```
df=pd.read_csv('Advertising.csv')
```

```
print(df.head())
print(df.shape)
```

In the below data we note that our target or dependent variable y is sales.
x1=Unnamed, x2=TV, x3=radio, x4=newspaper

```
   Unnamed: 0     TV  radio  newspaper  sales
0           1  230.1   37.8       69.2   22.1
1           2   44.5   39.3       45.1   10.4
2           3   17.2   45.9       69.3    9.3
3           4  151.5   41.3       58.5   18.5
4           5  180.8   10.8       58.4   12.9
(200, 5)
```

```
profile=ProfileReport(df)
profile.to_file('Report.Html')
```

Creates an HTML file with the profile details.

Step1: Find the relation ship between feature variable and Target variable

Here in the overview report we can say following EDA
Tv is highly correlated with **sales**
Radio is highly correlated with **sales**
There is no mention of relationship between Newspaper column and unnamed column with sales. (we need to see what relation it has in interaction plot-if no relation may need to think of removing this feature column)

Overview

Overview | Alerts 13 | Reproduction

Alerts

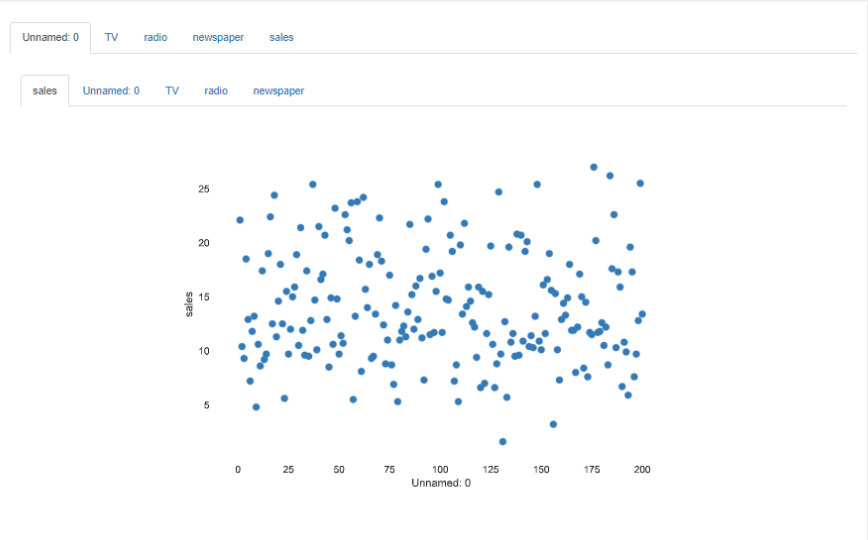
TV is highly correlated with sales	High correlation
radio is highly correlated with sales	High correlation
sales is highly correlated with TV and 1 other fields	High correlation
TV is highly correlated with sales	High correlation
radio is highly correlated with sales	High correlation
sales is highly correlated with TV and 1 other fields	High correlation
TV is highly correlated with sales	High correlation
sales is highly correlated with TV	High correlation
TV is highly correlated with sales	High correlation
radio is highly correlated with sales	High correlation
sales is highly correlated with TV and 1 other fields	High correlation
Unnamed: 0 is uniformly distributed	Uniform
Unnamed: 0 has unique values	Unique

Step2: Let us see interaction plot and identify the relationship between feature variables with target variable. And also see if there are any multicollinearity.

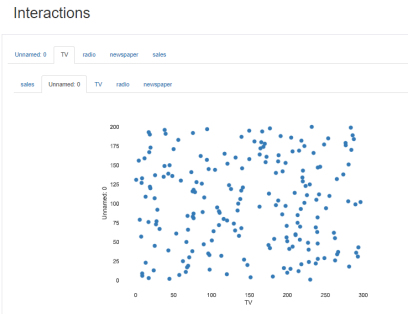
If there is a multicollinearity, then what happens is the model is going to start find the relationship between the independent x variables only and starts building biased model. What we want is the model should find the relationship between x and y variables not withing x variables. Hence we do not want multicollinear variables.

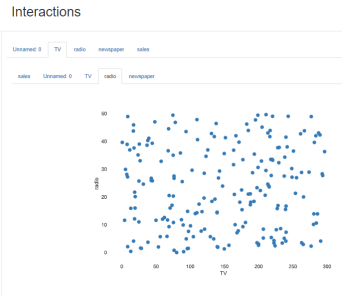
- ➔ Unnamed Vs Sales= No relation (SInce there is no relation ship with target variable as per my understanding this column can be removed entirely hence no need to check for multicollinearity for this column)

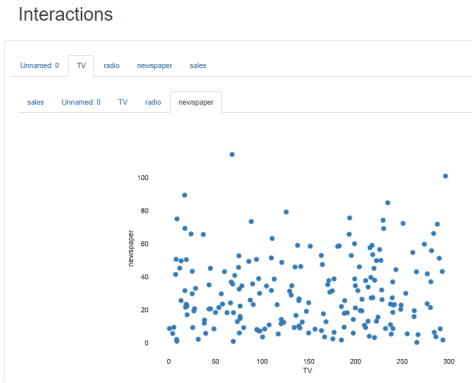
Interactions



- ➔ Tv Vs Sales- We can see there is a +ve linear relationship
- ➔ So check for if any multicollinearity.
- ➔ There is no multicollinearity exist with TV Vs any other x feature column, So keep this column for modelling

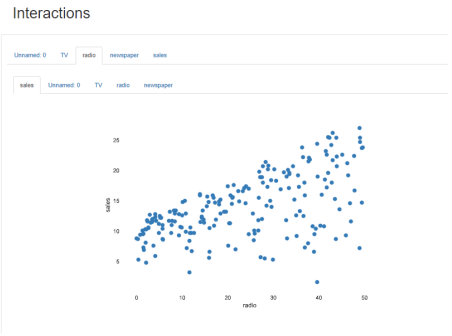


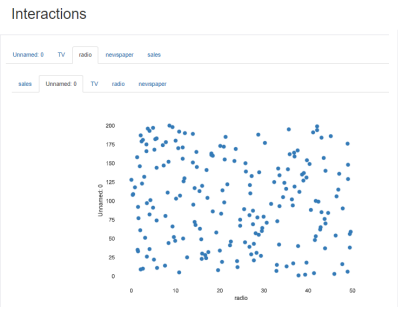


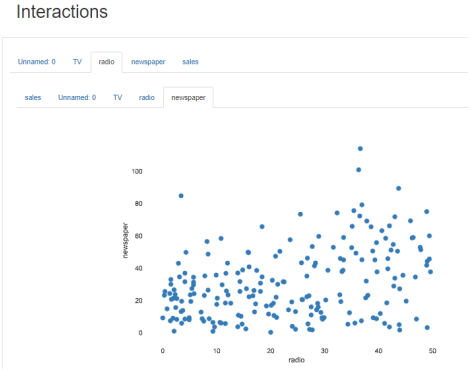


Radio Vs Sales

- ➔ We see there is a +ve linear relation ship between radio and target variable sales.
- ➔ So check for any multicollinearity
- ➔ There is no multicollinearity exist with radio Vs any other x feature column, So keep this column for modelling

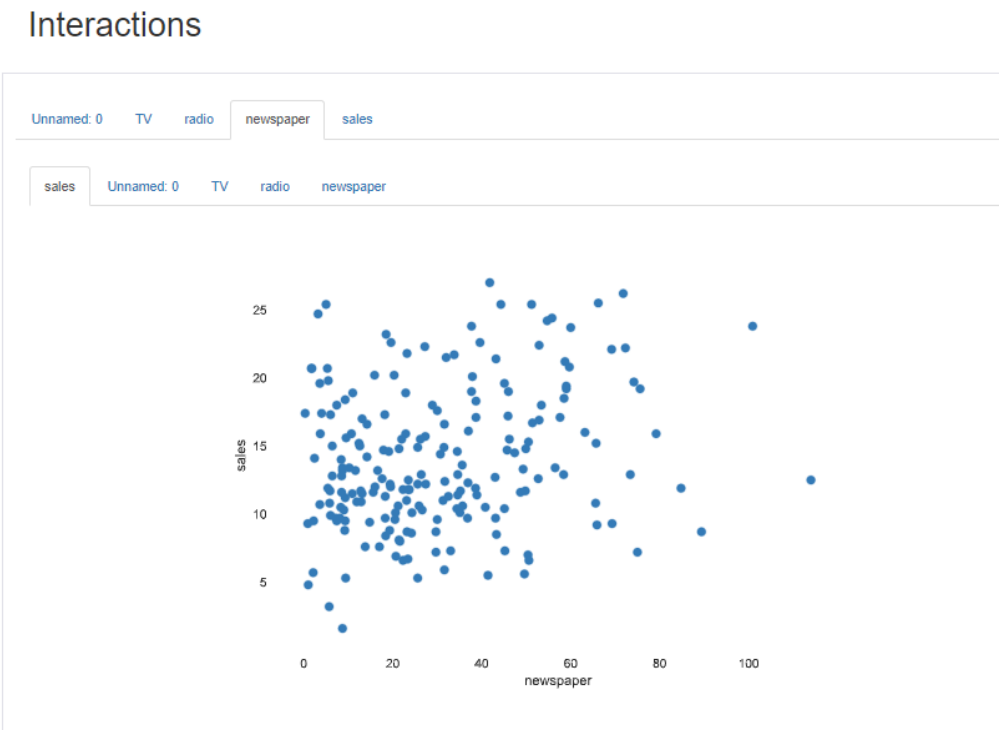






Newspaper Vs Sales

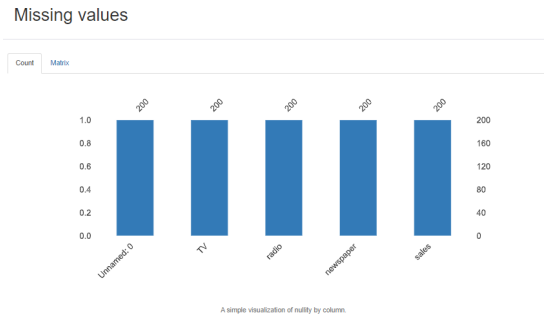
➔ Newspaper Vs Sales= No relation (SInce there is no relationship with target variable as per my understanding this column can be removed entirely hence no need to check for multicollinearity for this column)



Skewness is the numerical measure of the shape of the distribution. A skewness value greater than 1 or less than -1 indicates a highly skewed distribution. A value **between 0.5 and 1 or -0.5 and -1 is moderately skewed**. A value between -0.5 and 0.5 indicates that the distribution is fairly symmetrical.

Kurtosis is a another measure of symmetricity or shape Kurtosis is all about the tails of the distribution not the peaknen or flatten. It measures the tail-heaviness of the distribution.

Step3- Find out if there are any missing values
Here there are no missing values. Total number of values in each x variable is 200 only



We can save the html report inside the html file and render it for dashboarding purposes