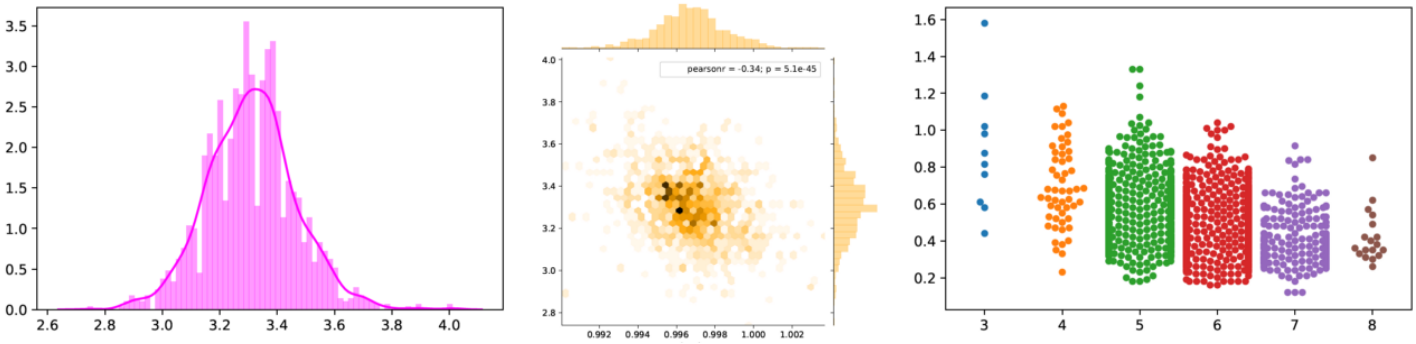


What is Exploratory Data Analysis?

Prasad Patil
Mar 23, 2018 · 6 min read

. . .

As I was contemplating what could be the maiden topic I should begin writing my blog with, in no time EDA popped up to my mind.Logically apt , isn't it ?! Why? You'll find out soon!

Intuition

Imagine your wolf pack decides to watch a movie you haven't heard of.There is absolutely no debate about that,it will lead to a state where you find yourself puzzled with lot of questions which needs to be answered in order to make a decision.Being a good chieftain the first question you would ask, what is the cast and crew of the movie? As a regular practice,you would also watch the trailer of the movie on YouTube.Furthermore,you'd find out ratings and reviews the movie has received from the audience.

Whatever investigating measures you would take before finally buying popcorn for your clan in theater,is nothing but what data scientists in their lingo call 'Exploratory Data Analysis'.

> Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it.

EDA explained using sample Data set:

To share my understanding of the concept and techniques I know,I'll take an example of white variant of Wine Quality data set which is available on UCI Machine Learning Repository and try to catch hold of as many insights from the data set using EDA.

To starts with,I imported necessary libraries (for this example pandas, numpy,matplotlib and seaborn) and loaded the data set.

Note : Whatever inferences I could extract,I've mentioned with bullet points.

```
In [2]: df = pd.read_csv('winequality-white.csv',sep=';')
        df.head()
```

Out[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

- Original data is separated by delimiter " ; " in given data set.
- To take a closer look at the data took help of " .head()"function of pandas library which returns first five observations of the data set.Similarly ".tail()" returns last five observations of the data set.

I found out the total number of rows and columns in the data set using ".shape".

```
In [3]: df.shape
Out[3]: (4898, 12)
```

- Dataset comprises of 4898 observations and 12 characteristics.
- Out of which one is dependent variable and rest 11 are independent variables — physico-chemical characteristics.

It is also a good practice to know the columns and their corresponding data types,along with finding whether they contain null values or not.

```
In [5]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

- Data has only float and integer values.
- No variable column has null/missing values.

The describe() function in pandas is very handy in getting various summary statistics.This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.

```
In [6]: df.describe()
```

Out[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

- Here as you can notice mean value is less than median value of each column which is represented by 50%(50th percentile) in index column.
- There is notably a large difference between 75th %tile and max values of predictors “residual sugar”,”free sulfur dioxide”,”total sulfur dioxide”.
- Thus observations 1 and 2 suggests that there are extreme values-Outliers in our data set.

Few key insights just by looking at dependent variable are as follows:

```
In [7]: df.quality.unique()
Out[7]: array([6, 5, 7, 8, 4, 3, 9], dtype=int64)
```

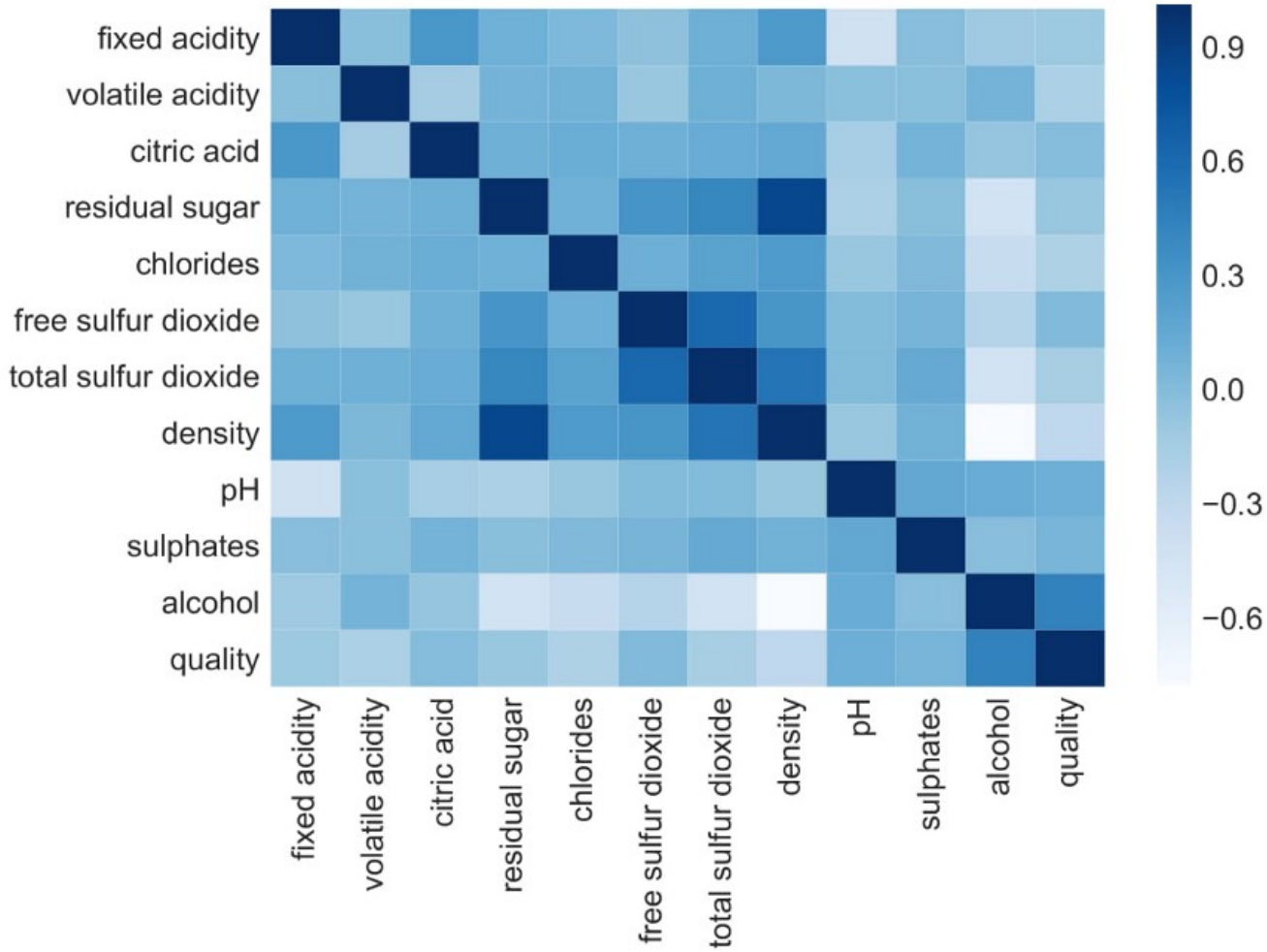
- Target variable/Dependent variable is discrete and categorical in nature.
- “quality” score scale ranges from 1 to 10;where 1 being poor and 10 being the best.
- 1,2 & 10 Quality ratings are not given by any observation. Only scores obtained are between 3 to 9.

```
In [8]: df.quality.value_counts()
Out[8]: 6    2198
        5    1457
        7     880
        8     175
        4     163
        3      20
        9       5
        Name: quality, dtype: int64
```

- This tells us vote count of each quality score in descending order.
- “quality” has most values concentrated in the categories 5, 6 and 7.
- Only a few observations made for the categories 3 & 9.

I got a good a glimpse of data. But that’s the thing with Data Science the more you get involved the harder it is for you to stop exploring.Let’s now explore data with beautiful graphs. Python has a visualization library ,Seaborn which build on top of matplotlib. It provides very attractive statistical graphs in order to perform both Univariate and Multivariate analysis.

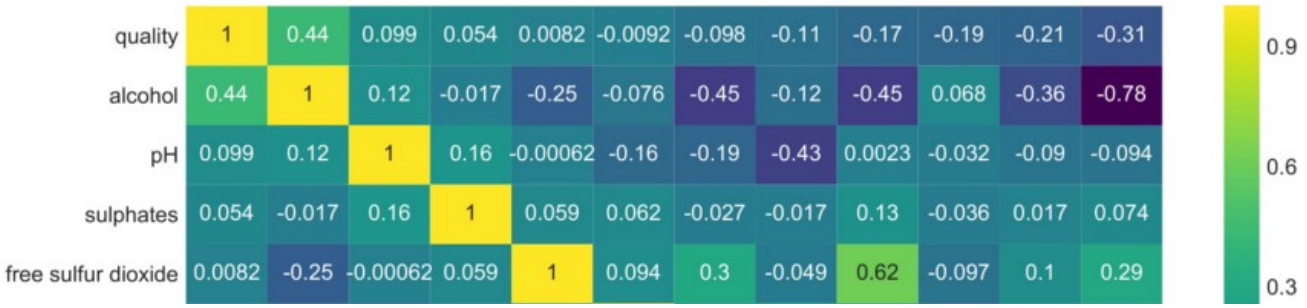
To use linear regression for modelling,its necessary to remove correlated variables to improve your model.One can find correlations using pandas ".corr()" function and can visualize the correlation matrix using a heatmap in seaborn.

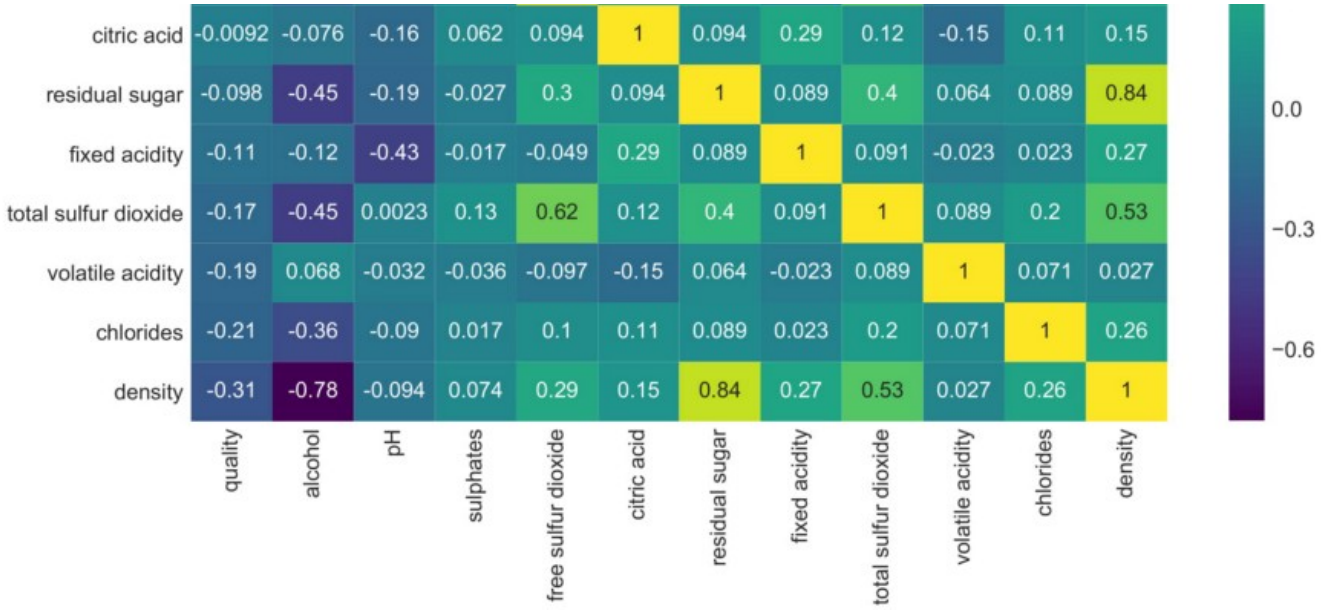


Heatmap

- Dark shades represents positive correlation while lighter shades represents negative correlation.
- If you set annot=True, you'll get values by which features are correlated to each other in grid-cells.

It's a good practice to remove correlated variables during feature selection.





Correlation Matrix

- Here we can infer that “density” has strong positive correlation with “residual sugar” whereas it has strong negative correlation with “alcohol”.
- “free sulphur dioxide” and “citric acid” has almost no correlation with “quality”.
- Since correlation is zero we can infer there is no linear relationship between these two predictors.However it is safe to drop these features in case you’re applying Linear Regression model to the dataset.

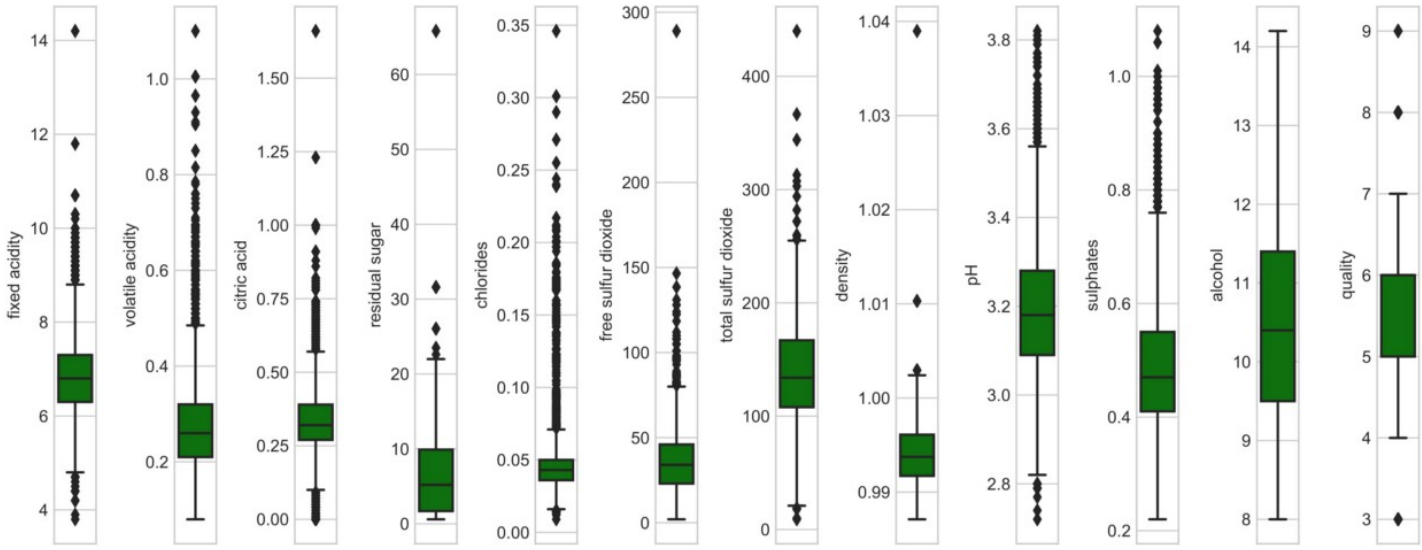
A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables.The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution.

The box plot (a.k.a. box and whisker diagram) is a standardized way of displaying the distribution of data based on the five number summary:

- Minimum
- First quartile
- Median
- Third quartile
- Maximum.

In the simplest box plot the central rectangle spans the first quartile to the third quartile (the interquartile range or IQR).

A segment inside the rectangle shows the median and “whiskers” above and below the box show the locations of the minimum and maximum.

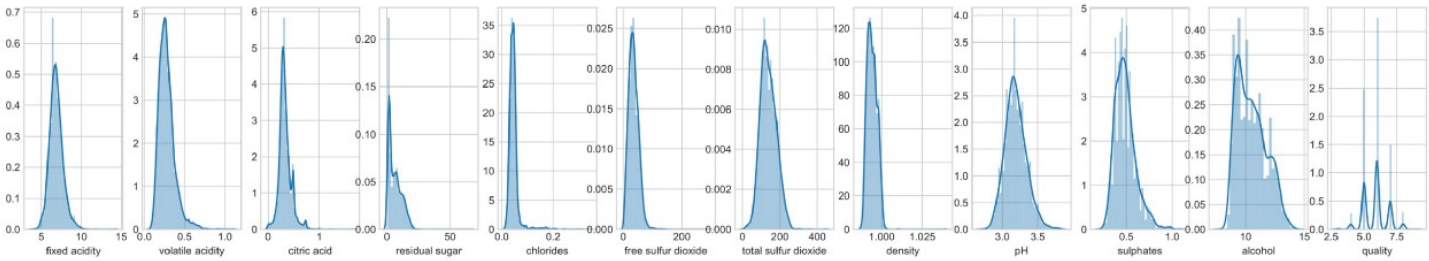


Boxplot

Outliers are either 3×IQR or more above the third quartile or 3×IQR or more below the first quartile.

- In our data set except “alcohol” all other features columns shows outliers.

Now to check the linearity of the variables it is a good practice to plot distribution graph and look for skewness of features. Kernel density estimate (kde) is a quite useful tool for plotting the shape of a distribution.



Distribution Plot

- “pH” column appears to be normally distributed
- remaining all independent variables are right skewed/positively skewed.

> *Lastly, to sum up all Exploratory Data Analysis is a philosophical and an artistical approach to guage every nuance from the data at early encounter.*

You can glance through my jupyter notebook **here** and try-test with different approaches, for eg. try out a pairplot and share what all inferences you could grab from it or if I failed to capture any useful information in my own approach,do share that too in comments.

Data Science Exploratory Data Analysis Machine Learning Data Visualization Pipeline