# Integrated Analysis - Decision Tree and K-means Clustering using Tableau & R

**Objective:-** The objective of this project is to create the Decision Tree and K-means clustering analysis using R and visualizing results through Tableau.

Below are the phases:

1)Data retrieval

2)Data pre-processing

3) Decision Tree using R

4) K-mean clustering using Tableau- R integration by invoking Rserve ()

5) Data Visualization in Tableau

# 1) Data Retrieval

1. Create a new directory and a R file and place the titanic.csv file in the new directory created. Read the csv file in the object called titanic as mentioned below.

```
titanic<-read.csv('titanic.csv')
```

The dataset is now in titanic data frame

2. Check the number of rows and columns as mentioned below.

Answer:

```
  dim(titanic)
[1] 891  12
```

# 2) Data Pre-processing

This step involves replacing missing value of the ages with their mean, adding age category column and replacing the label with a meaningful name

- What is the average age of the passangers?

```
meanAge<-sum(na.omit(titanic$Age))/length(na.omit(titanic$Age))
meanAge
```

```
Answer:
meanAge
[1] 29.69912
```

- Round off the age to their nearest integer.

```
titanic$Age[is.na(titanic$Age)]<-meanAge
titanic$Age<-round(titanic$Age)
```

- Create a new variable (vector) called Age Category and assign a category to every passenger in the dataset.

```
titanic$AgeCat[titanic$Age>=0&titanic$Age<=16]<-"0-16"
titanic$AgeCat[titanic$Age>=17&titanic$Age<=32]<-"17-32"
titanic$AgeCat[titanic$Age>=33&titanic$Age<=48]<-"33-48"
titanic$AgeCat[titanic$Age>=49&titanic$Age<=64]<-"49-64"
titanic$AgeCat[titanic$Age>=65]<-"65 and Above"
```

- Replace the integer value of 0 and 1 in the survivor variable (vector) with a meaningful labels.

```
titanic$Survived[titanic$Survived==0]<-"Not Survived"
titanic$Survived[titanic$Survived==1]<-"Survived"
```

- Convert the integer and character vectors to factor variables as mentioned in the screenshot for the other variables

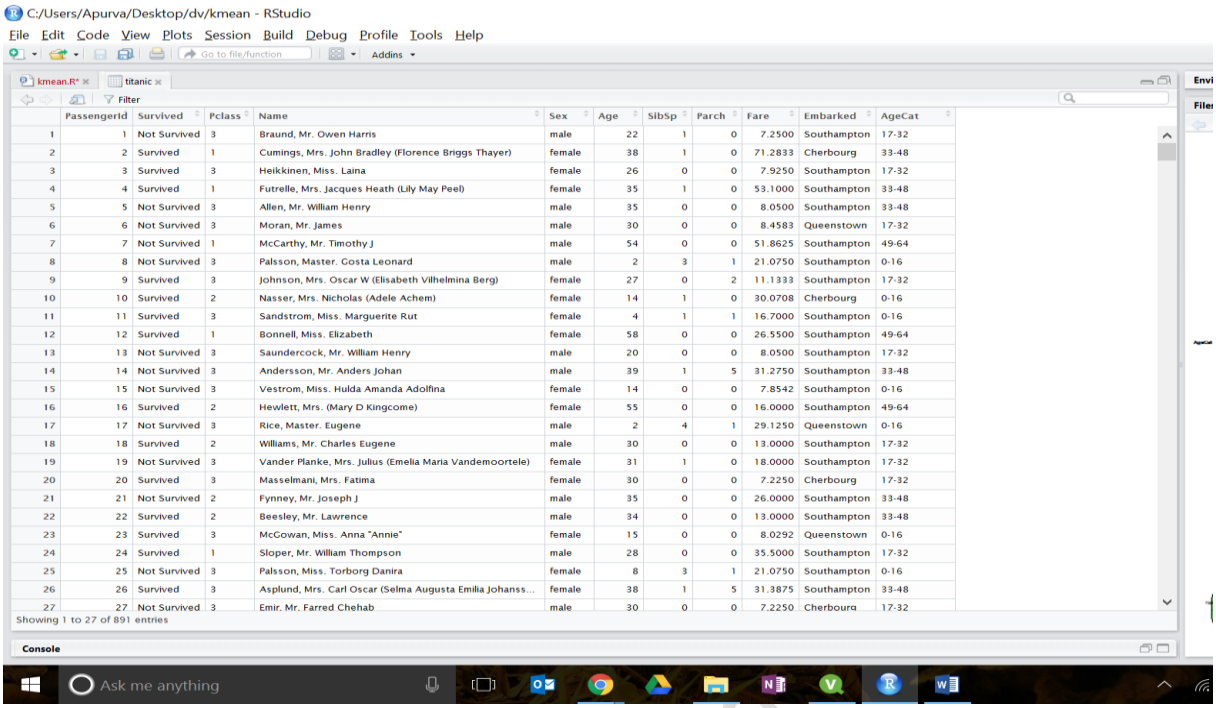
```
titanic$Pclass<-factor(titanic$Pclass)
titanic$AgeCat<-factor(titanic$AgeCat)
titanic$Survived<-factor(titanic$Survived)
titanic$Embarked<-as.character(titanic$Embarked)
titanic$Embarked[titanic$Embarked=="S"]<-"Southampton"
titanic$Embarked[titanic$Embarked=="C"]<-"Cherbourg"
titanic$Embarked[titanic$Embarked=="Q"]<-"Queenstown"
titanic$Embarked<-factor(titanic$Embarked)
```

- Remove the other redundant variables such as Ticket and Cabin from the titanic data frame.

```
titanic=titanic[c(-9,-11)]
```

- Crosscheck the processed values using the command as mentioned below.

```
View(titanic)
```




| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Fare | Embarked | AgeCat |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Not Survived | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | 7.2500 | Southampton | 17-32 |
| 2 | 2 | Survived | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 38 | 1 | 0 | 71.2833 | Cherbourg | 33-48 |
| 3 | 3 | Survived | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | 7.9250 | Southampton | 17-32 |
| 4 | 4 | Survived | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35 | 1 | 0 | 53.1000 | Southampton | 33-48 |
| 5 | 5 | Not Survived | 3 | Allen, Mr. William Henry | male | 35 | 0 | 0 | 8.0500 | Southampton | 33-48 |
| 6 | 6 | Not Survived | 3 | Moran, Mr. James | male | 30 | 0 | 0 | 8.4583 | Queenstown | 17-32 |
| 7 | 7 | Not Survived | 1 | McCarthy, Mr. Timothy J | male | 54 | 0 | 0 | 51.8625 | Southampton | 49-64 |
| 8 | 8 | Not Survived | 3 | Palsson, Master. Gosta Leonard | male | 2 | 3 | 1 | 21.0750 | Southampton | 0-16 |
| 9 | 9 | Survived | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27 | 0 | 2 | 11.1333 | Southampton | 17-32 |
| 10 | 10 | Survived | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14 | 1 | 0 | 30.0708 | Cherbourg | 0-16 |
| 11 | 11 | Survived | 3 | Sandstrom, Miss. Marguerite Rut | female | 4 | 1 | 1 | 16.7000 | Southampton | 0-16 |
| 12 | 12 | Survived | 1 | Bonnell, Miss. Elizabeth | female | 58 | 0 | 0 | 26.5500 | Southampton | 49-64 |
| 13 | 13 | Not Survived | 3 | Saundercock, Mr. William Henry | male | 20 | 0 | 0 | 8.0500 | Southampton | 17-32 |
| 14 | 14 | Not Survived | 3 | Andersson, Mr. Anders Johan | male | 39 | 1 | 5 | 31.2750 | Southampton | 33-48 |
| 15 | 15 | Not Survived | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 14 | 0 | 0 | 7.8542 | Southampton | 0-16 |
| 16 | 16 | Survived | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 55 | 0 | 0 | 16.0000 | Southampton | 49-64 |
| 17 | 17 | Not Survived | 3 | Rice, Master. Eugene | male | 2 | 4 | 1 | 29.1250 | Queenstown | 0-16 |
| 18 | 18 | Survived | 2 | Williams, Mr. Charles Eugene | male | 30 | 0 | 0 | 13.0000 | Southampton | 17-32 |
| 19 | 19 | Not Survived | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) | female | 31 | 1 | 0 | 18.0000 | Southampton | 17-32 |
| 20 | 20 | Survived | 3 | Masselmani, Mrs. Fatima | female | 30 | 0 | 0 | 7.2250 | Cherbourg | 17-32 |
| 21 | 21 | Not Survived | 2 | Fynney, Mr. Joseph J | male | 35 | 0 | 0 | 26.0000 | Southampton | 33-48 |
| 22 | 22 | Survived | 2 | Beesley, Mr. Lawrence | male | 34 | 0 | 0 | 13.0000 | Southampton | 33-48 |
| 23 | 23 | Survived | 3 | McGowan, Miss. Anna "Annie" | female | 15 | 0 | 0 | 8.0292 | Queenstown | 0-16 |
| 24 | 24 | Survived | 1 | Sloper, Mr. William Thompson | male | 28 | 0 | 0 | 35.5000 | Southampton | 17-32 |
| 25 | 25 | Not Survived | 3 | Palsson, Miss. Torborg Danira | female | 8 | 3 | 1 | 21.0750 | Southampton | 0-16 |
| 26 | 26 | Survived | 3 | Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johanss... | female | 38 | 1 | 5 | 31.3875 | Southampton | 33-48 |
| 27 | 27 | Not Survived | 3 | Emir, Mr. Farred Chehab | male | 30 | 0 | 0 | 7.2250 | Cherbourg | 17-32 |

# 3) Decision Tree

Decision tree works best for the categorical variables hence we will convert two variables SibSp and Parch into categorical variables and add the variables (vectors) to the data.

```
decision_tree<-titanic
SibSpCat= ifelse(decision_tree$SibSp >=3,">=3","<3")
decision_tree <-data.frame(decision_tree,SibSpCat)
decision_tree$SibSpCat <-as.factor(decision_tree$SibSpCat)
ParchCat= ifelse(decision_tree$Parch >=3,">=3","<3")
decision_tree <-data.frame(decision_tree,ParchCat)
decision_tree$ParchCat <-as.factor(decision_tree$ParchCat)
```

2) **Now we have to separate data into "training data" and "testing data".We will use training data to build the tree and test its accuracy using the testing data.**

Following is the code to separate the data into training data and testing data.

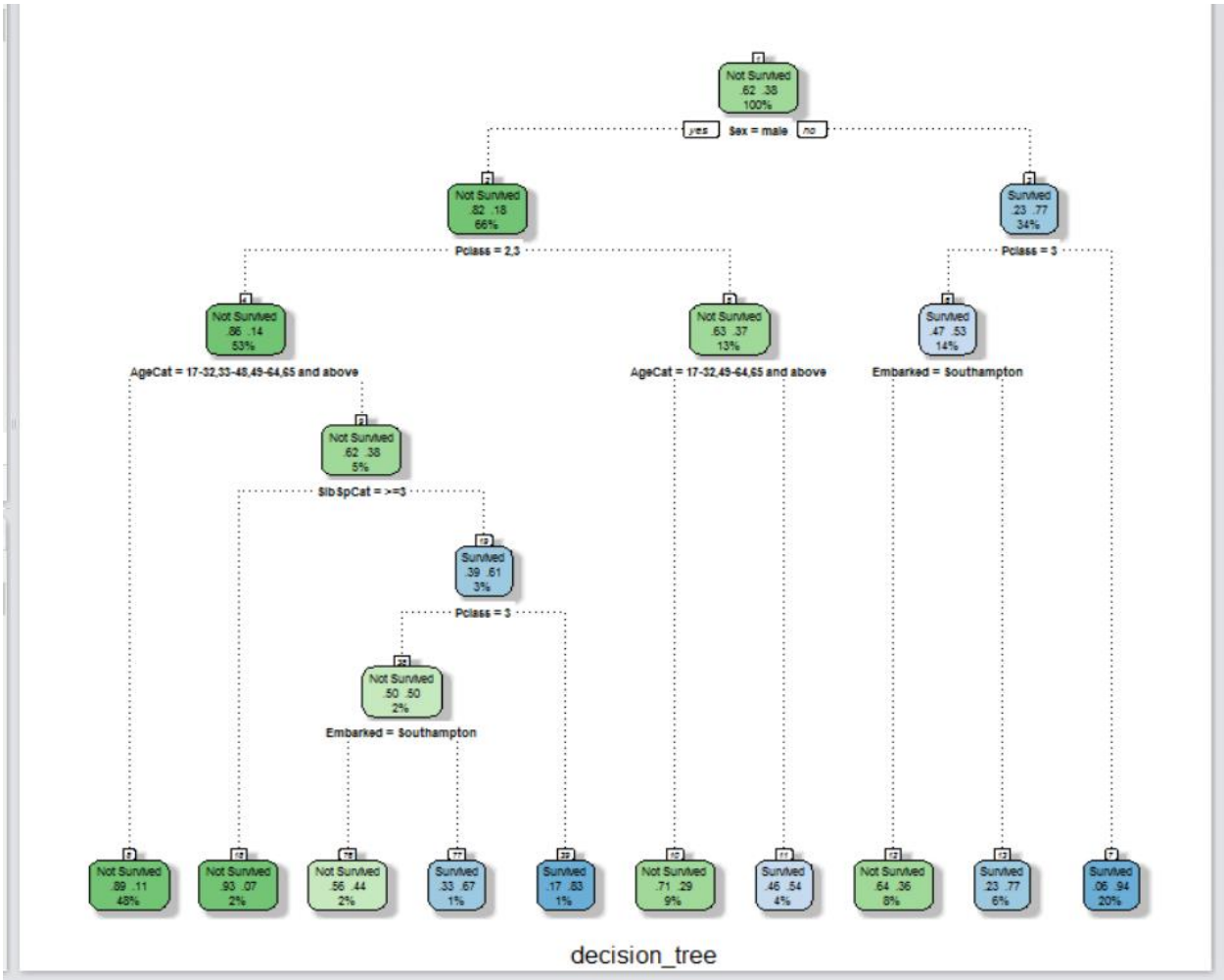
```
set.seed(1)

test = sample(1:nrow(decision_tree),nrow(decision_tree)/3)
train = -test
training_data = decision_tree[train,]
testing_data = decision_tree[test,]
testing_survived = decision_tree$Survived[test]
```

3) **Now the next step would be to build the tree**

Apurva Godghase

```
> tree_model=rpart(Survived ~ Pclass + Sex + AgeCat +Embarked +SibSpCat + ParchCat,data=training_data,method = "class",
control=rpart.control(minsplit=10,cp=0.00))
> fancyRpartPlot(tree_model,sub="decision_tree")
> tree_predict=predict(tree_model,testing_data,type="class")
>
>  mean(tree_predict != testing_survived)
```

Export the tree plot and save it as a PDF with a name **decision_tree_Plot**.




**Analysis:**

**1)What is the misclassification rate for the current tree model(up to 2 decimal places)?**

Answer:

| |
|---|
| *0.21885* |

**2)Which is the first  variable used for splitting?**

Answer:

Apurva Godghase

*Sex*

**3)What is the ratio of Survived: Not survived initially?**

Answer: 38:62

**4)What is the ratio of the Survived: not survived of Females?**

Answer: 77:23

**5)What is the ratio of Survived: Not Survived of the males who are from Pclass 1 ?**

Answer: 37:63

**6) Please list the top 6 variables from your decision tree in the order of importance**

Answer: **Survival, Sex,PClass,AgeCat,SibSpCat,Embarked**

# 4) K-means Clustering

The arguments with high probability in the Decision Tree result can be selected for the k-means clustering analysis to ensure the variables (vectors) that contribute to a higher survivability rate are selected for your clustering analysis.

Q1 ) Based on the decision tree from Part 3, focusing on the key decision variables (as obtained in Section 3, Answer 6) would be a good approach

Now that we know what variables to focus on from the above answer for our clustering, let's turn to the next part. Since the variables (vectors) selected in the above step are categorical variables and clustering requires continuous values, we will convert the below mentioned categorical variables to continuous

1) Sex
2) Embarked
3) Survived

Execute the below command to convert the above 3 categorical variables into continuous. The below code would create 3 new variables of type continuous and save them in the new file "titanicUpdated.csv".

```
titanicNew<-read.csv("E:/titanicNew.csv")
titanicUpdated<-titanicNew
SurvivedNum<-ifelse(titanicUpdated$Survived=="Not Survived",0,1)
titanicUpdated <-data.frame(titanicUpdated,SurvivedNum)

SexN<-ifelse(titanicUpdated $Sex=="male",1,0)
titanicUpdated <-data.frame(titanicUpdated, SexN)
```

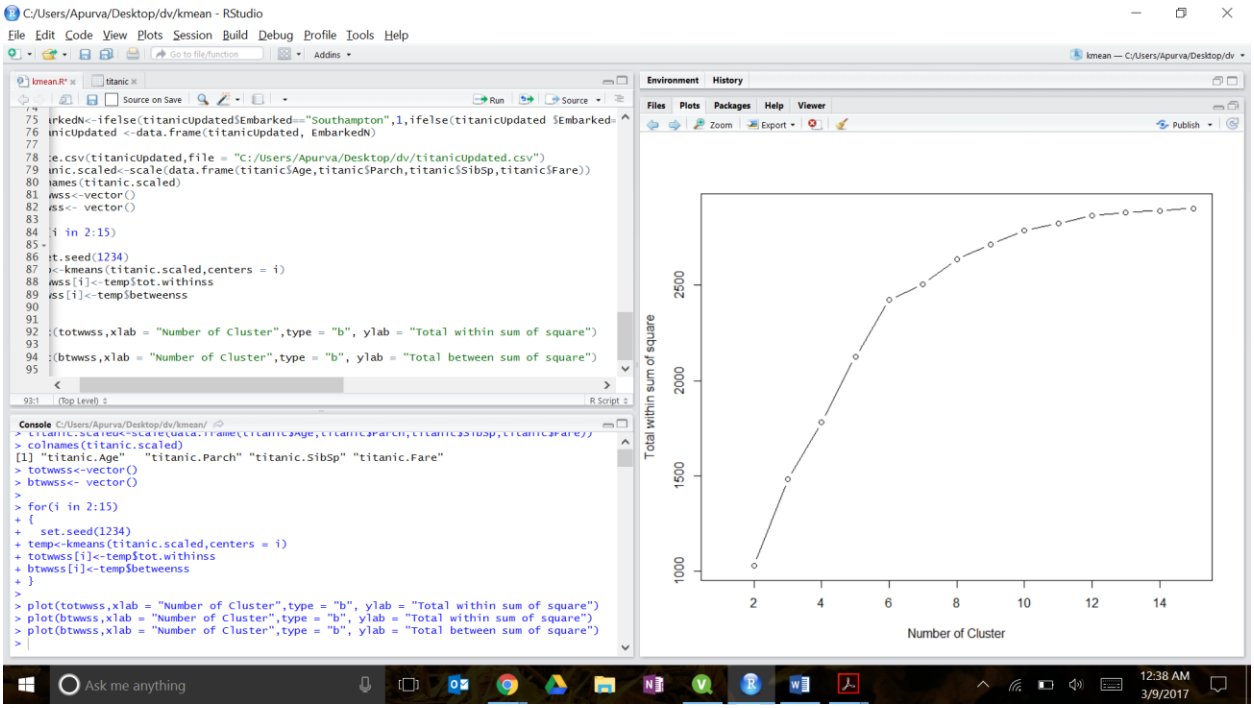
```
EmbarkedN<-ifelse(titanicUpdated$Embarked=="Southampton",1,ifelse(titanicUpdated
$Embarked=="Cherbourg",2,0))
titanicUpdated <-data.frame(titanicUpdated, EmbarkedN)

write.csv(titanicUpdated,file = "E:/titanicUpdated.csv")
```

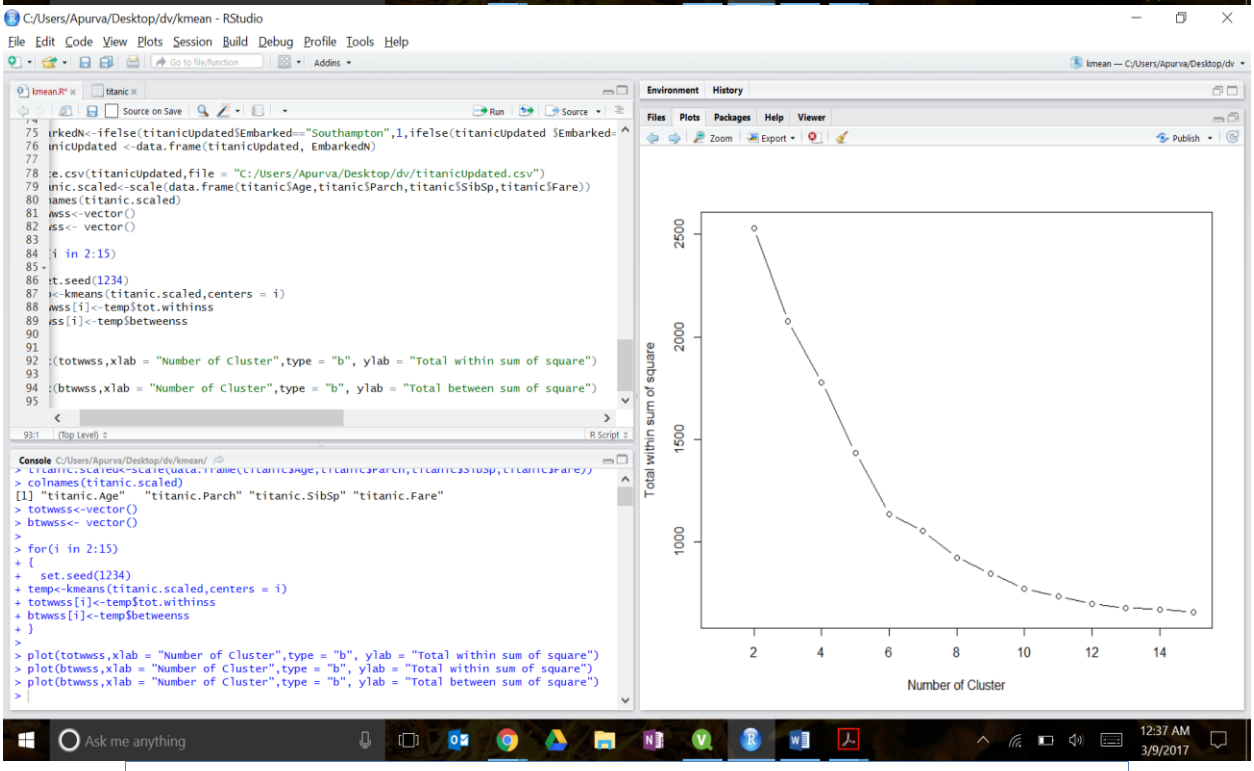
***Before the cluster analysis one has to select for the optimum number of clusters.***

2) Use the titanic dataset that we used in Section 3, for answering Q1, Q2 and Q3. Execute the below command to normalize the data and find the total number of clusters required.

```
titanic.scaled<-scale(data.frame(titanic$Age,titanic$Parch,titanic$SibSp,
titanic$Fare))
colnames(titanic.scaled)
totwss<-vector()
btwss<-vector()
for(i in 2:15)
{
  set.seed(1234)
  temp<-kmeans(titanic.scaled,centers=i)
  totwss[i]<-temp$tot.withinss
  btwss[i]<-temp$betweenss
}

plot(totwss,xlab="Number of Cluster",type="b",
ylab="Total Within Sum of Square")

plot(btwss,xlab="Number of Cluster",type="b",
     ylab="Total Between Sum of Square")
```

We need to keep adding clusters to the point where further addition of cluster won't do much of explanation of the variation. This is also the point where the slope of the curve changes suddenly and gives an angle to the graph. So we choose 6 clusters.




Now that ***we know what the optimum number of clusters to be used in our integrated analysis what clustering variables are important***? In the next section we will perform visual analysis integrating Tableau with R to visualize the clusters in Tableau.

Apurva Godghase

# 5) TABLEAU / R INTEGRATION

**Step 1: Initiate Rserve and Making Connection**

For us to be able to make a connection between Tableau and R, we need to do it through Rserve. Rserve itself is the server that is a program that responds to requests from clients. It listens for any incoming connections and processes incoming requests.

```
install.Packages("Rserve")
library(Rserve)
Rserve()
```

**Step 2: Open the titanicUpdated.csv file on Tableau**

**Step 3: Forming Clusters**

The variables “EmbarkedN”, “SexN” and “SurvivedNum” show up under the Measures column along with Age, Fare, Parch, Pclass etc.

**Parameter Creation:**
Created Paramers named “# of Clusters” and “*Seed” and set the value as 1234*

**Calculated Field creation:**

**********
```
Script_INT("
## Sets the seed
set.seed(.arg8[1])

## Studentizes the variables

age<-(.arg1-mean(.arg1))/sd(.arg1)
pclass<-(.arg2-mean(.arg2))/sd(.arg2)
embarkedn<-(.arg3-mean(.arg3))/sd(.arg3)
sex<-(.arg4-mean(.arg4))/sd(.arg4)
survived<-(.arg5-mean(.arg5))/sd(.arg5)
sibsp<-(.arg6-mean(.arg6))/sd(.arg6)

dat<-cbind(age,pclass,embarkedn,sex,survived,sibsp)

num<-.arg7[1]

## Creates the clusters
```

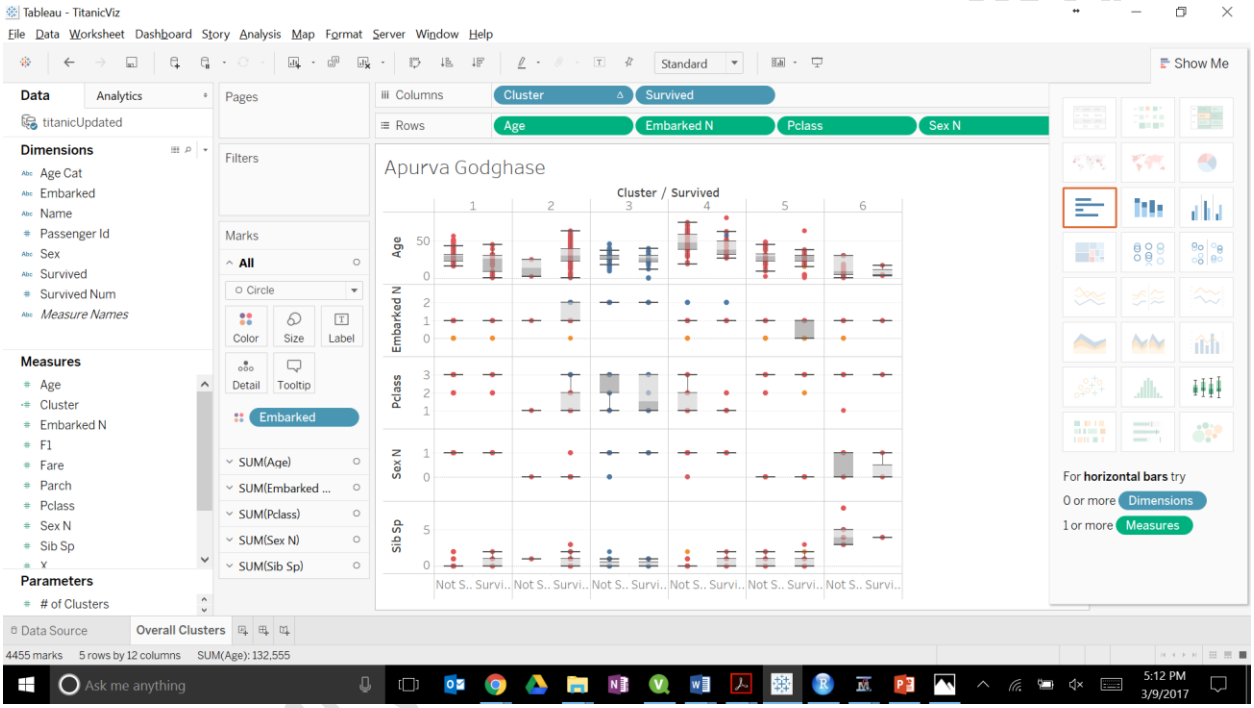
kmeans(dat,num)$cluster

",
SUM([Age]),SUM([Pclass]),SUM([EmbarkedN]),SUM([SexN]),SUM([SurvivedNum]),SUM([Sib Sp]),[# of Clusters],[Seed])

---

**Step 4: Visualization of Key variables across Clusters**

Objective: Visualize how different variables vary across different cluster.



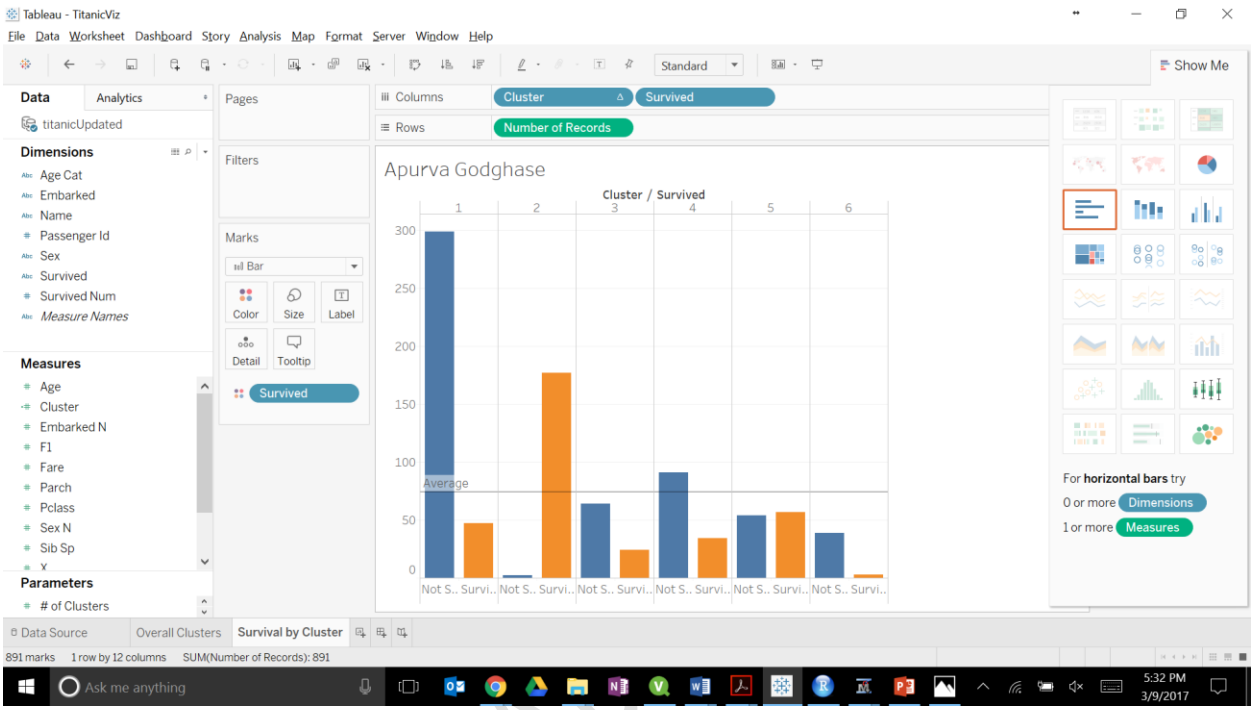

**Insights**

There are many passengers from age group 8-32 who survived and a large portion of them are from SouthHampton. There are many passengers from class 1 and 2 of cluster 4 who did not survive. Also, for the same cluster, there are large number of passengers having age between 38 and 59 who did not survive.

**Key Takeaway:** The key takeaway from this worksheet is that Sex is the most important variable in the formation of cluster as most of the clusters have predominantly one gender. This analysis is in line with our findings from the decision tree.

**Step 5: Survival by Cluster**

Objective: The objective of this sheet is to see which clusters (out of the 6) have highest survivability.

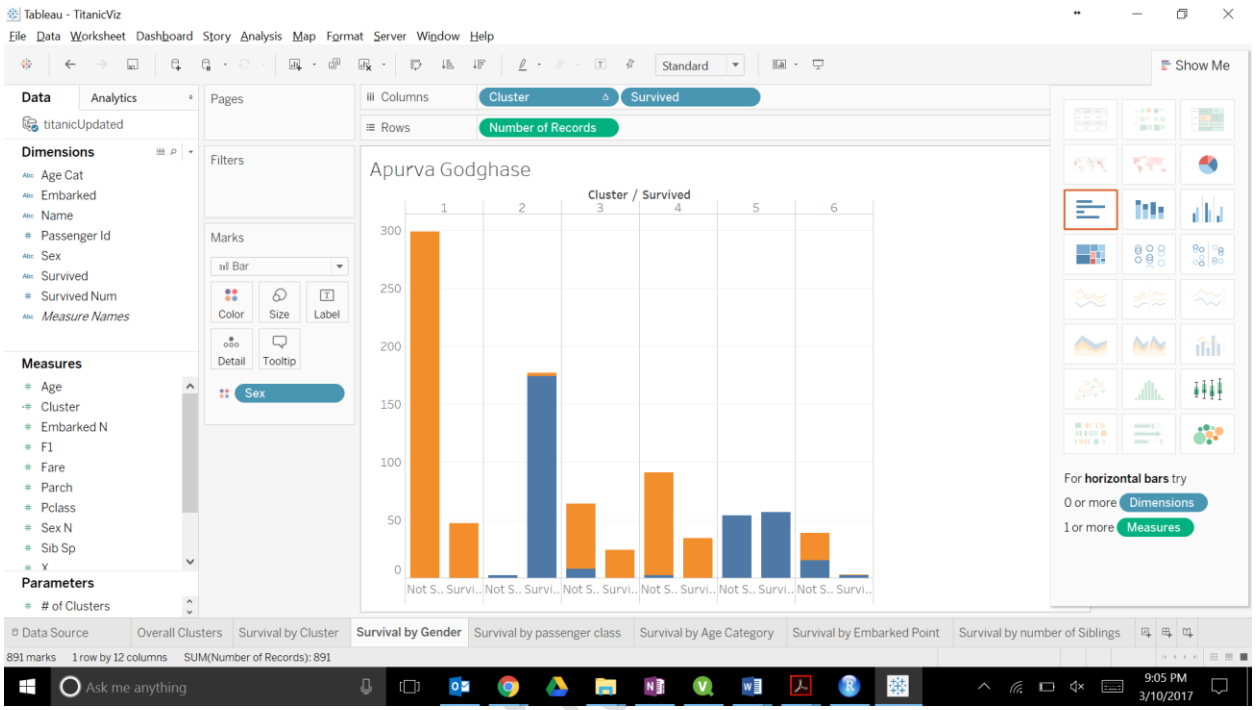



Cluster 2 and 5 has the highest survivability

**Key Takeaway:** The key takeaway from the above worksheet "**Survival by Cluster**" is that **only in case of two clusters**, the number of passengers survived either outweigh or compare to the number of passengers who didn't survive. Hence these clusters form the most important cluster for our analysis.

## Step 6: Survival by Gender

Objective: The objective of this section is to understand which is the best gender from a survivability perspective, in each of our top two clusters, identified in the previous step.




The ideal Gender (that has the best chance to survive) in your top two cluster.

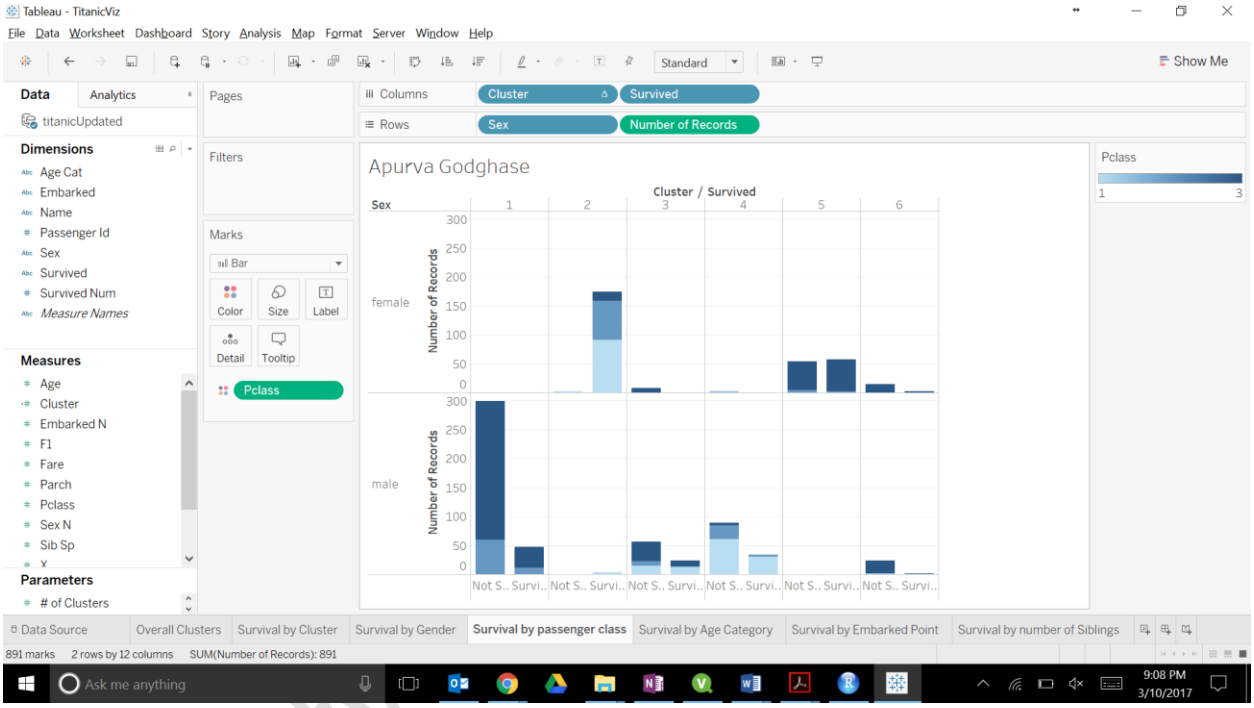
| | Cluster 2 | Cluster 5 |
|---|---|---|
| Ideal Gender | Female | Female |
| Ideal Passenger Class | | |
| Ideal Age Category | | |
| Ideal Embarked point | | |

**Key Takeaway:** The key takeaway from the above worksheet "**Survival by Gender**" is that in the two prominent clusters that we identified in previous steps, one gender clearly dominates the survivability. Note that cluster 1 is comprised of only male passengers.

**Step 7: Survival by Passenger Class**

Objective: The objective of this section is to understand which is the best gender/class combination from a survivability perspective, in each of our top two clusters.

| | Cluster 2 | Cluster 5 |
|---|---|---|
| Ideal Gender | Female | Female |
| Ideal Passenger Class | Class 1 | Class 3 |
| Ideal Age Category | | |
| Ideal Embarked point | | |
| Ideal number of siblings | | |

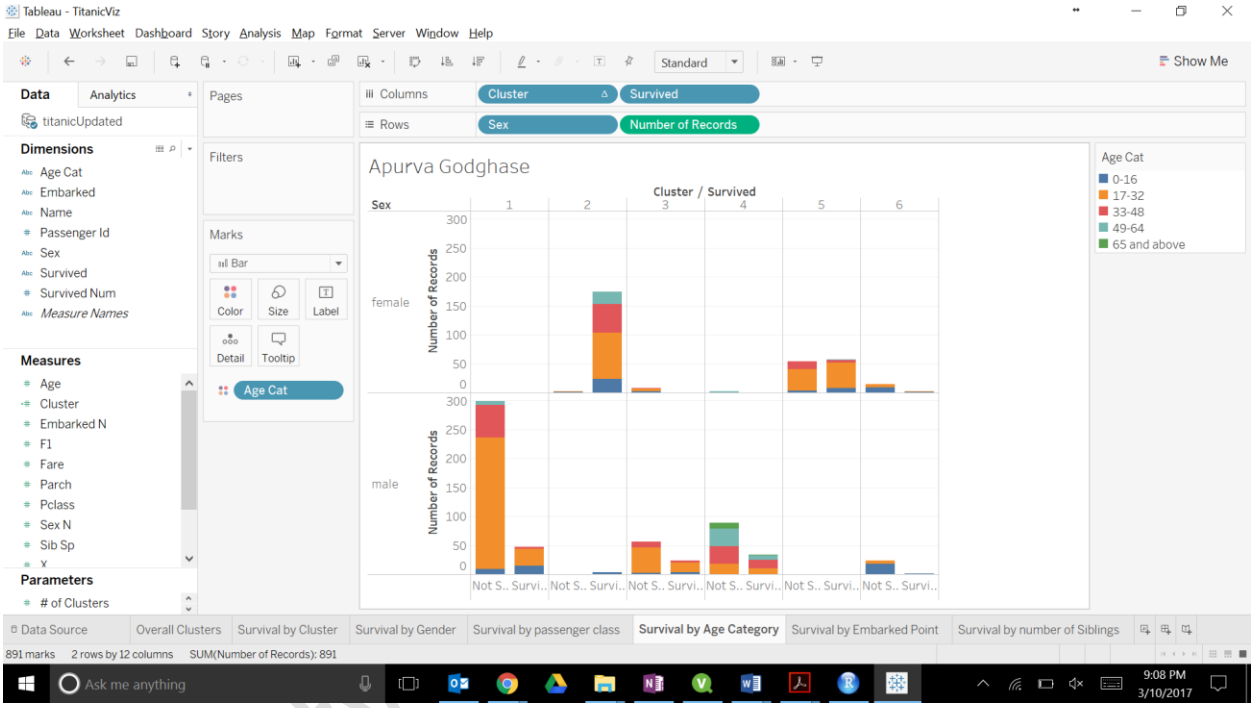



**Key Takeaway:** The key takeaway from the above worksheet "**Survival by Passenger class**" is that in some key clusters, most of the survivors belonged to a particular class.

**Step 8: Survival by Age Category**

Objective: The objective of this section is to understand which is the best gender/age category combination from a survivability perspective, in each of our top two clusters.

| | Cluster 2 | Cluster 5 |
|---|---|---|
| Ideal Gender | Female | Female |
| Ideal Passenger Class | Class 1 | Class 3 |
| Ideal Age Category | 17-32 | 17-32 |
| Ideal Embarked point | | |
| Ideal number of siblings | | |

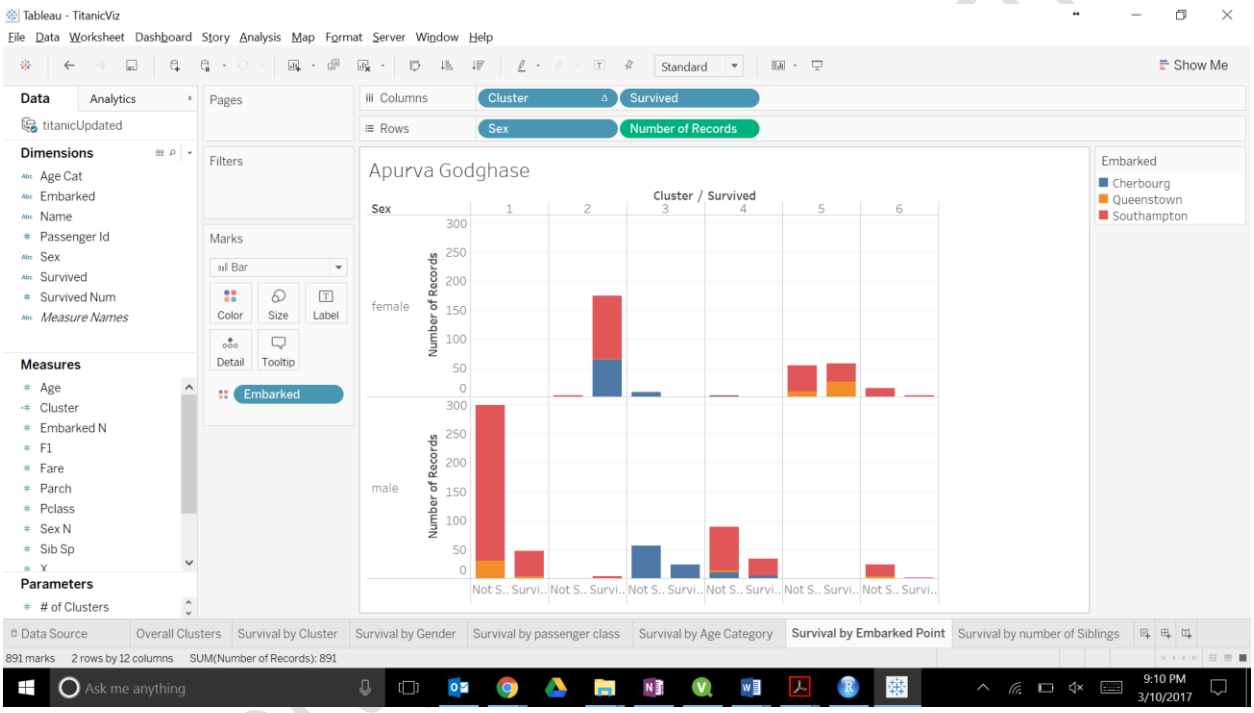



**Key Takeaway:** Most of the female survivors were from the age category 17-32 years. Most men who survived in cluster 1 also belonged to the age category 17-32 years. Overall, we can say that because majority of the passengers seem to fall it the age group 17-32 years their survivability salience is high.

## Step 9: Survival by Embarked Point

Objective: The objective of this section is to understand which is the best gender/embarked point combination from a survivability perspective, in each of the top two clusters.

| | <Cluster 1> | <Cluster 2> |
|---|---|---|
| Ideal Gender | Female | Female |
| Ideal Passenger Class | Class 1 | Class 3 |
| Ideal Age Category | 17-32 | 17-32 |
| Ideal Embarked point | Southhampton | Southhampton |
| Ideal number of siblings | | |

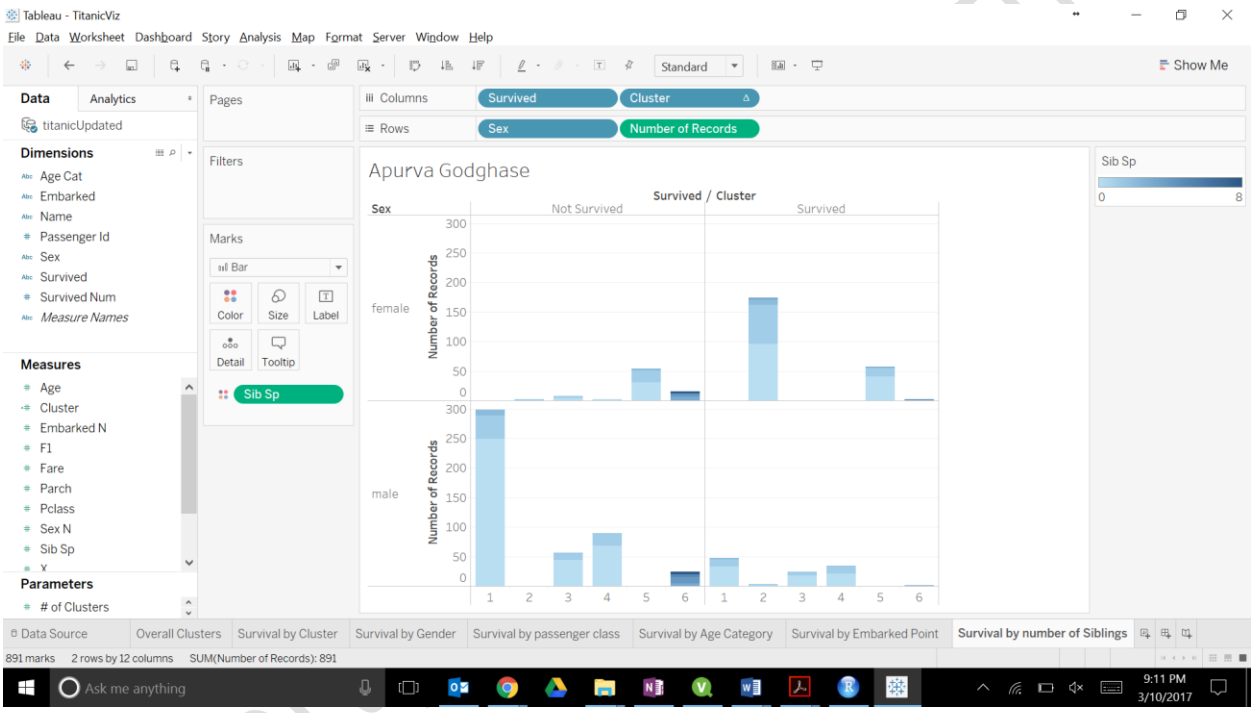



**Key Takeaway:** Most of the passengers embarked from Southampton. Amongst the survivors, majority of the passengers were from Southampton.

## Step 10: Survival by number of Siblings

Objective: The objective of this section is to understand which is the best gender/# of siblings combination from a survivability perspective, in each of the top two clusters.

| | <Cluster 1> | <Cluster 2> |
|---|---|---|
| Ideal Gender | Female | Female |
| Ideal Passenger Class | Class 1 | Class 3 |
| Ideal Age Category | 17-32 | 17-32 |
| Ideal Embarked point | Southhampton | Southhampton |
| Ideal number of siblings | 0 | 0 |




**Key Takeaway:** Majority of the passengers had zero siblings. But within cluster 6, for both male and female passengers, we see a mix of "# of siblings". These however constitutes a very small percentage of the total # of sibling count.

## Step 11: Summary

For Cluster 2,Females having 0 siblings, and of passenger Class 1, age Category between 17-32 ,having Embarked point Southhampton have best chances of survival.

And Cluster 5, Females having 0 Siblings, and of Passenger Class 3 with the ideal age category 17-32 having Embarked point "Southhampton" have best chances of survival