EXPLORATORY DATA ANALYSIS (EDA)

Preliminary step in data analysis to:

- Summarize main characteristics of the data
- Gain better understanding of the data set
- Uncover relationships between variables
- Extract important variables

DESCRIPTIVE STATISTICS

- Describe basic features of data
- Giving short summaries about the sample and measure of the data

Summarize statistics using pandas describe() method: df.describe()

	Unnamed: 0	symboling	normalized-losses	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke
count	201.000000	201.000000	164.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.00000(
mean	100.000000	0.840796	122.000000	98.797015	174.200995	65.889055	53.766567	2555.666667	126.875622	3.319154	3.256766
std	58.167861	1.254802	35.442168	6.066366	12.322175	2.101471	2.447822	517.296727	41.546834	0.280130	0.316049
min	0.000000	-2.000000	65.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	2.540000	2.070000
25%	50.000000	0.000000	NaN	94.500000	166.800000	64.100000	52.000000	2169.000000	98.000000	3.150000	3.110000
50%	100.000000	1.000000	NaN	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	3.310000	3.290000
75%	150.000000	2.000000	NaN	102.400000	183.500000	66.600000	55.500000	2926.000000	141.000000	3.580000	3.410000
max	200.000000	3.000000	256.000000	120.900000	208.100000	72.000000	59.800000	4066.000000	326.000000	3.940000	4.170000

The default setting of "describe" skips variables of type object. We can apply the method "describe" on the variables of type 'object' as follows: df.describe(include=['object'])

Summarize the categorical data is by using the value_counts() method:

For example:

	value_counts
drive-wheels	
fwd	118
rwd	75
4wd	8

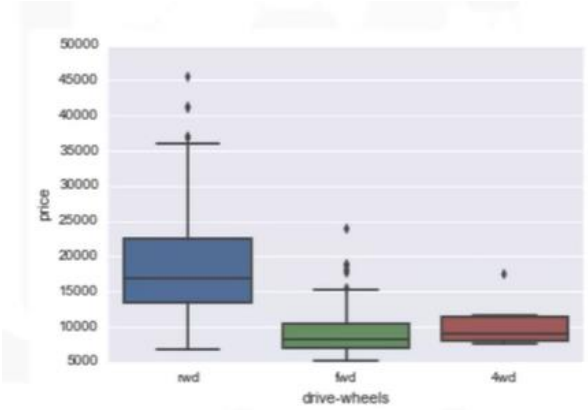
Drive_wheels_counts = df[“drive-wheels”].value_counts().to_frame() (since the data after extracted is series not data frame)

Drive_wheels_counts.rename(columns= {‘drive-wheels’:’value_counts’}, inplace= True)

Drive_wheels_counts.index.name = ‘drive-wheels’

Boxplot: show upper/lower extreme, upper/lower quartile, median, whisker, outlier

For example: sns.boxplot(x= “drive-wheels”, y= “price”, data= df)



Scatterplot:

Each observation represented as a point

Scatter plots show the relationship between 2 variables:

Predictor/independent variables on x-axis

Target/dependent variables on y-axis

For example:

y = df[“price”]

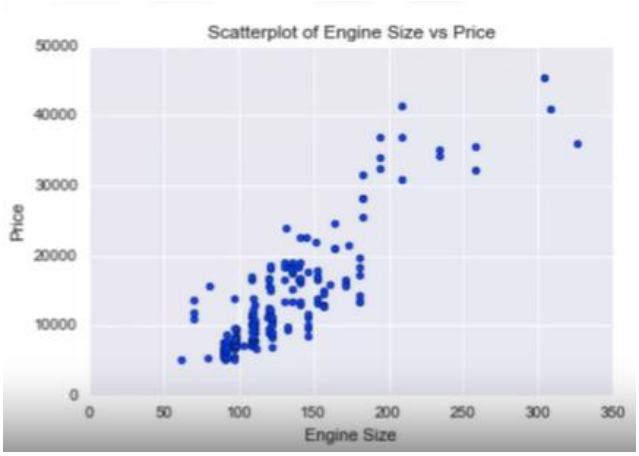
x= df[“engine-size”]

plt.scatter(x,y)

plt.title(“Scatterplot of engine size vs price”)

plt.xlabel(“Engine size”)

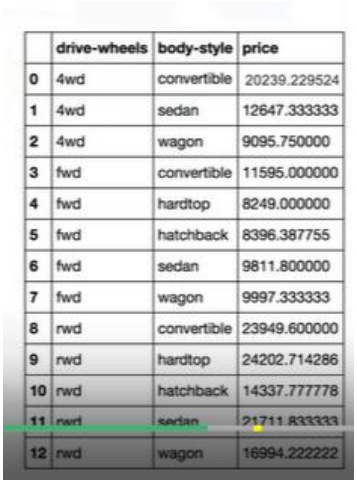
plt.ylabel(“Price”)



GROUPBY IN PYTHON

For example, let's group by the variable "drive-wheels". We see that there are 3 different categories of drive wheels. We use:

df['drive-wheels'].unique()



	drive-wheels	body-style	price
0	4wd	convertible	20239.229524
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

Use pandas dataframe.groupby() method:

Can be applied on categorical variables

Group data into categories

Single of multiple variables

For example:

df_test = df[[‘drive-wheels’, ‘body-style’, ‘price’]]

df_grp = df_test.groupby([‘drive-wheels’, ‘body-style’], as_index= False).mean()

Pandas method – Pivot()

One variable displayed along the columns and the other variable displayed along the rows

df.pivot = df_grp.pivot(index= ‘drive-wheels’, columns= ‘body-style’)

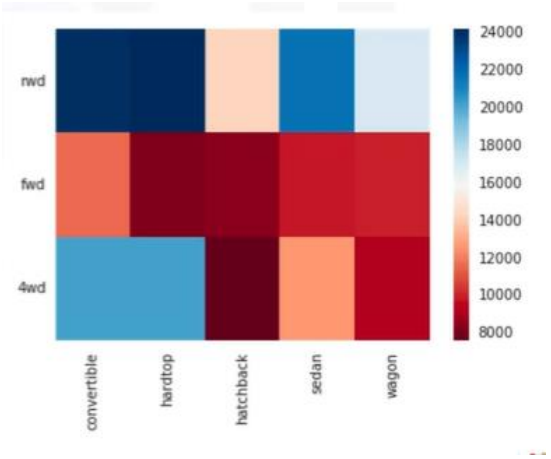
	price				
body-style	**convertible**	**hardtop**	**hatchback**	**sedan**	**wagon**
drive-wheels					
4wd	20239.229524	20239.229524	7603.000000	12647.333333	9095.750000
fwd	11595.000000	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.600000	24202.714286	14337.777778	21711.833333	16994.222222

Heatmap: Plot target variable over multiple variables

plt.pcolor(df_pivot, cmap= ‘RdBu’)

plt.colorbar()

plt.show()



CORRELATION: measure to what extent different variables is interdependent

For example:

Lung cancer → smoking

Rain → umbrella

Correlation does not imply causation

CASUSATION: the relationship between cause and effect between 2 variables

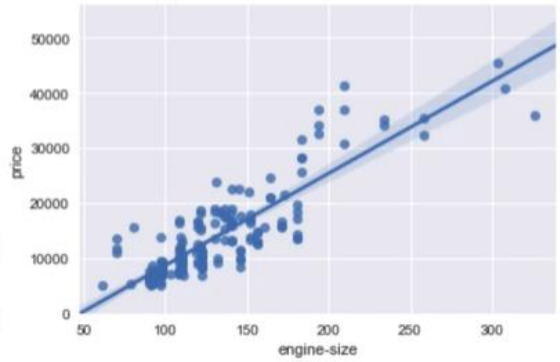
Correlation – positive linear relationship

Correlation between 2 features (engine-sized and price)

For example:

sns.regplot(x= “engine-size”, y= “price”, data= df)

plt.ylim(0,)

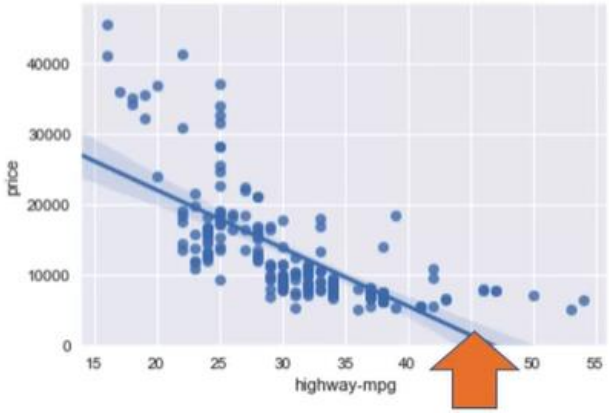


Correlation – negative linear relationship

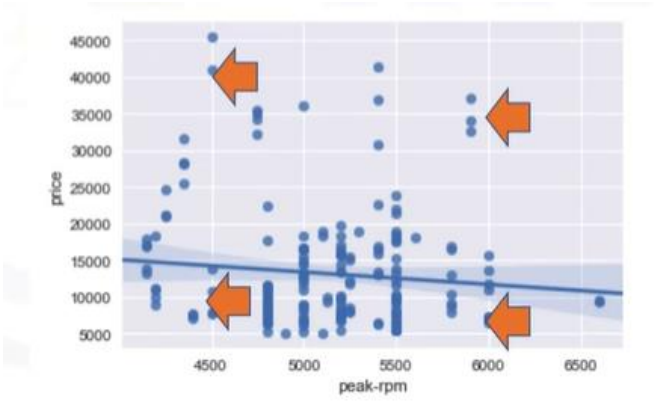
Correlation between 2 features (highway-mpg and price)

sns.regplot(x= “highway-mpg”, y= “price”, data= df)

plt.ylim(0,)



Correlation – Weak relationship



CORRELATION – STATISTICS

Pearson correlation: measure the strength of the correlation between 2 features

- Correlation coefficient
- P-value

Correlation coefficient:

Close to +1: large positive relationship

Close to -1: large negative relationship

Close to 0: no relationship

P-value:

P-value < 0.001: strong certainty in the result

P-value < 0.05: moderate certainty in the result

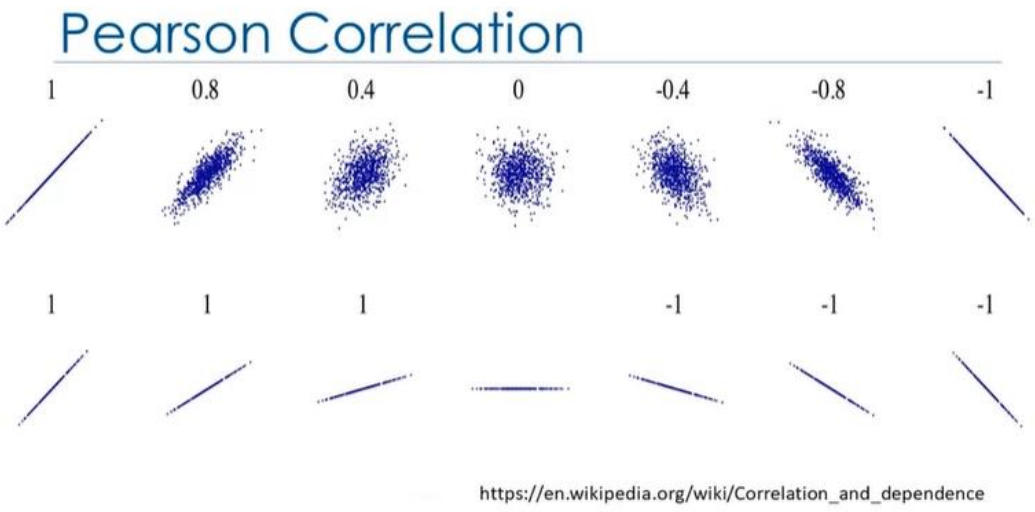
P_value < 0.1: weak certainty in the result

P-value > 0.1: no certainty in the result

Strong correlation:

Correlation coefficient close to 1 or -1

p-value less than 0.001



For example:

pearson_coef, p_value = stats.pearsonr(df[‘horsepower’], df[‘price’])

pearson correlation: 0.81

P-value: 9.35e-48

ASSOCIATE BETWEEN 2 CATEGORICAL VARIABLES: CHI-SQUARE

- Categorical variables
- We use the Chi-square test for association (denoted as $\chi 2$)
- The test is intended to test how likely it is that an observed distribution is due to chance
- The Chi-square tests a null hypothesis that the variables are independent

- The Chi-square does not tell you the type of relationship that exists between both variables; but only that a relationship exists

Categorical variables

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

Observed value

Expected value

aspiration / Fuel-type	Standard	Turbo	
Diesel	16.39	3.61	20
Gas	151.61	33.39	185
	168	37	

Row total * Column total / Grand total

IBM Developer SKILLS NETWORK

Categorical variables

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Degree of freedom = (row-1)*(column-1)

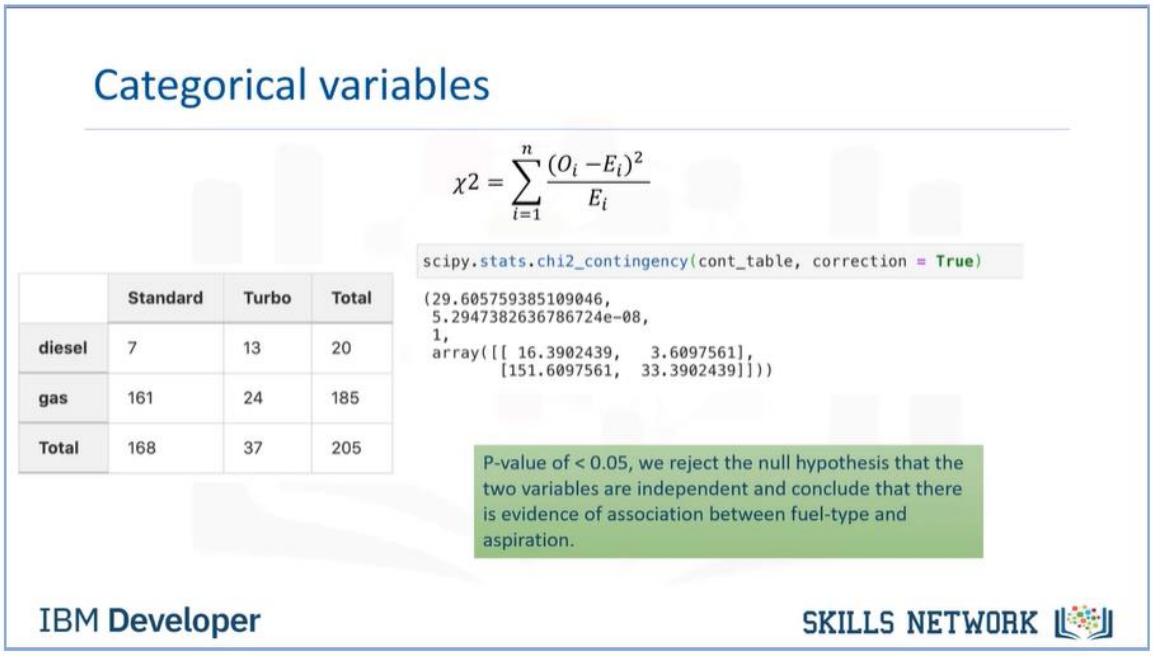
$$\chi 2 = 29.6$$

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2							
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92

P-value < 0.05, we reject the null hypothesis that the two variables are independent and conclude that there is evidence of association between fuel-type and aspiration.

IBM Developer SKILLS NETWORK



ANOVA: Analysis of Variance

The Analysis of Variance (ANOVA) is a statistical method used to test whether there are significant differences between the means of two or more groups. ANOVA returns two parameters:

F-test score: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means.

P-value: P-value tells how statistically significant our calculated score value is

If our price variable is strongly correlated with the variable we are analyzing, we expect ANOVA to return a sizeable F-test score and a small p-value.

Since ANOVA analyzes the difference between different groups of the same variable, the “groupby” function will come in handy. Because the ANOVA algorithm averages the data automatically, we do not need to take the average before hand. To see if different types of 'drive-wheels' impact ‘price’, we group the data.

For example:

f_val, p_val = stats.f_oneway(grouped_test2.get_group('fwd')['price'], grouped_test2.get_group('rwd')['price'], grouped_test2.get_group('4wd')['price'])

This is a great result with a large F-test score showing a strong correlation and a P-value of almost 0 implying almost certain statistical significance. But does this mean all three tested groups are all this highly correlated?

Let's examine them separately.

Fwd and rwd:

f_val, p_val = stats.f_oneway(grouped_test2.get_group('fwd')['price'], grouped_test2.get_group('rwd')['price'])