Question 1:-

Customers who cancelled the Policy:-

I am working for an insurance company. I want to cluster my customers.

Here are the predictors..

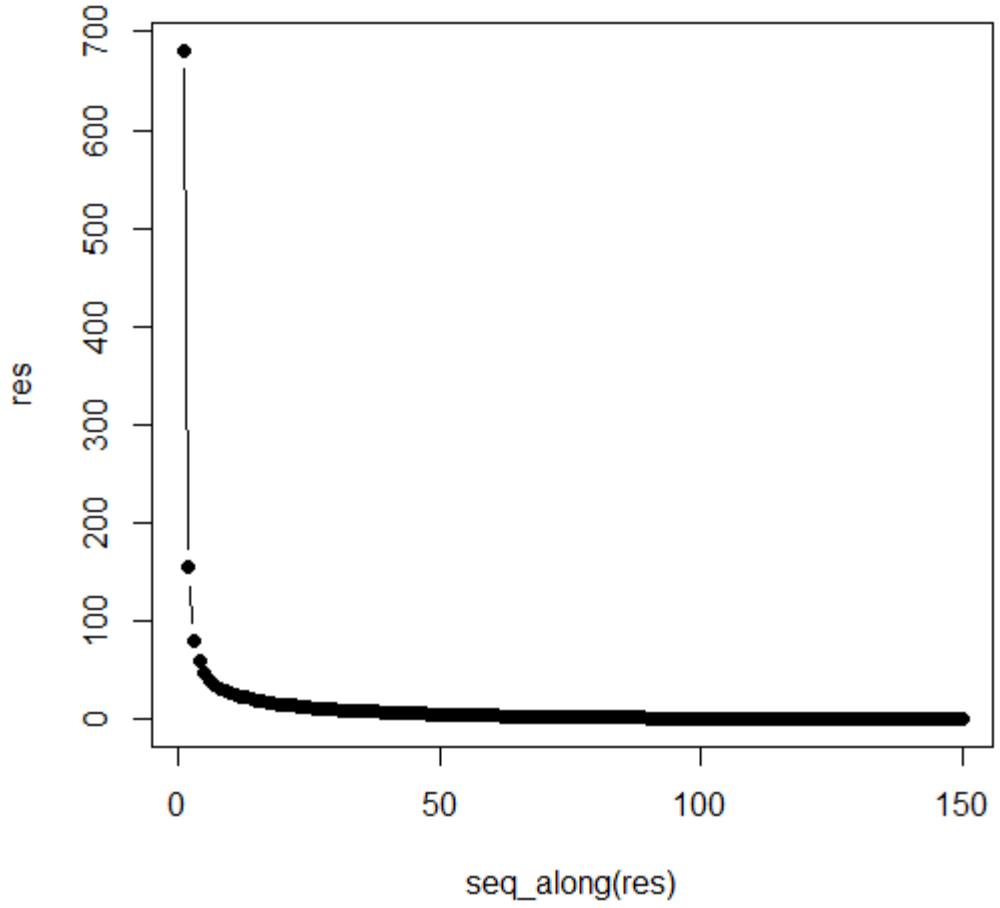
1. Number of missed payment
2. Number of times, signed on to Customer self-service portal in a given week.
3. Age group
4. Number of claims made in a year
5. Rating on customer feedback from survey.

The goal is to find clusters of customers and if any of the existing customers fall in that cluster than come up with targeted incentive plan and increase the retention.

Question 2:-

Finding the best K:- Value is 3

Elbow Graph:-



R Code for Best K:-

```
wss <- function(d) {
  sum(scale(d, scale = FALSE)^2)
}
wrap <- function(i, hc, x) {
  cl <- cutree(hc, i)
  spl <- split(x, cl)
  wss <- sum(sapply(spl, wss))
  wss
}
iris2 <- iris[, 1:4]  # drop Species column
cl <- hclust(dist(iris2), method = "ward.D")

res <- sapply(seq.int(1, nrow(iris2)), wrap, h = cl, x = iris2)
plot(seq_along(res), res, type = "b", pch = 19)
```

Finding K-means:-

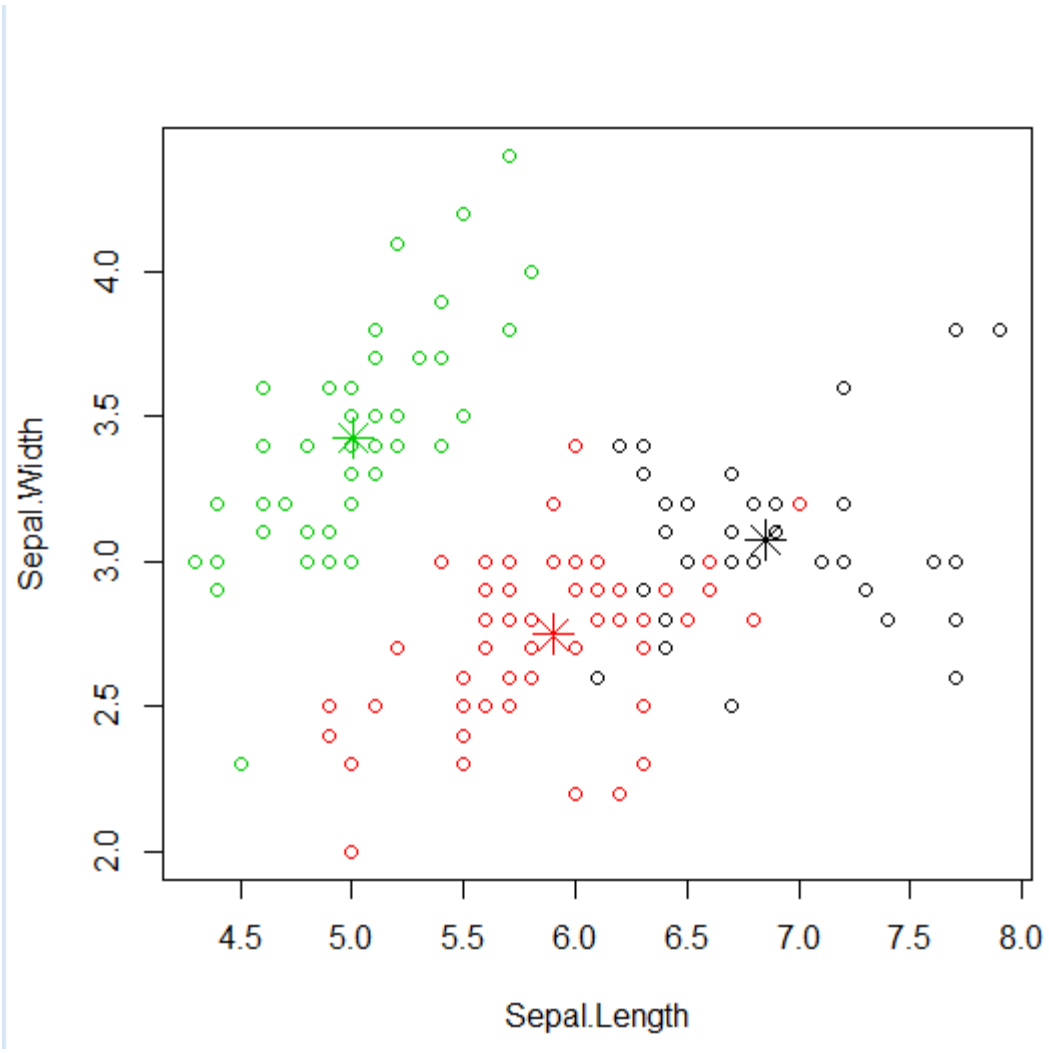
R-code:-

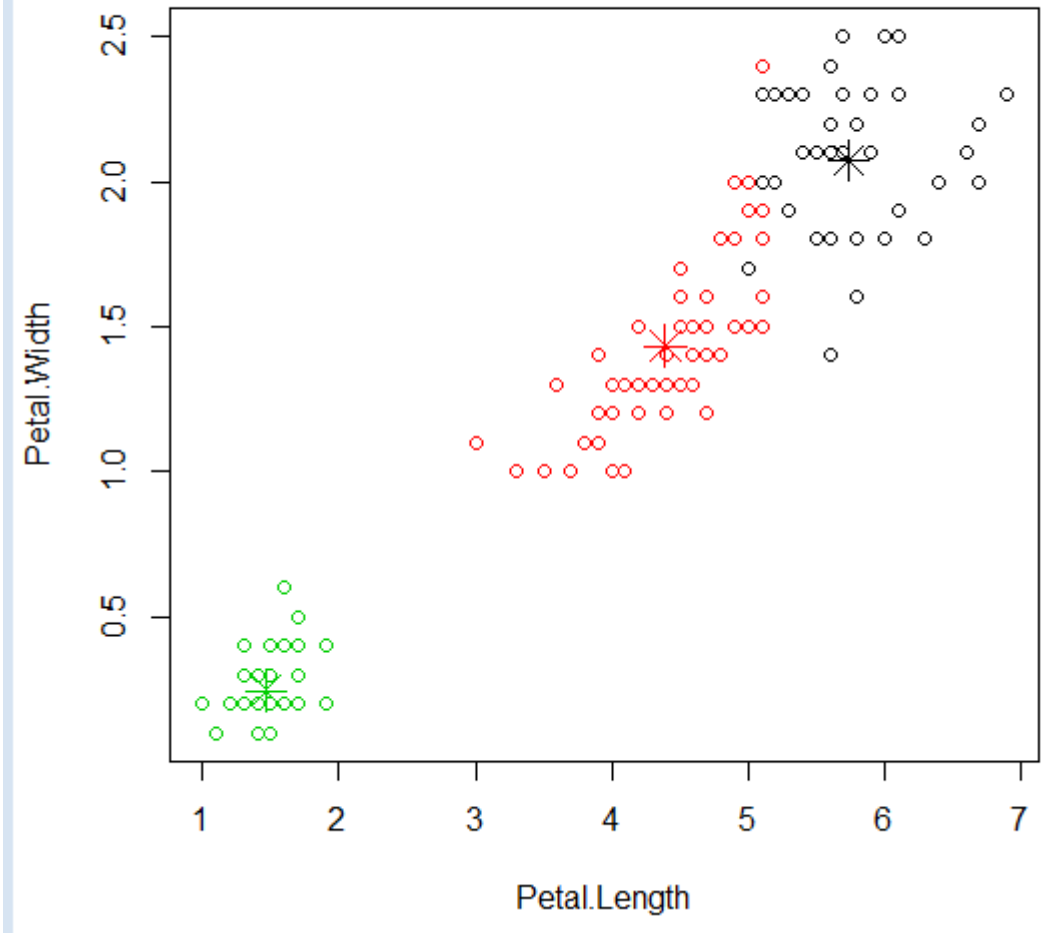
```
newiris <- iris
newiris$Species <- NULL
kc <- kmeans(newiris, 3)
print(kc)
```

Results:-

```
> source("gtkmeans1.R")
K-means clustering with 3 clusters of sizes 38, 62, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.901613    2.748387     4.393548    1.433871
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1
[112] 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1
[149] 1 2

Within cluster sum of squares by cluster:
[1] 23.87947 39.82097 15.15100
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
```

```
> table(iris$Species, kc$cluster)

              1  2  3
  setosa      0  0 50
  versicolor  2 48  0
  virginica  36 14  0
>
```


Sepal.Width
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal.Length


Petal.Width
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
Petal.Length

Question 3:-

1 and 46 are outliers for statistical reasons only. so the key assumption is we are simply looking for most extreme points which is alternative hypothesis

```
> grubbs.result <- grubbs.test(datacrime,type=11,two.sided=TRUE)
> grubbs.result

        Grubbs test for two opposite outliers

data:  datacrime
G = 3.31520, U = 0.88308, p-value < 2.2e-16
alternative hypothesis: 1 and 46 are outliers
```

Question 4:-

In the same example question1

Here are the predictors..

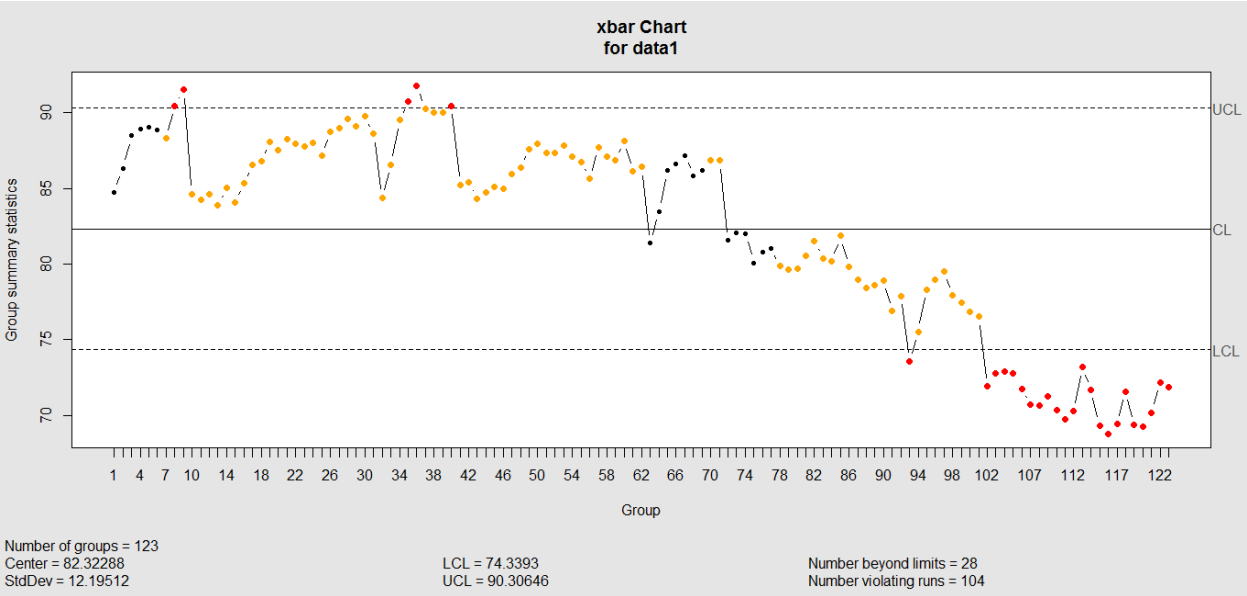
1. Number of missed payment
2. Number of times, signed on to Customer self-service portal in a given week.
3. Age group
4. Number of claims made in a year
5. Rating on customer feedback from survey.

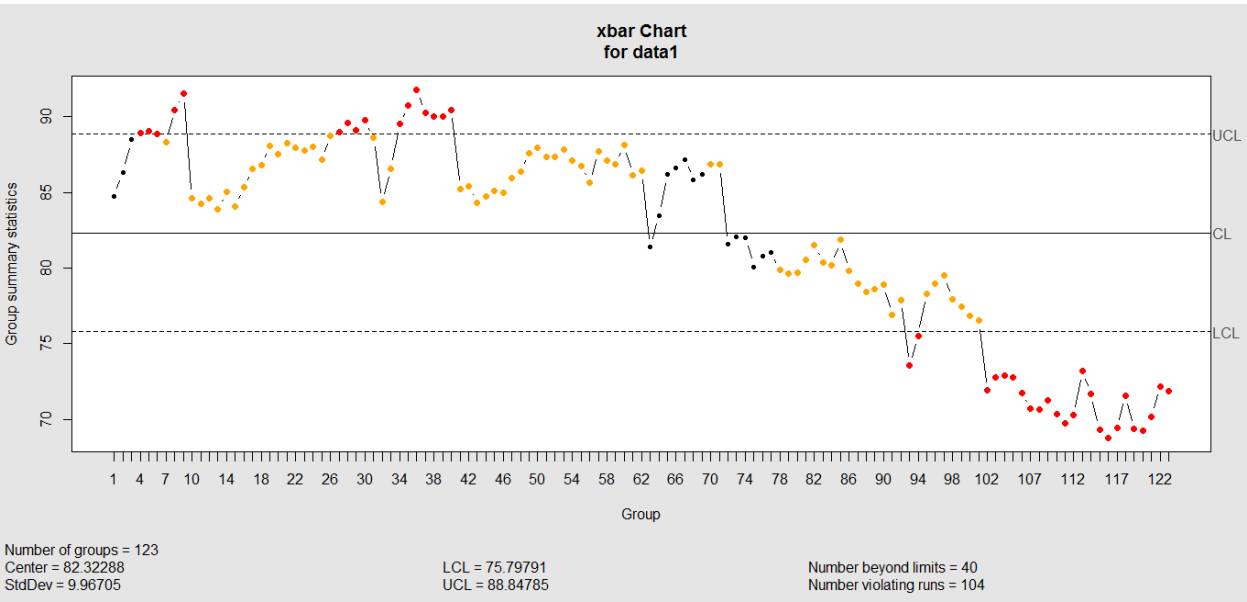
I want to find out if we give incentive to customer to log on to customer self-service portal will the cancellation come down.

Steps followed to get the change detection using CUSUM.

1. With the data that we have currently, plot a graph with Time T on x axis and number of new CSS accounts created.
2. Lets say μ is the mean of new CSS logons.
3. Identify using the formula $S_t = \max[0, S_{t-1} + (x_t - \mu)$
4. Then see because of new CSS accounts is cancellation is dropping.
5. Pull down the fluctuations by subtracting “C”.
6. Next step is to determine the value of “C” and “t” to quickly detect the change.
7. Now I have to trade off between early detection and late detection. Since this is not a mission critical value I want to be safe and increase the threshold so that I am very sure before I present it to my VP and the board. Higher threshold will give me confidence to show the results are consistent.
8. How would I choose the critical value ?
    a. I would have moderate C value as our policy renewals are even throughout the year.
    b. I will probably choose the UCL for the same.

Question 5:- 5.1





Based on the above chart

1. Summer officially ends at 102 row. Which is *October 9th.*
2. C will be 76 Degrees
3. Threshold will be 15. (3 degrees with occurrence 5 times)

Question 5:- 5.2

Based on the above graph Atlanta has become warmer in rows 26 to 42. Which translates to End of July and early August.