

**AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH (AIUB)**

**FACULTY OF SCIENCE AND TECHNOLOGY**

**Final ASSIGNMENT**

**Introduction To Data Science**

**Spring 2022-2023 Section: D**

**Submitted By**

# AL SHAKIB E ELAHI

# ID: 20-43665-2

# Department: CSE

**Supervised By**

# Tohedul Islam

# Assistant Professor

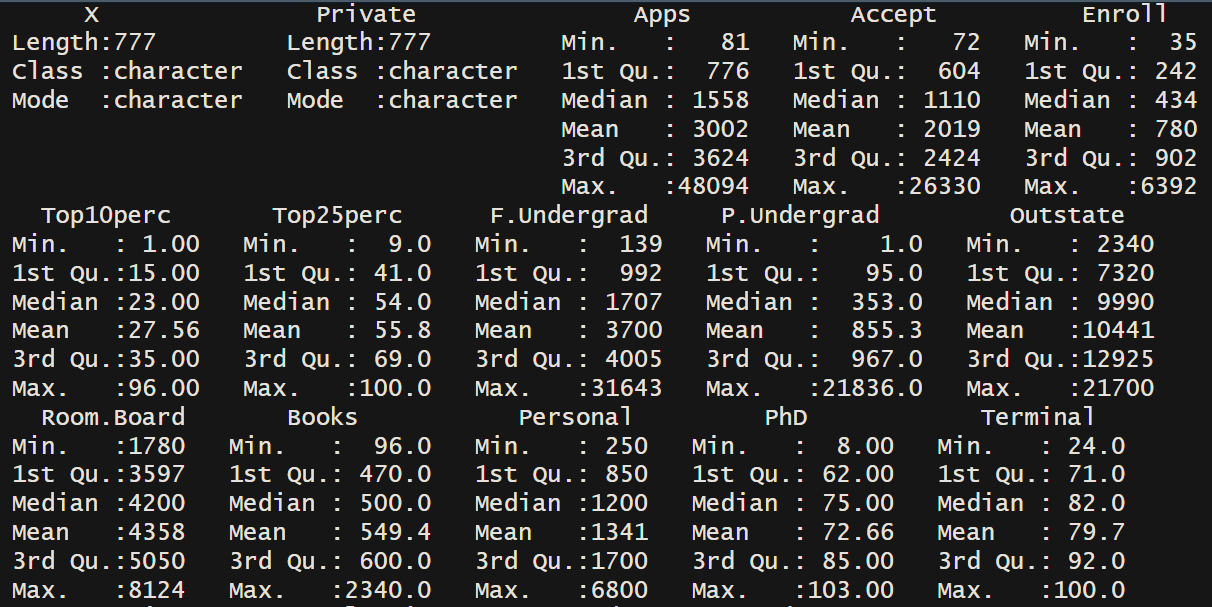
# Department of Computer Science

Date of Submission: **April 30, 2023**

1. **Dataset Description:**

* **Source:** [**https://www.kaggle.com/code/razzaswal/k-means-clustering-project-college-dataset/input**](https://www.kaggle.com/code/razzaswal/k-means-clustering-project-college-dataset/input)
* **The data frame has 777 observations followed by 18 variables and the names of the university.**
  + Private: A factor with levels No and Yes indicating private or public university
  + Apps: Number of applications received
  + Accept: Number of applications accepted
  + Enroll: Number of new students enrolled
  + Top10perc: Pct. new students from top 10% of H.S. class
  + Top25perc: Pct. new students from top 25% of H.S. class
  + F.Undergrad: Number of fulltime undergraduates
  + P.Undergrad: Number of parttime undergraduates
  + Outstate: Out-of-state tuition
  + Room.Board: Room and board costs
  + Books: Estimated book costs
  + Personal: Estimated personal spending
  + PhD: Pct. of faculty with Ph.D.’s
  + Terminal: Pct. of faculty with terminal degree
  + S.F.Ratio: Student/faculty ratio
  + perc.alumni: Pct. alumni who donate
  + Expend: Instructional expenditure per student
  + Grad.Rate: Graduation rate

1. **Summary:**

****

**Calendar

Description automatically generated**

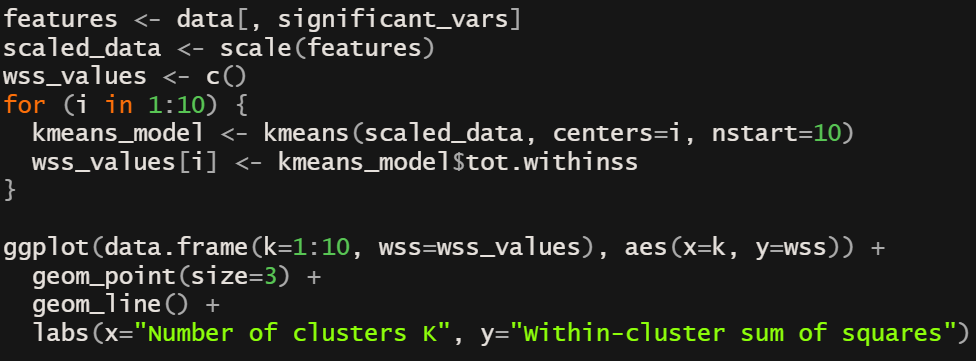
1. **Data Preprocessing:**

|  |  |
| --- | --- |
| **Note:**   * The `boxplot.stats()` function computes the median, quartiles, and outliers. `$out` attribute contains the values of any observations that are classified as outliers based on the default outlier detection rule. |  |

1. **Multivariant Analysis:**

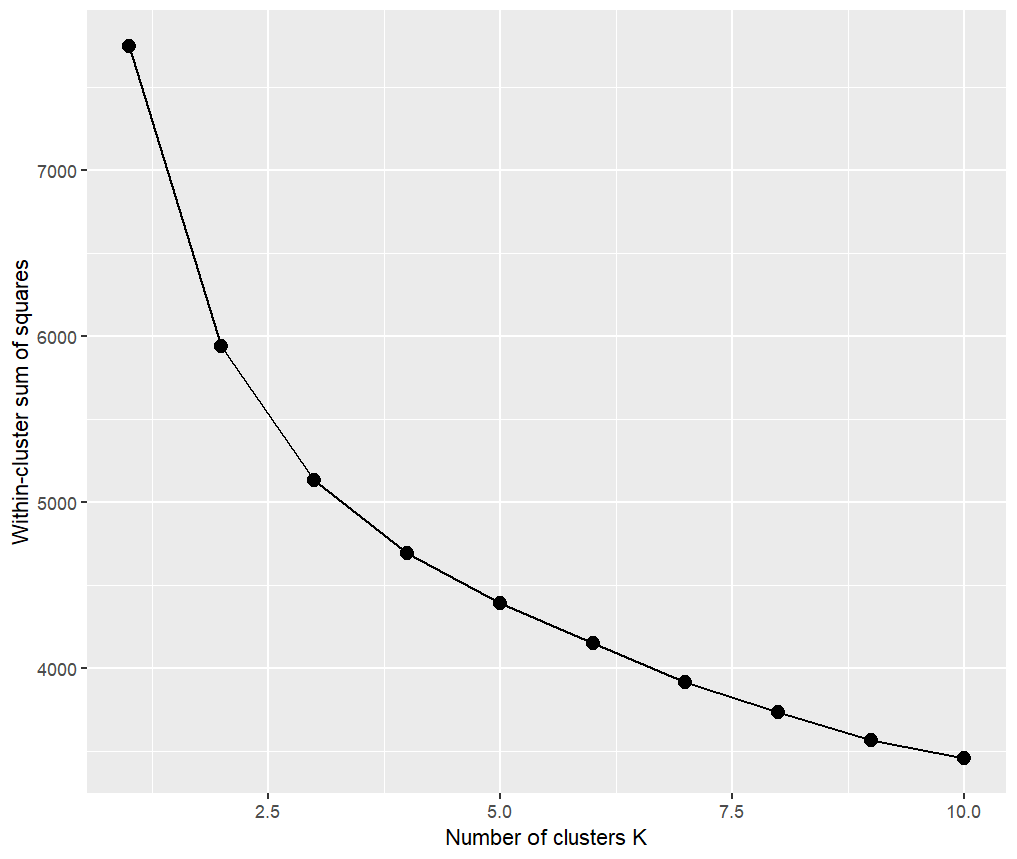
|  |  |
| --- | --- |
| **Text  Description automatically generated**  **Note:**   * `table()` function is used to create a contingency table showing the counts of observations * `chisq.test()` function is used to perform a chi-squared test of independence * `p.value` is one of the outputs returned by the chisq.test() function. |  |

1. **Elbow Method for Determining K value for K-means algorithm:**

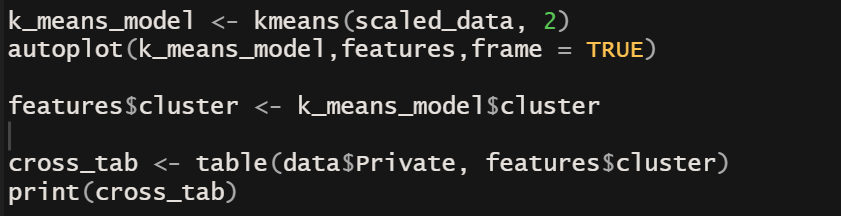
****

**Note:**

* kmeans() function is used to perform k-means clustering where in first parameter it takes a numeric matrix or data frame containing the data to be clustered. Then `centers` contains the number of clusters to be formed and `nstart` is the number of times the algorithm is run with different initial centers.
* `tot.withinss` is the total within-cluster sum of squares.

****

1. **Applying K-means Clustering Algorithm**

****

|  |  |
| --- | --- |
| **Text  Description automatically generated** |  |
|  |  |

**Note:**

* `autoplot` is a function from the `ggfortify` package that creates a `ggplot` object for visualizing clustering results.
* `k\_means\_model$cluster` is a vector of integers indicating the cluster assignment for each data point in the `scaled\_data` data frame.