## 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

Thomas Bakken

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

<https://github.com/xiaomeiX/TW1-visualization>

<https://github.com/lhaoSeattleu/TW1-visualization>

<https://github.com/thomasbakken-su/CPSC5310_TW1>

A summary of what you learned from the teamwork assignment:

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 could **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:

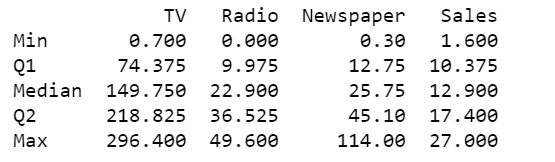
TV 149.75

Radio 22.90

Newspaper 25.75

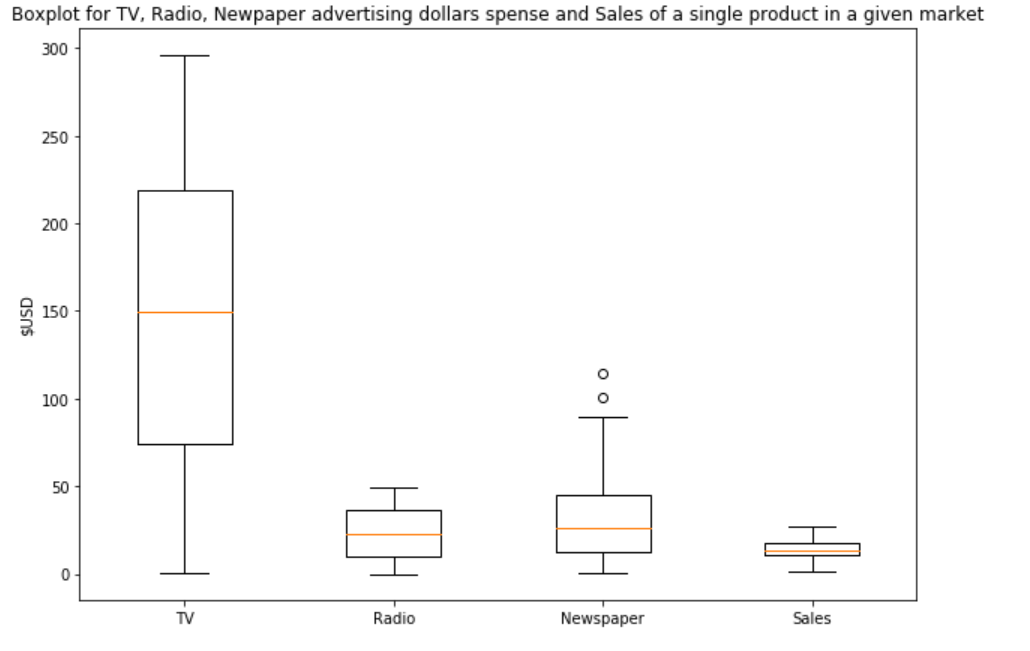
Sales 12.90

Data’s mean, std, min, q1, median, q3, and max are:

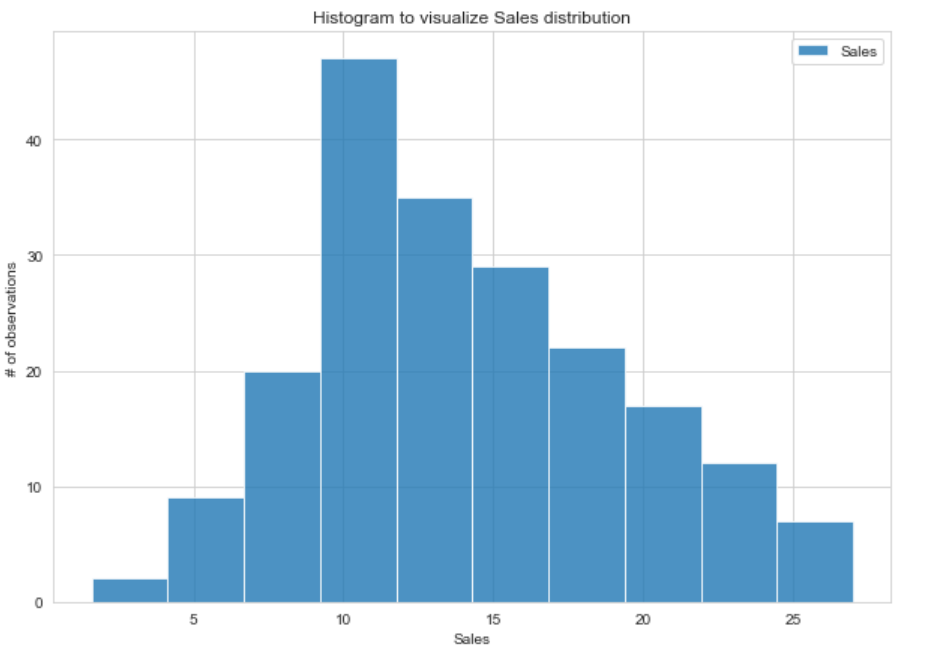


The Boxplot for TV, Radio, and Newpaper advertising dollars spent and sales of a single product in a given market shows:

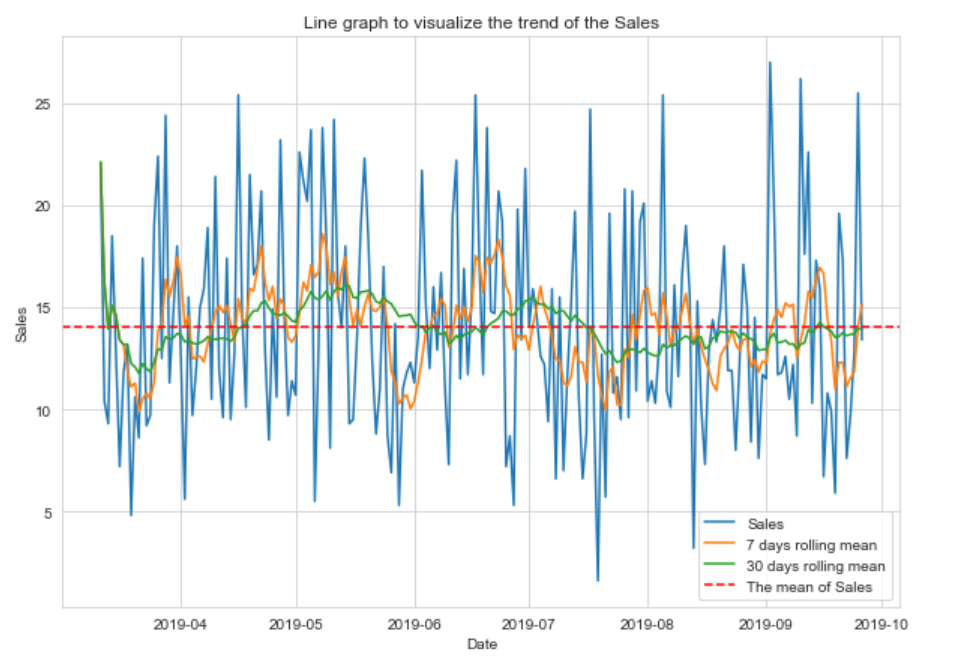
* The median, .25 percentile, .75 percentile, min and max match what we calculated before.
* Newpaper has two outliers which are larger than what we expected for 99% of data in gaussian or normal distribution



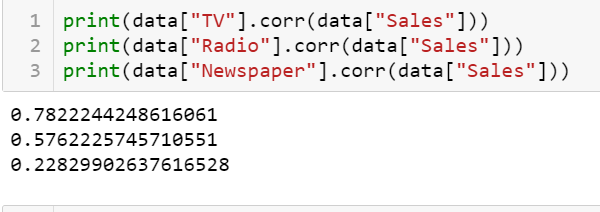
The Sales distribution histogram shows a normal distribution, and it is unimodal. The most frequent sales of a single product in a given market (in thousands of items) is around 10



The sales trend visualization shows a seasonality change throughout the year based on the 30 days rolling mean: the sale is relatively higher in the month of May, July, and September. The graph also fluctuates less as time goes on. The mean of sales is constant, which is 14.0225.



For each pair of Pearson's correlation coefficient, the TV-Sales = 0.7822244248616061, Radio-Sales = 0.5762225745710551, Newpaper-Sales = 0.22829902637616528. So, advertising dollars spent on TV have the strongest correlation with sales, and advertising dollars spent on newspapers has weakest correlation with Sales among these comparisons.



A scatter plot visualizes the correlation between TV-Sales, Radio-Sales and Newpaper-Sales. Scatter plots are used to observe relationships between variables such as correlation. It matches the Pearson's correlation coefficient result above.

