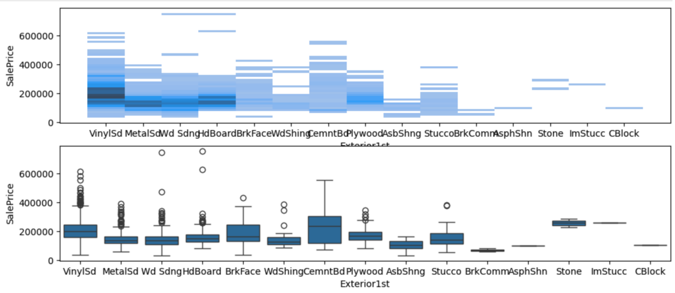
Visualization Plots Strength and Weakness

MEHDI VALINEJAD, *Student*



Box Cons 1.

Fig. 1. Box Cons 1

***Abstract*—Current report states the strength and weakness aspects of box, histogram, scatter and violin plots. the complete source code is available at https://github.com/TxCorpi0x/data_mining under asgn1 directory.**

***Index Terms*—box plot, histogram plot, scatter plot, violin plot, weakness, strength.**

I. Introduction

VISUALIZATION plots are means to visualize data to deduct analytical and statistical conclusions for dat mining and data science. There are different types of plots used for comparison such as Bar, Lollipop, Bullet, Dot, Range, Radial, Parallel, Radar, Waterfall and etx. or plots used for correlation analysis such as Heatmap, Bubble, Scatter, Hexagonal and etc. or Pie, Donut, hierarchy and etc, for hierarchical analysis.

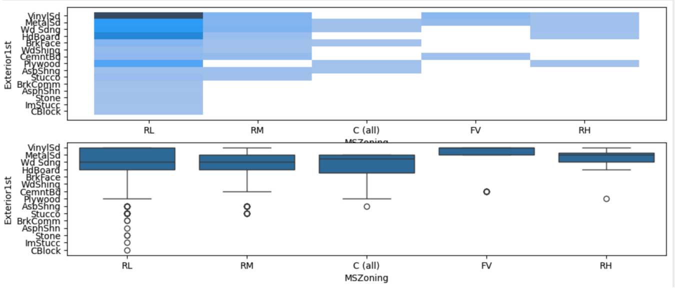
In the current report, Box, Scatter, Histogram and Violin plots are examined and implemented by Python to illustrate advantages and disadvantages of them in a visual and interactive environment with jupyter notebook.

II. Box

This mostly used to illustrate the distribution of data, it uses statistical information such as median and quartile to show the density and propagation of data in a perfect visualization.

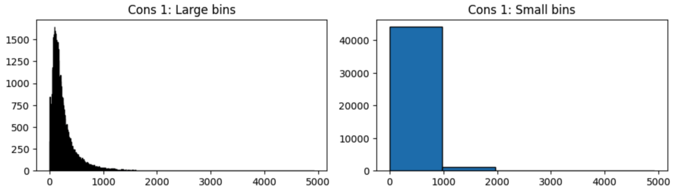
1.Strength

- Large datasets can be analyzed with box plots. median, upper quartile, lower quartile, min and the max of values can be used for a wide range of analysis and calculations.
- Distribution of data can be inspired quickly and simply.
- Comparing different types of data can be used get with distribution analysis between multiple box plots.
- Outliers can be detected simply by viewing the data outside of the min and max boundaries.



Box Cons 2.

Fig. 2. Box Cons 2



Histogram Cons 1.

Fig. 3. Histogram Cons 1

1.Weakness

- We can not extract exact values and detailed information from box plots. Box plots are not self explanatory by themselves so they need to be paired with other types of plots such as histograms. according to the following plot, the box plot does not show accurate data to be extracted as values. Fig(1)
- In the histogram, we can detect that the density of data is high for ”VinylSd” between 160000 and 200000, when we look at box plt, the data is shown between 180000 and 220000 and it is because of non gradient view of the box plot. Fig(1)
- No exact value is shown in the box plots and most of the visible data is related to statistics measures such as median. Fig(2)

III. Histogram

Histogram is useful if we want to show the distribution of a continuous data.

1.Strength

- Is acts outstanding in large datasets, this is by using the grouping (bins) as intervals.
- We can extract patterns and trends in data, symmetry, skewness and peaks can be detected easily.
- Outliers are shown according to the isolated bars showing in a histogram chart.

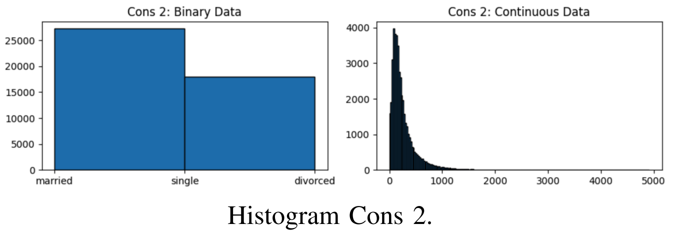


Fig. 4. Histogram Cons 2

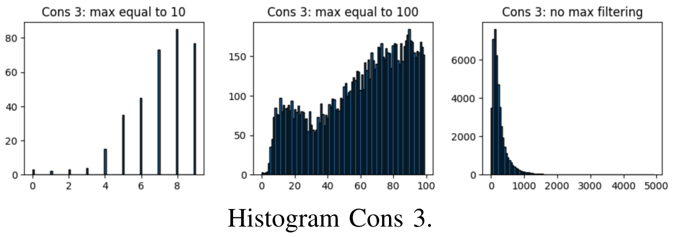


Fig. 5. Histogram Cons 3

- Quickly get insightful understanding of data.

1.Weakness

- Histogram is highly dependent to number of bins, in Cons 1 chart Fig(3), with small and large bins, for the bank load duration with a low bin (5) it is not simple to understand max occurrence of a same duration value extraction. in the right side we can not determine if the value is 0 or 1000.
- Using histogram for binary data or any data that is not continuous, will not get a good result to distinguish meaningful data. the data is grouped into ranges or intervals, the original data is lost and can price an exact value Fig(4).
- In the cons 3, changing the maximum included data makes the chart to be misleading, it is visible in the left and right chart that the distribution can not be detected correctly Fig(5).
- We can not gather useful information when we compare two different datasets to compare features using histogram, instead we can use bar charts.
- Detection of continuos and discrete variables is not possible.
- Detection of distribution type is possible but it is so difficult.
- All af the data should be in memory, so it is a high-end hardware for large dataset.

IV. Scatter

Histogram is a mathematical graph that uses Cartesian coordinates to show two variables within a data.

1.Strength

- Shows trends and relationships between two variables, to detect patterns and correlations in data we can use scattered plots.

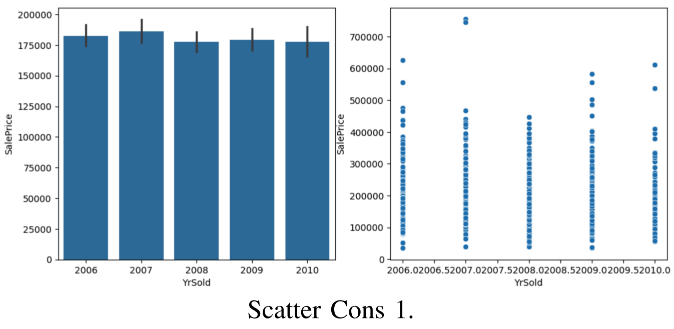


Fig. 6. Scatter Cons 1

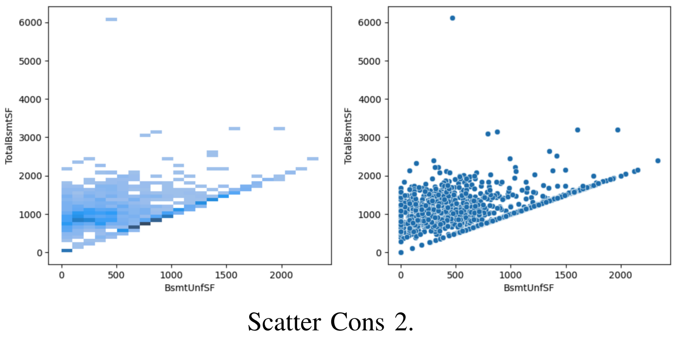
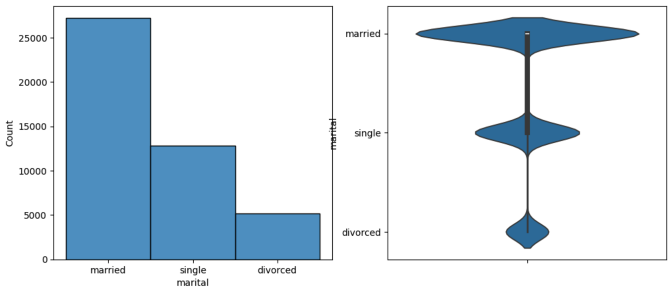


Fig. 7. Scatter Cons 2

- Is shows complete data points, illustrates distribution showing the maximum and the minimum of the values.
- Correlation detection including positive and negative correlation between variables.
- Accurate analysis can be done by scattered plots, this is because of high granularity of the data points.
- Outliers can be detected simply with a glance and we can distinguish between harsh outsider data to be assumed as outlier.

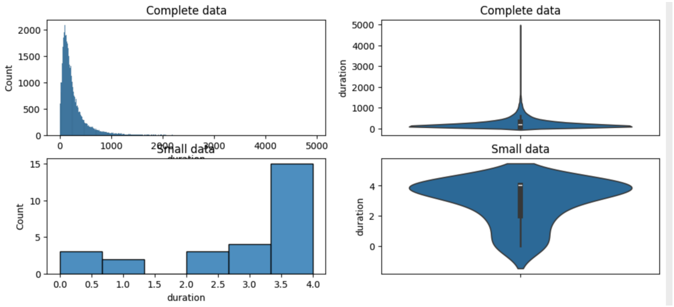
1.Weakness

- Scatter plot translates data for a subjective analysis, so different people may have different perceptions from the plot and data. In the following example, someone may assume values higher than 450000 as outliers and someone assumes 500000 as outlier. data is shown accurately but the detecting the density of each area of data is perceiving Fig(6).
- Scatter plot is meant to be used with continuous data. as it is visible we can not inspire any correlation between ”year built” and ”year sold”, also it should not be a relation but because the data is not continuous, we can not get a good estimation of these two variable with scatter plot Fig(6).
- Scatter plot shows correlation but it necessarily does not mean that there is a cause and effect relationship between two elements. scatter plot shows a relationship between ”total basement surface” versus ”uniform basement surface” but logically there is no cause and effect relationship between these two variable Fig(7).



Scatter Cons 1.

Fig. 8. Scatter Cons 1



Scatter Cons 2.

Fig. 9. Scatter Cons 2

V. Violin

Is a statistical graph to be used for comparison between probability distributions. compared to box plot, it shows more details related to the distribution.

1.Strength

- Compared to box plots, it shows more comprehensive view of the data's distribution, the peaks, valleys and tails makes viewer to detect the similarity between groups.
- To detect unusual clusters of data points by validating the shape and spread of curves, distinct groups can be detected simply.

1.Weakness

- Violin plot does not show a good result when there are differences between groups symmetry and skewness and shape. It is not easy to extract accurate information from the plot. For instance in the right plot, we can not detect the number of married, single or divorced. so we need to use violin plots in a combination with other plot such as bar plots or box plots Fig(8).
- Where there is not enough points to be fed into the violin plots, the accuracy of the violin plot drops for the small categories. In the Right side with small data, the plot is so misleading and we can not inspire any meaningful conclusion of it Fig(9).

Mehdi Valinejad Received the B.S. degree in industrial engineering from the Azad University (South Tehran Branch) in 2012, and is currently working Master's. degree at the University of Bahcesehir at Istanbul.