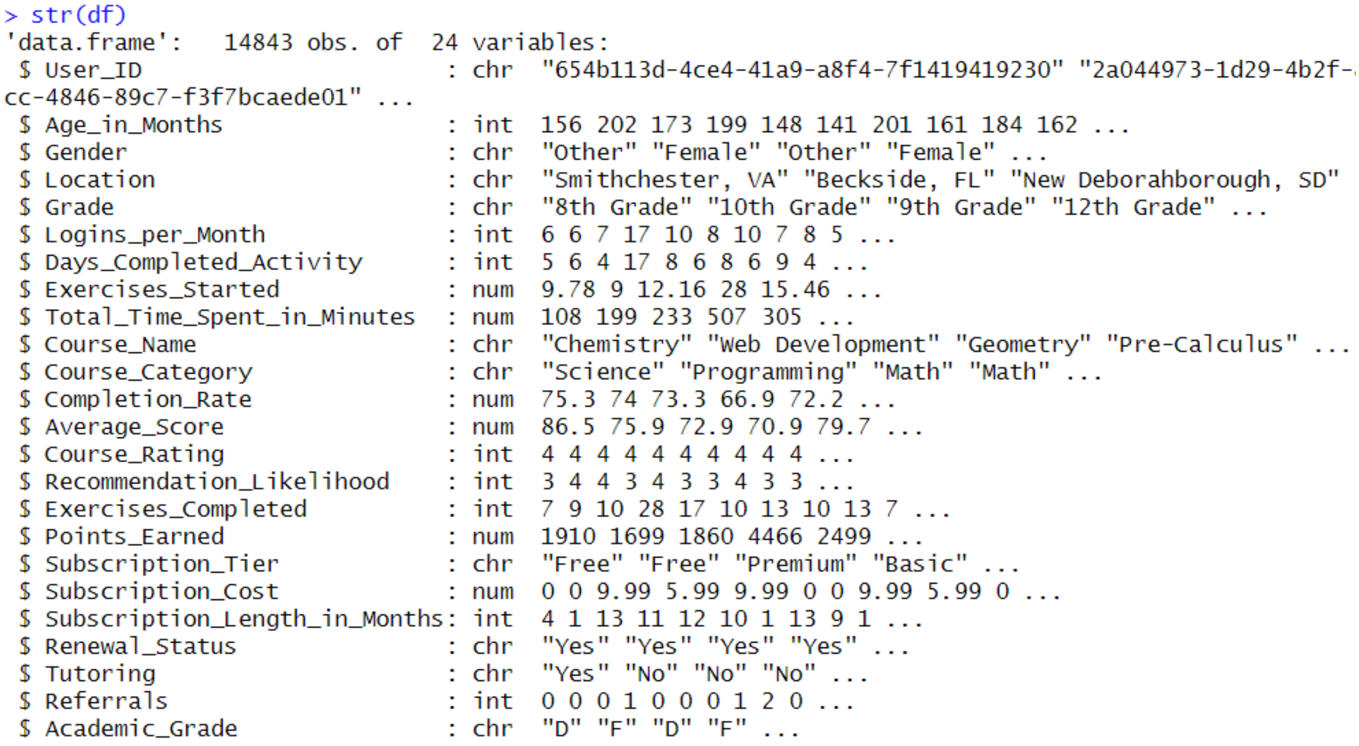
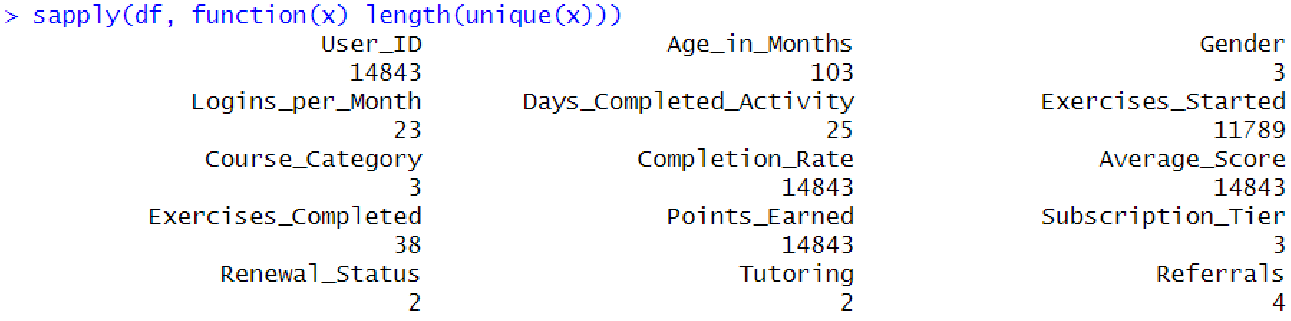
Process Followed:

The dataset chosen is that of a tutoring platform and it contains the following columns

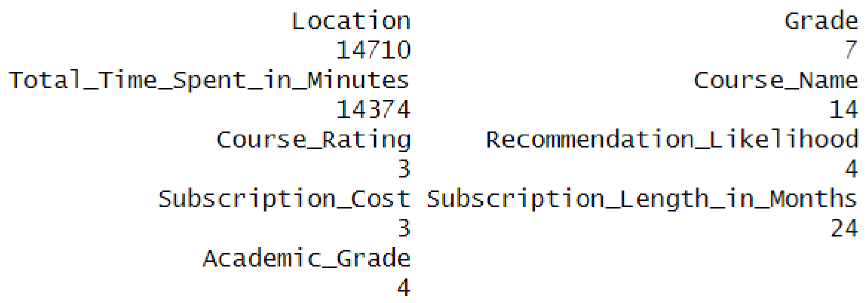


```
> str(df)
'data.frame':	14843 obs. of  24 variables:
 $ User_ID                     : chr  "654b113d-4ce4-41a9-a8f4-7f1419419230" "2a044973-1d29-4b2f-
cc-4846-89c7-f3f7bcaede01" ...
 $ Age_in_Months               : int  156 202 173 199 148 141 201 161 184 162 ...
 $ Gender                      : chr  "Other" "Female" "Other" "Female" ...
 $ Location                    : chr  "Smithchester, VA" "Beckside, FL" "New Deborahborough, SD"
 $ Grade                       : chr  "8th Grade" "10th Grade" "9th Grade" "12th Grade" ...
 $ Logins_per_Month            : int  6 6 7 17 10 8 10 7 8 5 ...
 $ Days_Completed_Activity     : int  5 6 4 17 8 6 8 6 9 4 ...
 $ Exercises_Started           : num  9.78 9 12.16 28 15.46 ...
 $ Total_Time_Spent_in_Minutes : num  108 199 233 507 305 ...
 $ Course_Name                 : chr  "Chemistry" "Web Development" "Geometry" "Pre-Calculus" ...
 $ Course_Category             : chr  "Science" "Programming" "Math" "Math" ...
 $ Completion_Rate             : num  75.3 74 73.3 66.9 72.2 ...
 $ Average_Score               : num  86.5 75.9 72.9 70.9 79.7 ...
 $ Course_Rating               : int  4 4 4 4 4 4 4 4 4 4 ...
 $ Recommendation_Likelihood   : int  3 4 4 3 4 3 3 4 3 3 ...
 $ Exercises_Completed         : int  7 9 10 28 17 10 13 10 13 7 ...
 $ Points_Earned               : num  1910 1699 1860 4466 2499 ...
 $ Subscription_Tier           : chr  "Free" "Free" "Premium" "Basic" ...
 $ Subscription_Cost           : num  0 0 9.99 5.99 9.99 0 0 9.99 5.99 0 ...
 $ Subscription_Length_in_Months: int  4 1 13 11 12 10 1 13 9 1 ...
 $ Renewal_Status              : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Tutoring                    : chr  "Yes" "No" "No" "No" ...
 $ Referrals                   : int  0 0 0 1 0 0 0 1 2 0 ...
 $ Academic_Grade              : chr  "D" "F" "D" "F" ...
```

Checking for the number of unique values of the various columns to decide relevance of analysis of that column



```
> sapply(df, function(x) length(unique(x)))
                    User_ID               Age_in_Months                      Gender 
                      14843                         103                           3 
           Logins_per_Month     Days_Completed_Activity           Exercises_Started 
                         23                          25                       11789 
            Course_Category             Completion_Rate               Average_Score 
                          3                       14843                       14843 
        Exercises_Completed               Points_Earned           Subscription_Tier 
                         38                       14843                           3 
             Renewal_Status                    Tutoring                   Referrals 
                          2                           2                           4 
```

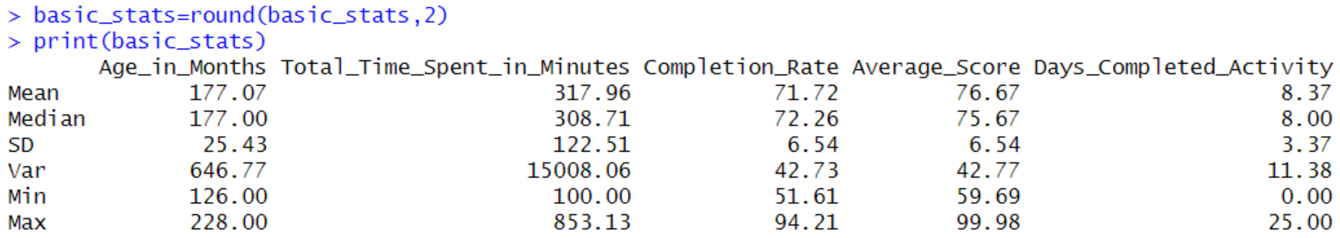


```
                   Location                         Grade
                      14710                             7
Total_Time_Spent_in_Minutes                   Course_Name
                      14374                            14
              Course_Rating     Recommendation_Likelihood
                          3                             4
          Subscription_Cost Subscription_Length_in_Months
                          3                            24
             Academic_Grade
                          4
```

User_ID is unique for all the cases and has no role in the analysis. Location too has many values, so can be ignored. Subscription_Tier and Subscription_Cost suggest the same ordinal variables so one of them can be ignored. (Cost and Months is ignored here) Referrals did not prove to be useful as well. Excercises started does not show any kind of continuous data so is ignored in this case. Course_Category proves to be more useful than Course_Name int this case so the same chosen.

Amongst all the remaining variables, comparatively more important variables that are appropriate for univariate analysis are chosen from each category of variable type. (numeric – discrete and continuous, categorical, ordinal) Univariate analysis on each of these variables is carried out and relevant graphs are plotted.

Statistics:



```
> basic_stats=round(basic_stats,2)
> print(basic_stats)
       Age_in_Months Total_Time_Spent_in_Minutes Completion_Rate Average_Score Days_Completed_Activity
Mean          177.07                      317.96           71.72         76.67                    8.37
Median        177.00                      308.71           72.26         75.67                    8.00
SD             25.43                      122.51            6.54          6.54                    3.37
Var           646.77                    15008.06           42.73         42.77                   11.38
Min           126.00                      100.00           51.61         59.69                    0.00
Max           228.00                      853.13           94.21         99.98                   25.00
```

1. Age_in_Months

- Mean = 177.07, Median = 177: distribution is almost symmetric (mean=median).
- SD = 25.43: ages are spread about 25 months around 177 (14-15 years).
- Range = 126 to 228: min age 10.5 years, max 19 years.

The dataset has students from middle school to college level, around 14-15 years.

2. Total_Time_Spent_in_Minutes

- Mean = 317.96, Median = 308.71: central tendency is similar, not skewed.
- SD = 122.51: some students spend much more or less than average.
- Range = 100 to 853: huge variation. Some spend very little (100 min), others spend 14 hrs total.

There is a wide variation in engagement time.

3. Completion_Rate (%)

- Mean = 71.72, Median = 72.26: students complete 72% of activities on average.
- SD = 6.54: most students are clustered around 65-78%.
- Range = 51.61 to 94.21: no extreme failures, but few complete almost everything.

Students are reasonably consistent, with most completing 2/3 to 3/4 of assigned tasks.

4. Average_Score

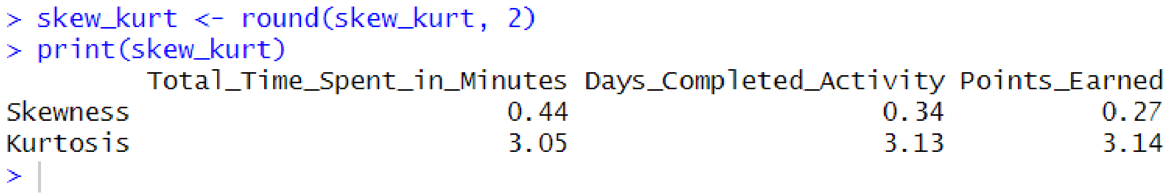
- Mean = 76.67, Median = 75.67: average performance around 76%.
- SD = 6.54: most students are within 70-83%.
- Range = 59.69 to 99.98: no one scores below 60 (minimum pass level), some nearly perfect.

The group is performing well overall, not too many weak students.

5. Days_Completed_Activity

- Mean = 8.37, Median = 8.00: on average, students completed activities on 8 days.
- SD = 3.37: some are more irregular.
- Range = 0 to 25: some never did it, others very consistent.

This reflects engagement habits.



```
> skew_kurt <- round(skew_kurt, 2)
> print(skew_kurt)
          Total_Time_Spent_in_Minutes Days_Completed_Activity Points_Earned
Skewness                         0.44                    0.34          0.27
Kurtosis                         3.05                    3.13          3.14
> 
```

1. Total_Time_Spent_in_Minutes

- Skewness = 0.44: slightly positively skewed, a few students spend much more time than average.
- Kurtosis = 3.05: close to normal distribution, moderate tails, not extremely peaked or flat.

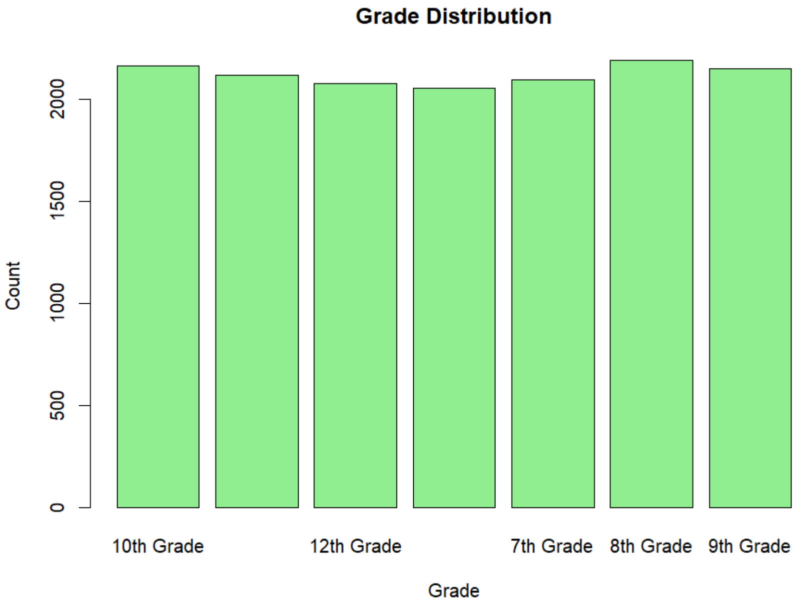
2. Days_Completed_Activity

- Skewness = 0.34: slightly positively skewed, a few students complete many more days than most.
- Kurtosis = 3.13: roughly normal distribution, moderate tails.

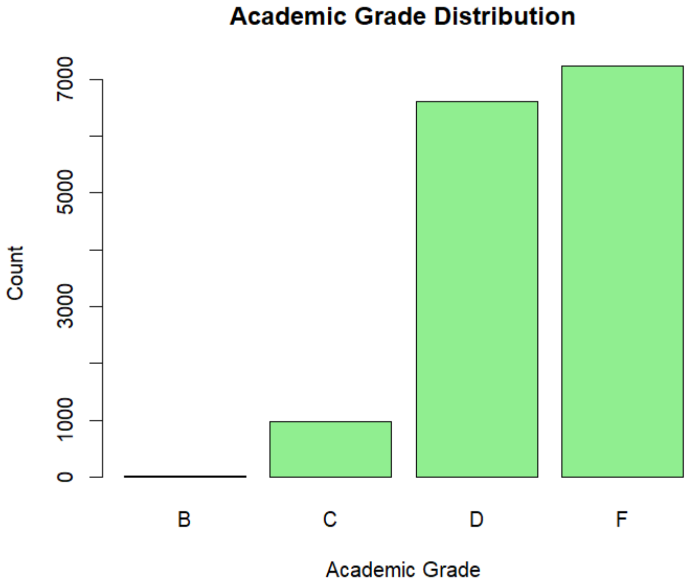
3. Points_Earned

- Skewness = 0.27: very slight positive skew, few students earn very high points.
- Kurtosis = 3.14: approximately normal distribution, tails are moderate.

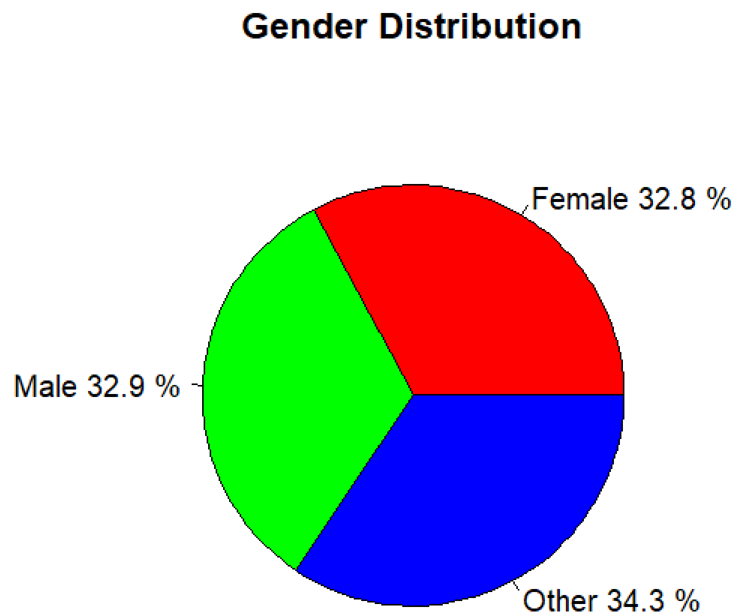
Plots:



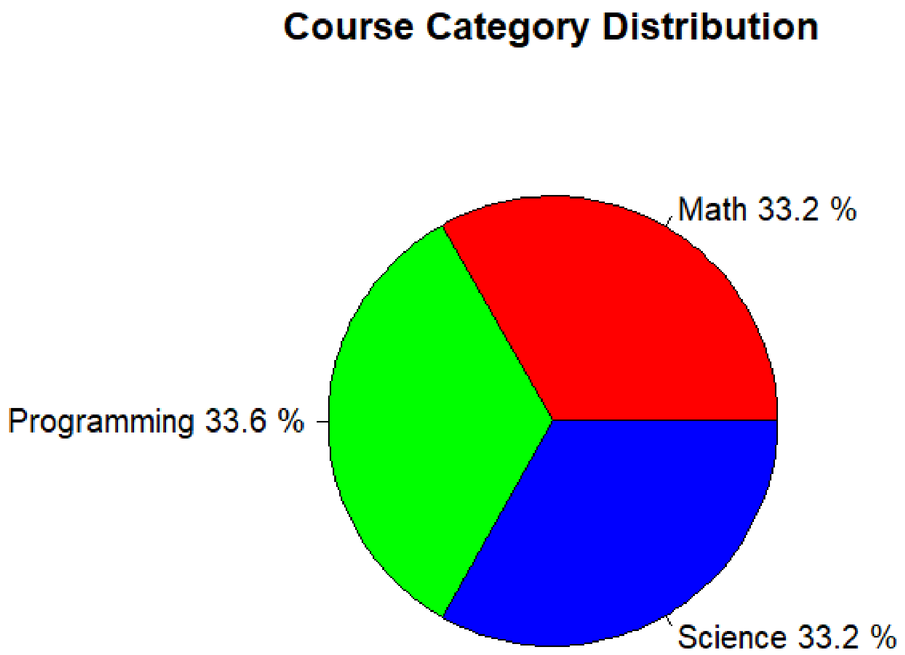
This shows the distribution of class grades of the students present in the dataset.



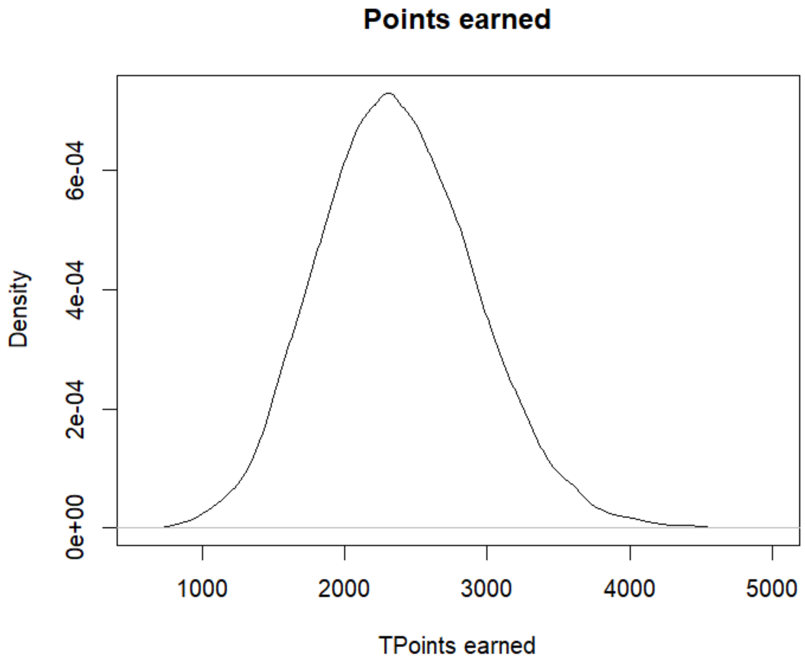
This academic grade distribution shows that the data consists of most of the students either failed or with D grade.



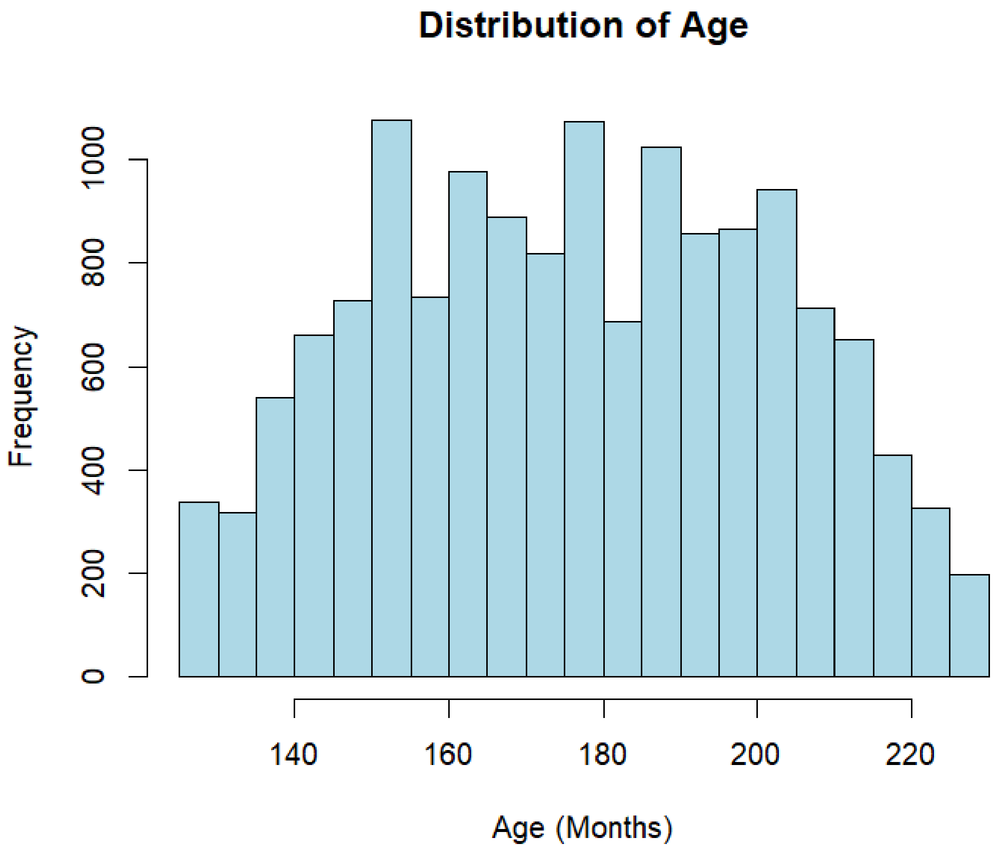
Equal division in the 3 categories of this categorical variable - gender. Population of other gender is slightly more.



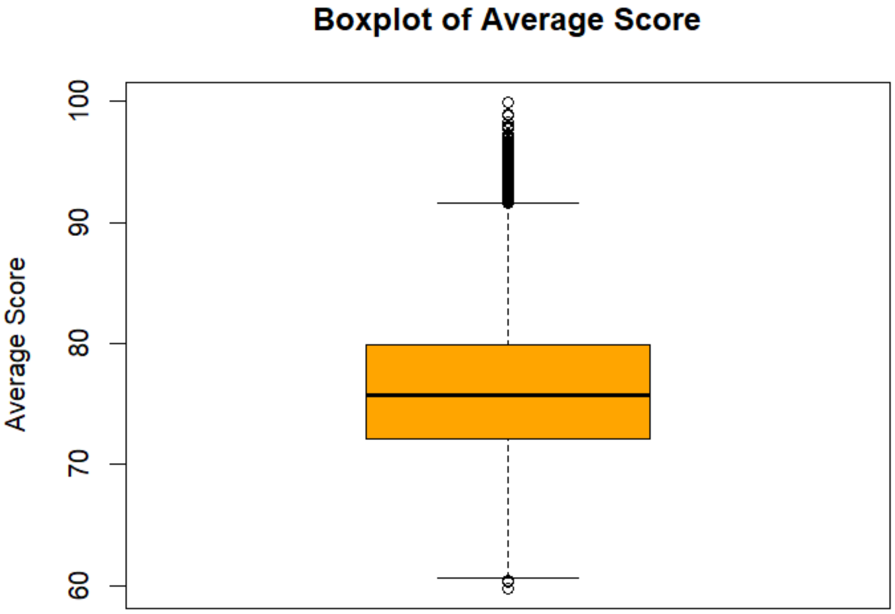
Equal division in the 3 categories of this categorical variable – Course_Category. Students enrolled in programming courses are slightly more.



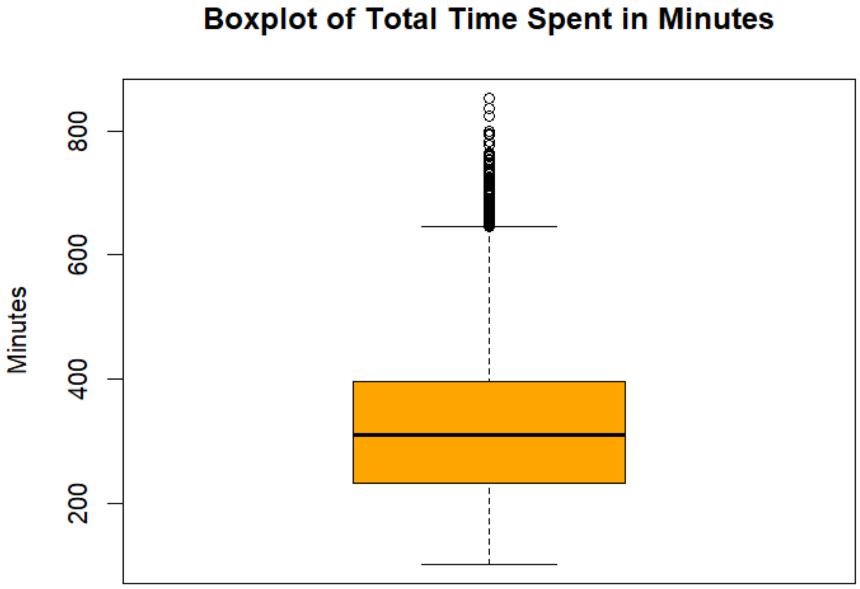
This density plot of Points earned by the students on the tutoring platform suggests that major density of population has earned about 2500 points and earned points range from about 500-4500.



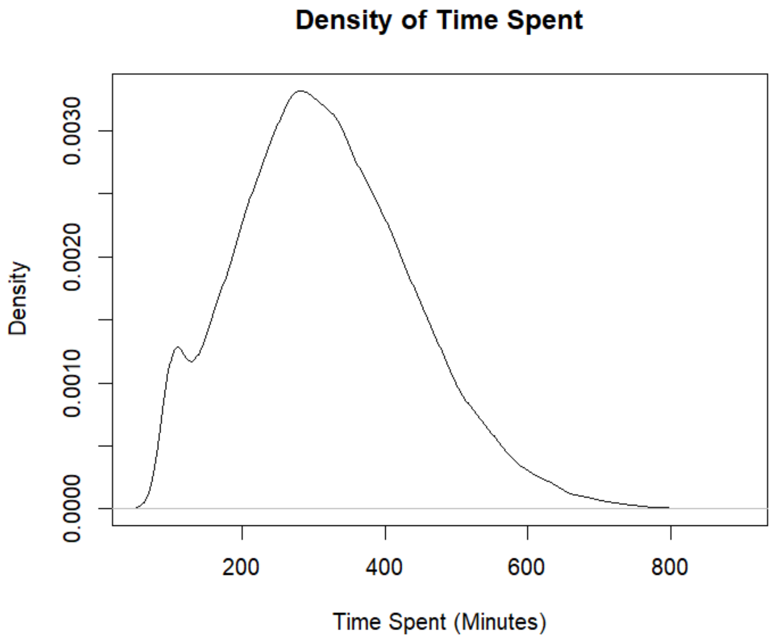
This histogram suggests continuous variation in the ages and unequal distribution of students in different monthly divided age groups.



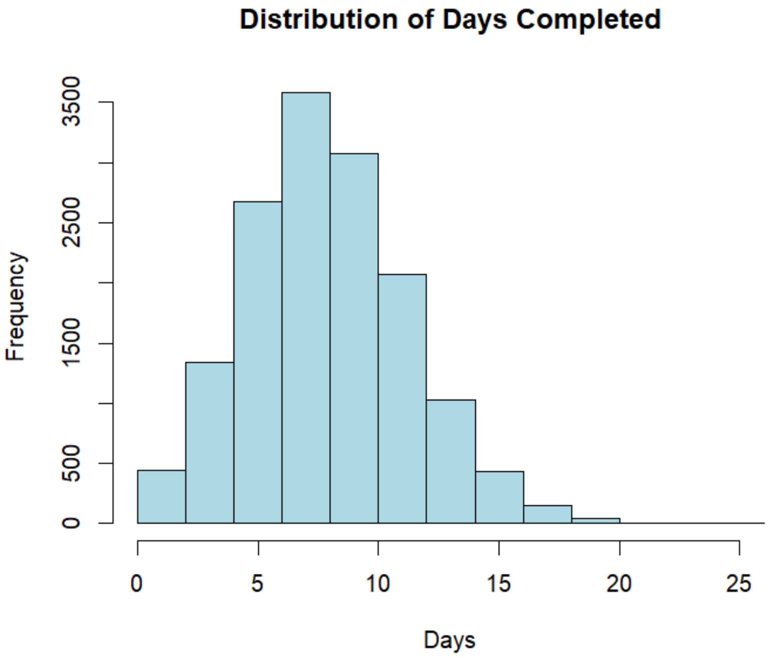
The IQR of Avg scores is 7. This plot also indicates good number of outliers



The IQR of Time spent is 180 minutes. This plot also indicates good number of outliers with time spent more than 600 minutes



The left skewness is seen in the above density plot, also indicating the outliers. Slightly positively skewed, a few students spend much more time than average.



The left skewness is seen in the above histogram of distribution of days completed, slightly positively skewed, a few students complete many more days than most.