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

**Student Name: Student Id:**

**Date:**

**Note: You must use the following conventions to name objects/systems created in this exercise.**

**Objective:-**  The objective of this exercise is to get a hands on experience on Decision Tree and K-means clustering analysis using R and visualizing results through Tableau.

Below are the objectives of the exercise

1)Data retrieval

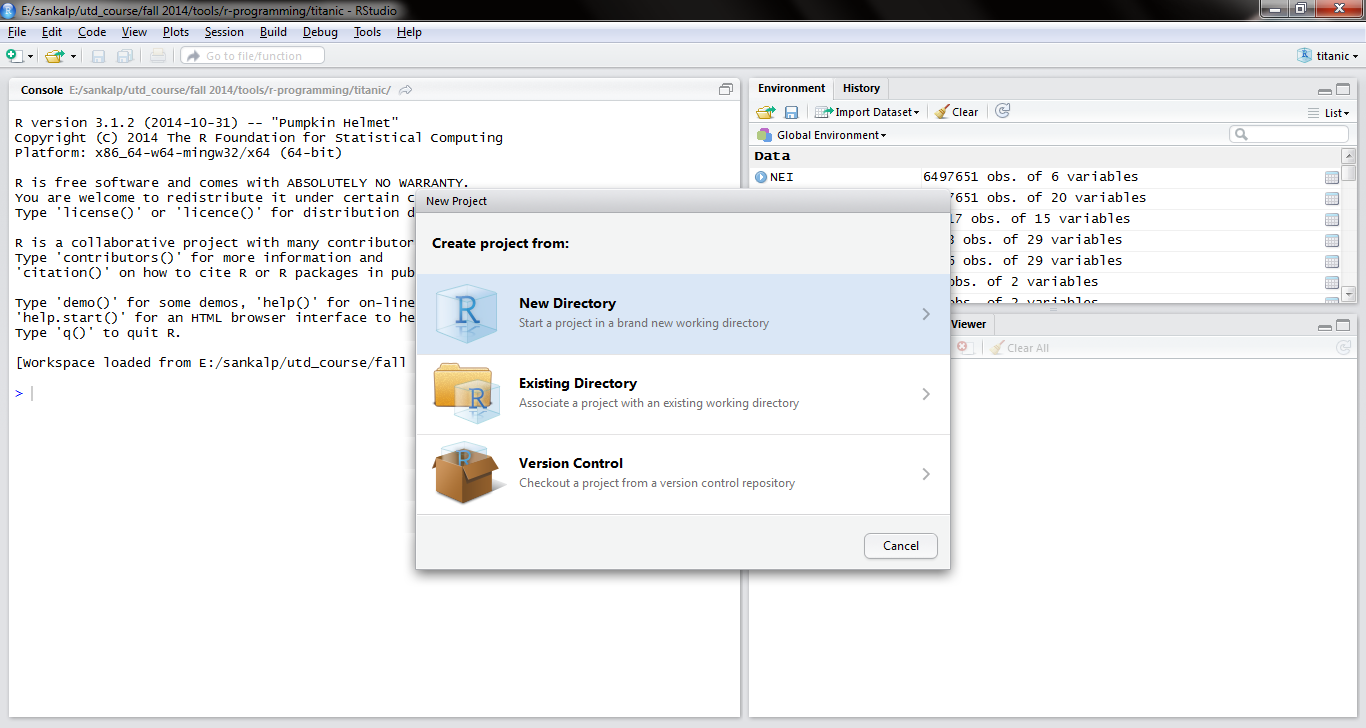
2)Data pre-processing

3) Decision Tree using R

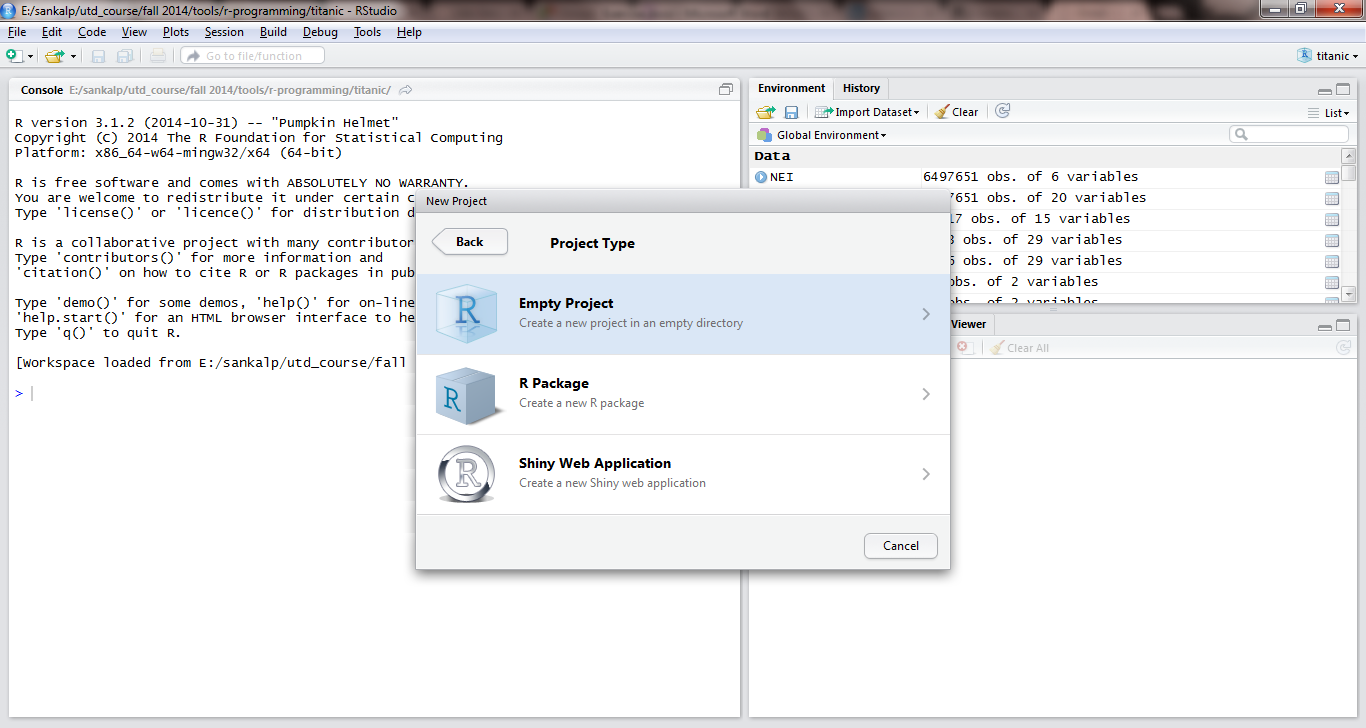
4) K-mean clustering using Tableau- R integration by invoking Rserve ()

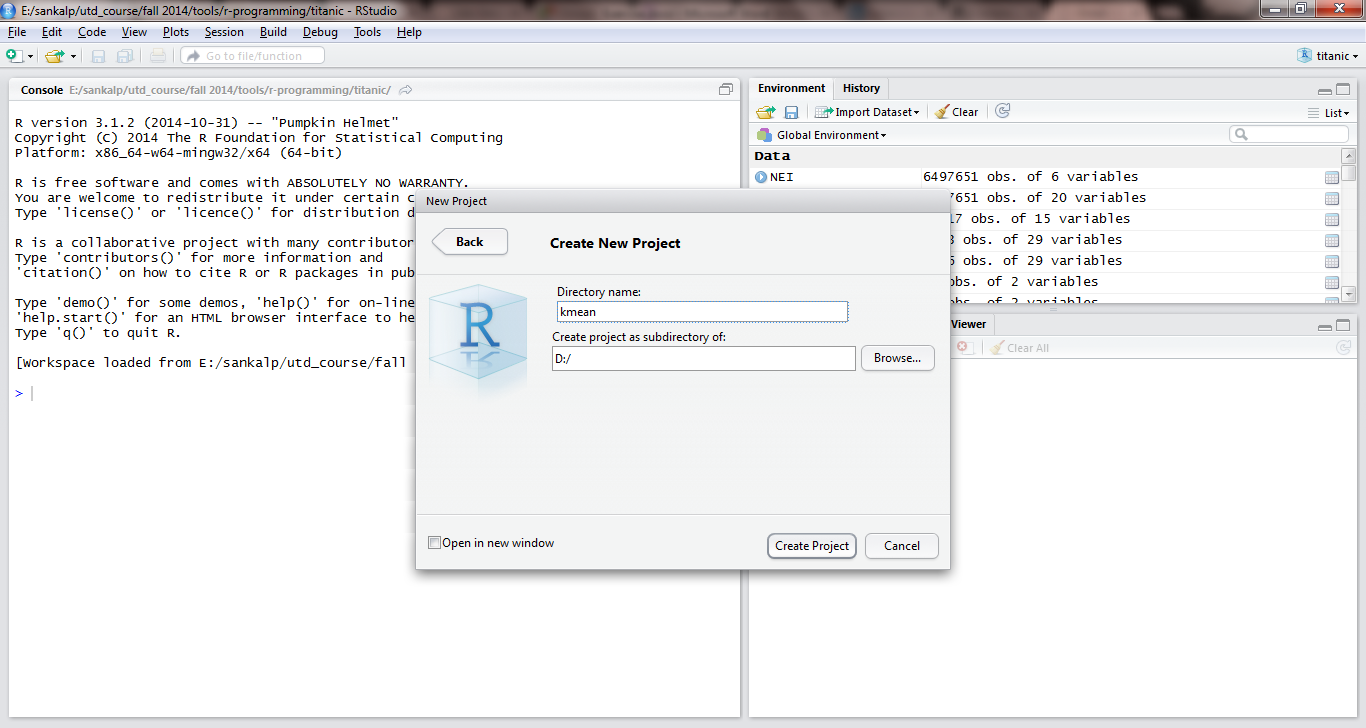
**1) Data Retrieval**

1)Open a new project in R studio using the below screenshots



2)Select Empty Directory



3)Name this directory to Kmean 

4)Save the new project in this new working directory

5)Create a new R file and name it to **KmeanSYYXXX.R**

6)Place the titanic.csv file in the new directory created in step 3

7)Read the csv file in the object called titanic as mentioned below.



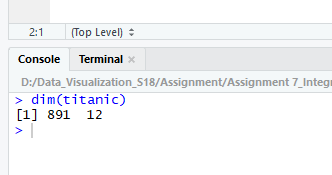
The dataset is now in titanic data frame

8)Execute the command to check the number of rows and columns as mentioned below.



9)Write the output of the above executed command.

Answer:



**2) Data Pre-processing**

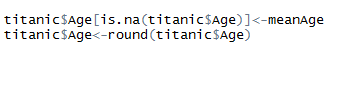
This steps involves replacing missing value of the ages with their mean, adding age category column and replacing the label with a meaningful name

1) Execute the below command to calculate the average age of the passengers using the below snapshot

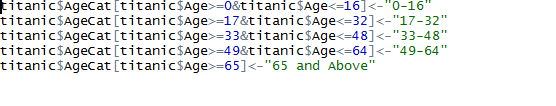


2) What is the average age that you get?

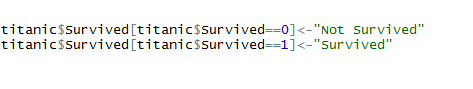
Answer: 29.6991

3) Execute the below commands to round off the age to their nearest integer.

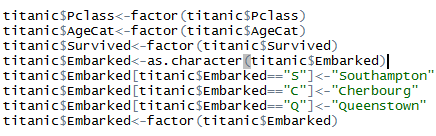
4) Create a new variable (vector) called Age Category and assign a category to every passenger in the dataset.



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



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



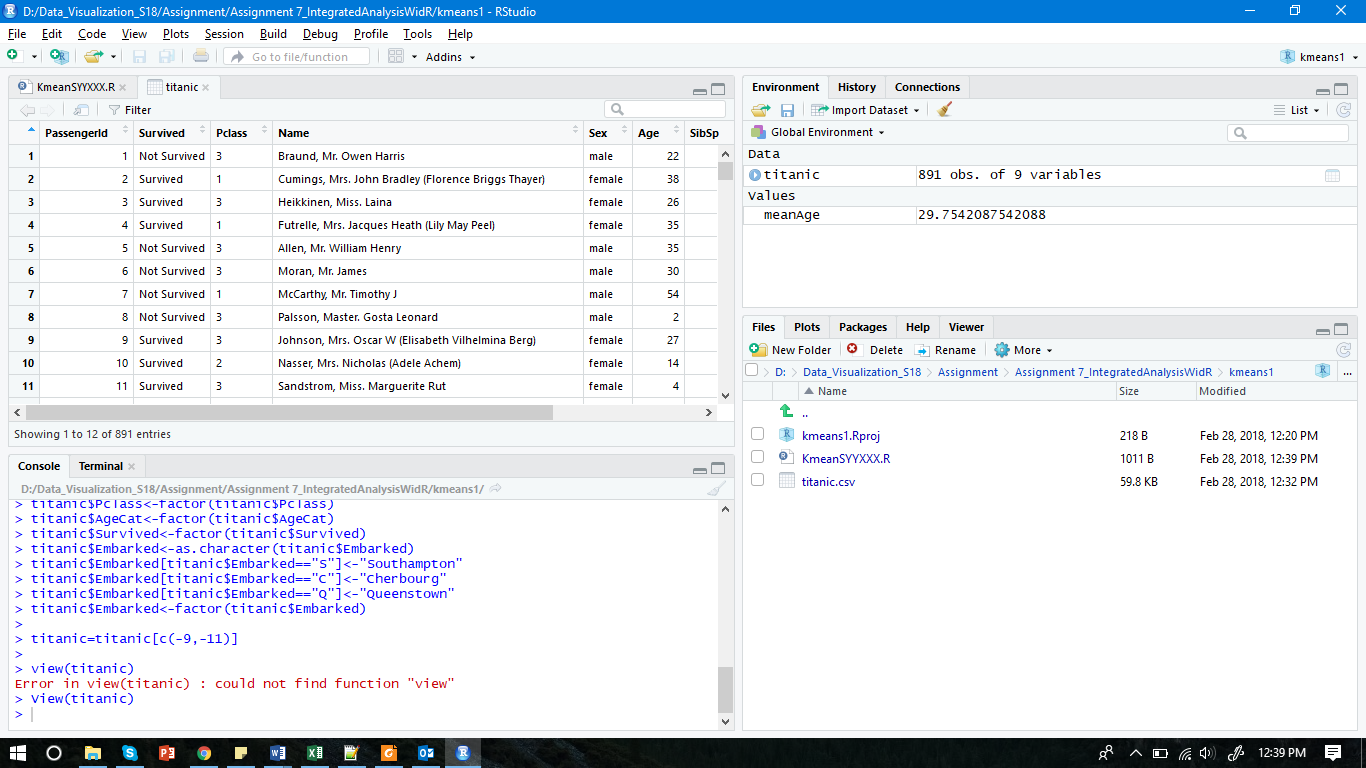
7) Remove the other redundant variables such as Ticket and Cabin from the titanic data frame.



8) Crosscheck the processed values using the command as mentioned below.



9) **Paste the screenshot** of the above executed command.

Answer: 

**Save the processed titanic dataset**

write.csv (titanic, file = "E:/titanicNew.csv")

**3) Decision Tree**

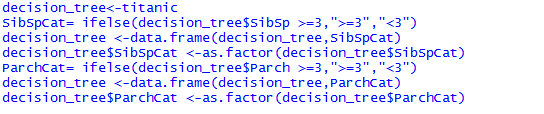
**Run Decision tree for the preprocessed Data**

If your R session is open from the previous section then skip to step 1**.** Else, please reopen your R window and type the following code and then proceed to step 1.

titanic<-read.csv(“E:/titanicNew.csv”)

**Steps:**

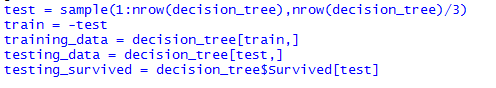
1. Enter the code below to run the Decision tree on the Titanic data. Decision tree works best for the categorical variables hence we will convert two variables SibSp and Parch into categorical variables and add the variables (vectors) to the data.



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.**

Use the following code to separate the data into training data and testing data.



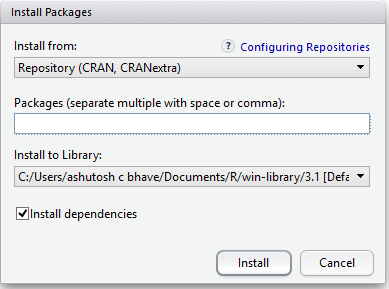
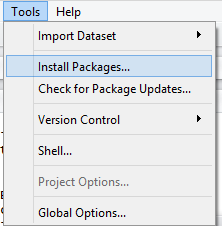


3) **Now the next step would be to build the tree**

To build the tree we need to install **rpart** package and **rattle** package and then import it. You can install package by going to

Tools->install packages->

Then writing the name of the package **rpart**.



The next step would be to import the package.

Use the following code to do so.



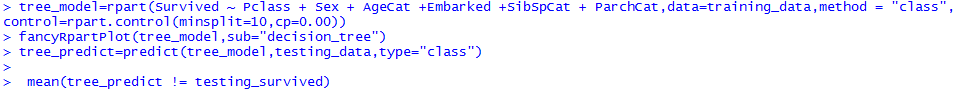
Follow the same steps for package **rattle**

**Note:** If you get any error, you will need to install RGtk2, rpart.Plot and RColorBrewer library package

Once you install the package, import it using run library(rpart) , library(rattle) etc.

**4) Next step would be building the tree using the training data and calculating the misclassification rate between the predicted values and actual values for testing data.**

Use the following code to generate the tree and get the misclassification rate.



This will give you the misclassification rate and a tree plot.

Export the tree plot and save it as a PDF with a name **decision\_tree\_Plot**. Paste the screenshot of the decision tree below. Note that the screenshot should be readable.

Answer:

**Question:-**

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

**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) 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: All variables OR Key Variables.

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

We will use the same titanic processed file that we saved in 2nd Part of the Assignment.

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)

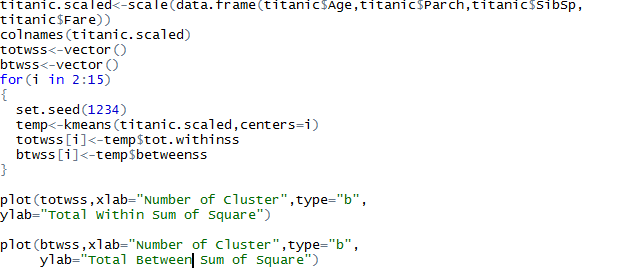
EmbarkedN<-ifelse(titanicUpdated$Embarked=="Southampton",1,ifelse(titanicUpdated $Embarked=="Cherbourg",2,0))

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.



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

**(Hint:** 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)

Answer:

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

Answer:

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 you will perform visual analysis integrating Tableau with R to visualize the clusters in Tableau.

**5) TABLEAU / R INTEGRATION**

You will have already used Tableau for previous assignments. In this part of the exercise we will

leverage the modeling power of R along with the visualization power of Tableau to perform our

exercise in integrated analysis.

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

Type the following code on your R window

install.Packages("Rserve")

library(Rserve)

Rserve()

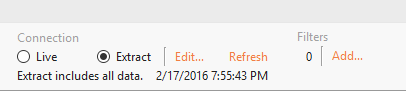
In order for us to be able to make a connection between Tableau and R, we need to do it through Rserve. Rserve itself is the server that is a program that responds to requests from clients. It listens for any incoming connections and processes incoming requests.

1. Open Tableau
2. Go to the Help menu and select “Manage R Connection”.
3. Enter a server name of “Localhost” (or “127.0.0.1”) and a port of “6311”.
4. Click on the “Test Connection” button to make sure everything runs smoothly. You should see a successful message. Click OK to close.

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

Open the file “titanicUpdated.csv” on Tableau. Save your Tableau file as TitanicViz.twbx

On the “Data Source” tab on Tableau please select the Connection as “Extract”. Screenshot is shown below for reference.



**Step 3: Forming Clusters**

Click on the New Worksheet in Tableau. On this worksheet you would see all the dimensions and measures in your current dataset represented in Tableau.

Q 5) Paste a screenshot of the current set of dimensions and measures. Ensure that the variables “EmbarkedN”, “SexN” and “SurvivedNum” shows up under the Measures column along with Age, Fare, Parch, Pclass etc.

Answer:

**Parameter Creation:**

Right click on the empty space within the Dimension window and select “Create Parameter”. Name the parameter as “# of Clusters”. Select the data type as Integer and enter the “*Ideal number of Clusters*”

that you obtained as part of this exercise under Current value cell. Radio button for “All” would be

enabled under allowable values.

Q6) Please paste screenshot once you populate all fields for this parameter.

Answer:

Similarly create another parameter called “Seed” and set the value as 1234. Now you would see two

parameters under your parameter window of your worksheet.

Q7) Paste a screenshot once you populate all fields for this “Seed” parameter.

Answer:

**Calculated Field creation:**

1. Right click on the empty space within the Dimension window and select “Create Calculated Field”.
2. In the window that opens up, type the name as “Cluster” and type the below piece of code

within the window and click OK.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

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])

Note: The highlighted part that you see above are the actual variable names used in your dataset. If your dataset has any other names for those measures, please use those names here, else you would

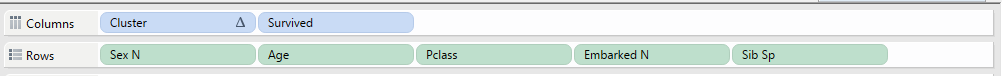
see an error.

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

Objective: The objective of this step is to visualize how different variables vary across different cluster.

Step:

1. Name the worksheet “Overall Clusters”.
2. Pull the Cluster field and Survived field on to the Columns. Please ensure you pull the Survived field from the Dimensions and not the one in Measure.
3. Ensure the sequence of the field is as below:



1. Pull the other variables SexN, Age, Pclass, EmbarkedN and SibSp on to the rows field.
2. Click on “Box and Whisker plot” from the ShowMe tab.
3. Drag “Embarked” field from Dimensions to Color.
4. Make sure the “Cluster” variable is set to Discrete and Survived field is set to “Dimension”. These two fields will appear in blue under Columns tab.
5. Go to worksheet – Show Title. Add your name to the title.

Q8) Paste the screenshot of the worksheet “Overall Clusters” below

Answer:

Q9) Provide your insights from cluster 1.

Answer:

Q10) Provide your insights from cluster 4.

Answer:

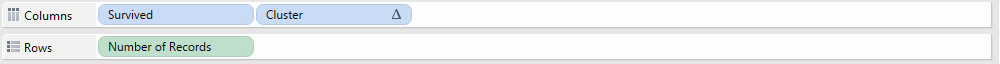
**Key Takeaway:** The key takeaway from this worksheet is that Sex is the most important variable in the formation of cluster as most of the clusters have predominantly one gender. This analysis is in line with our findings from the decision tree.

**Step 5: Survival by Cluster**

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

Task:

1. Click on new Worksheet and name it “Survival by Cluster”.
2. Drag the variables “Survived”(Dimension) and “Cluster” to the Columns field and set the “Cluster” to Discrete.
3. Drag the “Number of Records” variable to rows.
4. Ensure the sequence of the field is as shown below:



1. Select bar chart as visualization.
2. Also, select the variable “Survived” to Color section under “Marks” pane.
3. Go to worksheet – Show Title. Add your name to the title.
4. Right click on the vertical axis and select “Add reference line”
5. On the “Add Reference Line” screen, select Scope as “Entire table”, Value = “Sum of records” and “average” and click OK.

Q 11) Paste the screenshot of your worksheet. Ensure that you have added a Title and your name is contained in the title. Titles without names shall not be considered for grading.

Answer:

Q 12) From the bar chart, could you name two clusters where survivability is the highest?

Answer:

Going forward for our analysis, our focus would be on these two clusters.

Q 13) Enter the top two clusters you’ve identified and enter them against <Cluster 1> and <Cluster 2>. Leave the other rows blank for now.

|  |  |  |
| --- | --- | --- |
|  | <Cluster 1> | <Cluster 2> |
| Ideal Gender |  |  |
| Ideal Passenger Class |  |  |
| Ideal Age Category |  |  |
| Ideal Embarked point |  |  |
| Ideal number of siblings |  |  |

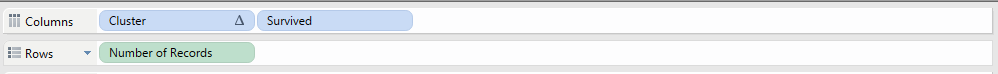
**Key Takeaway:** The key takeaway from the above worksheet “**Survival by Cluster**” is that **only in case of two clusters**, the number of passengers survived either outweigh or compare to the number of passengers who didn’t survive. Hence these clusters form the most important cluster for our analysis.

**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.

Task:

1. Click on new Worksheet and name it “Survival by Gender”.
2. Select Cluster and Survived (Dimension) into the columns. Make sure they are discrete.
3. Select Number of records in the rows
4. Ensure the sequence of the field is as shown below:



1. Drag “Sex” (Dimension) to Color section under Marks pane.
2. Go to worksheet – Show Title. Add your name to the title.

Q14) Based on your findings, enter the ideal Gender (that has the best chance to survive) in your top two cluster.

|  |  |  |
| --- | --- | --- |
|  | <Cluster 1> | <Cluster 2> |
| Ideal Gender | <Enter the gender that has the best survivability in this cluster> | <Enter the gender that has the best survivability in this cluster> |
| Ideal Passenger Class |  |  |
| Ideal Age Category |  |  |
| Ideal Embarked point |  |  |
| Ideal number of siblings |  |  |

Q15) Paste your screenshot of the Tableau worksheet below. Make sure it carries the Title with your name.

Answer:

Q16) If you are Female, which cluster should you belong, to ensure higher probability of survival?

Answer:

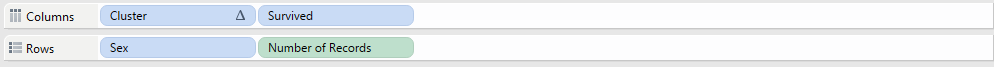
**Key Takeaway:** The key takeaway from the above worksheet “**Survival by Gender**” is that in the two prominent clusters that we identified in previous steps, one gender clearly dominates the survivability. Note that cluster 1 is comprised of only male passengers.

**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.

Task:

1. Click on new Worksheet and name it “Survival by Passenger Class”.
2. Select Cluster and Survived (Dimension) into the columns. Make sure they are discrete.
3. Select Sex (Dimension) and Number of records in the rows
4. Ensure the sequence of the field is as shown below



1. Drag “Pclass” to Color section under Marks pane.
2. Go to worksheet – Show Title. Add your name to the title.

Q17) Based on your findings, enter the ideal Gender/Passenger Class (that has the best chance to survive) in your top two cluster.

|  |  |  |
| --- | --- | --- |
|  | <Cluster 1> | <Cluster 2> |
| Ideal Gender |  |  |
| Ideal Passenger Class | <Please write the ideal class for the gender in above cell and above cluster.> | <Please write the ideal class for the gender in above cell and above cluster.> |
| Ideal Age Category |  |  |
| Ideal Embarked point |  |  |
| Ideal number of siblings |  |  |

Q18) Paste your screenshot of the Tableau worksheet below. Ensure you have the title with your name on it.

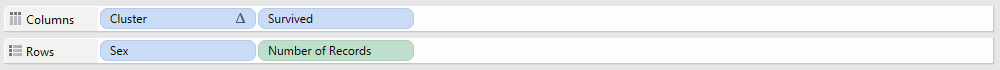
**Key Takeaway:** The key takeaway from the above worksheet “**Survival by Passenger class**” is that in some key clusters, most of the survivors belonged to a particular class.

**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.

Task:

1. Click on new Worksheet and name it “Survival by Age Category”
2. Select Cluster and Survived (Dimension) into the columns. Make sure they are discrete.
3. Select Sex (Dimension) and Number of records in the rows
4. Ensure the sequence of the field is as shown below



1. Drag Age Category (Dimension) to Color section under Marks pane.
2. Go to worksheet – Show Title. Add your name to the title.

Q19) : Based on your findings, enter the ideal Gender/Age Category Class (that has the best chance to survive) in your top two cluster.

|  |  |  |
| --- | --- | --- |
|  | <Cluster 1> | <Cluster 2> |
| Ideal Gender |  |  |
| Ideal Passenger Class |  |  |
| Ideal Age Category | <Please write the ideal age category for the gender in above cell and above cluster.> | <Please write the ideal age category for the gender in above cell and above cluster.> |
| Ideal Embarked point |  |  |
| Ideal number of siblings |  |  |

Q20) Paste your screenshot of the Tableau worksheet below. Ensure you add the title with your name.

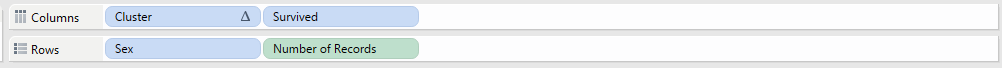
**Key Takeaway:** Most of the female survivors were from the age category 17-32 years. Most men who survived in cluster 1 also belonged to the age category 17-32 years. Overall, we can say that because majority of the passengers seem to fall it the age group 17-32 years their survivability salience is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

Task:

1. Click on new Worksheet and name it “Survival by Embarked Point”.
2. Select Cluster and Survived (Dimension) into the columns. Make sure they are discrete.
3. Select Sex (Dimension) and Number of records in the rows
4. Ensure the sequence of the field is as shown below:



1. Drag Embarked variable (Dimension) to Color section under Marks pane.
2. Go to worksheet – Show Title. Add your name to the title.

Q21) Output: Based on your findings, enter the ideal Gender/Embarked point (that has the best chance to survive) in your top two cluster.

|  |  |  |
| --- | --- | --- |
|  | <Cluster 1> | <Cluster 2> |
| Ideal Gender |  |  |
| Ideal Passenger Class |  |  |
| Ideal Age Category |  |  |
| Ideal Embarked point | <Please write the embarked point for the gender in above cell and above cluster.> | <Please write the embarked point for the gender in above cell and above cluster.> |
| Ideal number of siblings |  |  |

Q22) Paste your screenshot of the Tableau worksheet below. Ensure you have the title added with your name.

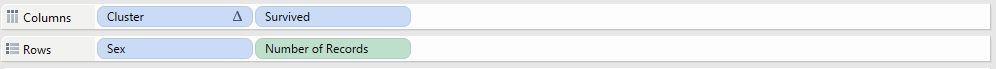
**Key Takeaway:** Most of the passengers embarked from Southampton. Amongst the survivors, majority of the passengers were from Southampton.

**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.

Task:

1. Click on new Worksheet and name it “Survival by number of Siblings”.
2. Select Cluster and Survived (Dimension) into the columns. Make sure they are discrete.
3. Select Sex (Dimension) and Number of records in the rows
4. Ensure the sequence of the field is as shown below:



1. Drag Sib Sp (Measure) to Color section under Marks pane.
2. Go to worksheet – Show Title. Add your name to the title.

Q23) Output: Based on your findings, enter the ideal Gender/# of Sib Sp (that has the best chance to survive) in your top two cluster.

|  |  |  |
| --- | --- | --- |
|  | <Cluster 1> | <Cluster 2> |
| Ideal Gender |  |  |
| Ideal Passenger Class |  |  |
| Ideal Age Category |  |  |
| Ideal Embarked point |  |  |
| Ideal number of siblings | <Please write the # of siblings for the gender in above cell and above cluster.> | <Please write the # of siblings for the gender in above cell and above cluster.> |

Q24) Paste your screenshot of the Tableau worksheet below. Ensure you have the title with your name on it.

Answer:

**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**

At this stage, you should have enough information based on the above tabular data, to profile the passengers that had the best chance of survivability in your top 2 cluster.

Q22) Please put the profile of passengers from each of the top two clusters in words. Use the above filled in table to form your passenger profile. Please write two separate sentences explaining each cluster profile.

Answer:

**Attach assignments in elearning**

1. Attach the **assignment document (only solutions to questions)**

in Microsoft Word

1. Attach all R program files
2. Attach the Tableau File (.twbx)