

# Exercise 7: Integrated Analysis - Decision Tree and K-means Clustering using Tableau & R

**Student Name: Ravi Dawar**

**Date: 3/2/2018**

## 1) Data Retrieval

Q9) Write the output of the command: `dim(titanic)`

**Answer:** 891 12 which means 891 Rows and 12 Columns

## 2) Data Pre-processing

Q2) What is the average age that you get?

**Answer 2:** 29.69912

**Answer 3:**

The screenshot displays the RStudio interface with the Titanic dataset loaded. The Environment pane shows the `titanic` data frame with 891 observations and 11 variables. The console shows the following R code and its output:

```
D:\Data Visualization\Assignments\Assignment 7\kmean - RStudio
> dim(titanic)
[1] 891 12
> titanic$agecat[is.na(titanic$age)] <- "0-10"
> 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$age<=67] <- "49-64"
> titanic$agecat[titanic$age>=68 & "65 and Above"
> titanic$survived[titanic$survived==0] <- "Not Survived"
> titanic$survived[titanic$survived==1] <- "Survived"
> titanic$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"] <- "Queens town"
> titanic <- titanic[c(-9,-11)]
> View(titanic)
```

The Environment pane shows the `titanic` data frame with 891 observations and 11 variables. The Values pane shows the mean age: 29.69912.

The R Documentation pane shows the `character` (Base) and `Character Vectors` section, describing the `as.character` function.

Dr/Data Visualization/Assignments/Assignment 7/kmean - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Environment History Connections

Data  
titanic 891 obs. of 11 variables  
Values  
meanAge 29.6991176470588

Files Plots Packages Help Viewer

R Character Vectors  
Find in Topic  
character (base)

R Documentation

Character Vectors

Description  
Create or test for objects of type "character".

Usage  
character(length = 0)  
as.character(x, ...)  
is.character(x)

Arguments  
length A non-negative integer specifying the desired length. Double values will be coerced to integer; supplying an argument of length other than one is an error.  
x object to be coerced or tested.  
... further arguments passed to or from other methods.

Details  
as.character and is.character are generic; you can write methods to handle specific classes of objects; see [help\("methods"\)](#). Further, for as.character, the default method calls as.vector, so dispatch is first on methods for as.character and then for methods for as.vector.  
as.character represents real and complex numbers to 15 significant digits (technically the compiler's setting of the ISO C constant DBL\_D10).

Showing 1 to 36 of 891 entries

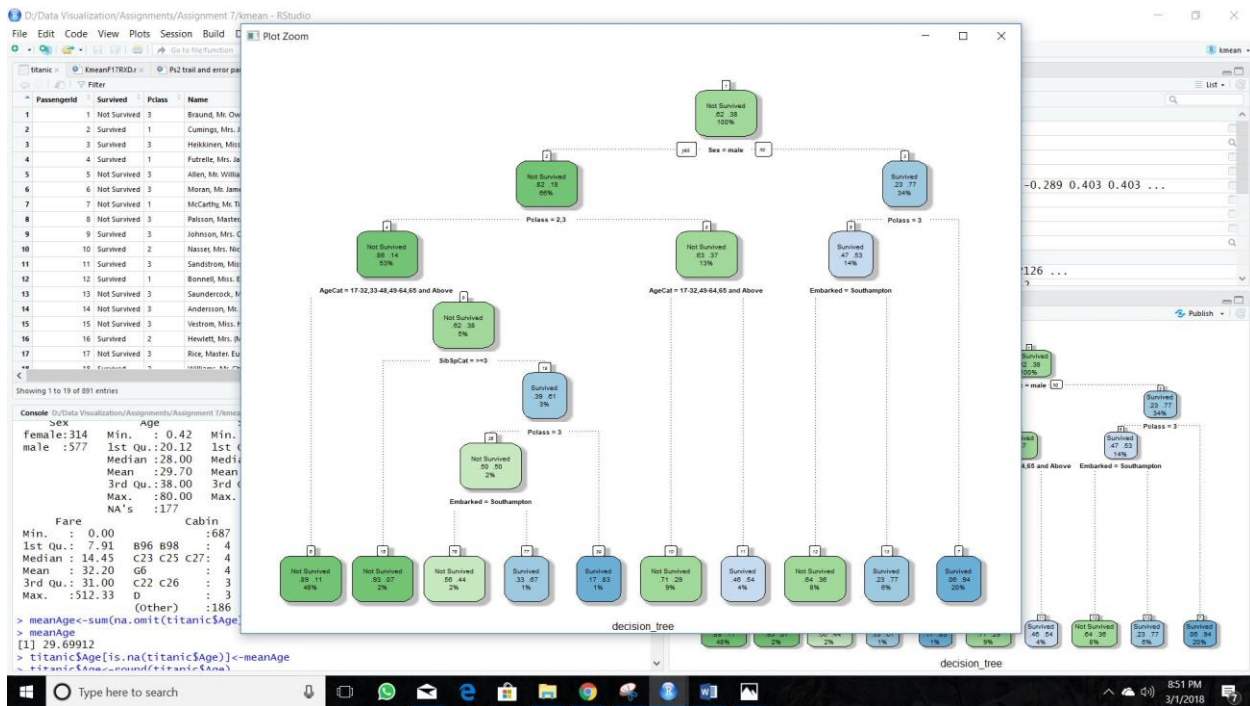
Console

Type here to search

### 3) Decision Tree

Q4) Screenshot of decision Tree is below

**Answer:**



Question:-

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

**Answer:** 0.21

2)Which is the first variable used for splitting?

**Answer:** 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:**

1. Sex
2. PClass
3. AgeCat
4. Embarked
5. SibSpcat
6. ParchCat

## 4) K-means Clustering

Q1 ) Based on the decision tree from Part 3) of the exercise, do you feel that using all the variables to cluster the passengers would be a good strategy or focusing on the key decision variables (as obtained in Section 3, Answer 6) would be a good approach? Select one of the two choices below.

**Answer 1)** Considering all the key variables, would be a good approach for clustering passengers.

3) Write the desired number of cluster you selected for the analysis

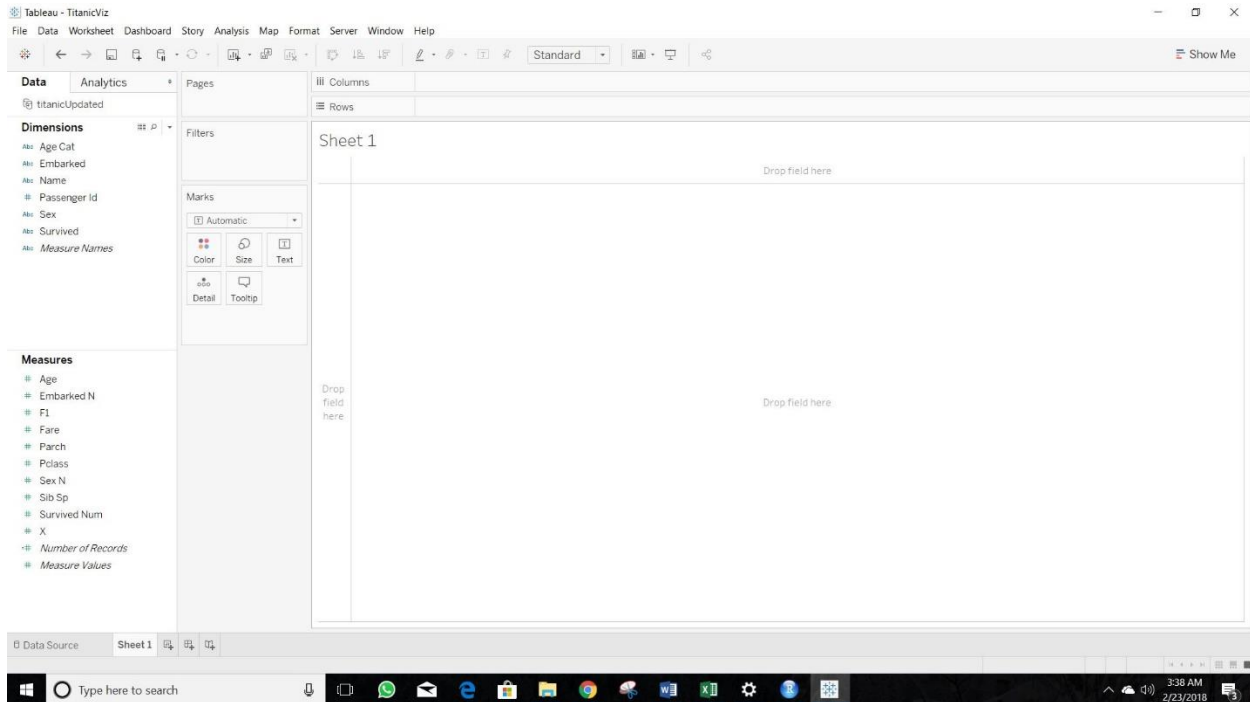
**Answer:** 6

#### 4) Paste the screenshot of the two plot

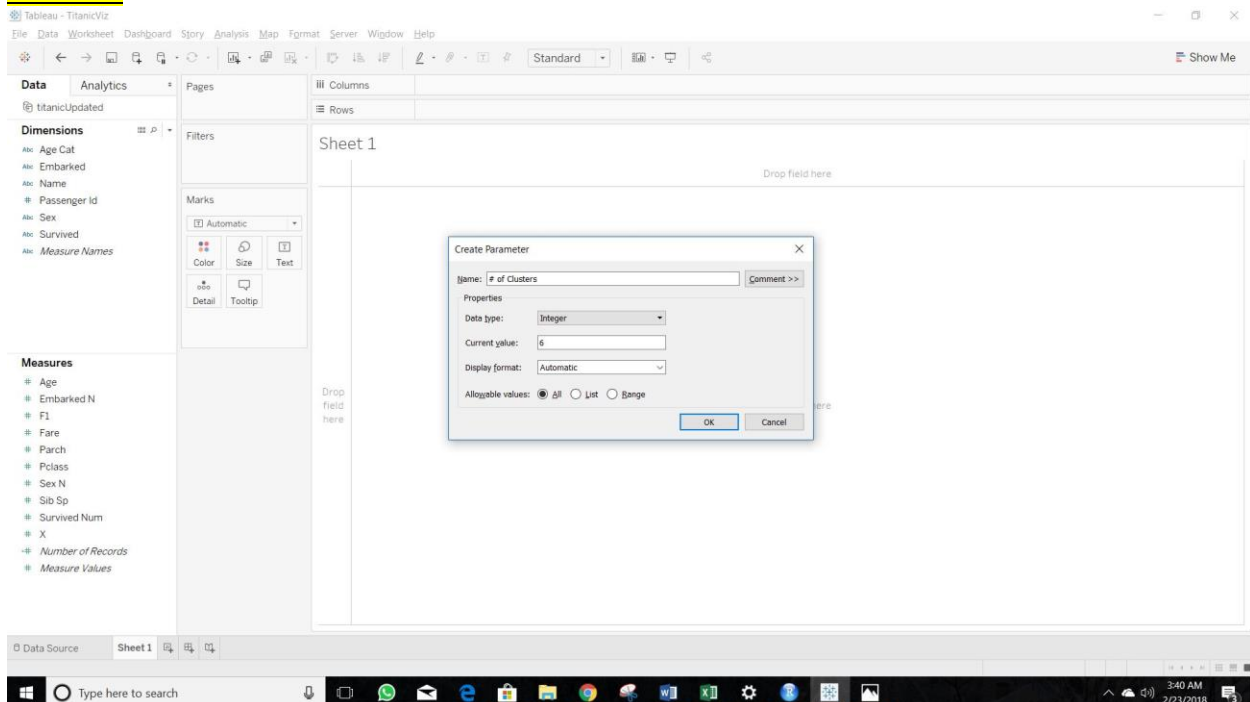


## 5) TABLEAU / R INTEGRATION

### Answer 5:



### Answer 6:





## Answer 7:

The screenshot shows the Tableau Desktop interface with a 'Create Parameter' dialog box open. The dialog box has the following fields and options:

- Name:** Seed
- Properties:**
  - Data type:** Integer
  - Current value:** 1,234
  - Display format:** Automatic
- Allowable values:**
  - ☒ All
  - ☐ List
  - ☐ Range

The background shows the Tableau interface with 'Sheet 1' and various panes like Dimensions, Measures, and Parameters.

## Answer 8:

The screenshot shows the Tableau Desktop interface with a dashboard titled 'Overall Clusters - Ravi Dawar'. The dashboard displays a grid of 30 small charts (5 rows by 6 columns) showing the distribution of various measures across different clusters and survival status. The columns are labeled 'Cluster / Survived' and the rows are labeled 'Sex N', 'Age', 'Pclass', 'Embarked N', and 'Sib Sp'. The legend indicates three clusters: Cherbourg (purple), Queenstown (yellow), and Southampton (green). The dashboard also shows a 'Seed' parameter value of 1,234.

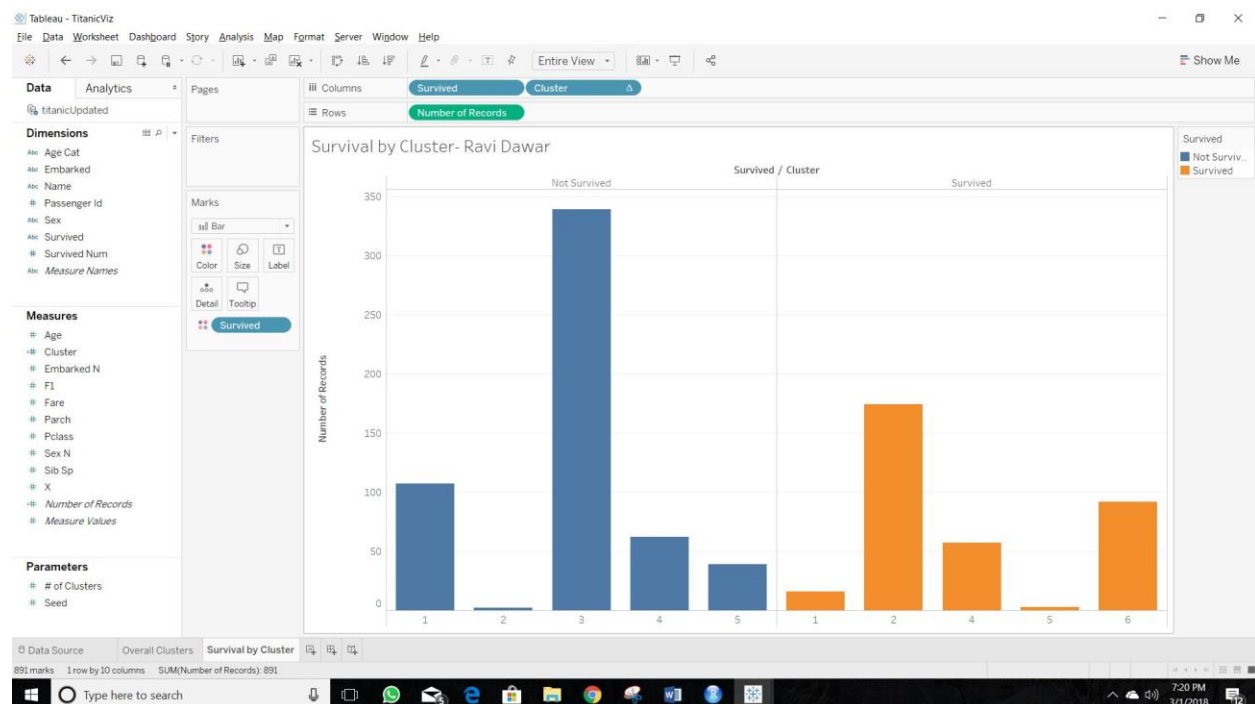
### Answer 9:

On analyzing Cluster 1 we can infer that all male passenger survived. No passenger survived below age 40 and above age 18 and those belonging from Pclass 3. It has passengers who embarked for all three Cities but passengers who embarked for Queenstown did not survive. Maximum number of siblings of passengers are 2 in this cluster.

### Answer 10:

On analyzing Cluster 4 we can infer that there are only female passengers. Passengers' age range is from 0 to 48 with outlier 1 and 63. No passenger survived below age 15 and above age 2. There was no passenger in Pclass 1. It has passengers who embarked for all three cities but passengers who embarked for Cherbourg did not survive. Maximum number of siblings of passengers who survived are 3 in this cluster.

### Answer 11:



**Answer 12:** Cluster 2 and Cluster 6 have highest survivability.

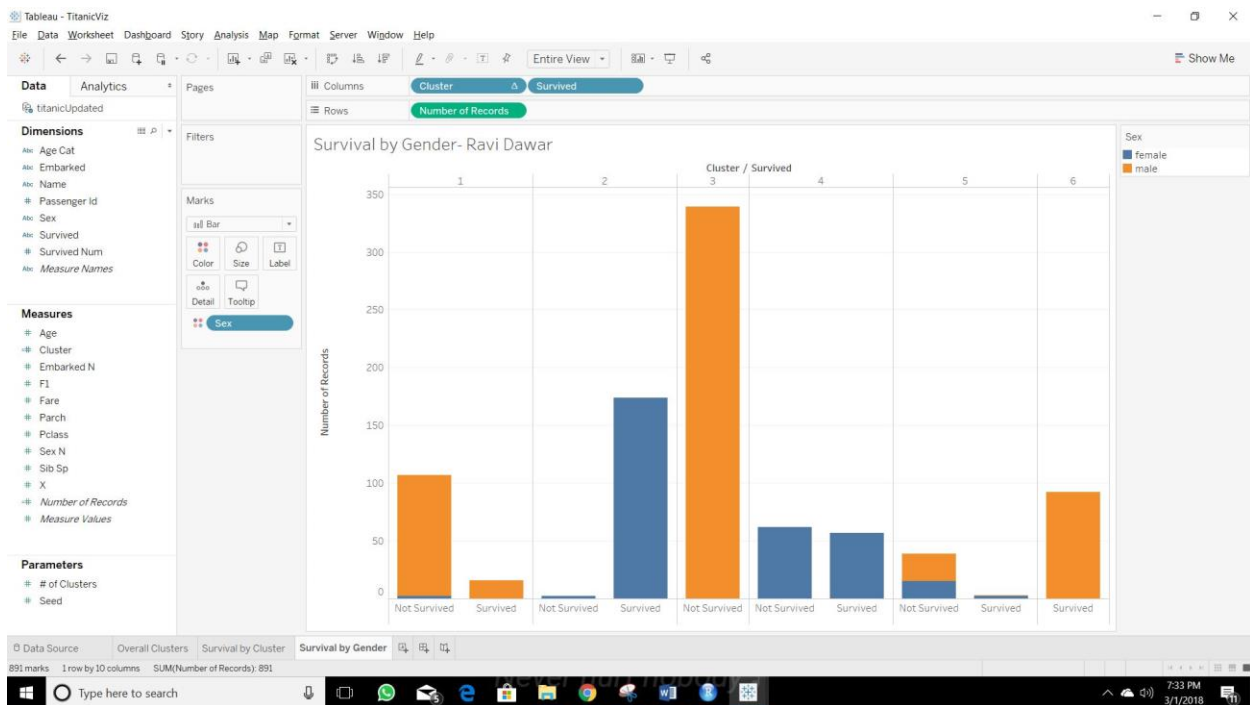
### Answer 13:

	Cluster 2	Cluster 6
Ideal Gender		
Ideal Passenger Class		
Ideal Age Category		
Ideal Embarked point		
Ideal number of siblings		

**Answer 14:**

	Cluster 2	Cluster 6
Ideal Gender	Female	Male
Ideal Passenger Class		
Ideal Age Category		
Ideal Embarked point		
Ideal number of siblings		

**Answer 15 :**



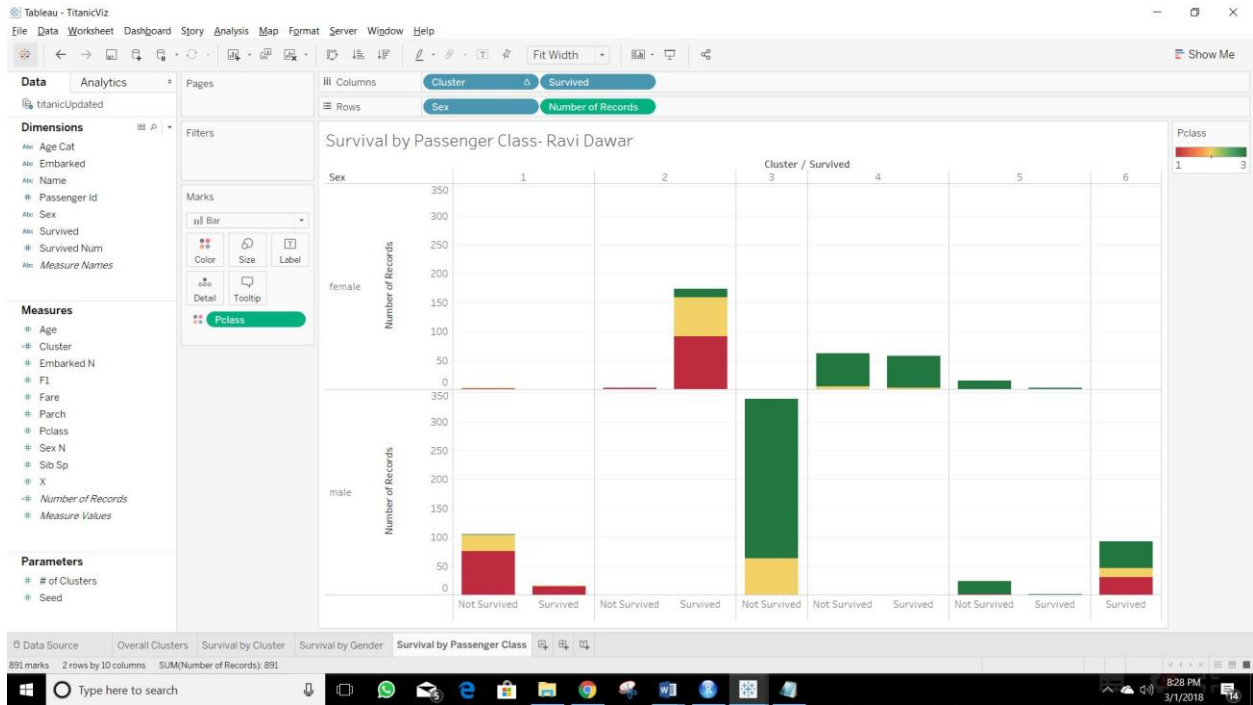
**Answer 16:** Cluster 2

**Answer 17:**

	Cluster 2	Cluster 6
Ideal Gender	Female	Male
Ideal Passenger Class	PClass 1	PClass 3
Ideal Age Category		
Ideal Embarked point		
Ideal number of siblings		



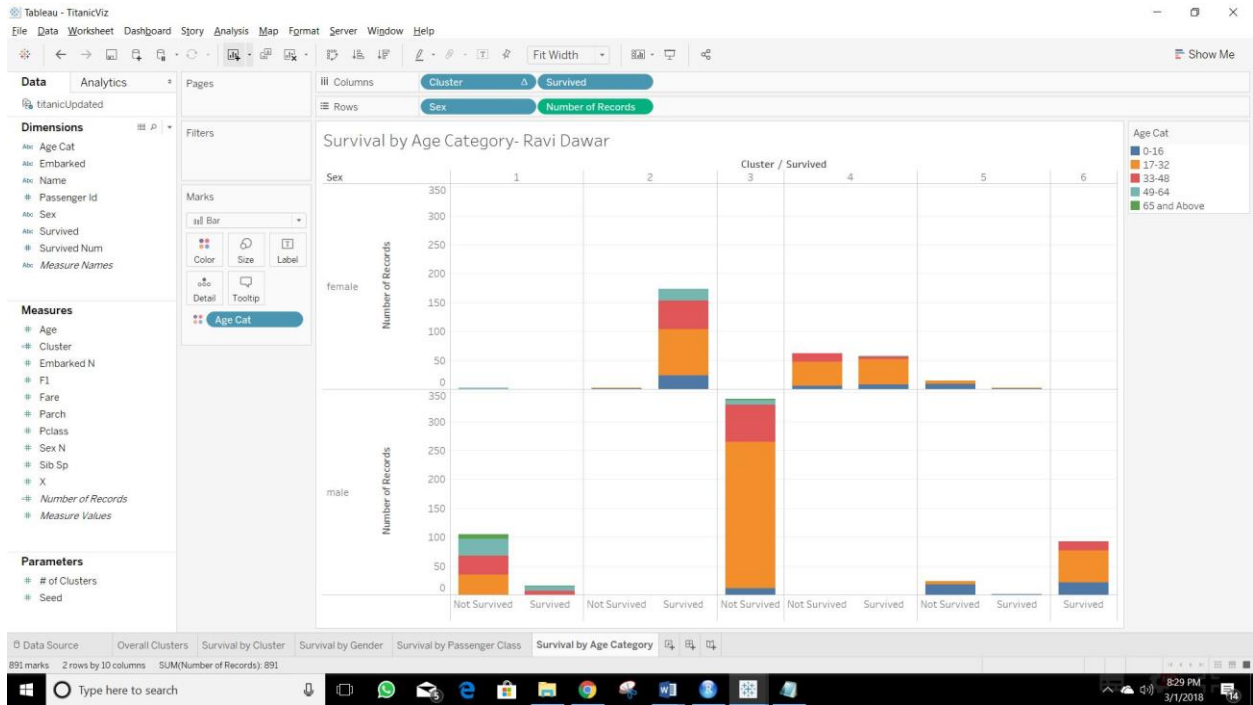
### Answer 18:



### Answer 19:

	Cluster 2	Cluster 6
Ideal Gender	Female	Male
Ideal Passenger Class		
Ideal Age Category	17-32	17-32
Ideal Embarked point		
Ideal number of siblings		

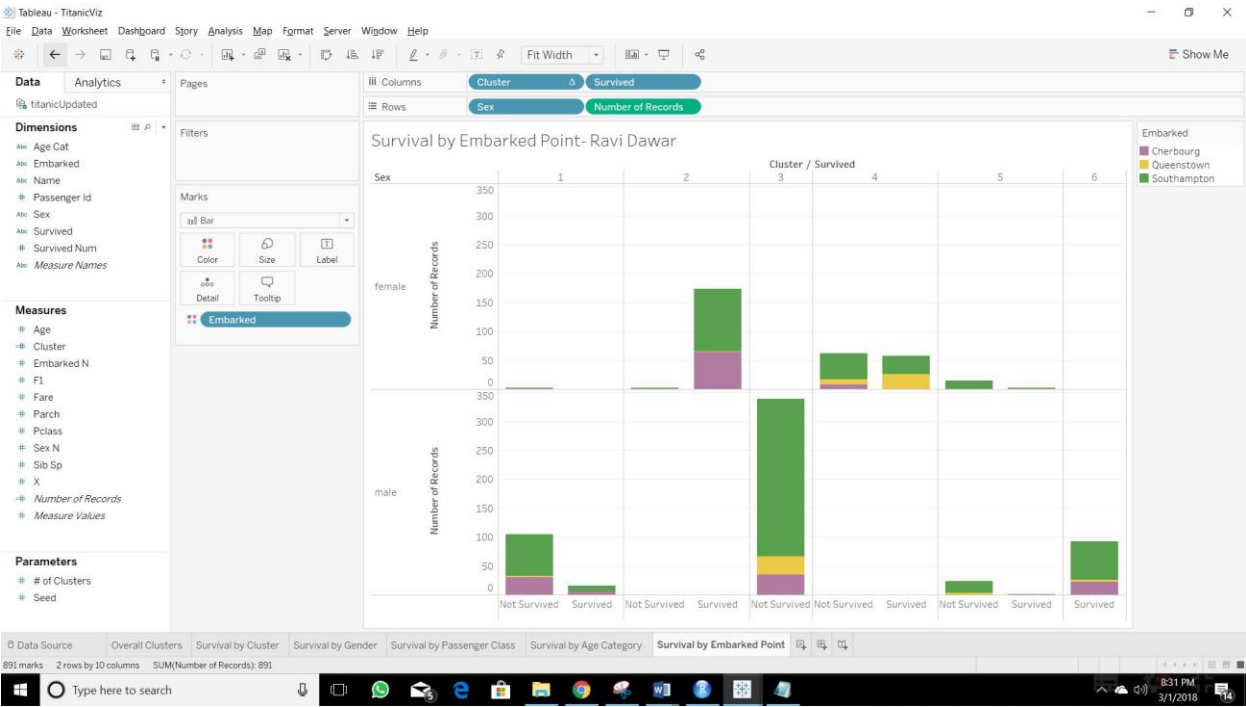
## Answer 20:



## Answer 21 :

	Cluster 2	Cluster 6
Ideal Gender	Female	Male
Ideal Passenger Class		
Ideal Age Category		
Ideal Embarked point	Southampton	Southampton
Ideal number of siblings		

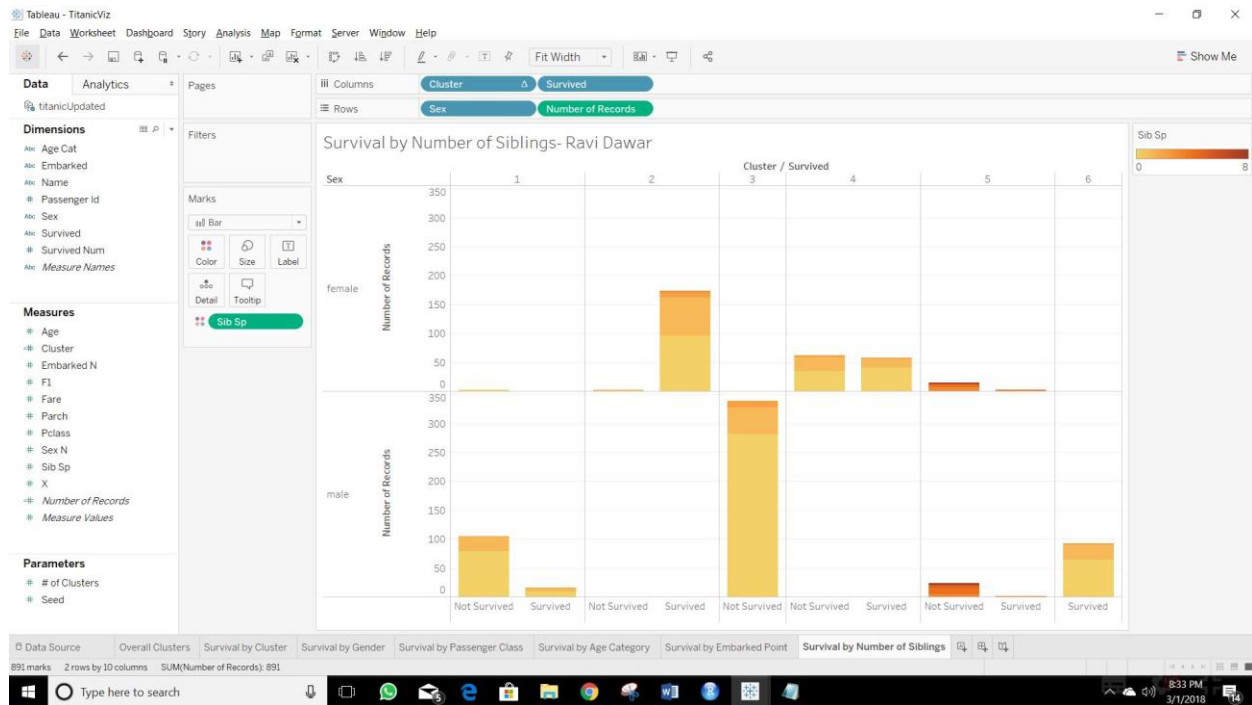
Answer 22:



Answer 23:

	Cluster 2	Cluster 6
Ideal Gender	Female	Male
Ideal Passenger Class		
Ideal Age Category		
Ideal Embarked point		
Ideal number of siblings	0	0

## Answer 24 :



## Answer 22:

Top two clusters which have the best chance of survivability in comparison to other cluster are:

**Cluster 2:** Ideal passenger profile is Female passenger which belongs to age range between 17 to 32, travelling in Pclass 1 who embarked from Southampton with no siblings.

**Cluster 6:** Ideal passenger profile is Male passenger which belongs to age range between 17 to 32, travelling in Pclass 3 who embarked from Southampton with no siblings.