

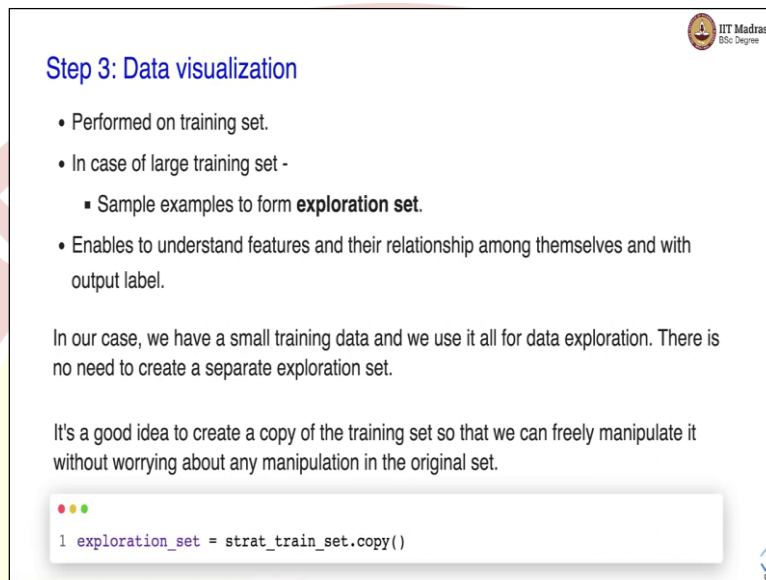
IIT Madras

ONLINE DEGREE

B. Sc in Programming and Data Science
Diploma Level
Dr. Ashish Tendulkar
Indian Institute of Technology – Madras

Data Visualization

(Refer Slide Time: 00:10)



Step 3: Data visualization

- Performed on training set.
- In case of large training set -
 - Sample examples to form **exploration set**.
- Enables to understand features and their relationship among themselves and with output label.

In our case, we have a small training data and we use it all for data exploration. There is no need to create a separate exploration set.

It's a good idea to create a copy of the training set so that we can freely manipulate it without worrying about any manipulation in the original set.

```
1 exploration_set = strat_train_set.copy()
```

Namaste welcome to the next video of machine learning practice course. In this video we will study the next step of end-to-end machine learning project that is data visualization. Data visualization is performed on training set and in case of large training set we usually sample examples to form an exploration set. Data visualization enables us to understand the features and the relationship among themselves and with the output label.

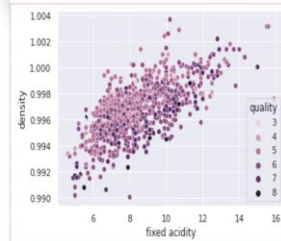
In our case since we have a small amount of training data will use all of it for data exploration. In this case there is no need to create a separate exploration set. It is a good idea to create a copy of the training set so, that we can freely manipulate it without worrying about any manipulation in the original data.

(Refer Slide Time: 01:07)

Scatter Visualization

With seaborn library:

```
1 sns.scatterplot(x='fixed acidity', y='density', hue='quality',  
2 data=exploration_set)
```



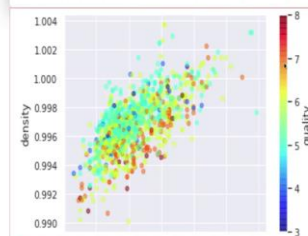
So, there are different visualization techniques that we can use the first one is the scatter visualization. So, we can use the seaborn library for doing the scatter visualization and there is a scatter plot function in the library that we can use. You have to predict we have to specify the x-axis and the y-axis and the data for visualization. So, the scatter plot prints the data in form of such a graph where we have fixed acidity on the x axis and density on the y axis.

And each point over here is a training point and each point has been assigned a colour based on the quality of the wine.

(Refer Slide Time: 02:01)

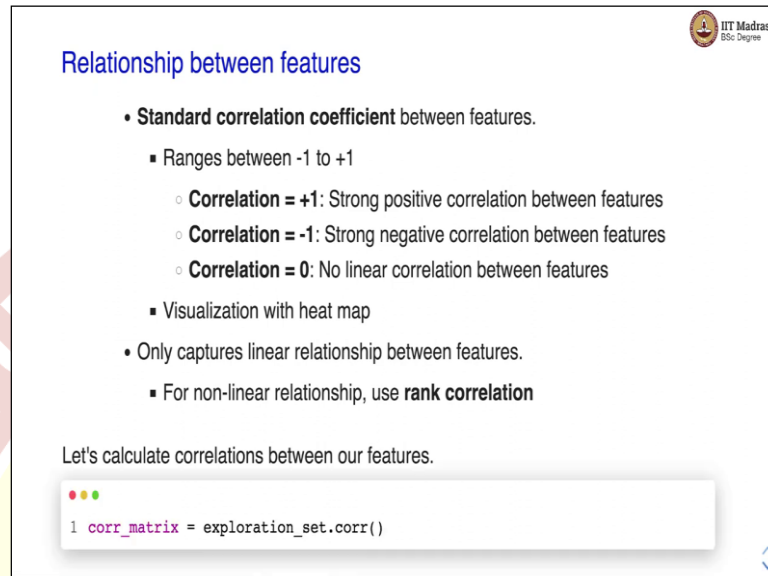
With matplotlib:

```
1 exploration_set.plot(kind='scatter', x='fixed acidity', y='density', alpha=0.5,  
2 c="quality", cmap=plt.get_cmap("jet"))
```



Another library that we can use is matplotlib for the same purpose. In matplotlib we can show the same kind of data with slightly different look at fill as well as you know by using some kind of a colour coded strip for denoting the quality.

(Refer Slide Time: 02:27)



Relationship between features

- **Standard correlation coefficient** between features.
 - Ranges between -1 to +1
 - **Correlation = +1**: Strong positive correlation between features
 - **Correlation = -1**: Strong negative correlation between features
 - **Correlation = 0**: No linear correlation between features
 - Visualization with heat map
- Only captures linear relationship between features.
 - For non-linear relationship, use **rank correlation**

Let's calculate correlations between our features.

```
1 corr_matrix = exploration_set.corr()
```

Apart from visualizing individual features we can also study relationship between features. One of the relationship is a standard correlation coefficient. It ranges between -1 to +1 correlation of +1 denotes a very strong positive correlation between the features correlation of -1 denotes a very strong negative correlation between the features and correlation of 0 means that there is no linear correlation between the features.

We can visualize correlation with a heat map. So, note that standard correlation coefficient only captures the linear relationship between the features. If you suspect or if you believe that there is a non-linear relationship between the features, you can use rank correlation for that purpose. Let us calculate correlation between our features. We can use core function for calculating the correlations between the feature.

(Refer Slide Time: 03:31)

Let's check features that are correlated with the label, which is quality in our case.

```
1 corr_matrix['quality']
```

fixed acidity	0.107940
volatile acidity	-0.383249
citric acid	0.210802
residual sugar	0.003710
chlorides	-0.120231
free sulfur dioxide	-0.048291
total sulfur dioxide	-0.194511
density	-0.193009
pH	-0.052063
sulphates	0.228050
alcohol	0.481197
quality	1.000000

Name: quality, dtype: float64

Notice that **quality** has strong positive correlation with **alcohol** content [0.48] and strong negative correlation with **volatile acidity** [-0.38].

Let us check out the features that are correlated with the label and in our case the label is the quality of the wine. So, you can see that. So, if I plot this correlation matrix with respect to the label these are correlation of different features with the label. You can see that sulfate has correlation coefficient of 0.22 alcohol has correlation coefficient of 0.48 which is probably the strongly correlated features with the quality of the wine.

Whereas volatile acidity is probably the strong negatively correlated features with the quality of the wine.

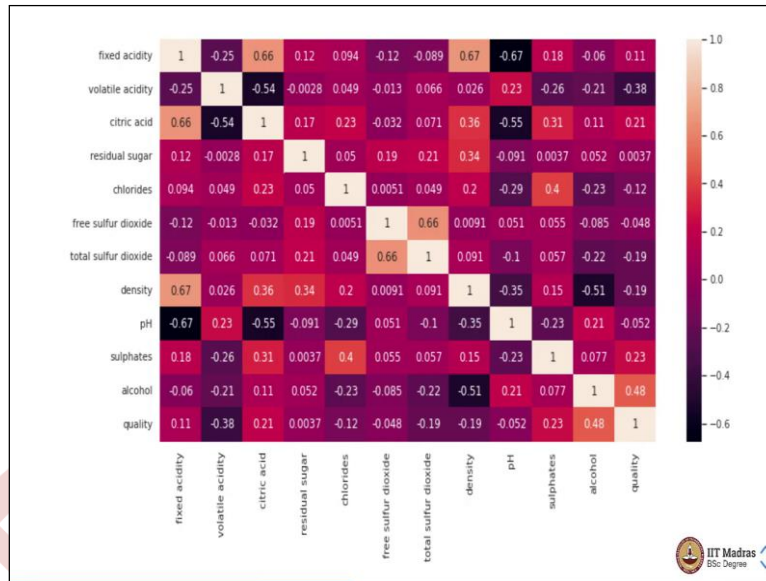
(Refer Slide Time: 04:19)

Let's visualize correlation matrix with heatmap:

```
1 plt.figure(figsize=(14,7))
2 sns.heatmap(corr_matrix, annot=True)
```

We can visualize the correlation matrix with heat map. So, there is a heat map function in seaborn library that we can use for that purpose.

(Refer Slide Time: 04:29)



So, here you can see all correlation coefficients in one go in form of a heat map. So, what you see is this heat map which is some kind of a symmetric matrix there are features on y-axis and then there are features also in the x-axis and what you see is each cell calculates correlation between the features. So, here you can see that fixed acidity and citric acid are strongly correlated with correlation coefficient of 0.66.

And you know you can see the colour coding that we have used for representing this correlation on the side on this colour bar here on the side. As we move towards correlation of wine the colour becomes fender and as we move towards a smaller or lower correlation the colour becomes darker. In this way you can quickly spot strongly positively correlated features as well as negatively correlated features.

So, these are some of the examples of negatively correlated features for example fixed acidity and pH's are negatively correlated.

(Refer Slide Time: 05:43)

You can notice:

- The correlation coefficient on diagonal is +1.
- Darker colors represent negative correlations, while fainter colors denote positive correlations. For example
 - citric acid and fixed acidity have strong positive correlation.
 - pH and fixed acidity have strong negative correlation.

Another option to visualize the relationship between the feature is with scatter matrix.

	fixed acidity	volatile acidity	citric acid
fixed acidity	1	-0.25	0.66
volatile acidity	-0.25	1	-0.54
citric acid	0.66	-0.54	1
residual sugar	0.12	-0.0028	0.17
chlorides	0.094	0.049	0.23
free sulfur dioxide	-0.12	-0.013	-0.032
total sulfur dioxide	-0.089	0.066	0.071
density	0.67	0.026	0.36
pH	-0.67	0.23	-0.55
sulphates	0.18	-0.26	0.31
alcohol	-0.06	-0.21	0.11
quality	0.11	-0.38	0.21

So, you can notice that the correlation coefficient on diagonal is one. The darker colors represent negative correlations while ventricular denote positive correlations. For example, citric acid and fixed acidity have strong positive correlation with correlation coefficient of 0.66. pH and fixed acidity have strong negative correlation which is correlation of -0.67. Another option to visualize the relationship between the features is with scatter matrix.

(Refer Slide Time: 06:20)



So, there is a scatter matrix function in pandas plotting library that we can use. So, scatter matrix function also plots the relationship between features in form of scatter plots. On diagonal you see some kind of histograms they denote the distribution of the individual feature. For example, this particular histogram shows the show that shows the distribution of the citric acid feature. Whereas this particular scatter plot shows the relationship

between citric acid and pH so, this is a scatter plot of citric acid pH versus citric acid and so on.

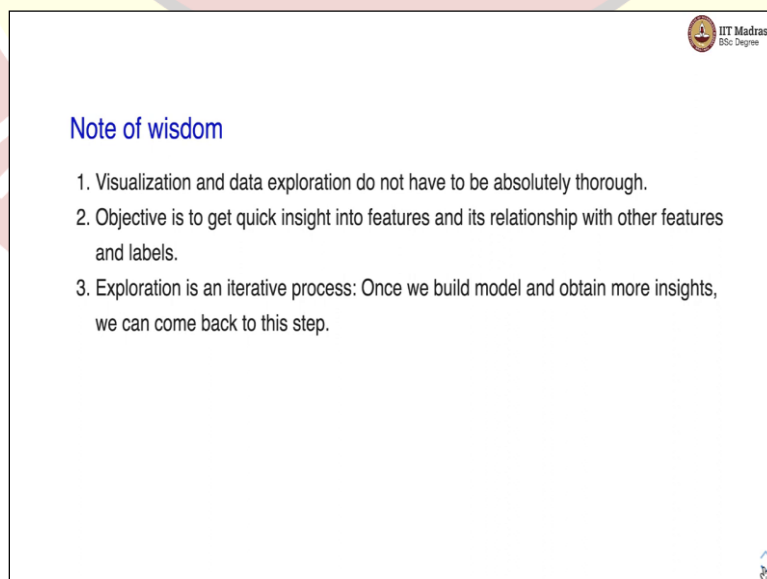
For the convenience of visualization, we are showing only a small number of attributes on the screen otherwise this would have been a very large plot which cannot be accommodated on this particular screen.

(Refer Slide Time: 07:19)



So, similar analysis can be carried out with combined features, features that are derived from the original features.

(Refer Slide Time: 07:31)



There is a small note of wisdom for all of your visualization and data exploration do not have to be absolutely thorough. The objective here is to get quick insight into the features

and its relationship with other features and the labels. Exploration is an iterative process once we build the model we obtain more insights and we can come back to the exploration step if necessary. That is all from the data visualization process. In the next step we will look at how to prepare the data for the training.

