Data Quality Inspection

Objectives

In this reading assignment you'll be able to:

- Inspect data using data profiling and visualization
- Examplify data profiling along with extracting metadata information
- Explain the use case of barplot, histogram, boxplot, scatterplot, and heatmap

Prerequisites required

- ML Level One

Data Quality Inspection

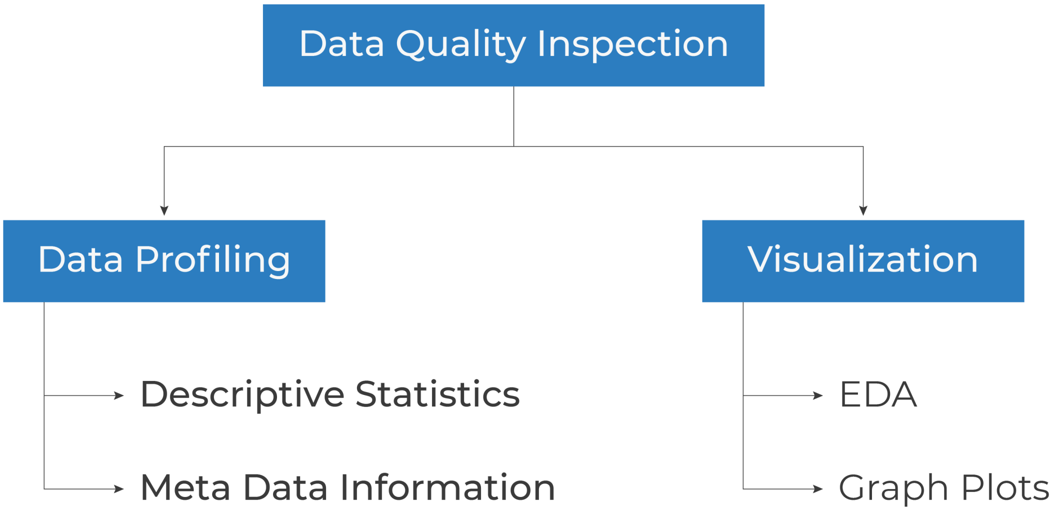

Figure 1: Methods of Data Quality Inspection

The two common approach of data quality inspection is: Data Profiling and Data visualization.

Data Profiling

Data profiling is a process of analyzing, understanding content, and interrelationships on the raw data to characterize the information embedded within a data set. Data profiling collects statistics and meta-information about available data that provide insight into the content of data sets.

In data profiling, you find descriptive statistics of data. Descriptive statistics include those that summarize the central tendency, dispersion, and shape of a dataset's distribution, excluding NaN values. You can use the results of data profiling to formulate a hypothesis about the data features.

In Machine learning, you find descriptive statistics of your dataset using Pandas dataframe.describe() (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html); dataframe.describe() method returns:

	RAD	TAX
count	506.000000	506.000000
mean	9.549407	408.237154
std	8.707259	168.537116
min	1.000000	187.000000
25%	4.000000	279.000000
50%	5.000000	330.000000
75%	24.000000	666.000000
max	24.000000	711.000000

In data profiling, you find descriptive statistics of data like count, mean, minimum value, percentiles, maximum values, etc. Use dataframe.describe() function to do data profiling. dataframe.describe() method returns:

1. Count:
   - Count is the number of non-missing data points for each variable. In the given, there are 506 rows.
2. Mean:
   - Mean is average. The calculation of the mean incorporates all values in the data. If you change one value, mean changes. In a normal distribution, the average represents the center of data points. However, It doesn't happen in a skewed distribution. The mean value of the RAD column is 9.54.
3. Standard deviation(Std):
   - Standard deviation tells you how to spread out the data is. It is a measure of how far each observed value is from the mean. In any distribution, about 95% of values will be within two standard deviations of the mean. The standard deviation of the RAD column is 8.707259. It also gives variance in data.
4. Minimum value(min):
   - Min is the minimum value in a column.
5. Percentile:
   - Percentile is a number where a certain percentage of scores fall below that number. In the RAD column, 50% of the data falls below value 5.
6. Maximum value (max):
   - Max is the extreme or maximum value in a column.

Meta-data information

Let's say the following table is your data frame. You want to extract meta-data information from this data.

In [ ]:

```
import pandas as pd
import numpy as np
data = {
    'f1':['x','y','x','x'],
    'f2':[0.2, np.nan, 0.21,0.29],
    'f3':[11.2,21.1, 12.2,13],
    'f4': pd.Categorical ([1,2,1,1]),
    'f4':[12,11,7,9]
}
df = pd.DataFrame(data)
df
```

Out[16]:

	f1	f2	f3	f4
0	x	0.20	11.2	12
1	y	NaN	21.1	11
2	x	0.21	12.2	7
3	x	0.29	13.0	9

Let's see the metadata information of this dataset using `dataframe.info().`

In [ ]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   f1      4 non-null      object 
 1   f2      3 non-null      float64
 2   f3      4 non-null      float64
 3   f4      4 non-null      int64  
dtypes: float64(2), int64(1), object(1)
memory usage: 256.0+ bytes
```

Let's interpret the results.

- The data frame is of pandas.core.frame.DataFrame type.
- The index ranges from zero to three with a total of four entries.
- There are total 5 columns namely `f1,f2,f3,f4, and f5.`
- `f1` is of an object type, `f2 and f3` is of float64 type, `f4` is category type, and `f5` is of int64 type.
- There are four not null entries on `f1,` similarly three not null values on `f2, etc.` You can find missing data using the not null count. However, other methodologies also exist to detect null values or missing values. I will discuss them in the next chapter.
- Dataframe is taking 256 bytes of memory.

Along with meta-data information, you can also find details about the object and category type features using `dataframe. describe(include="O")` method.

```
In [ ]:
```

```
df.describe(include='O')
```

```
Out[18]:
```

```
         f1
count    4
unique   2
top      x
freq     3
```

In this dataset, f1 is of object type there are two unique values with top, repeating being x with frequency three. There are 4 not null rows count.

Visualization

You can also use a visualization tool to visualize some information data profiling provides because visualization makes it display large data content in a figure, which will help hypothesis formulation about the data. Let's start with a barplot.

Bar Plot

The bar plot shows the relationship between a numeric and a categorical variable. In the bar plot, the categorical variable is represented as a bar, and the size of the bar represents its numeric value. The given bar diagram shows the number of occurrences of the value counts of different car classes. Car classes are the categorical variables and number of occurrences of each class in a numerical variable.

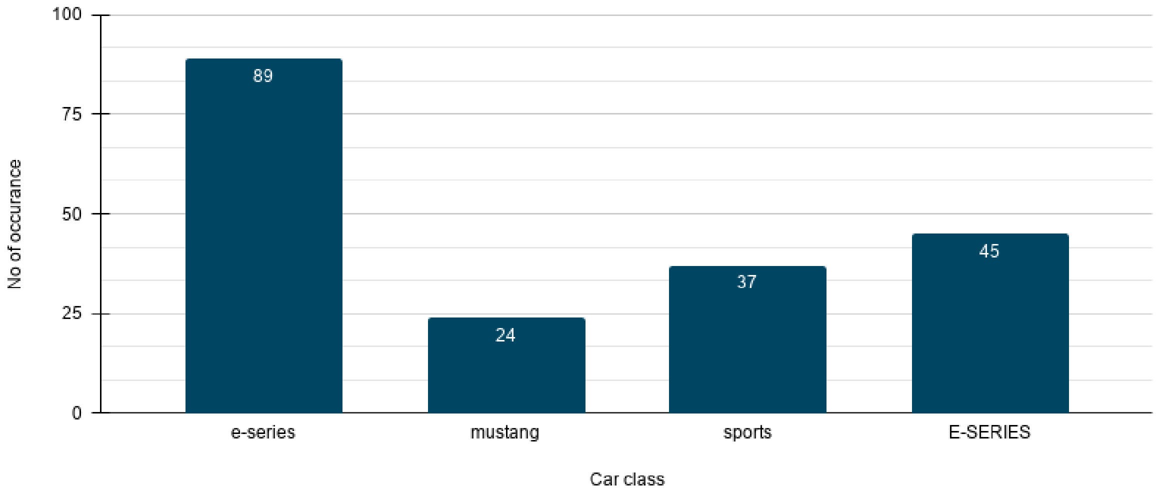

Figure 2: Bar Plot

The given bar diagram shows the number of occurrences of the value counts of different car classes. Car classes are the categorical variables and number of occurrences of each class in a numerical variable.

- The e-series car class occurs 89 times and hence is the highest occurring car class.

- The lowest occurring car class is the Mustang. It occurs 24 times.

The bar diagram looks strange because the lowercase e-series and the Uppercase E-SERIES are the same entity. Due to inconsistent capitalization, the actual count of e-series is now shown in the figure.

Histogram

A histogram is a representation of the distribution of numerical data that gives a discretized display of value frequency. The data points are split into discrete, evenly spaced bins, and the number of data points in each bin are plotted.

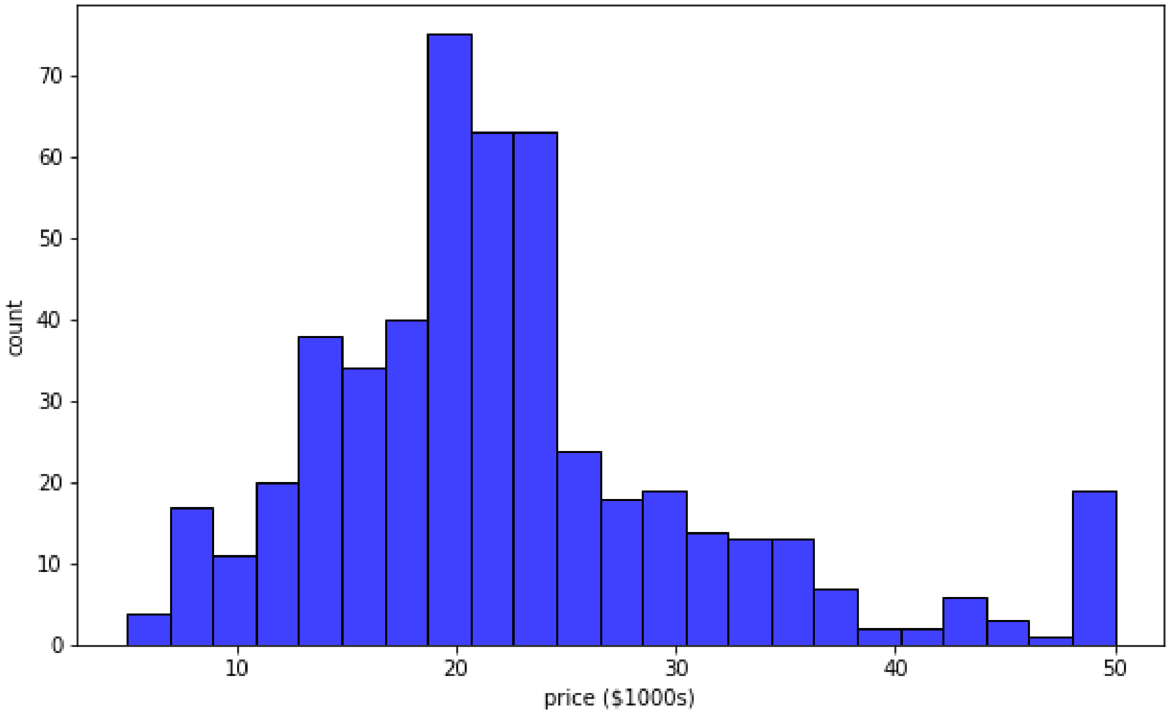

Figure 3: Histogram of Price column

In the above figure, the histogram contains the frequency distribution of the `price($1000s)` column. Most of the price value lies in between approx `17` to `25` thousand. Price Values between `40-50` are outliers.

Boxplot

Boxplots are a standardized way of displaying the distribution of data based on a five number summary: `minimum`, `first quartile (Q1)`, `median(Q2)`, `third quartile (Q3)`, and `maximum.`

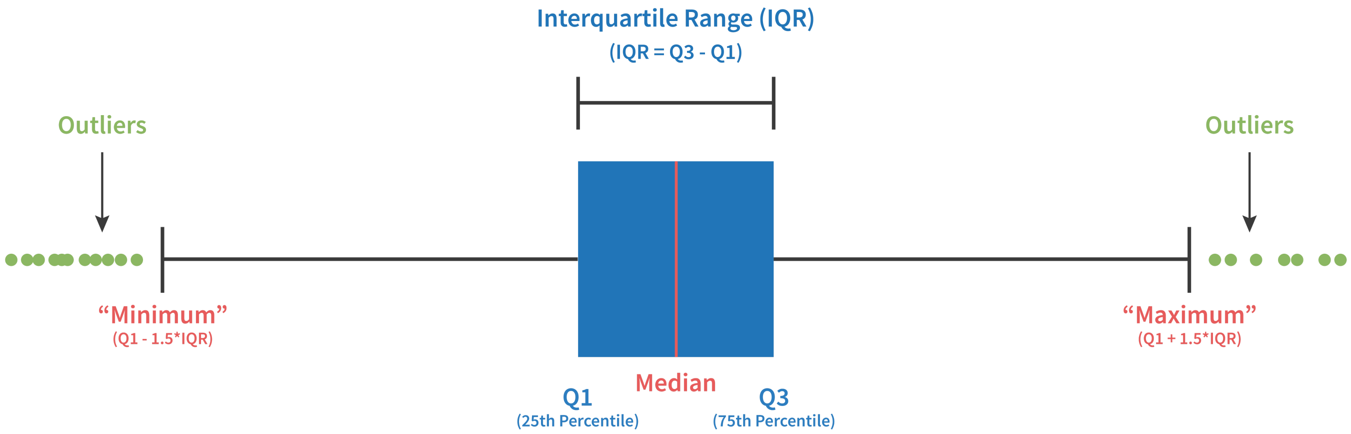

Figure 4: Boxplot

In the above figure green color points are outliers. Most of the value lies inside interquartile range(IQR). Median is a middle value which is given by Q2.

Scatter Plot

A scatterplot shows the relationship between two numerical variables. For example, following scatterplot shows the `price` and on x-axis and `LSTAT` is on y-axis:

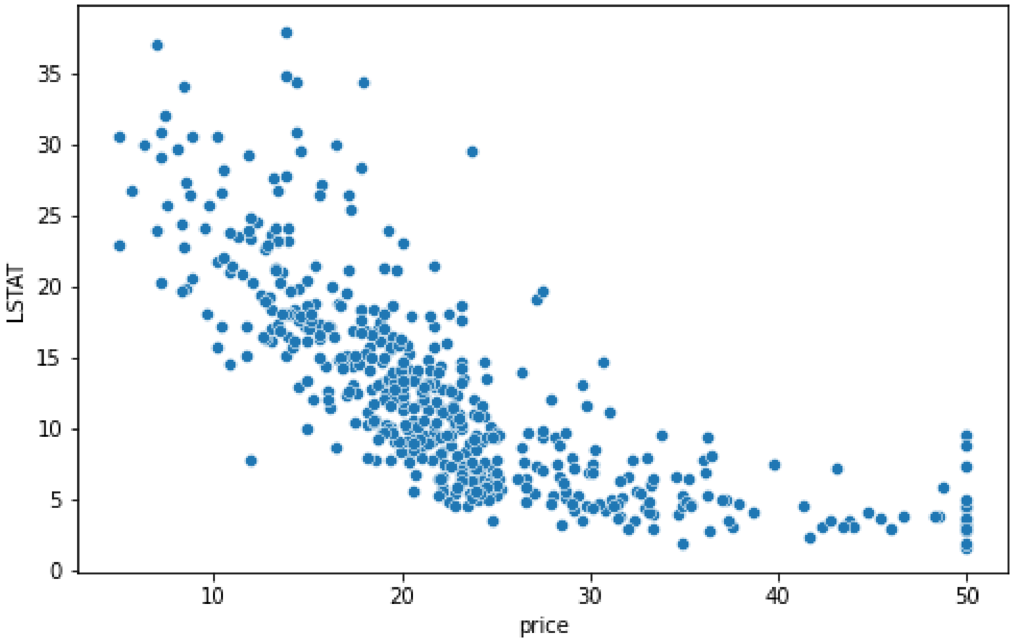

Figure 5: Scatterplot

Scatterplot describes the correlation between the two variables. In the above figure, as `price` increases, the value of `LSTAT` decreases. It means `price` and `LSTAT` are negatively correlated.

Heatmap

A heatmap is a 2D graphical representation of data where the individual values that are contained in a matrix are represented as colors. The following diagram is a heatmap of correlation between the variables.

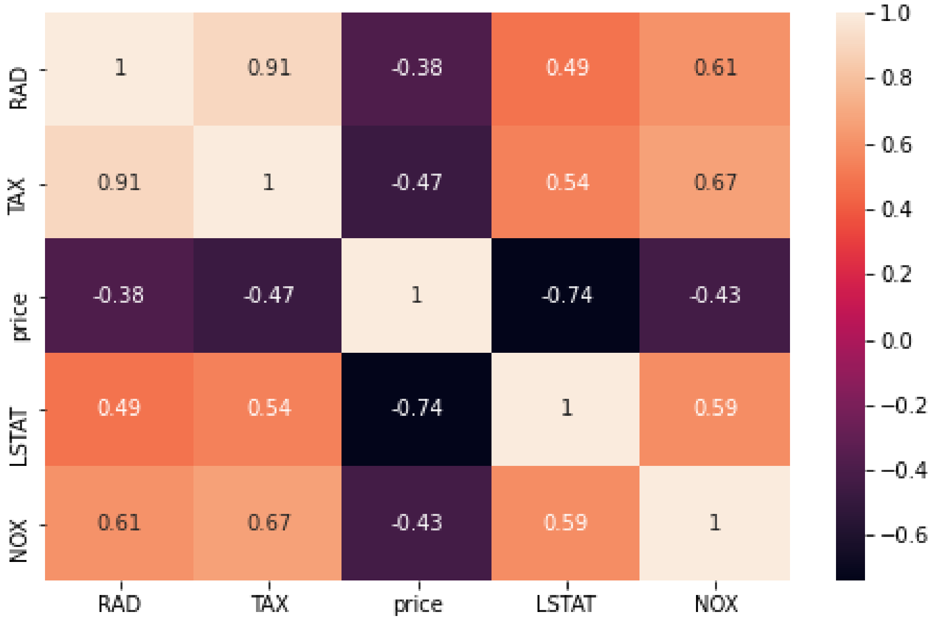

Figure 6: Heatmap

In the above heatmap, all diagonal value is one. It's because the correlation between two identical variables is one. The LSTAT and the price column has a -0.74 value of correlation. It is a negative correlation. TAX and RAD column is positively correlated, and their value of correlation is 0.91. The takeaway of this heatmap is that you cannot use TAX and RAD as the independent variable in the linear regression problem because it violates the linear regression assumption. Linear regression assumes that the independent variable should not be highly correlated. You should drop one column, either TAX or RAD. The dropped column is irrelevant.

Heatmap is also useful when you want to find NaN or missing values.

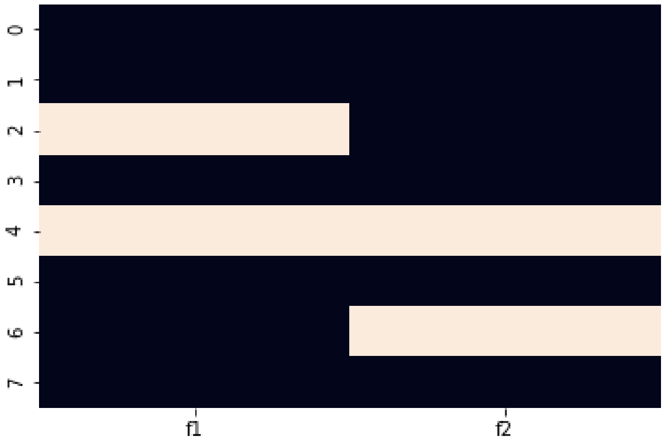

Figure 7: Heatmap of Missing data

The given heatmap shows:

- f1 contains the missing value in 2nd and 4th-row index while f2 consists of the missing value at 4th and 6th-row index.

Key Takeaways

- Data profiling of numerical variable returns descriptive statistics of data, like count, min, max, percentiles, standard deviations.
- Metadata includes data types of variables, not the null count of variables, number of rows and columns in data set, memory usage of data.
- Bar plot enables precise comparison between different categories.
- The histogram is used to visualize the distribution of the data.
- Heatmap is used to visualize correlation and missing data.
- Scatter Plot is used to visualize the correlation between the two numeric variables.