## TW1 Exploratory Data Analysis

**Part 2**:

Submit a summary of your learning to Canvas. Your document should include:

Full names of your team members who work on the assignment.

Xiaomei Xie

Lili Hao

URL links to the notebook of each student on GitHub repo.

[xiaomeiX/TW1-visualization (github.com)](https://github.com/xiaomeiX/TW1-visualization)

A summary of what you learned from the teamwork assignment.

We first collect the data and observe the dataset.

What are the features?

* **TV:** advertising dollars spent on TV for a single product in a given market (in thousands of dollars)
* **Radio:** advertising dollars spent on Radio
* **Newspaper:** advertising dollars spent on Newspaper

What is the response?

* **Sales:** sales of a single product in a given market (in thousands of items)

What else do we know?

* Because the response variable is continuous, this is a **regression** problem.
* There are 200 **observations** (represented by the rows), and each observation is a single market.

Data have 200 rows × 4 columns, the 4 columns are: TV, Radio, Newspaper, Sales

The means of each columns are:

TV 147.0425

Radio 23.2640

Newspaper 30.5540

Sales 14.0225

The median of each columns are:

TV 149.75

Radio 22.90

Newspaper 25.75

Sales 12.90

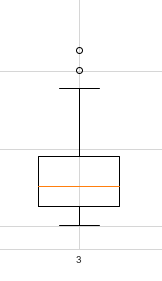
We can use data.describe() to see data’s mean, std, min, q1, median, q3, max.

We can import matplotlib.pyplot to plot the boxplot. Through boxplot, The spacings between the different parts of the box indicate the degree of dispersion (spread) such as min, q1, mean, q3, and max. From boxplot we can see skewness in the data and show outliers. In the newspaper column has outlier.

We can use data.corr() to find each pair of Pearson's correlation coefficient.

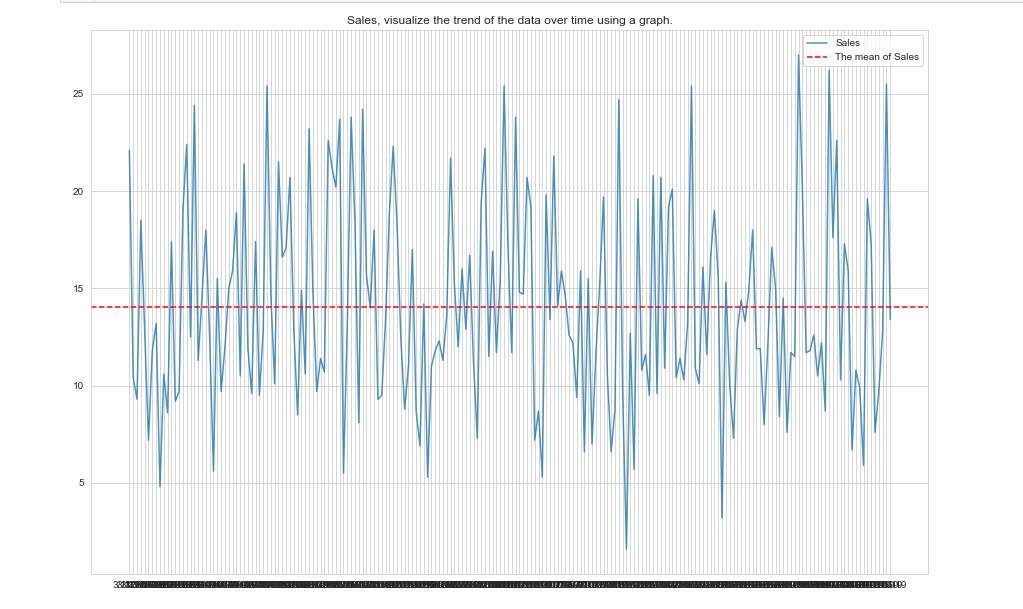
TV-Sales = 0.782224, Radio-Sales = 0.576223, Newspaper-Sales = 0.228299….

We can get conclusion the TV vs Sales have strong correlation coefficient.

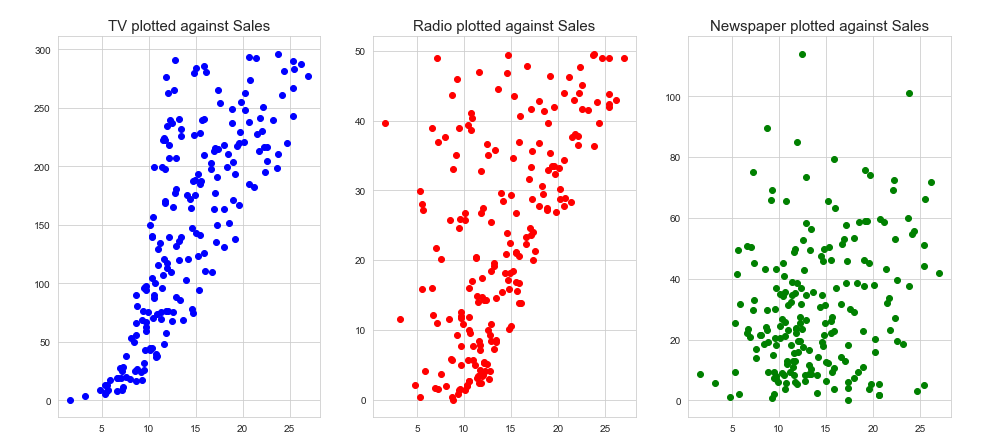
Visualize distribution of the data for sales using histogram, which is Unimodal and is normal distribution, the pattern is the bell-shaped curve known as the "normal distribution."

Boxplot, show Newpaper has two outlier.

'Sales, visualize the trend of the data over time using a graph' tell us, the sales of a single product in a given market are changed day by day, a stationary time series is one with statistical properties such as mean, where variances are all constant over time. However the mean of sales is constant, which is 14.0225.



A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal as Sales and vertical axis indicates values for an individual data point such as TV, Radio, Newspaper. Scatter plots are used to observe relationships between variables such as correlation.



From ‘TV plotted against Sales’, we can tell TV-Salves has strong correlation and positive relationship. However The newspaper plotted again Sales show, those variables has week strong relationship.