**ITA0448 – STATISTICS WITH R PROGRAMMING**

**NAME : RISHIKUMAR V**

**REG NO : 192121080**

**DAY 4**

**ASSESSMENT PART 2**

**1. Suppose that the data for analysis includes the attribute age. The age values for the data**

**tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,**

**33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. What is the median?**

**code :**

**ages <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)**

**sorted\_ages <- sort(ages)**

**median\_age <- median(sorted\_ages)**

**output:**

**[1] 25**

**2. Suppose that the data for analysis includes the attribute age. The age values for the data**

**tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,**

**33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.**

**Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?**

**code:**

**ages <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)**

**Q1 <- quantile(ages, 0.25)**

**Q3 <- quantile(ages, 0.75)**

**> Q1**

**25%**

**20**

**> Q3**

**75%**

**35**

**3.Load iris Dataset which is inbuilt in R .explore the dataset in terms of dimension and**

**summary statistics (2M)**

**code:**

**data(iris)**

**dim(iris)**

**[1] 150 5**

**head(iris)**

**summary(iris)**

**Sepal.Length Sepal.Width Petal.Length Petal.Width Species**

**Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa :50**

**1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor:50**

**Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica :50**

**Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199**

**3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800**

**Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500**

**4.Find the categorical column data and convert that to factor form, also find the number of**

**rows for each factors in dataset. (2)**

**iris$Species <- as.factor(iris$Species)**

**table(iris$Species)**

**setosa versicolor virginica**

**50 50 50**

**5. Find mean of numeric data in dataset based on Species group. and plot Bar chart (use**

**ggplot ) to interpret same (8m)**

**library(dplyr)**

**library(ggplot2)**

**dataset <- read.csv("my\_dataset.csv")**

**species\_means <- dataset %>%**

**group\_by(Species) %>%**

**summarize(mean = mean(NumericData))**

**ggplot(species\_means, aes(x = Species, y = mean)) +**

**geom\_bar(stat = "identity") +**

**labs(title = "Mean Numeric Data by Species",**

**x = "Species",**

**y = "Mean Numeric Data")**

**library(ggplot2)**

**data(iris)**

**6.Draw a suitable plot which summaries statistical parameter of Sepal.Width based on**

**Species group(6m)**

**ggplot(iris, aes(x = Species, y = Sepal.Width, fill = Species)) +**

**geom\_boxplot() +**

**labs(x = "Species", y = "Sepal Width", title = "Box plot of Sepal Width by Species")**

**7. Draw a suitable plot to find the skewness of the data for Sepal.Width and print the**

**comment about skewness. (6m)**

**library(ggplot2)**

**data(iris)**

**ggplot(iris, aes(x = Sepal.Width)) +**

**geom\_histogram(aes(y = ..density..), bins = 20, color = "black**

**8.Draw ggplot2 scatterplot showing the variables Sepal.Length and Petal.Length grouped by**

**the three-level factor “Species”. (6m)**

**library(ggplot2)**

**data(iris)**

**ggplot(iris, aes(x = Sepal.Length, y = Petal.Length, color = Species)) +**

**geom\_point() +**

**labs(x = "Sepal Length", y = "Petal Length", color = "Species")**