**Discussions:**

**Q2**

**Describe attribute types for each attribute in the Titanic data set.**

[**https://www.kaggle.com/c/titanic/data**](https://www.kaggle.com/c/titanic/data)

PassengerId - Numeric

Pclass - Ordinal

Name - Nominal

Sex - Nominal

Age - Numeric

SibSp - Numeric

Parch - Numeric

Ticket - Nominal

Fare - Numeric

Cabin – Nominal

Embarked – Nominal

**Q3**

**Convert PassengerId to nominal data type in R. Show the result of the conversion.**

mydata <- read.csv("~/Desktop/titanic-train.csv",header = TRUE)  
mydata$PassengerId <- factor(mydata$PassengerId)

str(mydata)

**Q4**

**Answer the following questions based on the Titanic data:**

**How many values are missing for each variable?**

**How do you handle the missing values in each variable?**

**Show the R code that you use to find and deal with missing values.**

length(which(is.na(titanic$Parch)))

titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age, na.rm = TRUE)

PassengerId - 0

Pclass - 0

Name - 0

Sex - 0

Age - 714 - Replace with average

SibSp - 0

Parch - 0

Ticket - 0

Fare - 0

Cabin - 687 - Replace with "NA"

Embarked - 2 - Remove or replace with "NA

**Q5**

**How do you summarize the central tendency and data spread of ordinal variables like the Pclass variable in the Titanic data?**

Ordinal can be treated as numeric or nominal

freq <- table(t$Pclass)

> freq  
  1   2   3  with values 186 173 355 respectively.

Q6 - Show a box plot and histogram of "Fare" in third class (Pclass==3).

Did you find any outliers from the visualization?

Crosstab "Embarked" and "Survived."

Is there correlation between where people embarked and their survival? Why?

Yes, many outliers, about 52 - as seen in the boxplot. All customers who paid more than about 27.125 (upper whisker).

This is the crosstab for third class

    0   1

  C  41  25

  Q  45  27

  S 286  67

Here, there is a definite correlation, as people who boarded from "S" were far more unlikely to survive as compared to the other ports.

**Q7**

**Aggregate the average fare for men and for women in the Titanic data set.**

Male  - 25.52

Female - 44.48

male = titanic[titanic$Sex=='male',]mean(male$Fare)  
female = titanic[titanic$Sex=='female',]mean(female$Fare)

**Q8**

**Choose a numeric variable (e.g. Age, Fare) in the Titanic data set and transform it using discretization. Show the R code and result.**

age <- cut(data$Age, breaks = c(0,10,20,30,40,50,60,Inf), labels = c("Child","Teen","Twenties","Thirties","Fourties","Fifties","Old"))

> age  [1] Twenties Thirties Twenties Thirties Thirties <NA>     Fifties  Child    Twenties [10] Teen     Child    Fifties  Teen     Thirties Teen     Fifties  Child    <NA>     [19] Thirties <NA>     Thirties Thirties Teen     Twenties Child    Thirties <NA>     [28] Teen     <NA>     <NA>     Thirties <NA>     <NA>     Old      Twenties Fourties

**Q9**

**Produce a random sample of 100 examples out of the Titanic data. Produce a systematic sample of 100 examples out of the Titanic data. Show your R code.**

rsample<- titanic[sample(1:nrow(titanic),100,replace = FALSE),]s

sample<-titanic[seq(1,nrow(titanic),10),]