condition of the material on the exterior
Foundation	Type of foundation
BsmtQual	Height of the basement
BsmtCond	General condition of the basement
BsmtExposure	Walkout or grade level basement walls
BsmtFinType1	Quality of the basement finished area
BsmtFinSF1	Finished area of the basement in square feet

Dataset Description

Variable	Description
BsmtFinType2	Quality Of second finished area (if present)
BsmtFinSF2	Finished square feet
BsmtUnfSF	Unfinished square feet of basement Area
TotalBsmtSF	Total square feet of basement area
Heating	Type of heating
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system
1stFlrSF	First Floor square Feet
2ndFlrSF	Second floor square feet
LowQualFinSF	Low quality finished square feet (all floors)
GrLivArea	Above grade (ground) living area square Feet
BsmtFullBath	Basement full bathrooms
BsmtHalfBath	Basement half bathrooms

Dataset Description

Variable	Description
FullBath	Full bathroom above grade
HalfBath	Half bathrooms above grade
Bedroom	Number of bedrooms above basement level
Kitchen	Number of kitchens
KitchenQual	Kitchen quality
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
Functional	Home functionality rating
Fireplaces	Number of fireplaces
FireplaceQu	Fireplace quality
GarageType	Garage location
GarageYrBlt	Year garage was built
GarageFinish	Interior finish of the garage
GarageCars	Size of the garage in car capacity
GarageArea	Size of the garage in square feet
GarageQual	Garage quality

Dataset Description

Variable	Description
GarageCond	Garage condition
PavedDrive	Paved driveway
WoodDeckSF	Wood deck area in square feet
OpenPorchSF	Open porch area in square feet
EnclosedPorch	Enclosed porch area in square feet
3SsnPorch	Three season porch area in square feet
ScreenPorch	Screen porch area in square feet
PoolArea	Pool area in square feet
PoolQC	Pool quality
Fence	Fence quality
MiscFeature	Miscellaneous feature not covered in other categories
MiscVal	Value of miscellaneous feature
MoSold	Month sold
YrSold	Year sold
SaleType	Type of sale
SaleCondition	Condition of sale

Tasks to Perform

1) Download the “PEP1.csv” using the link given in the Feature Engineering project problem statement

2) For a detailed description of the dataset, you can download and refer to data_description.txt using the link given in the Feature Engineering project problem statement

Tasks to Perform

1) Import the necessary libraries

   1.1 Pandas is a Python library for data manipulation and analysis.

   1.2 NumPy is a package that contains a multidimensional array object and several derivative ones.

   1.3 Matplotlib is a Python visualization package for 2D array plots.

   1.4 Seaborn is built on top of Matplotlib. It's used for exploratory data analysis and data visualization.

2) Read the dataset

   2.1 Understand the dataset

   2.2 Print the name of the columns

   2.3 Print the shape of the dataframe

Tasks to Perform

2.4 Check for null values

2.5 Print the unique values

2.6 Select the numerical and categorical variables

3) Descriptive stats and EDA

3.1 EDA of numerical variables

3.2 Missing value treatment

3.3 Identify the skewness and distribution

3.4 Identify significant variables using a correlation matrix

3.5 Pair plot for distribution and density

Project Outcome

- The aim of the project is to help understand working with the dataset and performing analysis.
- This project will assess the data and prepares a fresh dataset for training and prediction
- To create a box plot to identify the variables with outliers

.

Submission Process

1. Complete the project in the Simplilearn lab
2. Complete each task listed in the problem statement
3. Take screenshots of the results for each question and the corresponding code
4. Save it as a document and submit it using the assessment tab
5. Tap the "Submit" button (this will present you with three choices)
6. Attach three files and then click "Submit“

Note: Be sure to include screenshots of the output

Thank You