Hackathon

Divya, Nupur, Ranadheer, Varun

Data Description:

The dataset consists of sales data of laptops across various stores for the month of January in the year 2008. The data is structured with 17 parameters (columns) and 9757 entries(rows). Laptop models with various specifications such as screen size, RAM, battery life, processor speeds, HD storage (hard disk storage capacity) along with their retail prices at 16 different stores have been presented. The description of different variants of the laptops are as follows:

1. *Configuration:* It is describing the configuration of the laptops purchased throughout the month of January.
2. *Customer Postcode:* It is the ZIP code of customer where he/she resides.
3. *Store Postcode:* It is the ZIP code of store where the laptop is available.
4. *Retail Price:* It is the price of the particular laptop sold at.
5. *Screen Size (Inches):* It is the dimension of screen size of laptops available at stores which is 15 inches only.
6. *Battery Life (Hours):* 3 different batteries are equipped to these models whose battery life is 4, 5 & 6 hours respectively depending on the model.
7. *RAM (GB):* Laptops are observed to have 2 RAM options i.e., 1GB & 2GB respectively.
8. *Processor Speed (GHzs):* There are 2 different processor speeds i.e., 1.5GHZ & 2GHZ.
9. *Integrated Wireless:* Laptops are categorized on the basis of functionality of being Wireless Yes/No.
10. *HD Size (GB):* Laptops are categorized on the basis of space of Hard Disk offered that is 40, 80, 120 and 300 GBs respectively.
11. *Bundled Applications:* Laptops are categorized on the basis of offered collective applications or not.
12. *Customer Store Distance:* It is a mathematical figure representing the distance between store and customer location.

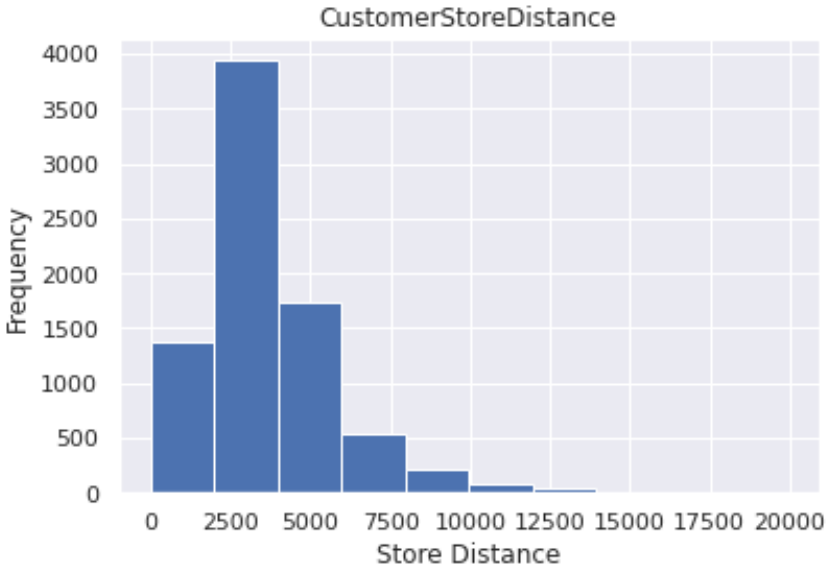
Interestingly enough, Though there is only 1 screen size irrespective of the variant being considered, there is a column dedicated to mention the screen size for every such model

Data Cleaning:

Date has been extracted from the date and placed into a column of its own in order to assess the changes in the pattern on a monthly basis. Column name ‘DATE’ has been capitalized for improved consistency across the data set. The position of the decimal point in the columns CustomerStoreDistance has been shifted by a few places for a better ease of understanding. Datasets like Wireless_y, Wireless_n, Bundle_n, Bundle_y etc. has been used for better understanding and improved visualization.

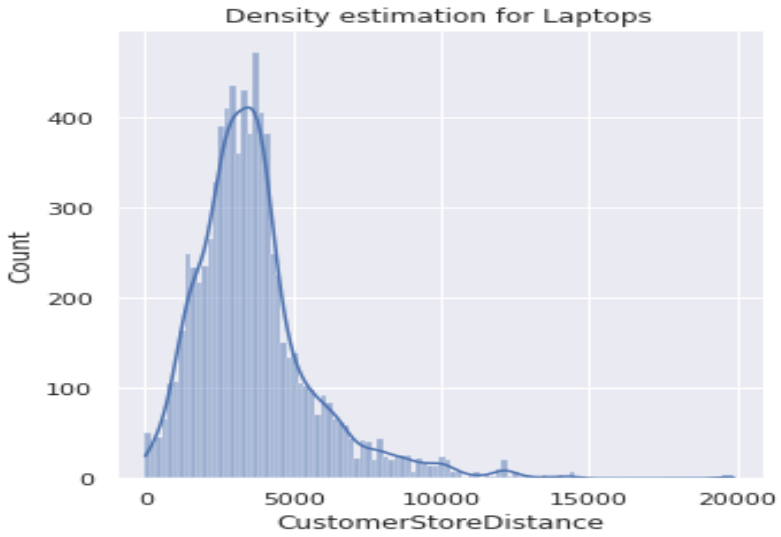
Understanding the Visualizations:

1. **Histogram-**



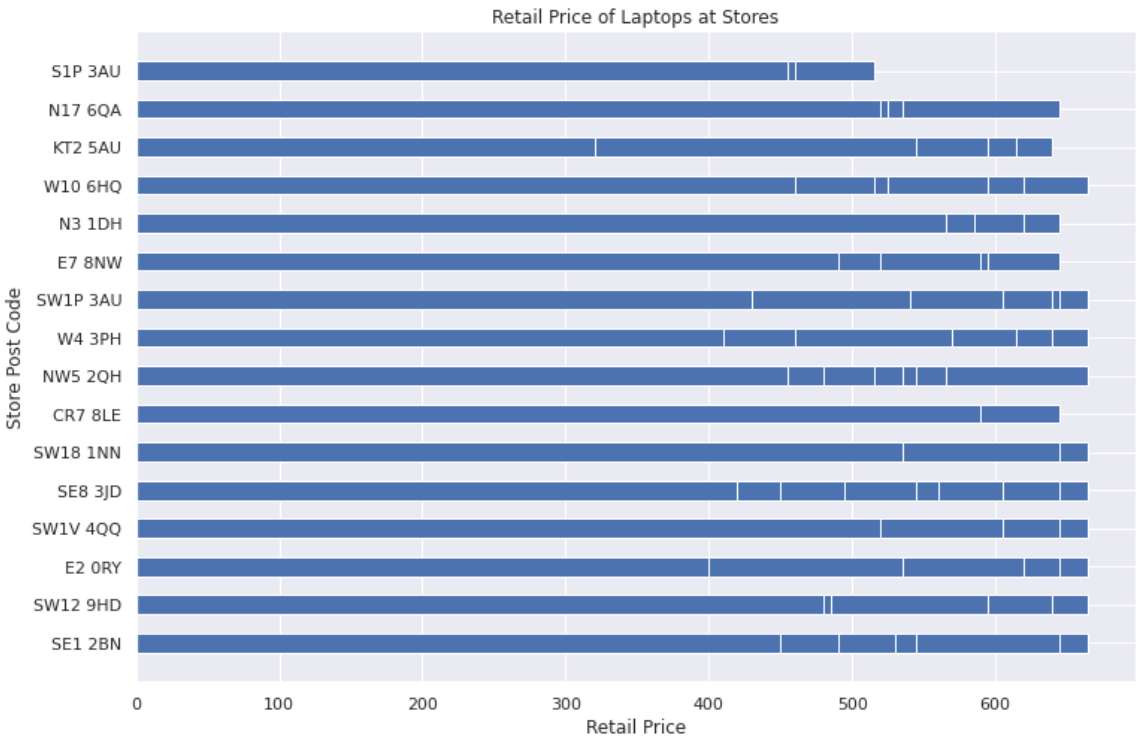
The above-mentioned Histogram is plotted on Customer Store Location, where x-axis is showing the distance of customer from stores and y-axis is having the frequency of the customers. The interesting insight from plot is almost half of the customers from the dataset fall in the distance range of 2000-3800 from store.

2. **Distribution Plot(displot)-**



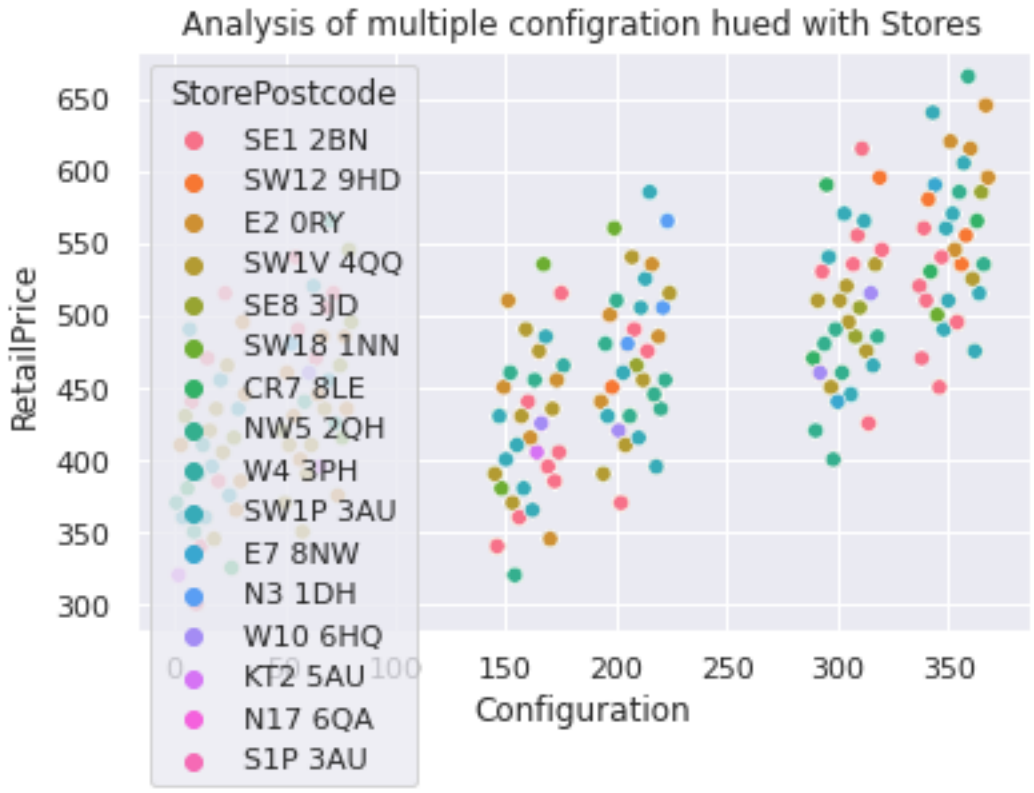
The above-mentioned plot is Distribution plot, taking Customer-Store Distance as base parameter with x- axis describing the distance between customer and the store and y-axis is counting the number of customers falling under that specified radar of distance. We have set kernel density estimation ‘True’ here.

3. Horizontal Bar Plot-



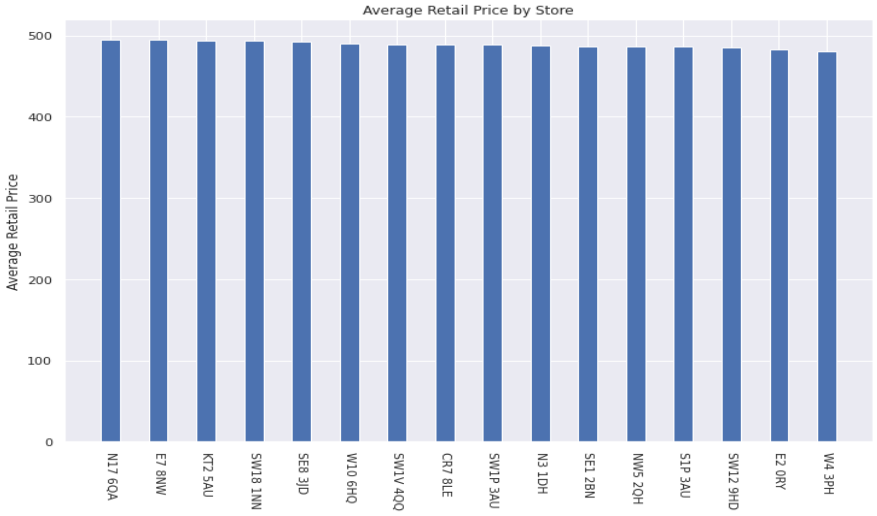
The above-mentioned horizontal bar chart is plotted for Store Postcode against Retail Price, in which y axis is covering labels for store postcodes while x- axis is involved in representing the retail price. While the highest retail price is observed in W10 6HQ postcode, the lowest price is observed in S1P.

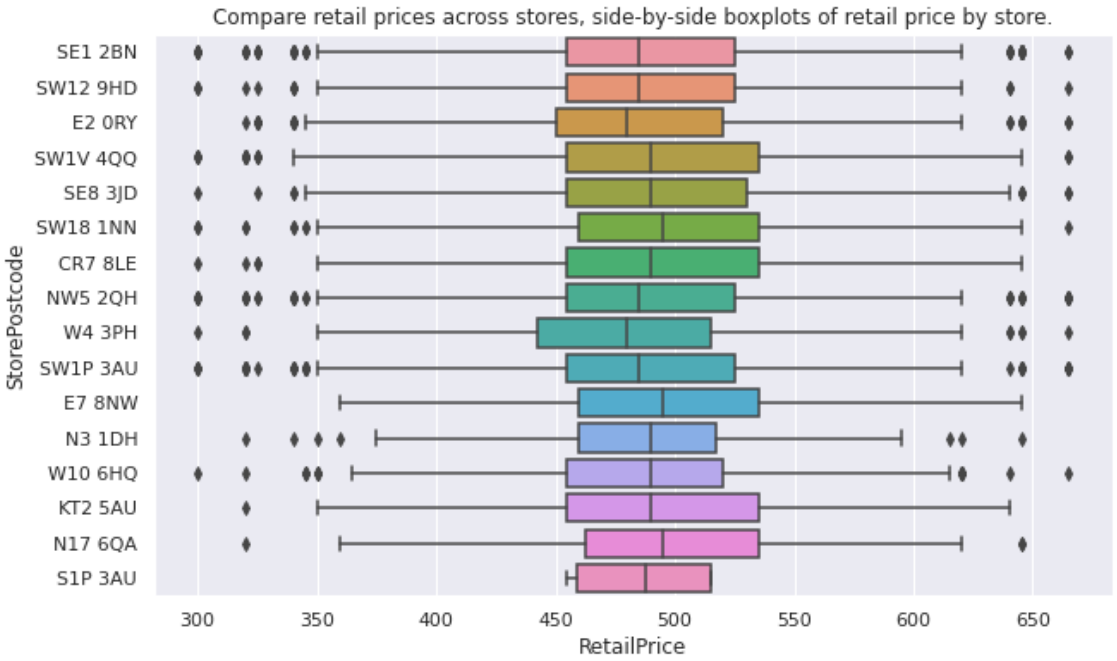
4. Scatter Plot-



This Scatter plot is describing the dataset on the regards of configuration at x-axis against the retail price over y- axis, having different color codes for each unique store's postcode under the hue argument of the scatter plot.

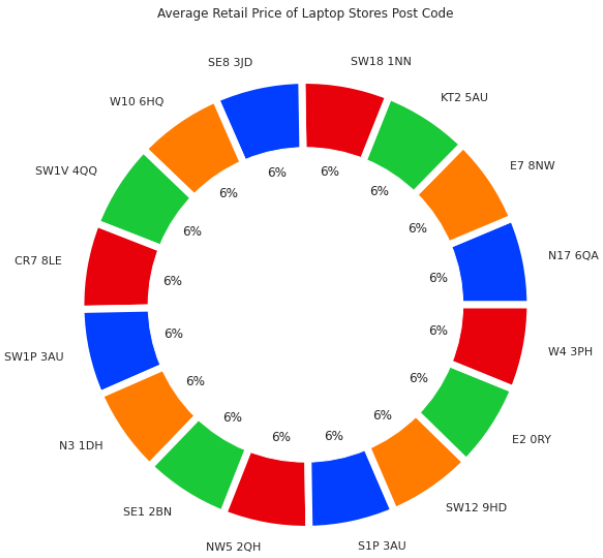
5. Barchart and Boxplot-





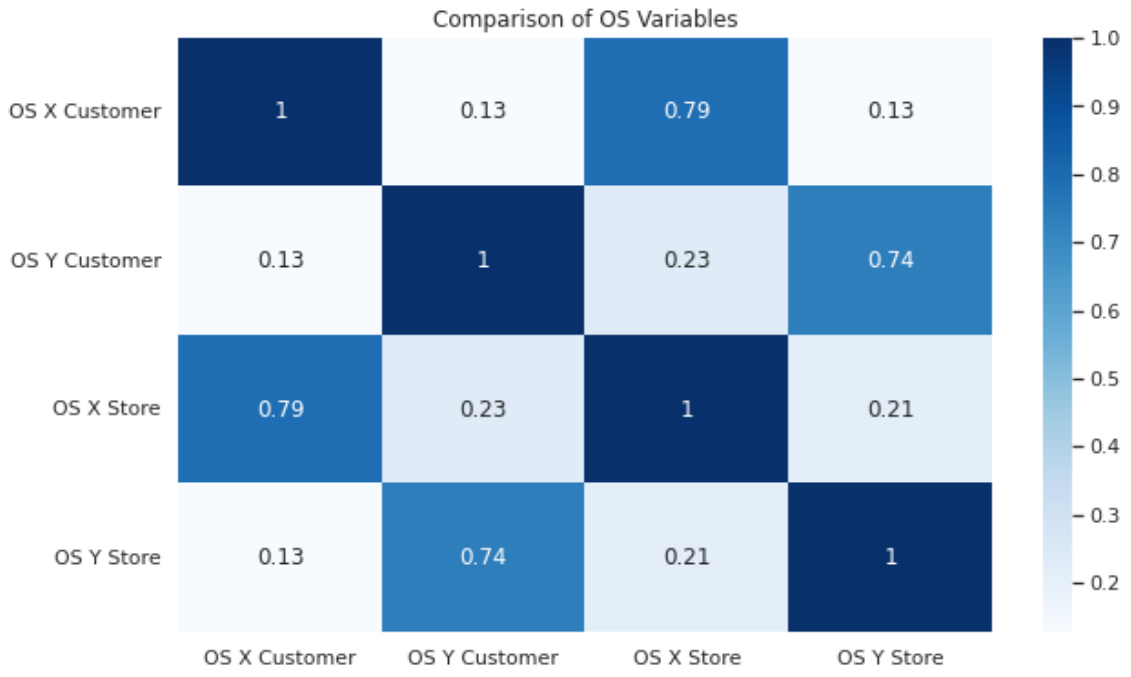
The above-mentioned bar chart is used for visualizing the Store's postcode on x-axis with the average retail price of the laptop. On a broad vision, store N17 6Q A has the highest average and store W4 3PH has the lowest average. Similarly, the boxplot of Store's postcode on the ordinate and the retail price of laptops respectively on the abscissa of the plot. Significant outliers are scrutinized in SE1 2BN, NW5 2HQ and SW1P 3AU respectively while S1P 3AU is free of outliers under the boxplot, Interquartile range is not varying much among stores and median values are also close enough.

6. Donut Chart-



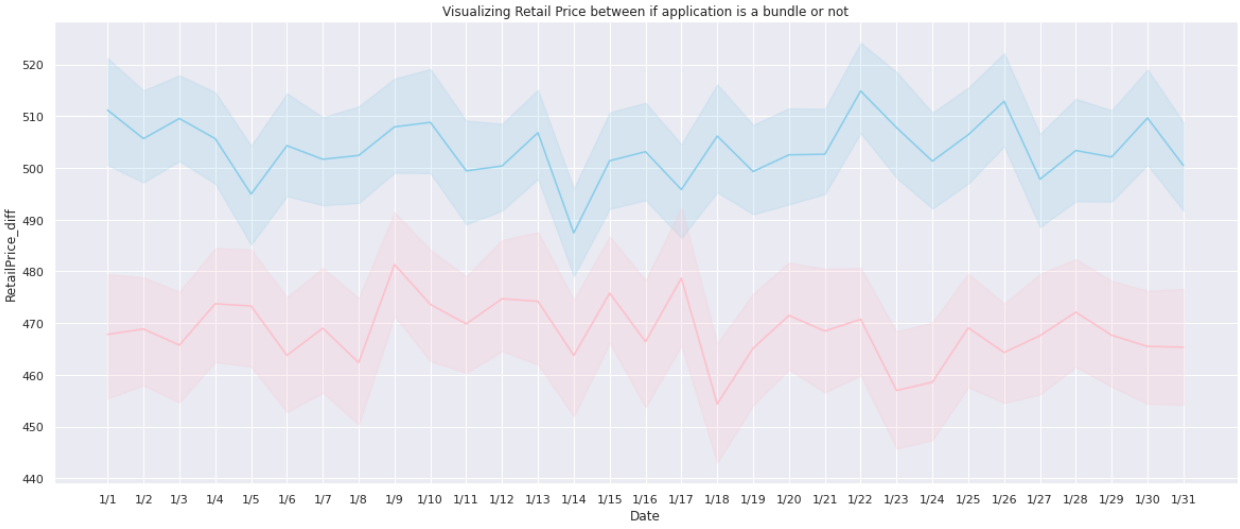
The Donut chart here is representing the percentile distribution of laptops falling under the category of particular store postcode on the ground of average retail price which is calculated by grouping the postcodes respectively.

7. Heatmap-



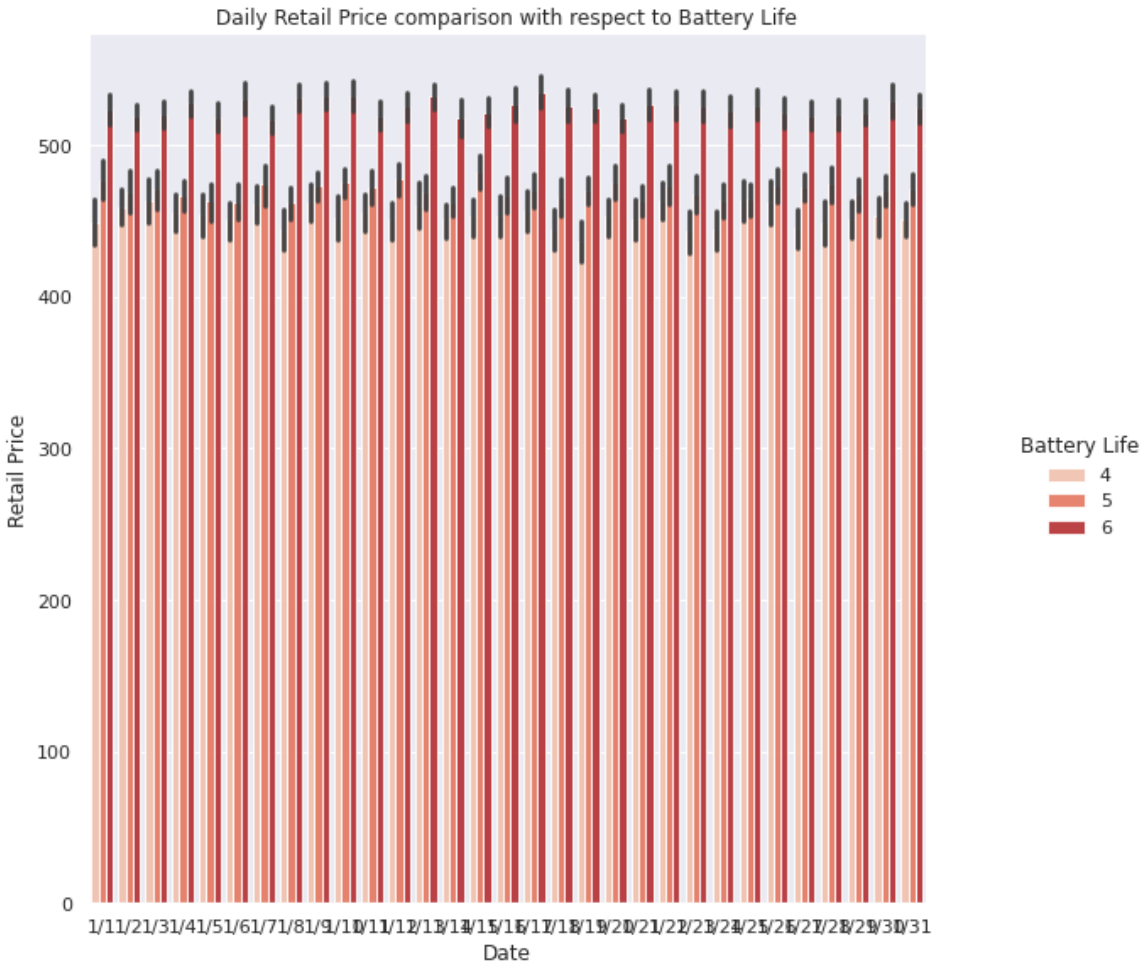
So basically, OS customer X and Y respectively and OS Store X and Y are perfectly serving as the coordinates of start and endpoints of the customer and store locations. The best they can be visualized is using the Heatmap on the pointer scale of 1.0 having divisions of 0.1 pointers.

8. Normalizing with periodical change in Line Plot-



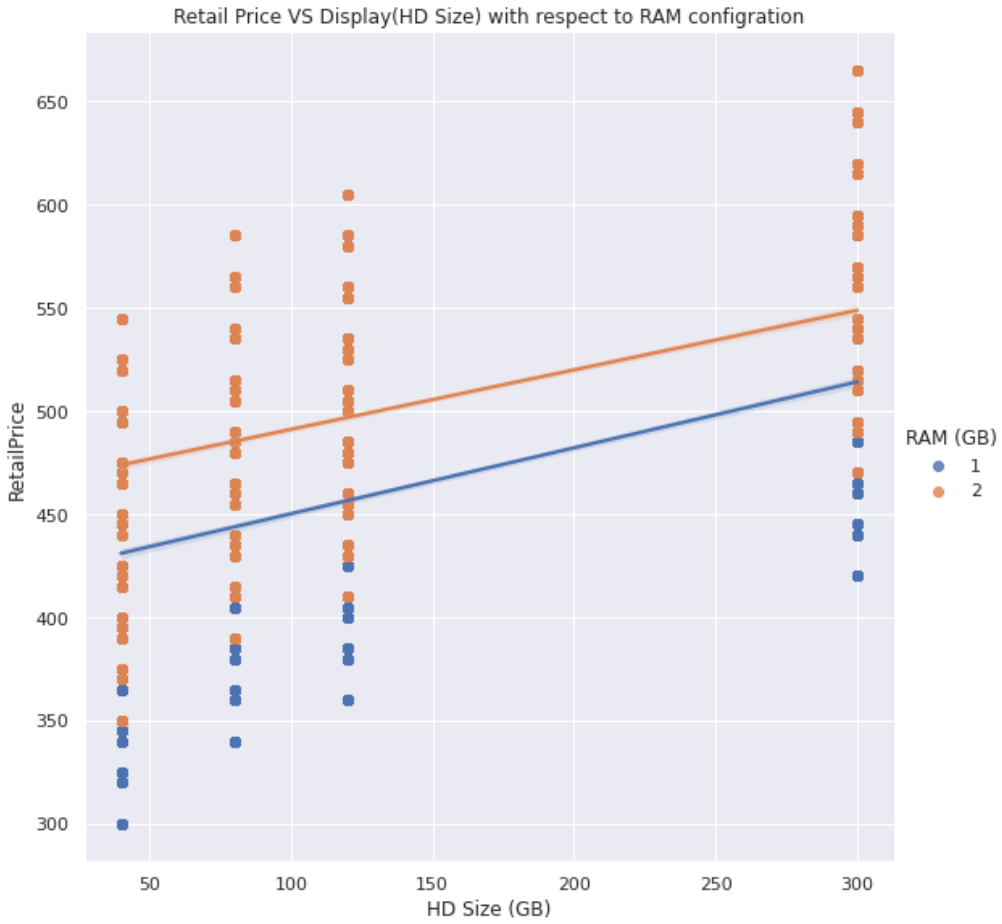
So, applying normalizing on the ground of percentage change on Retail Price of Laptops taking both bundled application with skyblue color and non-bundled application with red color lines are visualized. Since Bundled Application offer better functionality, the retail price of the one having it was always higher than non- bundled one throughout the month of January.

9. **Catplot-**



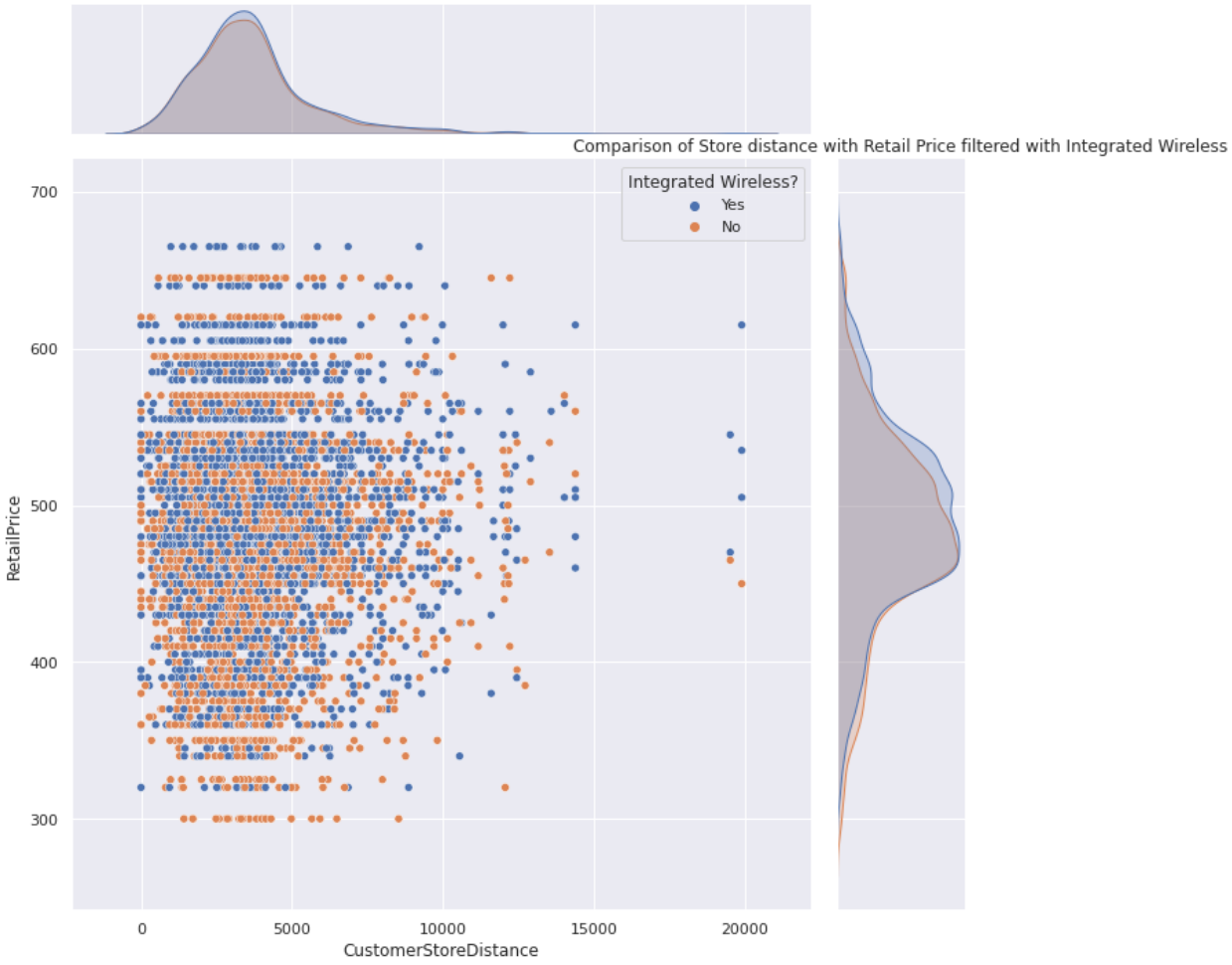
The above-mentioned chart is catplot, visualizing the Retail Price of laptops across the Sales dataset, based on supportive battery hours compatibility. The Price is observed against the date on which the laptop is sold. The but obvious insight is the inclination of customer towards the longer battery life laptops that is 6 hours category.

10. **Lineplot (FacetGrid)-**



The above-mentioned chart is simultaneously visualizing the configuration of Hard Disk Space on x-axis with the retail price of laptops on y-axis including the RAM variant available to the customer respectively. The direct relation that can be drawn from the chart is that the retail price of laptops is directly proportional to the better RAM and Hard Disk availability.

11. **Jointplot-**



The above-mentioned joint plot is the distribution of laptops sold in January on the grounds of retail price against the customer store distance with offering integrated wireless functionality.

12. Stacked Area Plot-

The above mentioned two plots are visualization of Retail Price categorized on the ground of being wired and wireless integrations with configurations of the respective laptops across the dataset. Clearly spikes in retail price is observed when the configuration gets better with the laptops.

Conclusion-

The laptop sales data set was used to compare the sales attributes of various models of laptops, across 16 different stores, with different sets of technical features. Various plots have been laid out systematically to analyze the varying trends in the sales based on the varying parameters. Major trends were captured using the graphical tools like histogram, scatter plot, donut chart and heat map to name a few. An important inference from various plots is that laptops of 3 particular RAM, processor & battery life specifications stood out the most to the customers as they were sold in highest numbers.