Quiz-Course 7. IBM Data Analysis with Python (Coursera)

Week 1 Practice Quiz.

How many columns does the dataset have

- **26**
- 205

Week 1 Practice Quiz: Python Packages for Data Science

1.What is a Python library?

- **A collection of functions and methods that allows you to perform lots of actions without writing your code.**
- A file that contains data.

Week 1 Practice Quiz: Importing and Exporting Data in Python

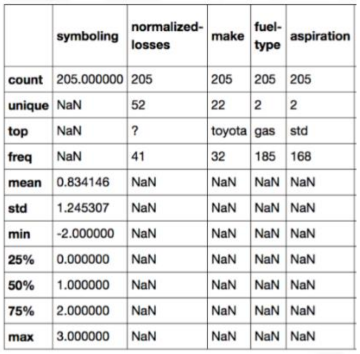
1. What does the following method do to the dataframe? `df : df.head(12)`

- **Show the first 12 rows of dataframe.**
- Shows the bottom 12 rows of dataframe.

Week 1- Practice Quiz: Getting Started Analyzing Data in Python

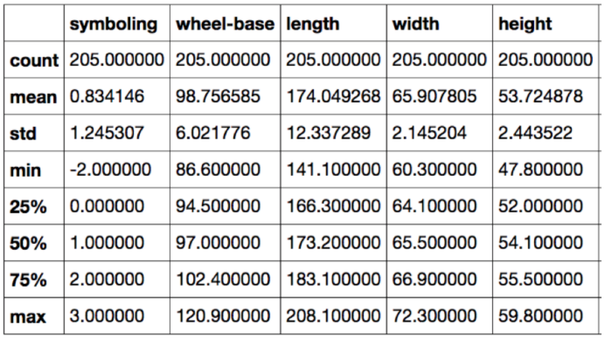
1.What is the correct output of? `df.describe(include = "all")`

a. It shows the statistic summary for all numerical and categorical variables



	symboling	normalized-losses	make	fuel-type	aspiration
count	205.000000	205	205	205	205
unique	NaN	52	22	2	2
top	NaN	?	toyota	gas	std
freq	NaN	41	32	185	168
mean	0.834146	NaN	NaN	NaN	NaN
std	1.245307	NaN	NaN	NaN	NaN
min	-2.000000	NaN	NaN	NaN	NaN
25%	0.000000	NaN	NaN	NaN	NaN
50%	1.000000	NaN	NaN	NaN	NaN
75%	2.000000	NaN	NaN	NaN	NaN
max	3.000000	NaN	NaN	NaN	NaN

b. It shows the statistic summary for all numerical variables



	symboling	wheel-base	length	width	height
count	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878
std	1.245307	6.021776	12.337289	2.145204	2.443522
min	-2.000000	86.600000	141.100000	60.300000	47.800000
25%	0.000000	94.500000	166.300000	64.100000	52.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000
75%	2.000000	102.400000	183.100000	66.900000	55.500000
max	3.000000	120.900000	208.100000	72.300000	59.800000

Week 1- Graded Quiz: Importing Datasets

What does csv stand for?

- **Comma Separated Values**
- Car Sold values
- none of the above

2.Select the libraries you will use for this course?

- **matplotlib**
- **pandas**
- **scikit-learn**
- TensorFlow

3.What task does the following command perform? `df.to_csv("A.csv")`

Save df to a CSV file named "A.csv" in the current working directory.

4. We have the list `headers_list: [headers_list = ['A', 'B', 'C']`

We also have the dataframe df that contains three columns. What is the correct syntax to replace the headers of the dataframe df with values in the list headers_list?

- **`df.columns =  headers_list`**
- `df.head()`
- `df.tail()`

5.How would you generate descriptive statistics for all the columns for the dataframe **df?**

`df.describe(include = "all")`

Week 2 – Practice Quiz- Dealing with Missing Values in Python

1.How would you access the column ”body-style" from the dataframe df?

- `df[ "body-style"]`
- `df==”bodystyle"`

2.What is the correct symbol for missing data?

- **nan**
- no-data

Week 2- Practice Quiz- Formatting

1. How would you cast each element in the column 'price' to an integer?

- `df['price'] = int (df['price'])`
- **`df['price'] = df['price'].astype ('int')`**

Week 2- Practice Quiz- Data Normalization in Python

1. What is the maximum value for feature scaling?

1

> → We scale the feature to a value between 0 and 1 so the maximum value should be 1.

Week 2- Practice Quiz- Turning categorical variables into quantitative variables in Python

1.Why do we convert values of Categorical Variables into numerical values

- To save memory
- **Most statistical models cannot take in objects or strings as inputs**

Week 2- Graded Quiz- Data Wrangling

1.What task do the following lines of code perform?

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

- **calculate the mean value for the 'bore' column and replace all the NaN values of that column by the mean value**
- nothing; because the parameter inplace is not set to true
- 'horsepower'

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

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

3.What data type is the following set of numbers? 666, 1.1,232,23.12

- int
- **float**
- object

4.The following code is an example of:

```
(df["length"]-df["length"].mean())/df["length"].std()
```

- simple feature scaling
- min-max scaling
- **z-score scaling (standardization)**

5.The following code is an example of:

```
df["length"]= df["length"]/df["length"].max()
```

- **simple feature scaling**
- min-max scaling
- z-score scaling (standardization)

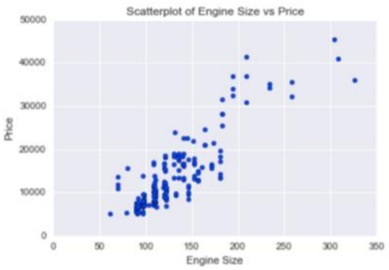
6.Consider the two columns 'horsepower', and 'horsepower-binned'; fromdf; how many categories are there in the 'horsepower-binned' column?

3

	horsepower	**Horsepower-binned**
0	111.0	Medium
1	110.0	Medium
2	101.0	Low
3	182	High
4	121	Medium

Week 3- Practice Quiz- Descriptive Statistics

1. Which one about this is correct about the scatter plot below?

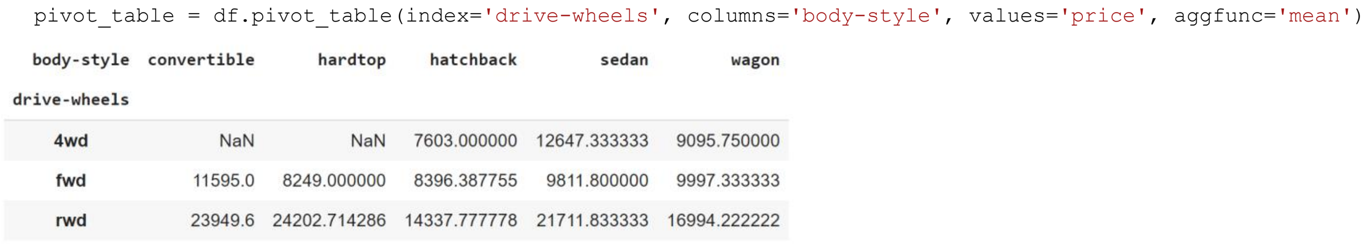


- **Positive linear relationship**
- Negative linear relationship

Week 3- Practice Quiz: GroupBy in Python

1. Which one is correct?

- **A pivot table has one variable displayed along the columns and the other variable displayed along the rows.**
- A pivot table contains statistical information for each column

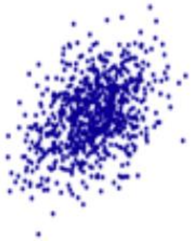
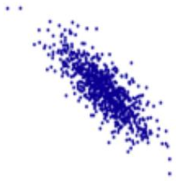


```
pivot_table = df.pivot_table(index='drive-wheels', columns='body-style', values='price', aggfunc='mean')
```

body-style	convertible	hardtop	hatchback	sedan	wagon
drive-wheels					
4wd	NaN	NaN	7603.000000	12647.333333	9095.750000
fwd	11595.0	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.6	24202.714286	14337.777778	21711.833333	16994.222222

Week 3- Practice Quiz: Correlation-Statistics

1. Select the plot with negative correlation

A (positive correlation)	**B (negative correlation)**
	

Week 3-Qraded Quiz- EDA

1.What happens if the method `describe` is applied to a dataframe with NaN values

- an error will occur
- all the statistics calculated using NaN values will also be NaN
- **NaN values will be excluded**

2.How would you use the `groupby` function to find the average "price" of each car based on "body-style" ?

- **`df[['price','body-style']].groupby(['body-style'],as_index= False).mean()`**
- `df.groupby(['price" ],as_index= False).mean()`
- `mean(df.groupby(['price','body-style'],as_index= False))`

2.What is the largest possible element resulting in the operation **`df.corr()`** ?

- 100
- 1000
- **1**

3.If the p-value of the Pearson Correlation is 1, then ...

- The variables are correlated
- The variables are not correlated
- **None of the above**

4.Consider the dataframe df;what method displays the first five rows of a dataframe?

- `df.describe()`
- **`df.head()`**
- `df.tail()`

5.What is the Pearson Correlation between variables X and Y, if X=Y?

- **1**
- -1
- 0

6.What is the Pearson Correlation between variables X and Y, if X=-Y?

- **-1**
- 1
- 0

7. If we have 10 columns and 100 samples, how large is the output of **`df.corr()`**?

- 10 X 100
- **10 X 10**
- 100 X 100

Week 4- Practice Quiz: Linear Regression and Multiple Linear Regression

1.Consider the following lines of code, what variable contains the predicted values :

```
from sklearn.linear_model import LinearRegression
lm=LinearRegression()
X = df[['highway-mpg']]
Y = df['price']
lm.fit(X, Y)
Yhat=lm.predict(X)
```

- Y
- X
- **Yhat**

2.Consider the following equation: what is the parameter b_0 (b subscript 0)

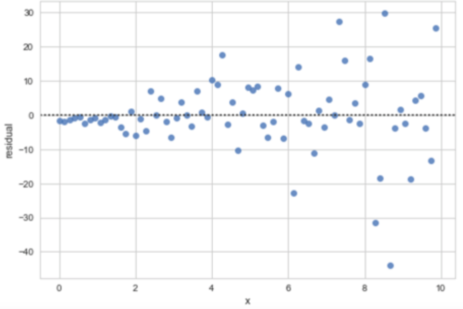
$$y = b_0 + b_1 x$$

- the predictor or independent variable
- the target or dependent variable
- **the intercept**
- the slope

→ b_0 is intercept; b_1 is slope

Week 4- Practice Quiz: Model Evaluation using Visualization

1.Consider the following **Residual Plot**. Is our linear model correct?



- Yes
- **No**

→ The variance of the residuals increases with x -> we expect residuals to be distributed evenly around the x-axis with similar variance

Week 4-Practice Quiz: Polynomial Regression and Pipelines

1.What is the order of the following Polynomial

$$\hat{Y} = b_0 + b_1\, x_1 + b_2(\, x_1)^2 + b_3(\, x_1)^3$$

- 1
- 2
- **3**

2.What functions are used to generate Polynomial Regression with more than one dimension

-
  ```
  f=np.polyfit(x,y,3)
  p=np.poly1d(f)
  ```

-
  ```
  pr=PolynomialFeatures(degree=2)
  pr.fit_transform([1,2], include_bias=False)
  ```

Week 4- 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)
```

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

2. Of the following answer values which one is the minimum value for R^2

- **0**
- 1
- 10

Week 4- Graded Quiz- Model Development

1.If the predicted function is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

The method is:

- Polynomial Regression
- **Multiple Linear Regression**

2.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**

1. **Standardize the data:** The first line of code defines a list of tuples called **Input**. The first tuple contains the string 'scale' and an instance of **StandardScaler()**, which is used for standardizing the data. The second tuple contains the string 'model' and an instance of **LinearRegression()**, which represents a linear regression model.
2. **Create a pipeline:** The second line of code creates a machine learning pipeline called **pipe**. This pipeline specifies that the data should first be standardized (scaled) using the **StandardScaler()** and then a linear regression model (**LinearRegression()**) should be applied.
3. **Fit the pipeline:** The third line of code fits (trains) the pipeline **pipe** on the feature matrix **Z** and the target variable **y**. This means that it will apply the scaler to the features and then train a linear regression model on the standardized features and target.
4. **Make predictions:** The fourth line of code uses the trained pipeline **pipe** to make predictions on the feature matrix **Z**. This means it will apply the same scaling transformation to the features in **Z** and then use the trained linear regression model to predict the target variable.

3.If X is a dataframe with 100 rows and 5 columns, and y is the target with 100 samples, and assuming all the relevant libraries and data have been imported, and the following line of code has been executed:

```
LR = LinearRegression()
LR.fit(X, y)
yhat = LR.predict(X)
```

How many samples does yhat contain?

- **100**
- 5
- 500

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

- -1
- **1**
- 0

5.Consider the following equation:

$$y = b_0 + b_1\, x$$

The variable y is what?

- The predictor or independent variable
- **The target or dependent variable**
- The intercept

6.What does the following line of code do? `lm = LinearRegression()`

- Fit a regression object lm
- Create a linear regression object
- Predict a value

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

- 0
- 1
- **2**

→ The Polynomial Features with degree=2 creates polynomial features up to the second degree. In other words, it generates not only the original features (degree 1) but also all possible combinations of the original features up to the second degree.

8.Which 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 linear Wavelets

→ In Polynomial linear regression, the relationship between the predictor variables and the target variable is not linear, but the model itself is still linear in terms of its parameters or coefficients. This is because the coefficients associated with the polynomial features are linear, which allows for the estimation of the model using linear regression techniques.

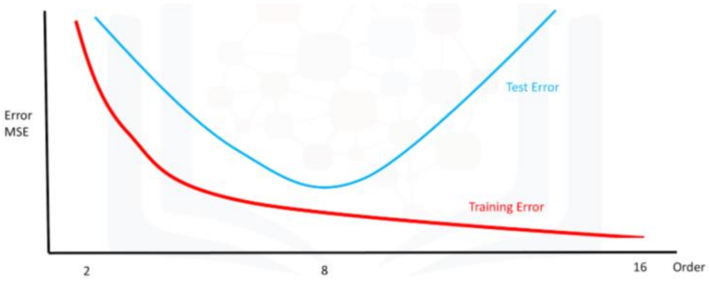
Week 5- Practice Quiz- Model Evaluation

1.What function randomly splits a dataset into training and testing subsets

- **train_test_split**
- cross_val_score
- cross_val_predict

Week 5- Practice Quiz: Overfitting, Underfitting and Model Selection

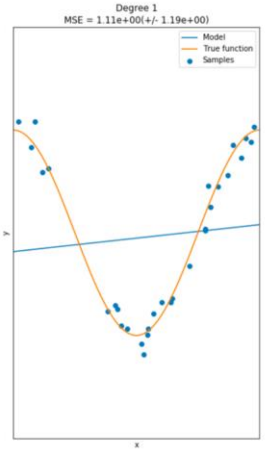
1.In the following plot, the vertical axis shows the mean square error and the horizontal axis represents the order of the polynomial. The red line represents the training error the blue line is the test error. Should you select the 16 order polynomial.



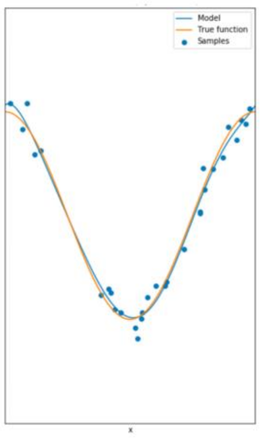
- **no**
- yes

→ We use the test error to determine the model error. For this order of the polynomial, the training error is smaller but the test error is larger.

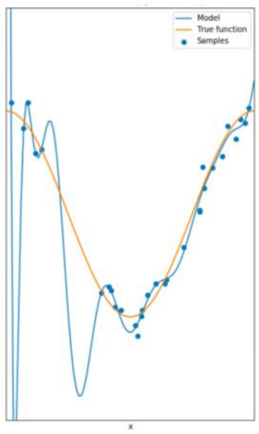
2.What model should you select?



a



b

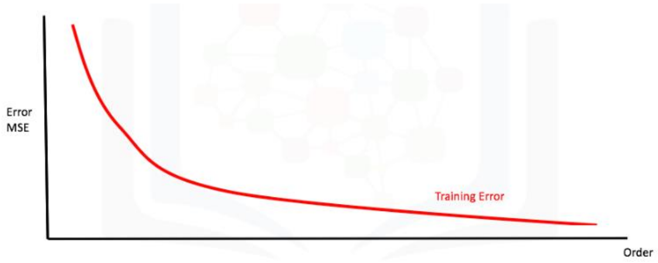


c

- a
- **b**
- c

→ a is not flexible enough, c is overfitting

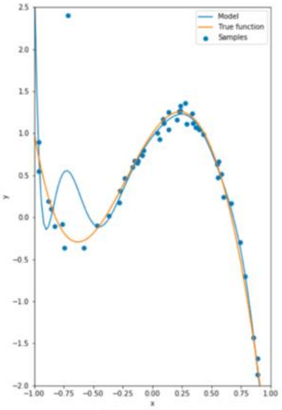
3.True or False, the following plot shows that as the order of the polynomial increases, the mean square error of our model decreases on the test data:



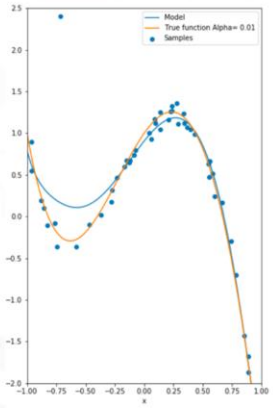
False - this plot shows the training error

Week 5- Practice Quiz- Ridge Regression

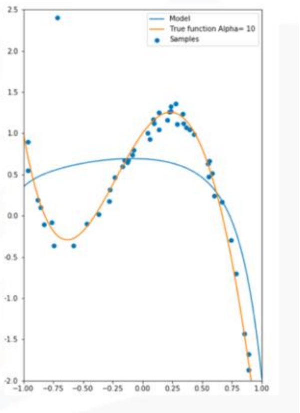
1. The following models were all trained on the same data, select the model with the highest value for alpha



a



b



c

- a
- b
- **c**

→ **c:** The model that exhibits the "**most" underfitting** is usually the model with the **highest parameter value for alpha**

→ a: The most overfitting-> lowest parameter value for alpha

Week 5- Graded Quiz- Model Refinement

1.What is the output of the following code?

```
cross_val_predict (lr2e, x_data, y_data, cv=3)
```

- **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

2.What dictionary value would we use to perform a grid search to determine if normalization should be used and for testing the following values of alpha? 1,10, 100

- `{'alpha': [1,10,100]}]`
- `alpha=[1,10,100]`
  `normalize=[True,False]`
- **`[{'alpha':[1,10,100],'normalize':[True,False]} ]`**

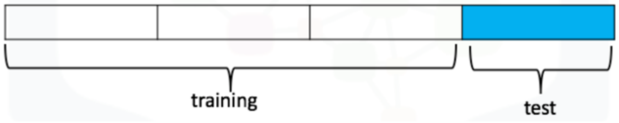
3.You have a linear model; the average R^2 value on your training data is 0.5, you perform a 100th order polynomial transform on your data then use these values to train another model. After this step, your average R^2 is 0.99; which of the following comments is correct?

- 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 better idea**

→ While increasing the polynomial order may significantly improve the model's performance on the training data (which is indicated by the increase in R^2 from 0.5 to 0.99), it doesn't necessarily mean that the model will perform better on unseen data (test data).

→ Overly complex models, such as high-order polynomial regression, can lead to overfitting, where the model fits the training data extremely well but fails to generalize to new, unseen data. Therefore, it's crucial to evaluate the model's performance on a separate test dataset to ensure that it's not overfitting and that it performs well in practice.

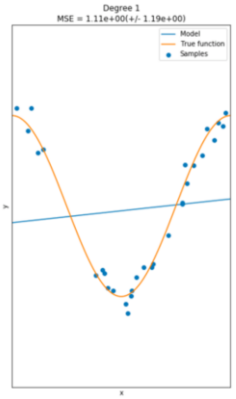
4.Consider the following diagram of 4 fold cross-validation. From the diagram how many folds are used for training?



- **3**
- 1
- 4

→ Each of the four sections represents one fold, and you would use 3 folds (or sections) for training and 1 fold for testing in each of the four iterations.

5.The following is an example of what?



- Overfitting
- Perfect fit
- **Underfitting**

6. 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?

- Your model is underfitting, so perform a polynomial transform
- Nothing. Your model performs flawlessly on your validation data
- **Your model is overfitting, so increase the parameter alpha**

Week 6- Final Exam

1.What does csvfile stand for?

- **comma separated values**
- car seller values

2.Scikit-learn is used for?

- **Statistical modelling including regression and classification.**
- Fast array processing.

3.What tells us the way the data is encoded?

- File path
- **Encoding scheme**
- Data path
  - → File path: tells you where the data is stored

4.What does the head() method return?

- It returns the data types of each column
- It returns the last five rows
- **It returns the first five rows**

5. What Python library is used for fast array processing?

- Matplotlib
- Scikit-learn
- **Numpy**

6.What library is primarily used for data analysis?

- **pandas**
- scikit-learn
- matplotlib

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

- `df.tail()`
- **`df.tail(10)`**
- `df.head()`

8.What is the function used to remove rows and columns with Null or NaN values?

- `replacena()`
- `removena()`
- **`dropna()`**

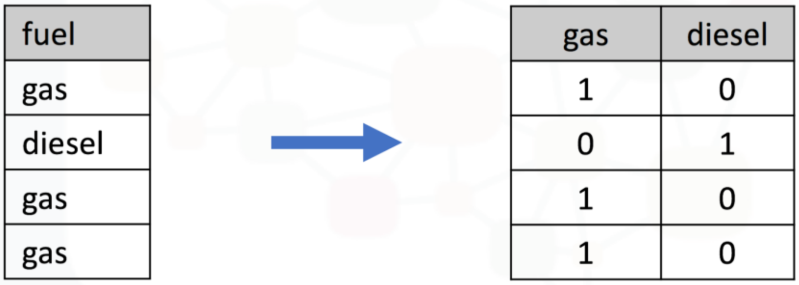
9. How would you multiply each element in the column df["c"] by 5 and assign it back to the column df["c"]

- **df["c"]=5*df["c"]**
- df["a"]=df["c"]*5
- 5*df["b"]

10.Consider the column "length"; select the correct code for z-score or standard score.

- df["length"] = (df["length"]-df["length"].min())/ (df["length"].max()-df["length"].min())
- **df["length"] = (df["length"]-df["length"].mean())/df["length"].std()**

11.Consider the following image: what is the name of the operation that transformed the column fuel into quantitative variables?



fuel
gas
diesel
gas
gas

gas	diesel
1	0
0	1
1	0
1	0

- Data standardization
- **One-hot encoding**

12.What function will change the name of a column in a dataframe?

- **rename()**
- exchange()
- replace()