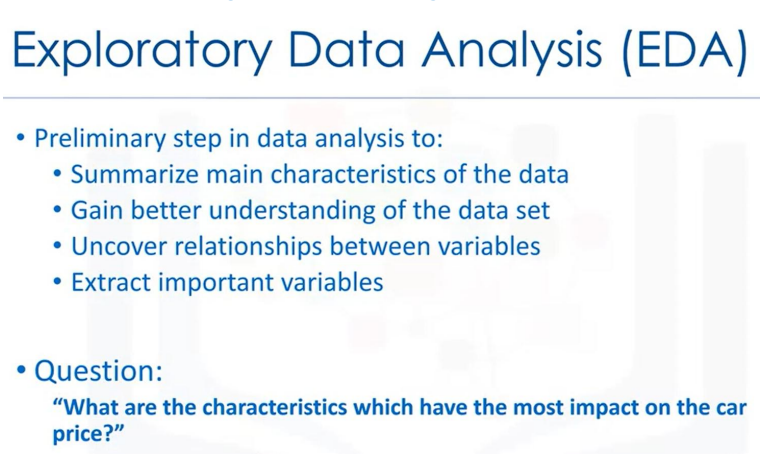
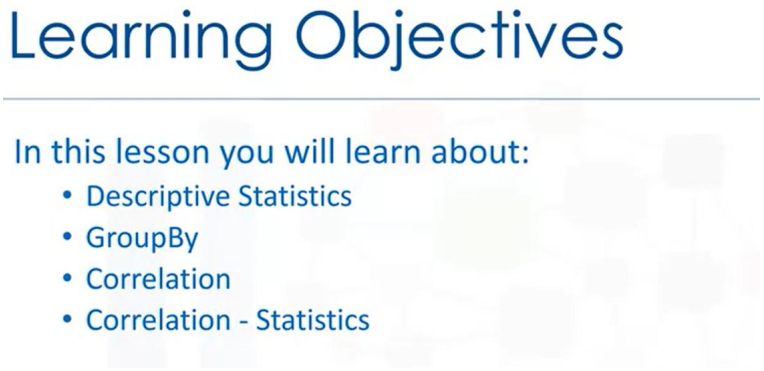
Week 3 - Exploratory Data Analysis

Learning Objectives

- Implement descriptive statistics
- Demonstrate the basics of grouping
- Describe data correlation processes
- Describe why and how to apply the Chi-Squared test

1. Exploratory Data Analysis





In this module we're going to cover the basics of exploratory data analysis using python.
Exploratory data analysis or in short **EDA** is an approach to analyze data in order to summarize main characteristics of the data gain better understanding of the data set, uncover relationships between different variables, and extract important variables for the problem we're trying to solve. The main question we are trying to answer in this module is what are the characteristics that have the most impact on the car price? We will be going through a couple of different useful exploratory data analysis techniques in order to answer this question, In this module you will learn about:

- **Descriptive statistics**, which describe basic features of a data set and obtains a short summary about the sample and measures of the data.
- **Basic of grouping data using group by** and how this can help to transform our data set,
- **Correlation** between different variables,
- **Advanced correlation**, where we'll introduce you to various correlation statistical methods namely pearson correlation and correlation heat maps

总结:

2. Descriptive Statistics

Descriptive Statistics- Describe()

- Describe basic features of data
- Giving short summaries about the sample and measures 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	201.000000
mean	100.000000	0.840796	122.000000	98.797015	174.200995	65.889055	53.766667	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

In this video, we'll be talking about Descriptive Statistics. When you begin to analyze data, it's important to first explore your data before you spend time building complicated models. One easy way to do so, is to calculate some Descriptive Statistics for your data. Descriptive statistical analysis helps to describe basic features of a data set, and obtains a short summary about the sample and measures of the data. Let's show you a couple different useful methods.

One way in which we can do this is by using the describe function in pandas. Using the describe function and applying it on your data frame, the ***describe*** function automatically computes basic statistics for all numerical variables. It shows the mean, the total number of data points, the standard deviation, the quartiles and the extreme values. Any NAN values are automatically skipped in these statistics. This function will give you a clear idea of the distribution of your different variables.

Descriptive Statistics - Value_Counts()

- summarize the categorical data is by using the **value_counts()** method

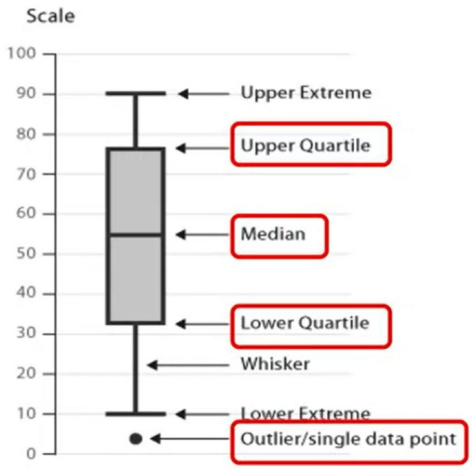
```
drive_wheels_counts=df["drive-wheels"].value_counts().to_frame()

drive_wheels_counts.rename(columns={'drive-wheels':'value_counts'}, inplace=True)
drive_wheels_counts
```

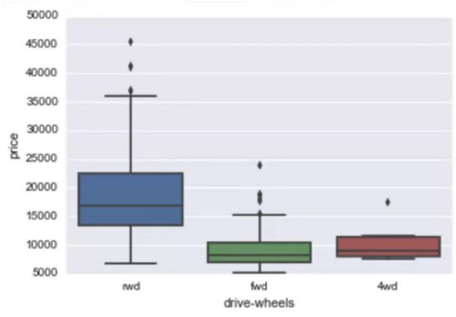
	value_counts
drive-wheels	
fwd	118
rwd	75
4wd	8

You could have also **categorical variables** in your data set. These are variables that can be divided up into different categories or groups, and have discrete values.
For example, in our data set we have the drive system as a categorical variable, which consists of the categories, forward wheel drive, rear wheel drive and four wheel drive. One way you can summarize the categorical data, is by using the function ***value_counts***. We can change the name of the column to make it easier to read. We see that we have 118 cars in the front wheel drive category. 75 cars in the rear wheel drive category, and 8 cars in the four wheel drive category.

Descriptive Statistics - Box Plots




```
sns.boxplot(x= "drive-wheels", y= "price",data=df)
```



Box plots are a great way to visualize numeric data, since you can visualize the various distributions of the data.

- The main features that the box plot shows, are the **median** of the data, which represents where the middle data point is.
- The **upper quartile** shows where the 75th percentile is. The **lower quartile** shows where the 25th percentile is. The data between the upper and lower quartile represents the **interquartile range**.
- Next you have the **lower and upper extremes**. These are calculated as 1.5 times the interquartile range, above the 75th percentile, and as 1.5 times the **IQR** below the 25th percentile.
- Finally, box plots also display **outliers** as individual dots that occur outside the upper and lower extremes.

With box plots, you can easily spot outliers, and also see the distribution and skewness of the data. Box plots make it easy to compare between groups. In this example, using box plot we can see the distribution of different categories of the drive wheels feature over price feature. We can see that the distribution of price between the rear wheel drive, and the other categories are distinct. But the price for front wheel drive and four wheel drive are almost indistinguishable.

总结:

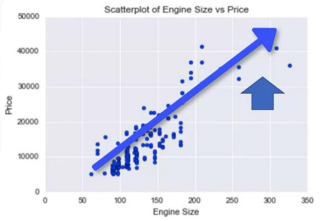
Descriptive Statistics - Scatter Plot

- Each observation represented as a point.
- Scatter plot Show the relationship between two variables.

1. Predictor/independent variables on x-axis.
2. Target/dependent variables on y-axis.

```
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")
```

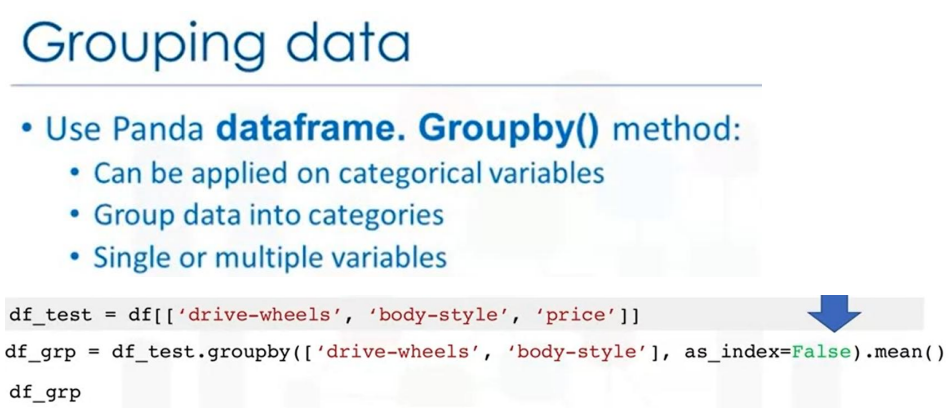


Often times we tend to see continuous variables in our data. These data points are numbers contained in some range. For example, in our data set price and engine size are continuous variables. What if we want to understand the relationship between engine size and price. Could engine size possibly predict the price of a car?
One good way to visualize this is using a **scatter plot**. Each observation in the scatter plot is represented as a point. This plot shows the relationship between two variables. The predictor variable, is the variable that you are using to predict an outcome. In this case our predictor variable is the engine size. The target variable is the variable that you are trying to predict. In this case, our target variable is the price. Since this would be the outcome. In a scatter plot, we typically set the predictor variable on the x-axis or horizontal axis, and we set the target variable on the y-axis or vertical axis.

In this case, we will thus plot the engine size on the x-axis and the price on the y-axis. We are using, the matplotlib functions scatter here, taking in x and y variable. Something to note is that it's always important to label your axes, and write a general plot title, so that you know what you're looking at. Now how is the variable engine size related to price? From the scatter plot, we see that as the engine size goes up, the price of the car also goes up. This is giving us an initial indication that there is a positive linear relationship between these two variables.

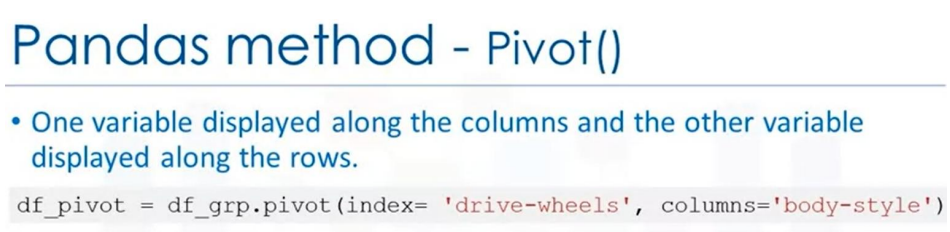
总结:

3. GroupBy in Python



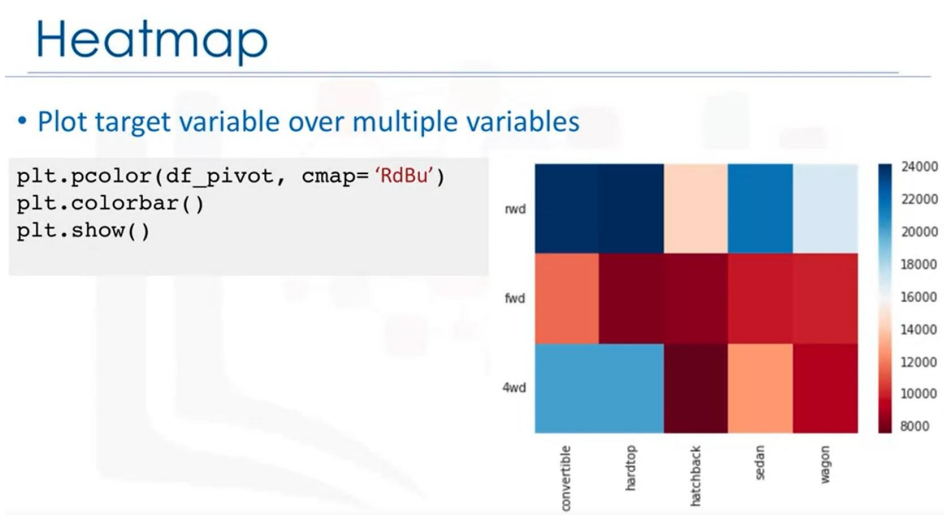
	drive-wheels	body-style	price
0	4wd	convertible	20239.229524
1	4wd	sedan	12647.333333
2	4wd	wagon	9095.750000
3	fwd	convertible	11595.000000
4	fwd	hardtop	8249.000000
5	fwd	hatchback	8396.387755
6	fwd	sedan	9811.800000
7	fwd	wagon	9997.333333
8	rwd	convertible	23949.600000
9	rwd	hardtop	24202.714286
10	rwd	hatchback	14337.777778
11	rwd	sedan	21711.833333
12	rwd	wagon	16994.222222

In this video, we'll cover the basics of grouping and how this can help to transform our dataset.
Assume you want to know, is there any relationship between the different types of drive system, forward, rear, and four-wheel drive, and the price of the vehicles? If so, which type of drive system adds the most value to a vehicle? It would be nice if we could group all the data by the different types of drive wheels and compare the results of these different drive wheels against each other.
In Pandas, this can be done using the group by method. The ***groupby*** method is used on categorical variables, groups the data into subsets according to the different categories of that variable. You can group by a single variable or you can group by multiple variables by passing in multiple variable names.
As an example, let's say we are interested in finding the average price of vehicles and observe how they differ between different types of body styles and drive wheels variables. To do this, we first pick out the three data columns we are interested in, which is done in the first line of code. We then group the reduced data according to drive wheels and body style in the second line. Since we are interested in knowing how the average price differs across the board, we can take the mean of each group and append it this bit at the very end of the line too. The data is now grouped into subcategories and only the average price of each subcategory is shown. We can see that according to our data, rear wheel drive convertibles and rear wheel drive hard hardtops have the highest value while four wheel drive hatchbacks have the lowest value. A table of this form isn't the easiest to read and also not very easy to visualize.



	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

To make it easier to understand, we can transform this table to a pivot table by using the **pivot method**.
In the previous table, both drive wheels and body style were listening columns. A pivot table has one variable displayed along the columns and the other variable displayed along the rows. Just with one line of code and by using the Panda's pivot method, we can pivot the body style variable so it is displayed along the columns and the drive wheels will be displayed along the rows. The price data now becomes a rectangular grid, which is easier to visualize.
This is similar to what is usually done in Excel spreadsheets.



In this example, we use pyplot's p color method to plot heat map and convert the previous pivot table into a graphical form. We specify the red-blue **color scheme**. In the output plot, each type of body style is numbered along the x-axis and each type of drive wheels is numbered along the y-axis. The average prices are plotted with varying colors based on their values. According to the color bar, we see that the top section of the heat map seems to have higher prices than the bottom section.

Another way to represent the pivot table is using a **heat map plot**. Heat map takes a rectangular grid of data and assigns a color intensity based on the data value at the grid points. It is a great way to plot the target variable over multiple variables and through this get visual clues with the relationship between these variables and the target.

总结:

4. Correlation

Correlation

What is Correlation?

- Measures to what extent different variables are interdependent.
- For example:
  - Lung cancer → Smoking
  - Rain → Umbrella
- Correlation doesn't imply causation.

In this video, we'll talk about the correlation between different variables. **Correlation** is a statistical metric for measuring to what extent different variables are interdependent. In other words, when we look at two variables over time, if one variable changes how does this affect change in the other variable.

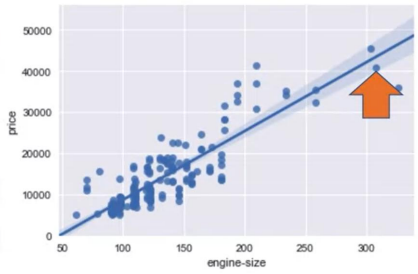
For example, smoking is known to be correlated to lung cancer since you have a higher chance of getting lung cancer if you smoke. In another example, there is a correlation between umbrella and rain variables where more precipitation means more people use umbrellas. Also, if it doesn't rain people would not carry umbrellas. Therefore, we can say that umbrellas and rain are interdependent and by definition they are correlated.

It is important to know that correlation doesn't imply causation. In fact, we can say that umbrella and rain are correlated but we would not have enough information to say whether the umbrella caused the rain or the rain caused the umbrella. In data science we usually deal more with correlation.

Correlation - Positive Linear Relationship

- Correlation between two features (engine-size and price).

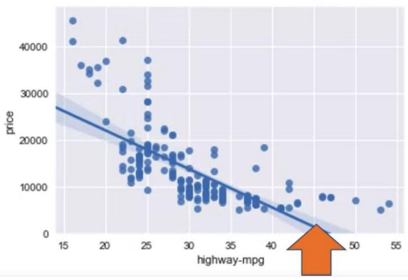
```
sns.regplot(x="engine-size", y= "price", data=df)
plt.ylim(0,)
```



Correlation - Negative Linear Relationship

- Correlation between two features (highway-mpg and price).

```
sns.regplot(x="highway-mpg", y= "price", data=df)
plt.ylim(0,)
```



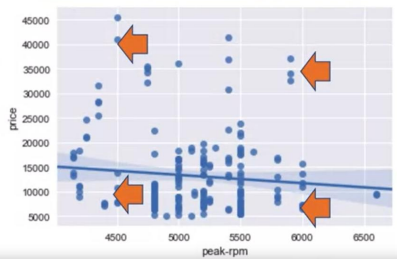
Let's look at the correlation between engine size and price. This time we'll visualize these two variables using a scatter plot and an added linear line called a regression line, which indicates the relationship between the two. The main goal of this plot is to see whether the engine size has any impact on the price. In this example, you can see that the straight line through the data points is very steep which shows that there's a positive linear relationship between the two variables. With increase in values of engine size, values of price go up as well and the slope of the line is positive. So there is a **positive correlation** between engine size and price. We can use seaborn.***regplot*** to create the scatter plot.

As another example, now let's look at the relationship between highway miles per gallon to see its impact on the car price. As we can see in this plot, when highway miles per gallon value goes up the value price goes down. Therefore there is a negative linear relationship between highway miles per gallon and price. Although this relationship is negative the slope of the line is steep which means that the highway miles per gallon is still a good predictor of price. These two variables are said to have a **negative correlation**.

Correlation - Negative Linear Relationship

- Weak correlation between two features (peak-rpm and price).

```
sns.regplot(x="peak-rpm", y= "price", data=df)
plt.ylim(0,)
```



Finally, we have an example of a **weak correlation**.

For example, both low peak RPM and high values of peak RPM have low and high prices. Therefore, we cannot use RPM to predict the values.

总结:

5. Correlation - Statistics

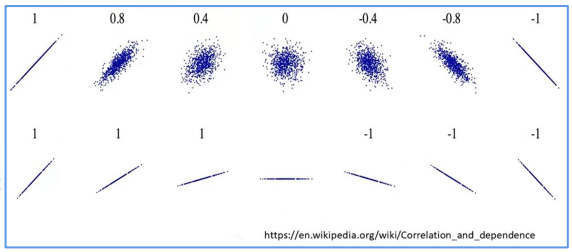
Pearson Correlation

- Measure the strength of the correlation between two 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



故这里：

Null hypothesis H_0: These two features are uncorrelated.

Alt hypothesis H_A: These two features are correlated. (p value越小，certainty越大)

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

One way to measure the strength of the correlation between continuous numerical variable is by using a method called **Pearson correlation**. Pearson correlation method will give you two values: the correlation coefficient and the P-value. So how do we interpret these values?

- For the correlation coefficient, a value close to 1 implies a large positive correlation, while a value close to negative 1 implies a large negative correlation, and a value close to zero implies no correlation between the variables.
- Next, the P-value will tell us how certain we are about the correlation that we calculated. For the P-value, a value less than.001 gives us a strong certainty about the correlation coefficient that we calculated. A value between.001 and.05 gives us moderate certainty. A value between 0.05 and 0.1 will give us a weak certainty. And a P-value larger than .1 will give us no certainty of correlation at all.

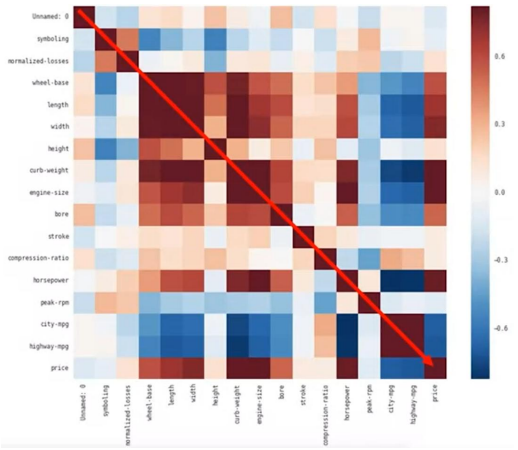
We can say that there is a strong correlation when the correlation coefficient is close to 1 or -1, and the P-value is less than.001. The following plot shows data with different correlation values.

Pearson Correlation- Example

```
pearson_coef, p_value = stats.pearsonr(df['horsepower'], df['price'])
```

- Pearson correlation: 0.81
- P-value：9.35 e-48

Correlation- Heatmap



In this example, we want to look at the correlation between the variable's horsepower and car price. See how easy you can calculate the Pearson correlation using the SI/PI stats package? We can see that the correlation coefficient is approximately.8, and this is close to 1. So there is a strong positive correlation. We can also see that the P-value is very small, much smaller than.001. And so we can conclude that we are certain about the strong positive correlation.

Taking all variables into account, we can now create a heatmap that indicates the correlation between each of the variables with one another. The color scheme indicates the Pearson correlation coefficient, indicating the strength of the correlation between two variables. We can see a diagonal line with a dark red color, indicating that all the values on this diagonal are highly correlated. This makes sense because when you look closer, the values on the diagonal are the correlation of all variables with themselves, which will be always 1. This correlation heatmap gives us a good overview of how the different variables are related to one another and, most importantly, how these variables are related to price.

总结：

6. Association between two 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.

Chi-Square Test for association

- 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.

In this video, we will learn how to find out if there is a relationship between two **categorical variables**. When dealing with the relationships between two categorical variables, we can't use the same correlation method for continuous variables, we will have to employ the use of chi square test for the association. The **Chi-square test** is intended to test how likely it is that an observed distribution is due to chance. It measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

Before we go into an example, let's go through some important points.

- The Chi-square test's null hypothesis is that the variables are independent. The test compares the observed data to the values that the model expects if the data was distributed in different categories by chance. Anytime the observed data doesn't fit within the model of the expected values, the probability that the variables are dependent becomes stronger, thus proving the null hypothesis incorrect.
- **The Chi-square does not tell you the type of relationship that exists between both variables only that a relationship exists**.

Categorical variables

- Is there an association between fuel-type and aspiration?

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

Observed value

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

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

$$\text{Expected value} = \frac{\text{Row total} \quad * \text{Column total}}{\text{Grand total}}$$

Expected value

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

We will use the cars dataset. Assuming we want to test the relationship between fuel-type and aspiration; these are categorical variables. It is either the fuel-type of the car is gas or diesel, and the aspiration is that either the car is standard or Turbo. To do this we will find the observed counts of cars in each category. This can be done by creating a **crosstab** using the pandas library. A **crosstab** is a table showing the relationship between two or more variables. When the table only shows the relationship between two categorical variables, a crosstab is also known as a **contingency table**.

In our case the crosstab or contingency table is shows us the counts in each category: a standard car with diesel fuel, a standard car with gas fuel, a turbo car with diesel fuel, or a turbo car with gas fuel. The formula for chi-square is given as follows The summation of the observed value i.e., the counts in each group minus the expected value all squared divided by the Expected value. Expected values are based on the given totals, that is what can we say individual cells would be if we did not know the observed values.

- To calculate the expected value of a standard car with diesel, We take the row total which is 20 multiplied by The column total 168 Divided by the Grand total of 205. This will give you 16.39, sixteen point three nine.
- If we do the same thing for Turbo cars with gas fuel, we will take Row Total 185 multiplied by Column total 37, and we divide by the Grand total 205 we get 33.39, thirty-three point three nine.
- If we repeat the same procedure for all of them we get these values.

If we took the row totals, column totals, and grand total we will get the same values as the totals as the observed values.

总结:

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.

故这里：

Null hypothesis H_0: These two variables are independent. (no association between them)

Alt hypothesis H_A: There is a association between these two variables.

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

```
scipy.stats.chi2_contingency(cont_table, correction = True)
```

```
(29.605759385109046,
 5.2947382636786724e-08,
 1,
 array([[ 16.3902439,   3.6097561],
        [151.6097561,  33.3902439]]))
```

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

Now going back to this formula, if we took a a summation of all the observed minus the expected values all squared divided by the expected value, we will get a chi-square value of 29.6.

On the chi-square table we check on the degree of freedom equals one, and find the value closest to twenty-nine point six, here we can see that twenty-nine point six will fall in between a p-value ~~less than 0.1 and 0.25. Therefore, we can say the p-value is greater than 0.1. Since the p-value is greater~~ less than 0.05, we reject the null alternative hypothesis that the two variables are independent and therefore we conclude that there is an association between fuel type and aspiration.

To do this in python we will use the ***chi square contingency*** function in the ***scipy.statistics*** package. The function will print out the chi-square test value twenty-nine point six and the second value is the p-value which is very close to 0 and a degree of freedom of 1. If you remember the chi-square table did not give an exact p-value but a range in which it falls, python will give the exact p-value. We can see the same results as our previous slides. It also prints out the expected values which we also calculated by hand since the p-value is close to zero, we reject the null hypothesis that the two variables are independent and conclude that there is evidence of association between fuel-type and aspiration.

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.

总结：