**Assignment -1 (Group 7)**

**2.1**  
a. Supervised Learning  
b. Supervise Learning  
c. Supervised\* Learning  
d. Unsupervised Learning  
e. Supervised Learning  
f. Supervised Learning  
g. Unsupervised Learning  
h. Unsupervised Learning

**2.2** *Describe the difference in roles assumed by the validation partition and the test partition.*

The *training data* are the data from which the classification or prediction algorithm “learns,” or is “trained,” about the relationship between predictor variables and the outcome variable.

Once the algorithm has learned from the training data, it is then applied to another sample of labeled data (the *validation data*) where the outcome is known but initially hidden, to see how well it does in comparison to other models. Validation Partition is the set that is used to improvise on the existing code or technique while Test Partition is the one that is live to public for use or for a pilot run. If many different models are being tried out, it is prudent to save a third sample, which also includes known outcomes (the *test data*) to use with the model finally selected to predict how well it will do. The model can then be used to classify or predict the outcome of interest in new cases where the outcome is unknown.  
 Validation set actually can be regarded as a part of training set, because it is used to build your model, neural networks or others. The training data (validation partition) set is used to construct a predictive relationship for all machine learning models. **A test set** is a set of data that is independent of the training data.

**The validation set** is used to compare their performances and decide to select a model among different models

**The test set** is used to obtain the performance characteristics such as accuracy (RMSE,...). **The test data set is not used in model building process**

**2.6** In fitting a model to classify prospects as purchasers or non-purchasers, a certain company

drew the training data from internal data that include demographic and purchase

information. Future data to be classified will be lists purchased from other sources,

with demographic (but not purchase) data included. It was found that “refund issued”

was a useful predictor in the training data. Why is this not an appropriate variable to

include in the model?

The data considered for future data is demographics and lists of purchases from other sources but not the refund issued. The reasons can be the below:  
- Refund can be issued for reasons like wrong size or damaged item, that doesn’t equate to not desiring the item. This refund can be followed by a purchase of the same item.  
- Refund issued could be negligibly small as compared to the purchased items; hence it merely makes a difference.

2.7 A dataset has 1000 records and 50 variables with 5% of the values missing, spread

randomly throughout the records and variables. An analyst decides to remove records

with missing values. About how many records would you expect to be removed?

Ans:

Since the dataset has 1000 records and 50 variables;  
  
It has 1000\*50=50000 values  
  
5% of 50000 => 2500 values are missing from the table.

Even with only 50 variables, if only

5% of the values are missing (spread randomly and independently among cases

and variables), almost 92% of the records would have to be omitted from the

analysis. (The chance that a given record would escape having a missing value is

0*.*9550 = 0*.077*)  
Number of records left will be 50000(.077) = 3850

An alternative to omitting records with missing values is to replace the missing

value with an imputed value (mean. median), based on the other values for that variable across

all records.

Maximum number of records that may contain these missing values = 1000 (If around 2.5 values missing from every record)  
Minimum number of records that may contain these missing values = 2500/50 = 50 (if each of these 50 records contains missing values for all its variables)  
  
So; the range is 50 records to be removed(best case scenario)->all 1000 records to be removed(worst case scenario)

2.8 Normalize the data in Table 2.17, showing calculations.

TABLE 2.17

**Age Income ($)**

25 49,000

56 156,000

65 99,000

32 192,000

41 39,000

49 57,000

The simplest way of doing this with your spreadsheet is as follows:

1. Calculate the mean and standard deviation of the values (raw scores) for the variable in question. In this case, we have obtained a mean of 98666.66 and a standard deviation of 62867.055

2. Subtract this mean score from each case's obtained score.

3. Divide this result by the standard deviation. (For instance, 8.5/12.5 = 0.68) This is a z score, and it is the basic form of standard score from which all others are derived. (That includes IQ scores, T scores, stanines, and so forth.) The mean score will be 0. Most of the values will be between -1 and +1; about 95% will be between -2 and +2. [Normalized values mentioned stated below]

|  |  |  |
| --- | --- | --- |
| Age | Income($) | **Normalized Values** |
| 25 | 49000 | **-0.790026821** |
| 56 | 156000 | **0.911977498** |
| 65 | 99000 | **0.0053023** |
| 32 | 192000 | **1.484614466** |
| 41 | 39000 | **-0.949092645** |
| 49 | 57000 | **-0.662774162** |
|  |  |  |

**Chapter 3**

**Problems 3.2**

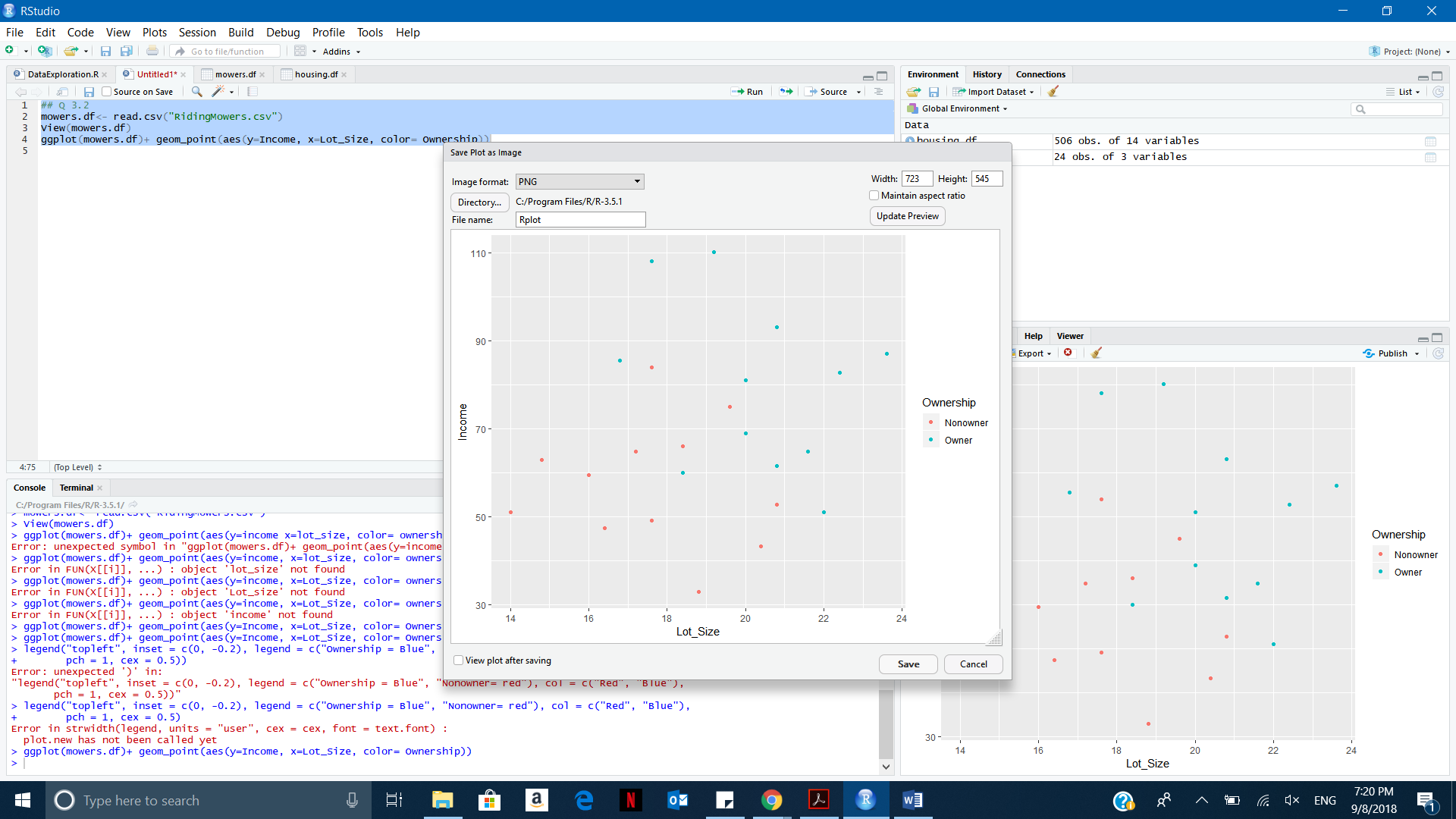
**Code:**

mowers.df<- read.csv("RidingMowers.csv")

View(mowers.df)

ggplot(mowers.df)+ geom\_point(aes(y=Income, x=Lot\_Size, color= Ownership))

**Output:**



**Problem 3.3A**

**Code:**

## Q3.3-A

laptop.df<- read.csv("LaptopSalesJanuary2008.csv")

View(laptop.df)

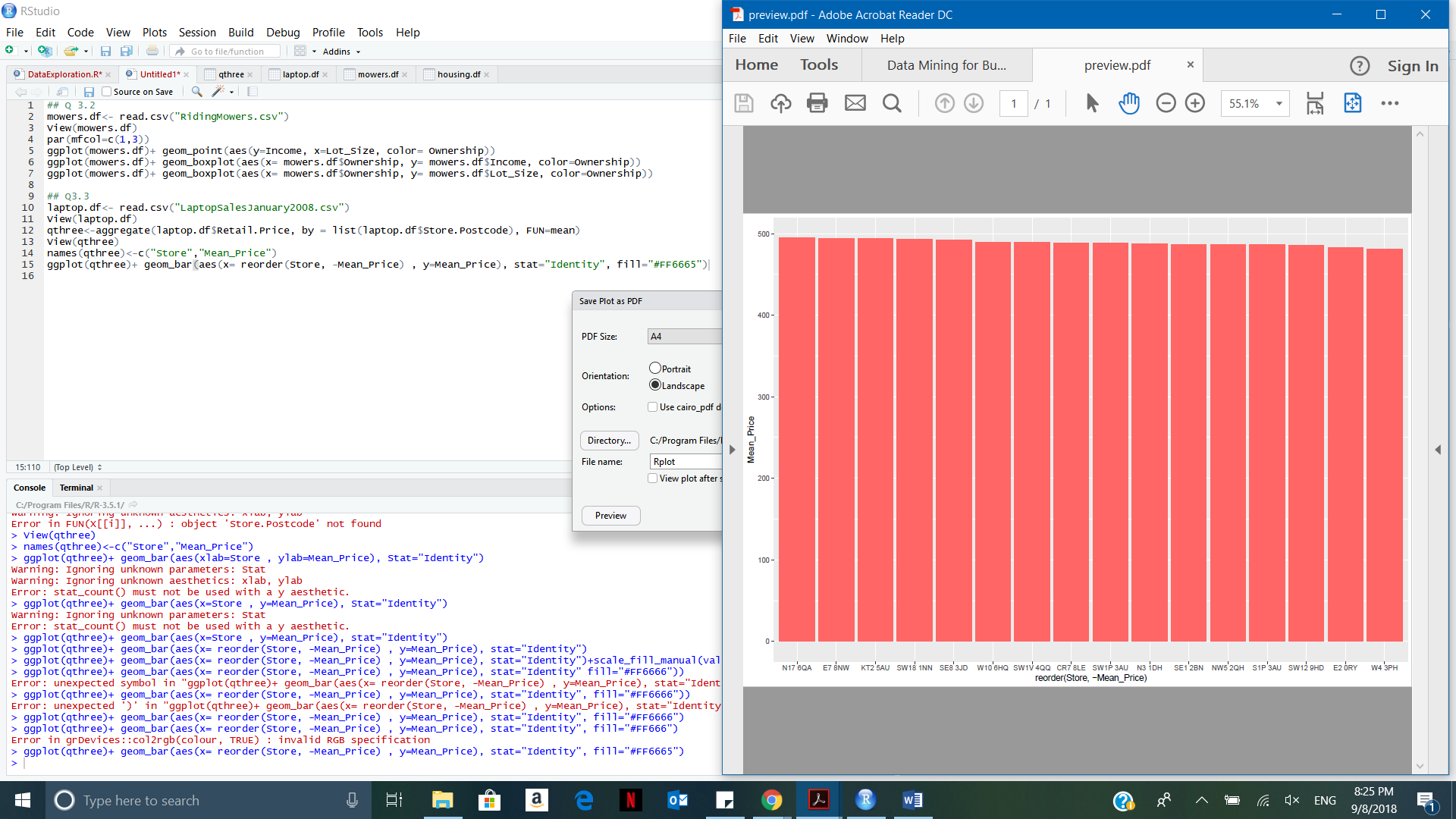
qthree<-aggregate(laptop.df$Retail.Price, by = list(laptop.df$Store.Postcode), FUN=mean)

View(qthree)

names(qthree)<-c("Store","Mean\_Price")

ggplot(qthree)+ geom\_bar(aes(x= reorder(Store, -Mean\_Price) , y=Mean\_Price), stat="Identity", fill="#FF6665", width = 0.6)

**Output:**



Store N176QA has the highest average retail price, whereas W43PH has the lowest average retail price.

**Problem 3.3B**

**Code:**

par(mfcol[1,12])

boxplot (sales.df$Retail.Price ~ sales.df$Store.Postcode, xlab = "Stores", ylab = "Retail Price",col=rainbow(20),las=2)

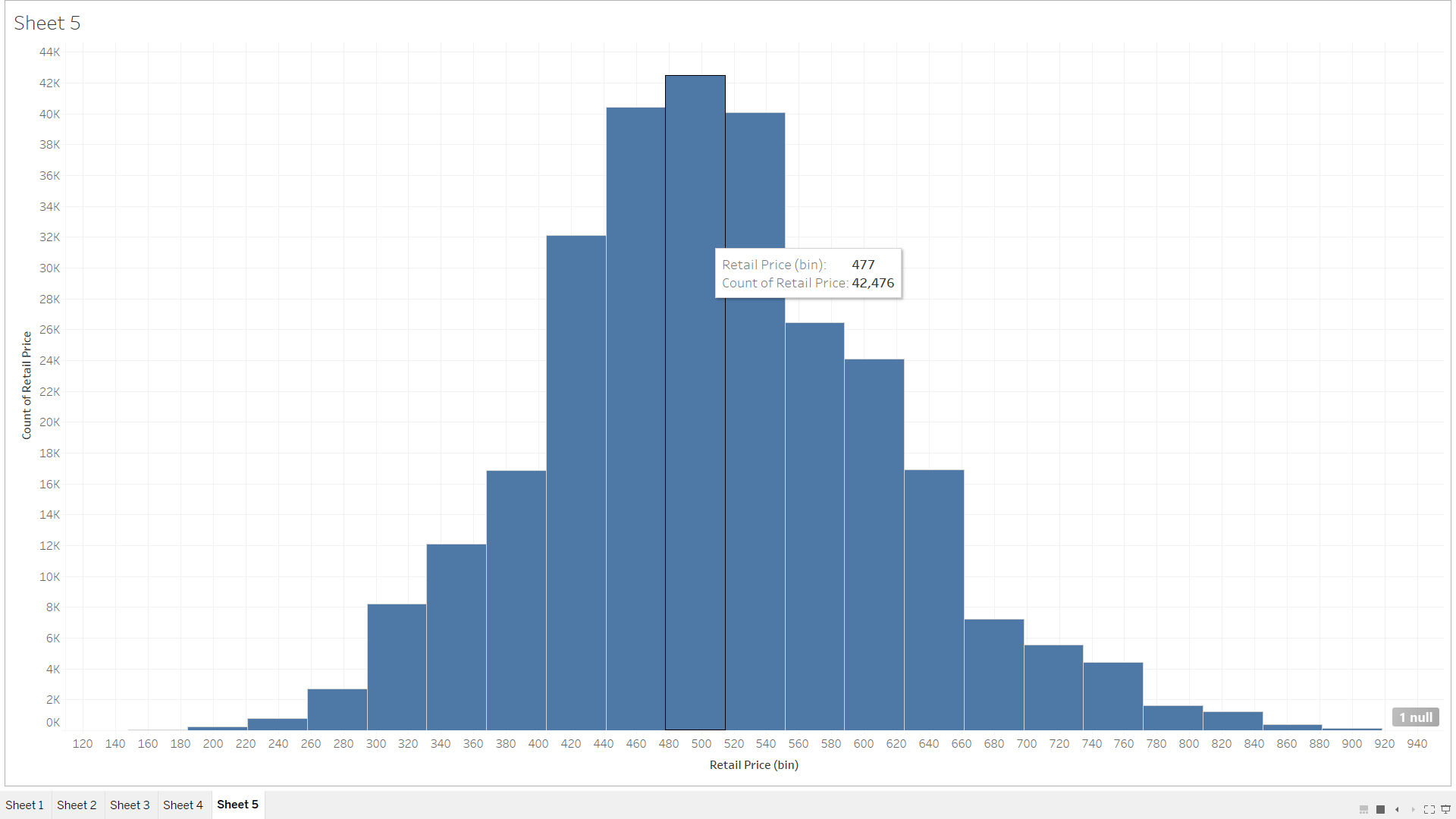
**Output:**

A screenshot of a cell phone

Description generated with high confidence

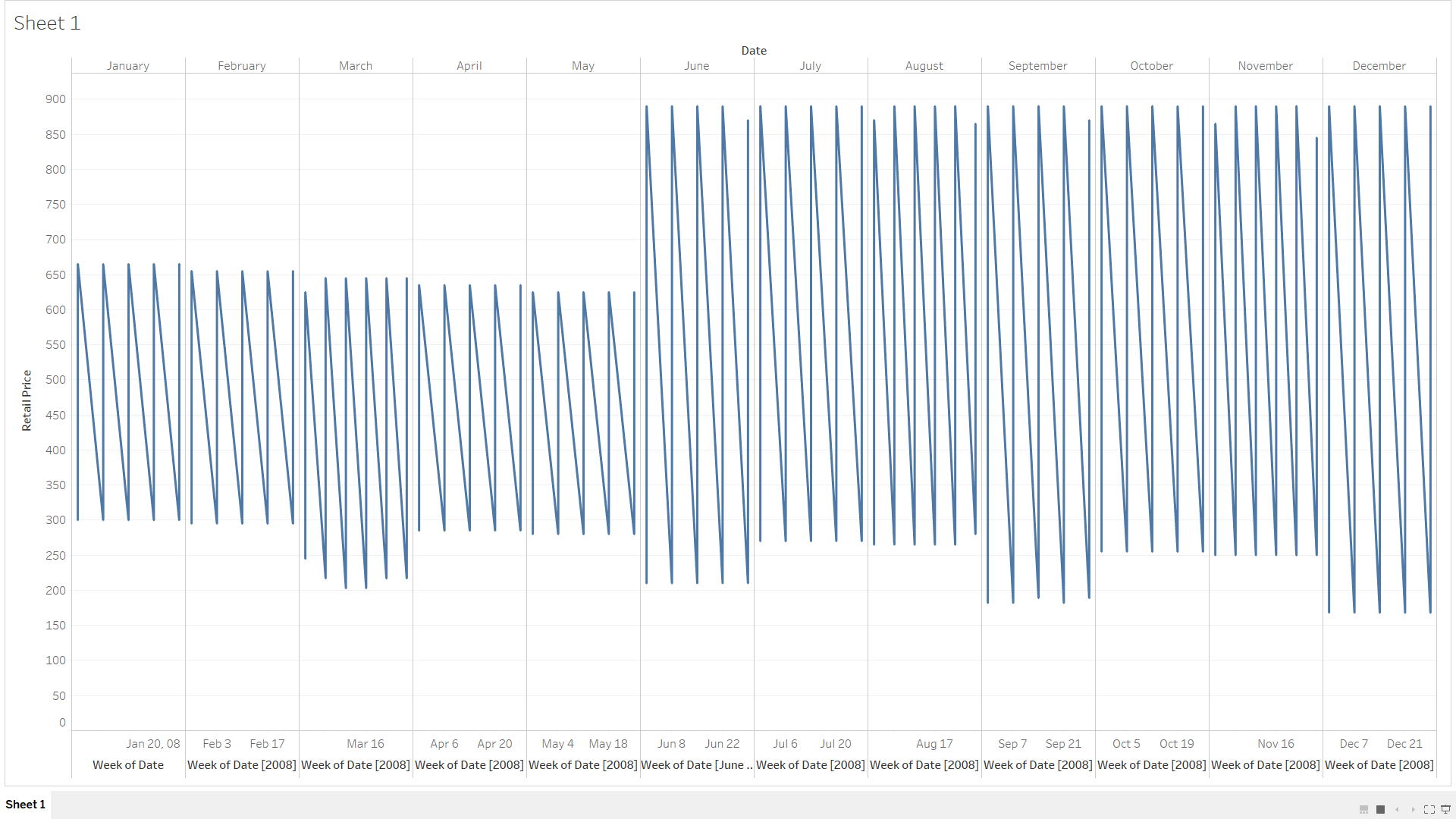
**Problem3.4**

1. **Price Questions**
2. **At what price are the laptops selling?**



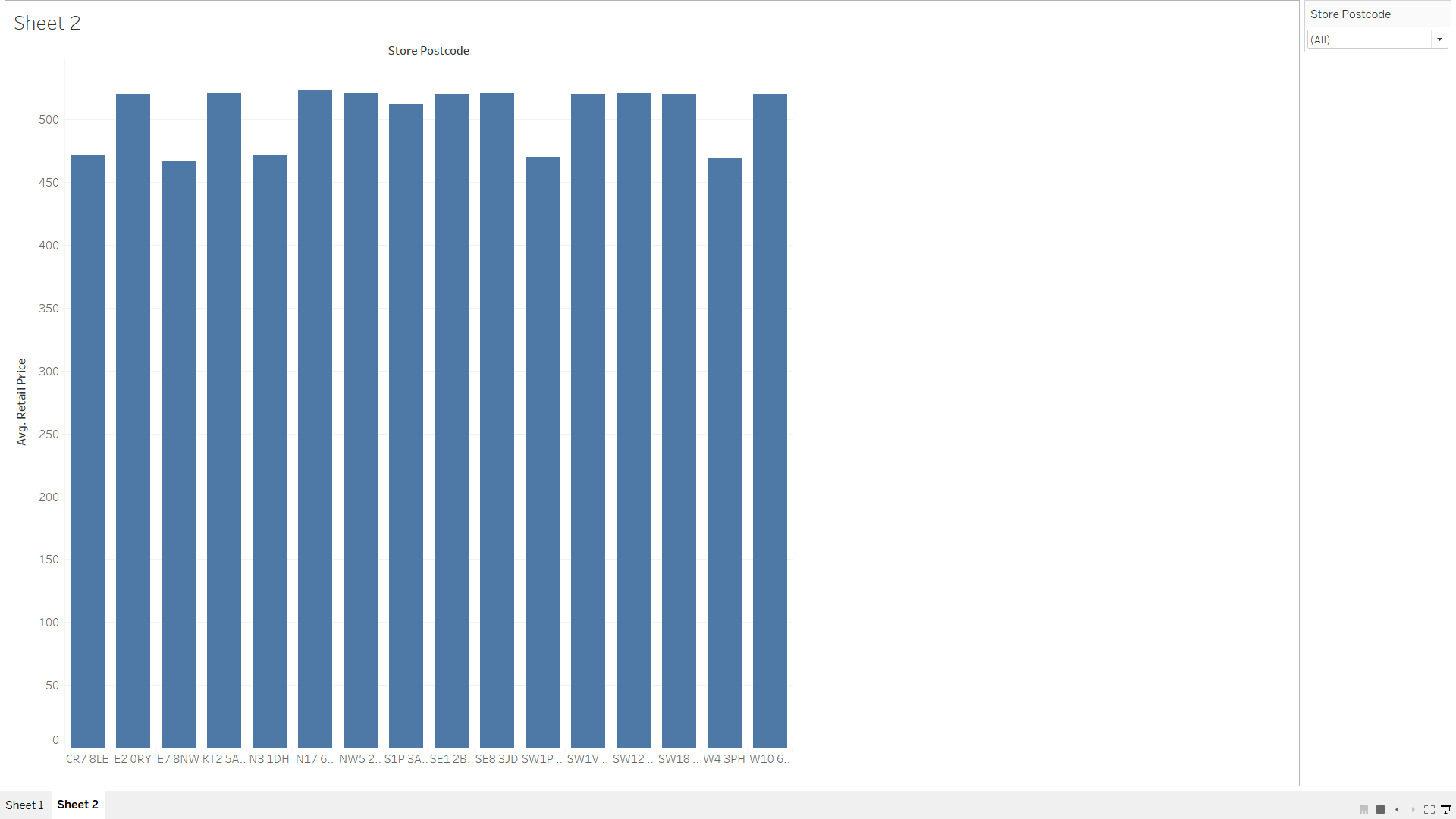
Majority laptops are selling between 450-500 range.

1. **Does price change with time? (Hint: Make sure that the date column is recognized as such. The software should then enable different temporal aggregation**



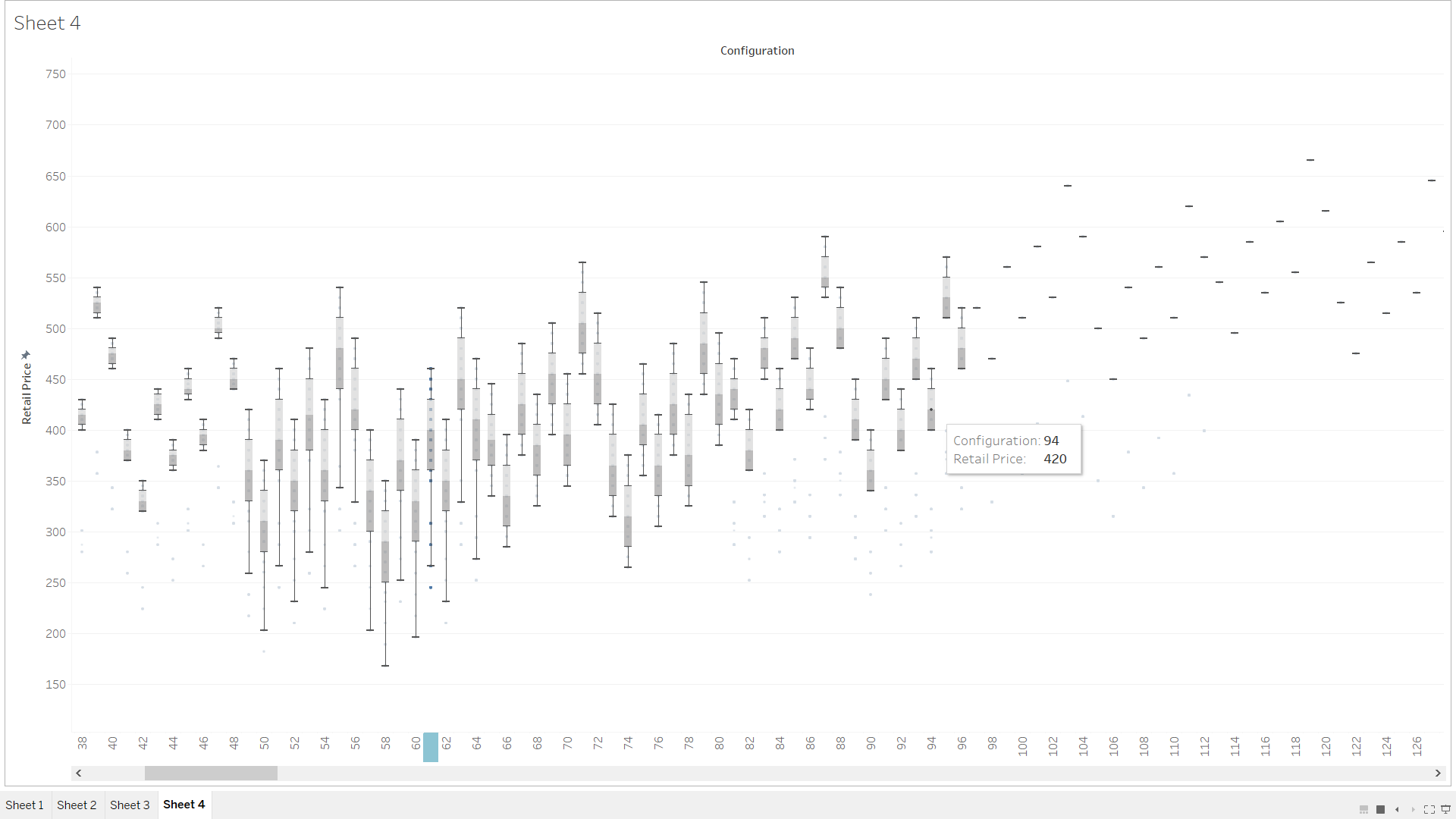
Yes, price varies with time. We can see that the retail prices are from June till December are as compared to prices from January to May. Also, the fluctuations in the prices after May are high as compared to the period between January May.

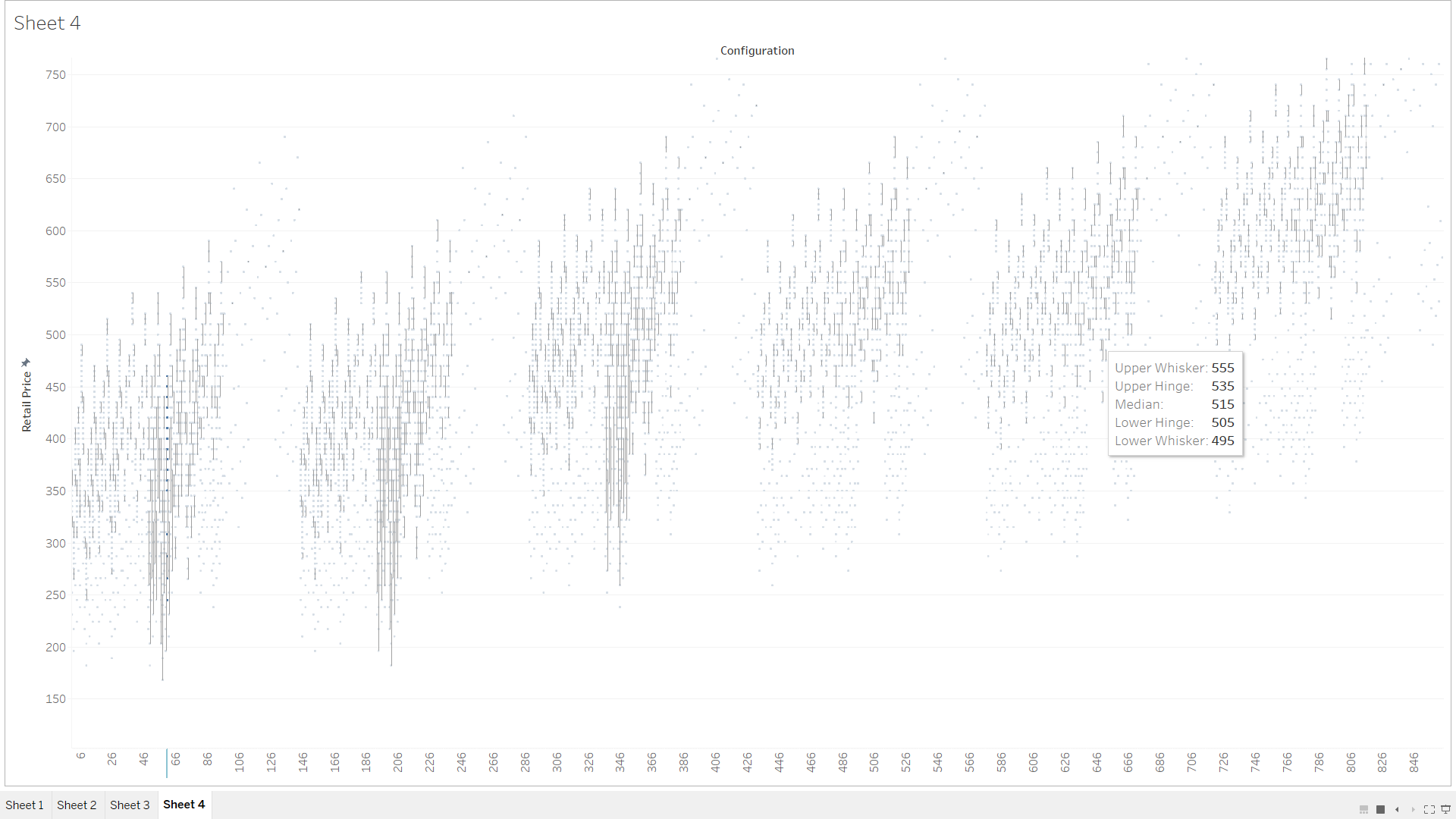
1. **Are prices consistent across retail outlets?**



Yes, Prices are consistent across stores.

1. **How does price change with configuration?**





Yes, Price increases as the configuration number goes high, But sales of the laptops with low configuration number are high.

1. **Location Question**
2. **Where are the stores and customers located?**

**A picture containing text

Description generated with high confidence**

1. **Which stores are selling the most?**

**A close up of text on a white background

Description generated with high confidence**

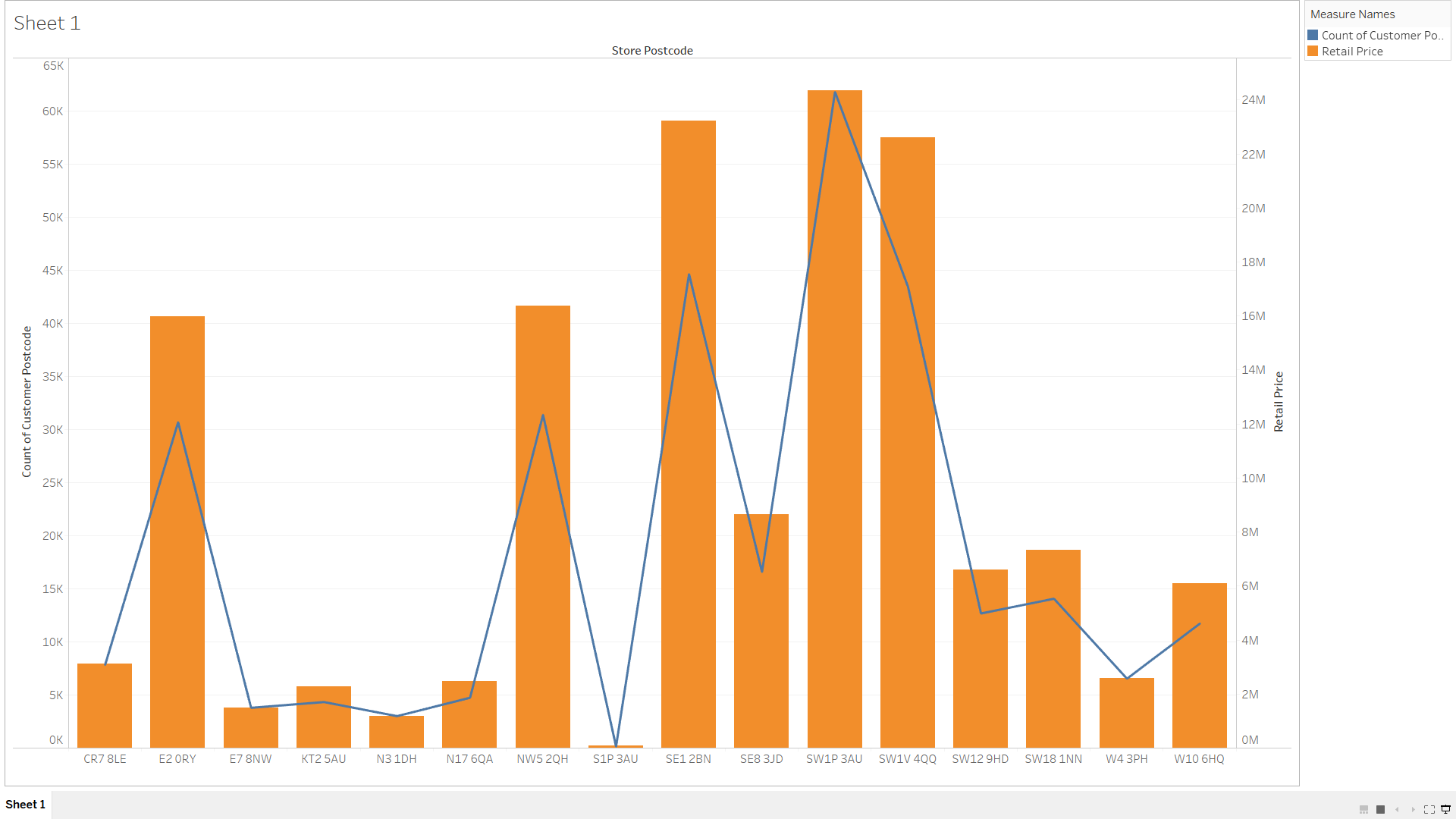
Store SW1P3AU has the highest sales

1. **How far would customers travel to buy a laptop?**

**A close up of a map

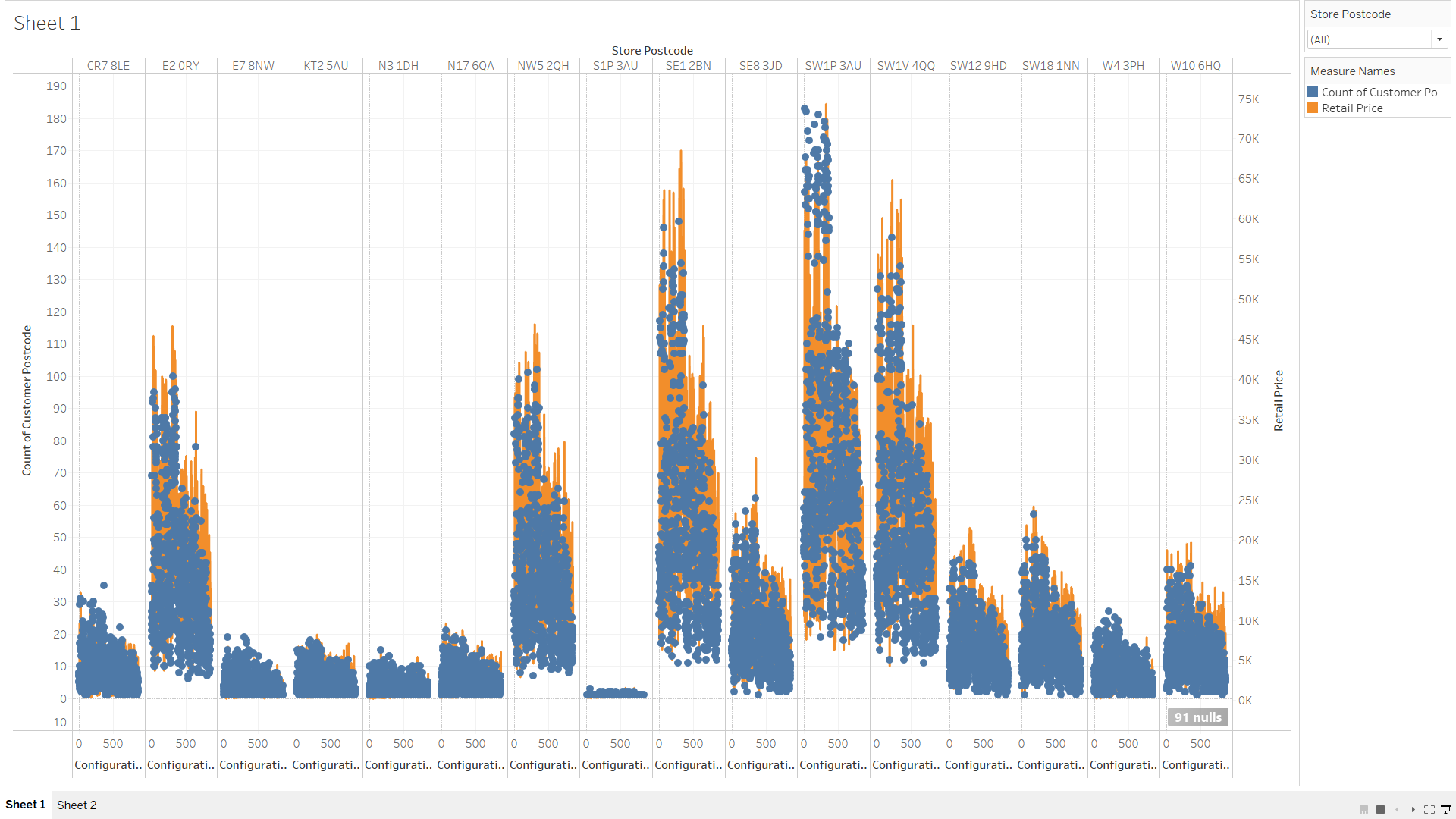
Description generated with high confidence**

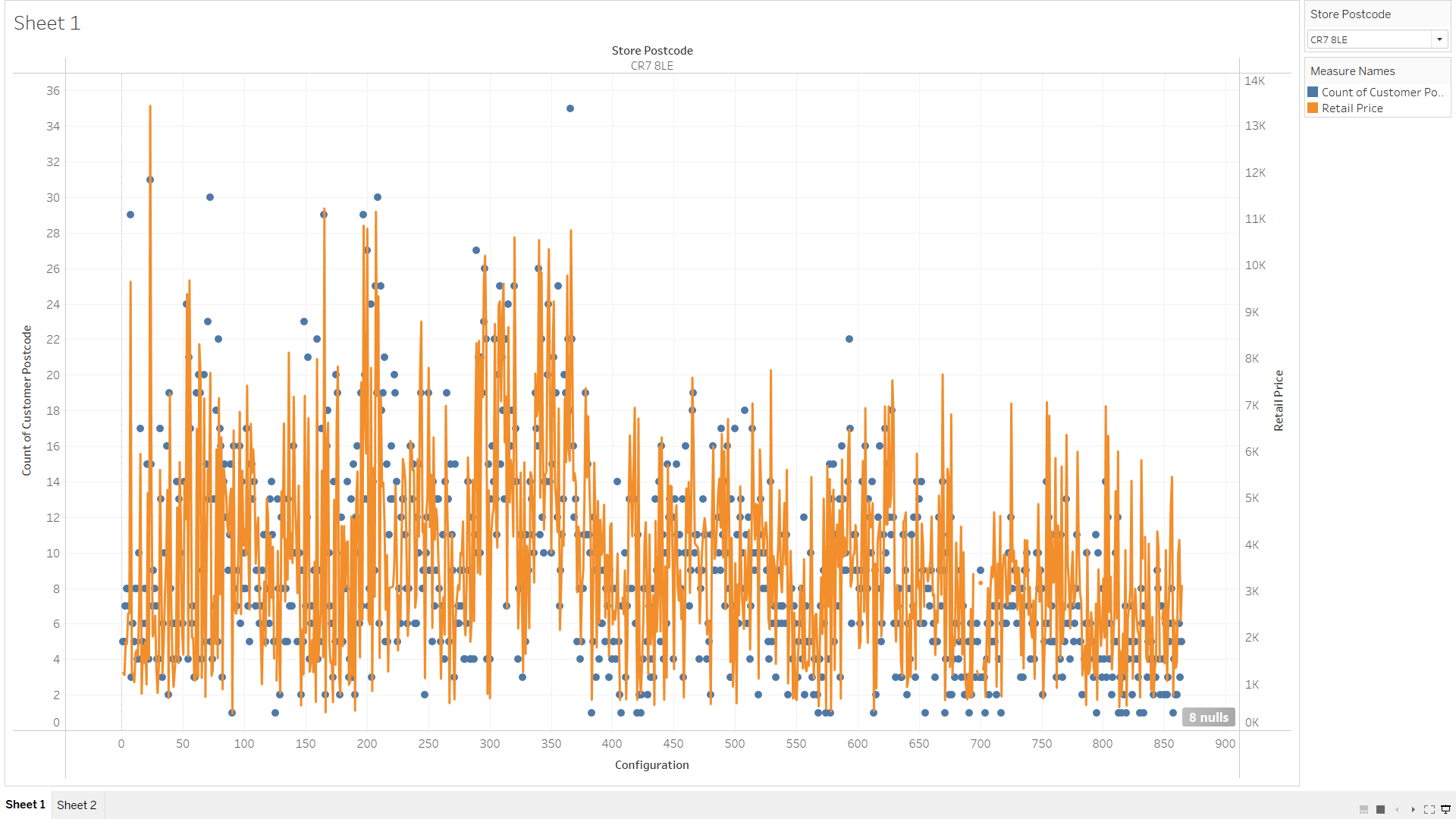
1. **Revenue Questions**
2. **How do the sales volume in each store relate to Acell’s revenues?**

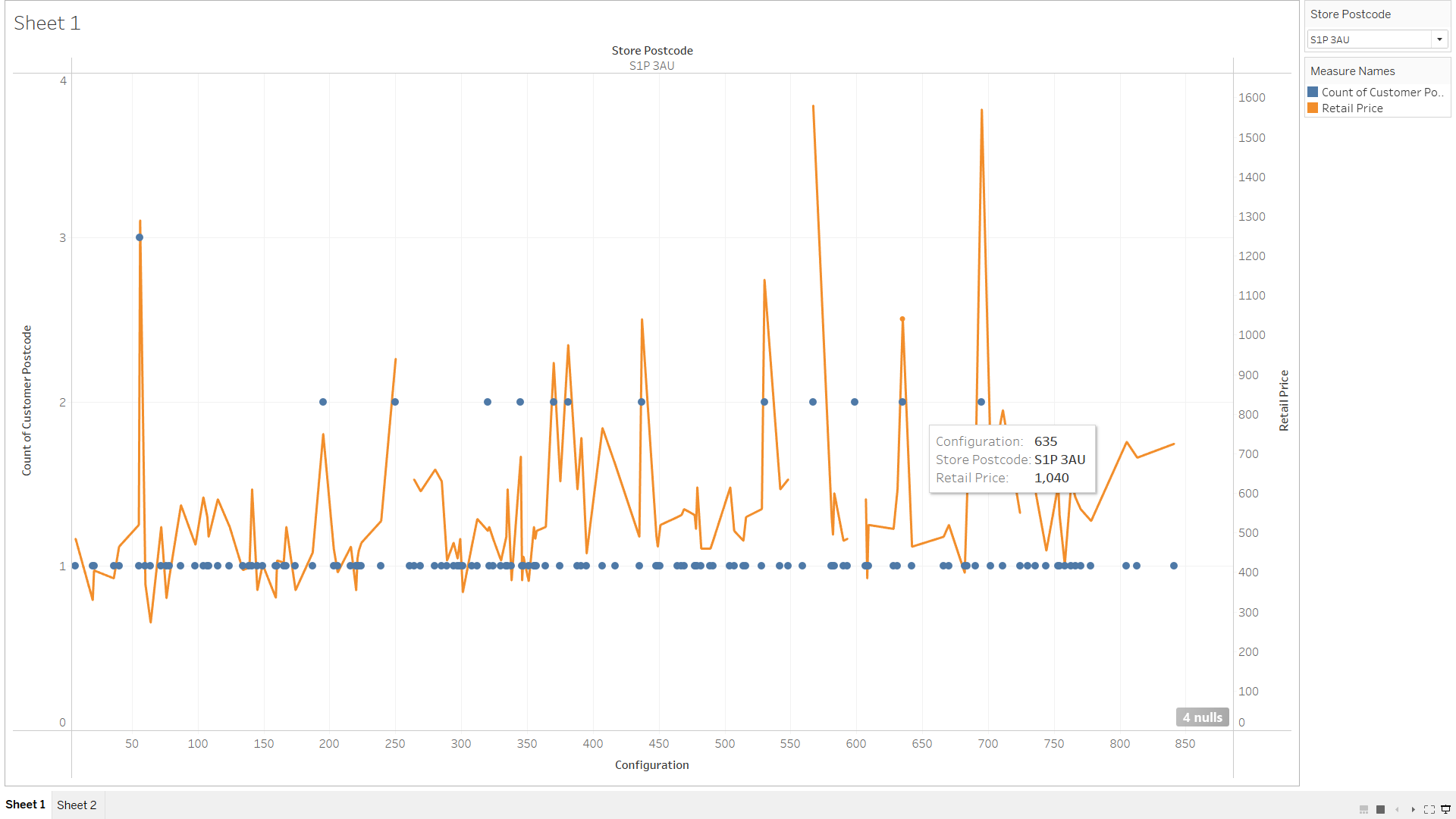


Sales volume of each store is positively correlated with acell’s revenue. As the store sales go up so does the revenue and vice-versa.

1. **How does this relationship depend on the configuration?**

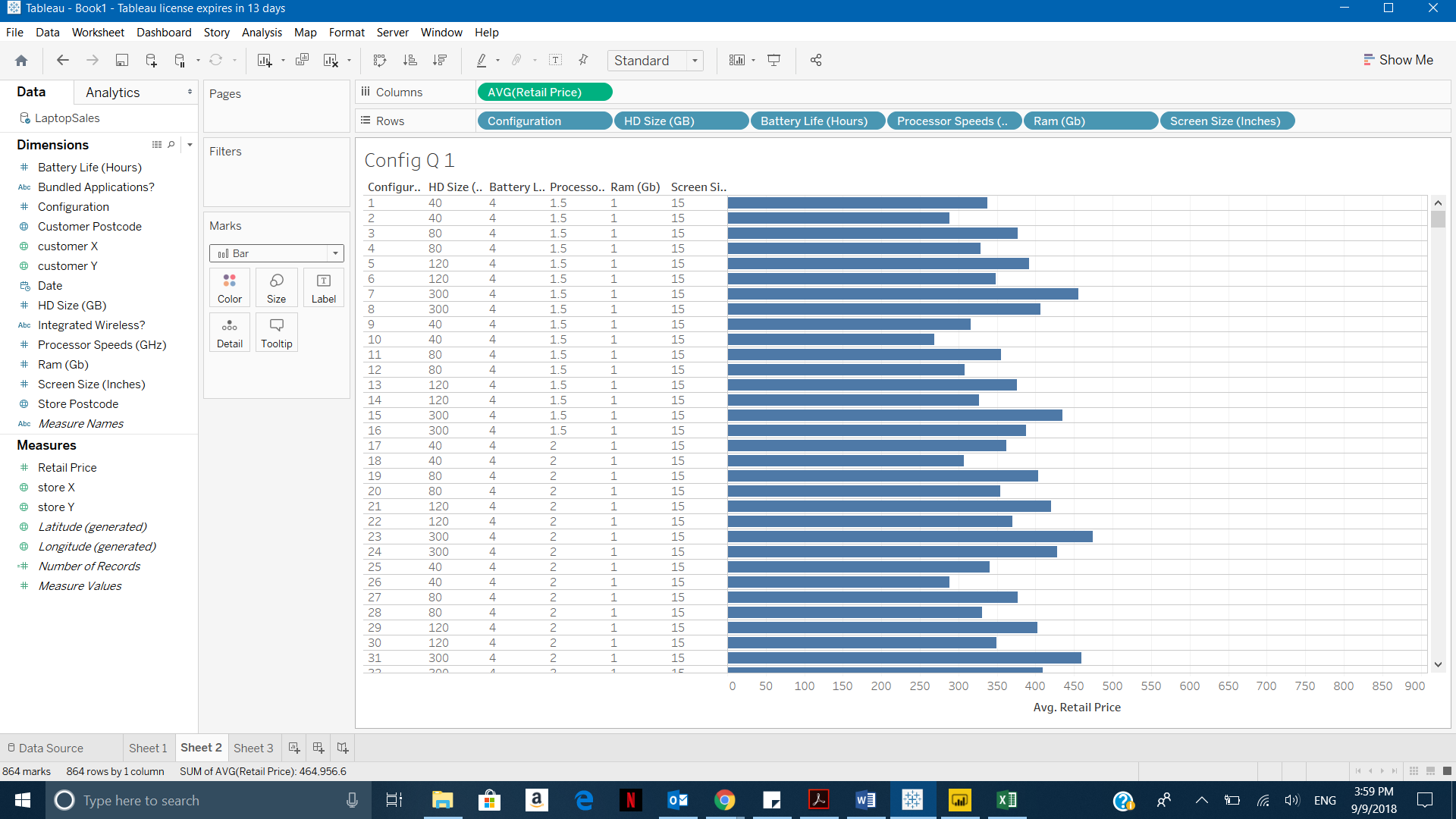


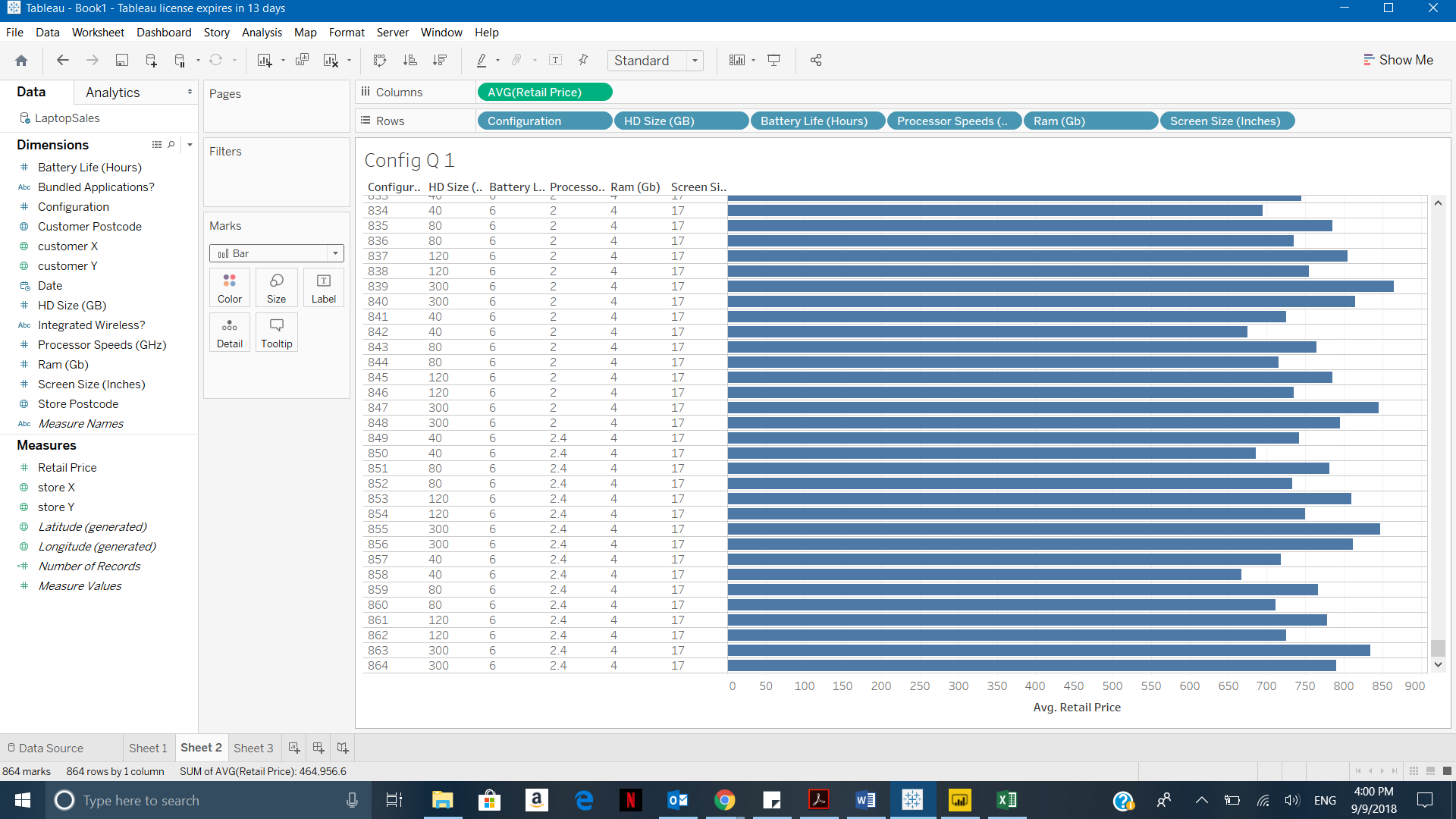




There no clear relation between Configuration, revenue and Volume of laptops

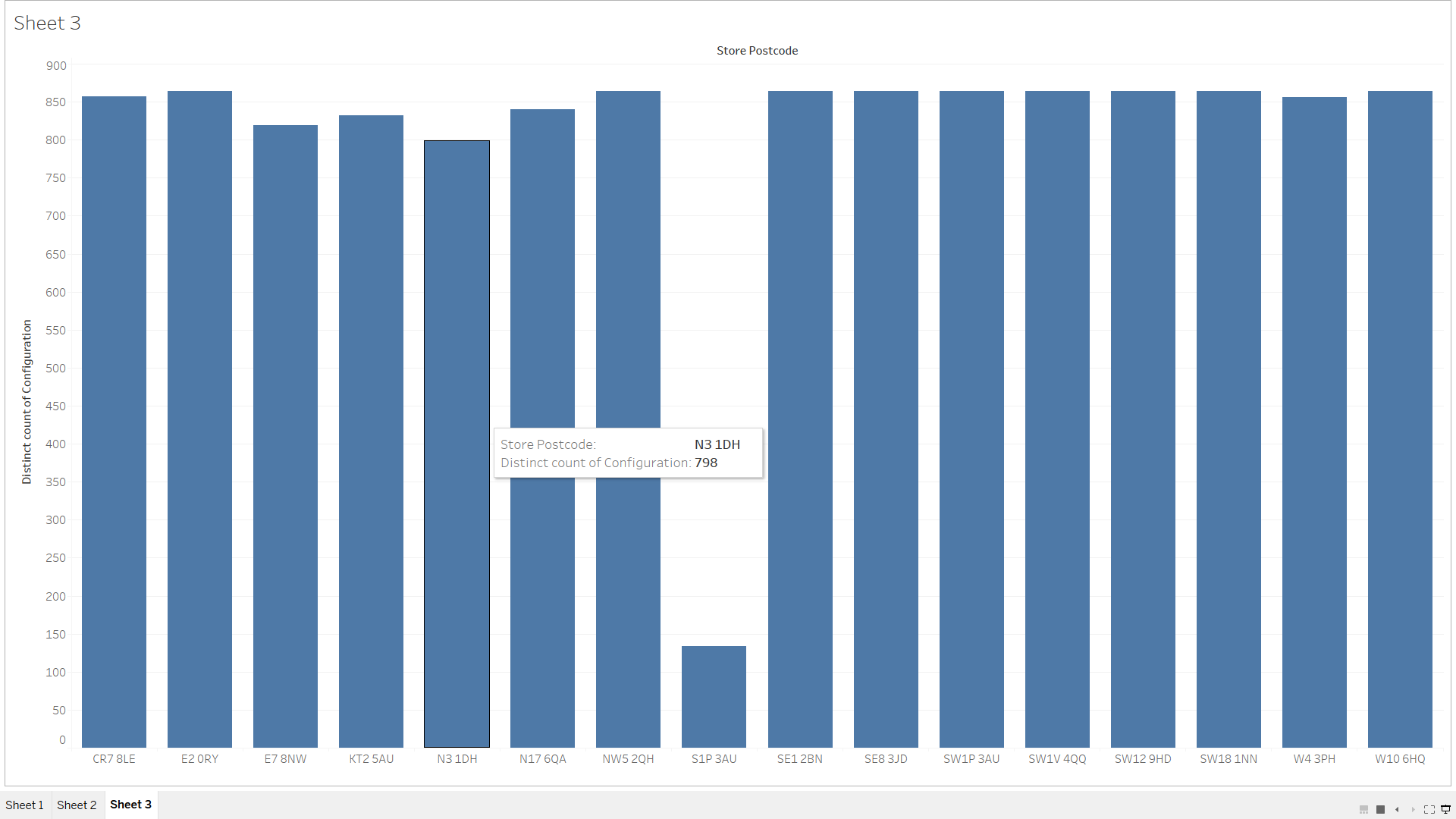
1. **Configuration Questions**
2. **What are the details of each configuration? How does this relate to price?**





The price Increases with the Config number.

1. **Do all stores sell all configurations?**



No, all stores Don’t sell laptops of all Configurations.

**Chapter 4**

**Problem 4.3**

**a.**

**Ans:** The categorical variables in the dataset are:

1. Fuel\_Type
2. Color
3. Model

**b.**

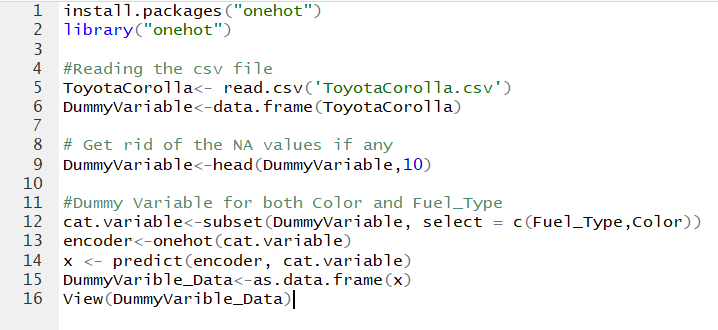
**Ans:** A dummy variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into [mutually exclusive](https://en.wikipedia.org/wiki/Mutually_exclusive_events) categories. A dummy variable can thus be thought of as a [truth value](https://en.wikipedia.org/wiki/Truth_value) represented as a numerical value 0 or 1.

**c.**

**Ans:** The number of dummy variables for a categorical variable with N categories is N or N-1.

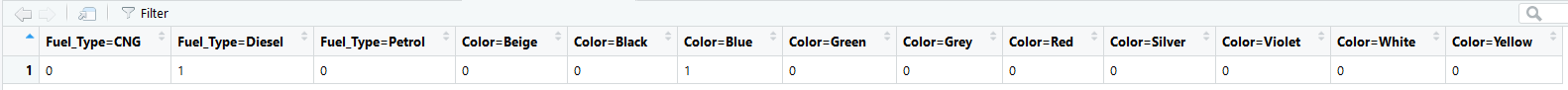
**d.**

**Ans:** The r code snippet to create dummy variables



**Explanation for Fuel\_Type and color**

The fuel\_type variable contains 3 values i.e. Diesel, petrol and CNG. By dividing the categorical variable into 3 dummy variables, we can increase a slight efficiency of the analysis.

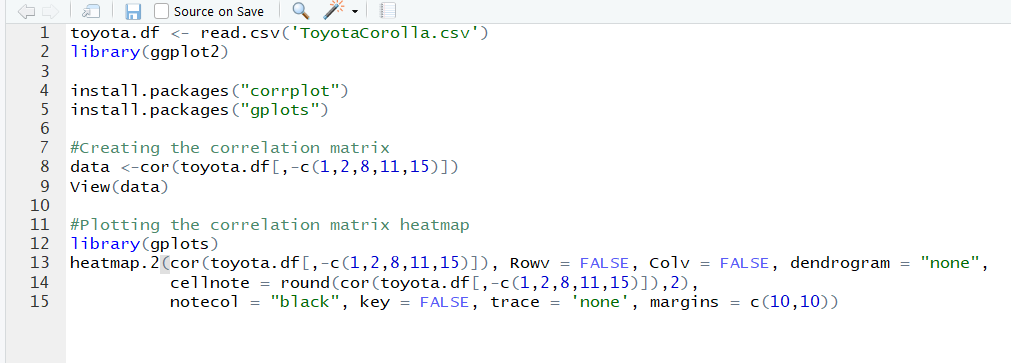


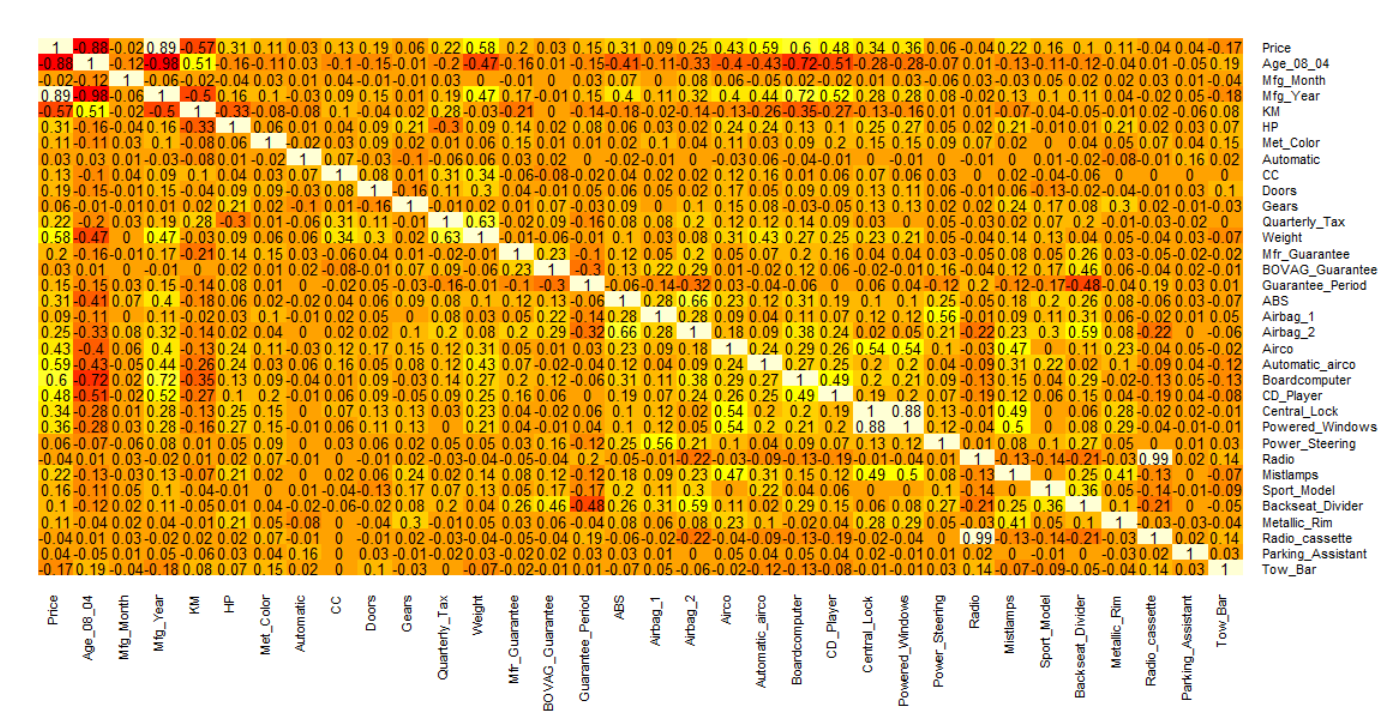
The figure shows that the selected record is a car with fuel\_type diesel as the Boolean value of the column Diesel is 1. The values in Petrol and CNG column is 0 which means a car is neither a petrol or CNG car.

Also, the Boolean value of color – Blue is 1 which means the car is blue. For all other color variables, the Boolean value is 0.

**e.**

**Ans:** R code snippet to create a correlation matrix

****

**Correlation Matrix**

As depicted in the heat map, the blocks with darker shades are tightly dependent. The correlation values in those cells are larger in values as compared to other cells.

Larger the value, more the variables are dependent on each other. A negative sign shows the inverse proportionality and the positive values shows direct proportionality.

For example: The value for Price-Mfg\_Year correlation is high (0.89) which means higher the Mfg\_Year higher the price.

Similarly, in case of Price-Age-08-04 correlation value (-0.88) is high in magnitude and tells us that these two variables are inversely proportional.

More the age, lower the price.