uday sai

4 Followers About

How to perform EDA which complements our Data Science story- Part 2

uday sai Apr 13, 2020 · 4 min read

In the previous article we have seen the objectives of EDA and theoretical concepts of implementing EDA. In this article we will proceed with the tools that are used in implementing EDA in python.

Note: This article is a continuation of my previous article, you can find my previous article on:

How to perform EDA which complements our Data Science story- Part 1

EDA is critical aspect of the data science projects. The main purpose of EDA is to analyze the data to get acquainted…

medium.comc

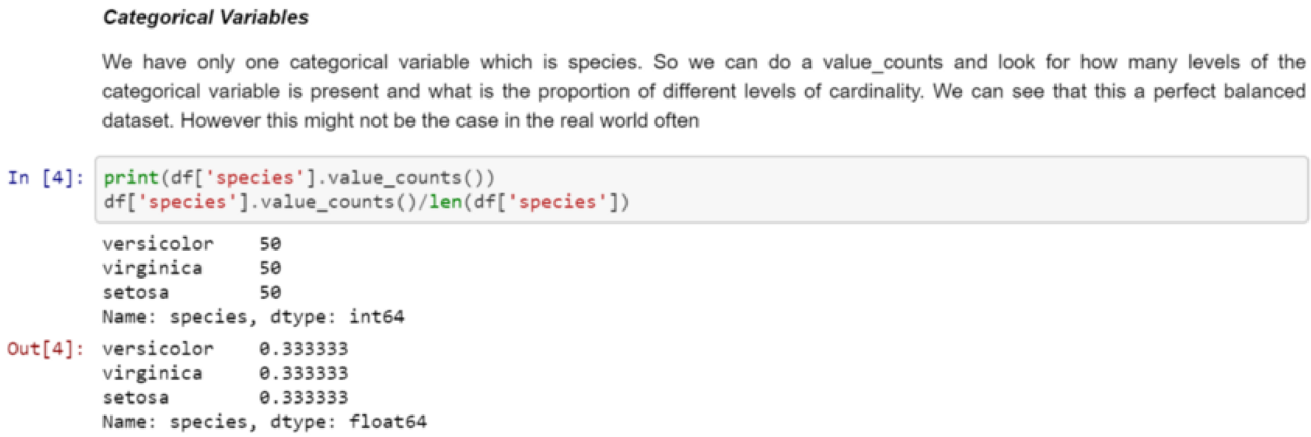
Let us explore the strategies in performing EDA with the help of some python libraries like numpy, pandas, matplotlib, seaborn and pandas-profiling. We can categorize the EDA as follows:

> *Univariate Non-Graphical EDA*
>
> *Univariate Graphical EDA*

> *Multivariate Non-Graphical EDA*
>
> *Multivariate Graphical EDA*

Univariate Non-Graphical EDA

The univariate non-graphical EDA is mostly associated with the descriptive statistics. For Categorical variable the descriptive statistics are the proportion of the data points associated with each level of cardinality in this categorical variable. In this process we will identify the 'rare' categories which are potential outliers in case of the categorical variable. The below is implementation of univariate non-graphical EDA on categorical variable on iris dataset:



Categorical Variables

We have only one categorical variable which is species. So we can do a value_counts and look for how many levels of the categorical variable is present and what is the proportion of different levels of cardinality. We can see that this a perfect balanced dataset. However this might not be the case in the real world often

```
In [4]: print(df['species'].value_counts())
        df['species'].value_counts()/len(df['species'])

        versicolor    50
        virginica     50
        setosa        50
        Name: species, dtype: int64
Out[4]: versicolor    0.333333
        virginica     0.333333
        setosa        0.333333
        Name: species, dtype: float64
```

Image: Univariate Non-graphical EDA on categorical variable

The below is implementation of univariate non-graphical EDA on numerical variable on iris dataset. This is done by df.describe() operation

Image: Univariate Non-graphical EDA on numerical variable

The following is code to implement outliers in the numerical variables:

Image: Code to detect outliers in numerical variables

After identification of outliers and rare categories in the variables, we will have a series of discussions with business analysts/ domain experts and thoroughly investigate the reasons for this outliers

Univariate Graphical EDA

The graphical EDA is mostly concerned with numerical variables to identify distribution of variables. The most common distributions that we encounter in Machine Learning field is Gaussian distribution. The other uncommon distributions are binomial, bernoulli, uniform, poisson, exponential distributions. However this cannot be case every time and variables need to be transformed by log, square-root, polynomial, box-cox transformations to make the variable follow normal distributions.

In addition to this we can use this graphs to identify Probability Distribution Function (PDF) and Cumulative Distribution Functions (CDF) for continuous variables, where as Probability Mass Functions (PMF) and Cumulative Distribution Functions (CDF) for discrete variables. These help us to identify the proportion of the variables that lie in a specific interval.

Image: Histograms for Numerical Variables

Image: CDF for numerical variables

Image: Box-Plots to detect outliers

Multivariate Non-Graphical EDA

The multivariate analysis is very useful in analyzing the association of features with the target variables. However, it is not possible visualize the multiple variables at a time so we will analyze two-variables at a time. This step is used to generate some insights of numerical variables for other categorical variables.

In this step we will detect the descriptive statistics for each level of cardinality in the target variable. This gives us a thorough idea on how the features vary for each level, and thus help us to generate some useful rules which are worthy in feature engineering and model selection.

Image: Non- Graphical Multivariate EDA

Multivariate Graphical EDA

The multivariate graphical EDA encompasses heatmap, jointplot, violinplots, pairplot, boxplot, countplot. We will be using seaborn library to implement these plots. The purpose of the heatmap is to generate correlations of the numerical variables. This gives us insights of the correlation between these variables and detect the strong variables which influence the target variable.

The boxplot, barplot are used to visualize the outliers, count of the variables with respect to a categorical variable (it can be explanatory or target variable). The pairplot is used to visualize scatter plots, histograms of all the numerical variables. The violin plots are used to visualize both outliers and distribution of the variables.

Image: Correlation Plot for numerical variables

Conclusion

In this article I have discussed about the tools used to perform EDA. In addition to the tools the most important step is to document all the insights and have a thorough understanding of data and discuss with domain experts any of the anomalies.

Please refer the code implementation of EDA on Iris dataset in the following link:

udaysai50/Full-Stack-Data-Science

Permalink Dismiss GitHub is home to over 40 million developers working together to host and review code, manage...

github.com

Eda Python3 Machine Learning Data Science Data Visualization

About Help Legal