Project_Descriptive_Analysis

Group 04

2024-07-22

```
#install.packages("readr")
#install.packages("dplyr")
#install.packages("summarytools")
#install.packages("tidyr")


# Load necessary libraries
library(readr)

## Warning: package 'readr' was built under R version 4.3.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(summarytools)

## Warning: package 'summarytools' was built under R version 4.3.3

library(tidyr)

## Warning: package 'tidyr' was built under R version 4.3.3

# Load the CSV file
data <- read_csv("C:/Users/HP/Desktop/kaleniya/3rd year/statistical_modeling/
online_course_engagement_data.csv")

## Rows: 9000 Columns: 9

## ── Column specification ────────────────────────────────────────────────────
──────

## Delimiter: ","
## chr (1): CourseCategory
```

```
## dbl (8): UserID, TimeSpentOnCourse, NumberOfVideosWatched, NumberOfQuizzes
Ta...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.
```

```r
# View the first few rows of the data
head(data)
```

```
## # A tibble: 6 × 9
##   UserID CourseCategory TimeSpentOnCourse NumberOfVideosWatched
##    <dbl> <chr>                     <dbl>                 <dbl>
## 1   5618 Health                     30.0                    17
## 2   4326 Arts                       27.8                     1
## 3   5849 Arts                       86.8                    14
## 4   4992 Science                    35.0                    17
## 5   3866 Programming                92.5                    16
## 6   8650 Health                     79.5                    12
## # ℹ 5 more variables: NumberOfQuizzesTaken <dbl>, QuizScores <dbl>,
## #   CompletionRate <dbl>, DeviceType <dbl>, CourseCompletion <dbl>
```

```r
# Calculate the percentage of missing values in each column
missing_percentage <- colSums(is.na(data)) / nrow(data) * 100

# Identify columns with more than 60% missing values
cols_to_remove <- names(missing_percentage[missing_percentage > 60])

# Remove columns with more than 60% missing values
data_cleaned <- data %>% select(-all_of(cols_to_remove))

# Replace missing values with mean values for the remaining columns
# Compute means for columns that are numeric
mean_values <- sapply(data_cleaned, function(col) {
  if (is.numeric(col)) {
    mean(col, na.rm = TRUE)
  } else {
    NA
  }
})

# Replace NA values with computed means
data_cleaned <- data_cleaned %>%
  mutate(across(where(is.numeric), ~ replace_na(., mean_values[cur_column()])
))

# View the cleaned data
head(data_cleaned)
```

```
## # A tibble: 6 × 9
##   UserID CourseCategory TimeSpentOnCourse NumberOfVideosWatched
##    <dbl> <chr>                     <dbl>                 <dbl>
## 1   5618 Health                     30.0                    17
## 2   4326 Arts                       27.8                     1
## 3   5849 Arts                       86.8                    14
## 4   4992 Science                    35.0                    17
## 5   3866 Programming                92.5                    16
## 6   8650 Health                     79.5                    12
## # ℹ 5 more variables: NumberOfQuizzesTaken <dbl>, QuizScores <dbl>,
## #   CompletionRate <dbl>, DeviceType <dbl>, CourseCompletion <dbl>
```

```
# save the cleaned dataset to a new CSV file
#write_csv(data_cleaned,"C:/Users/HP/Desktop/kaleniya/3rd year/statistical_mo
deling/Cleaned_Dataset.csv")

# Summary statistics for the entire dataset
summary(data_cleaned)
```

```
##      UserID     CourseCategory     TimeSpentOnCourse NumberOfVideosWatched
##  Min.   :   1   Length:9000        Min.   : 1.005    Min.   : 0.00
##  1st Qu.:2252   Class :character   1st Qu.:25.441    1st Qu.: 5.00
##  Median :4484   Mode  :character   Median :49.818    Median :10.00
##  Mean   :4499                      Mean   :50.164    Mean   :10.02
##  3rd Qu.:6751                      3rd Qu.:75.070    3rd Qu.:15.00
##  Max.   :9000                      Max.   :99.993    Max.   :20.00
##  NumberOfQuizzesTaken   QuizScores    CompletionRate       DeviceType
##  Min.   : 0.000       Min.   :50.01   Min.   : 0.00933   Min.   :0.0000
##  1st Qu.: 2.000       1st Qu.:62.28   1st Qu.:25.65361   1st Qu.:0.0000
##  Median : 5.000       Median :74.74   Median :50.26412   Median :1.0000
##  Mean   : 5.091       Mean   :74.71   Mean   :50.34015   Mean   :0.5007
##  3rd Qu.: 8.000       3rd Qu.:87.02   3rd Qu.:75.57249   3rd Qu.:1.0000
##  Max.   :10.000       Max.   :99.99   Max.   :99.97971   Max.   :1.0000
##  CourseCompletion
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3964
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
# Detailed descriptive statistics
# Descriptive statistics for all columns
descr(data_cleaned)
```

```
## Non-numerical variable(s) ignored: CourseCategory
```

```
## Descriptive Statistics
## data_cleaned
## N: 9000
##
```

```
##                     CompletionRate   CourseCompletion   DeviceType   Numbe
rOfQuizzesTaken
## ----------------- --------------- ----------------- ----------- -------
---------------
##               Mean            50.34               0.40         0.50
5.09
##            Std.Dev            28.95               0.49         0.50
3.16
##                Min             0.01               0.00         0.00
0.00
##                 Q1            25.65               0.00         0.00
2.00
##             Median            50.26               0.00         1.00
5.00
##                 Q3            75.57               1.00         1.00
8.00
##                Max            99.98               1.00         1.00
10.00
##                MAD            36.97               0.00         0.00
4.45
##                IQR            49.92               1.00         1.00
6.00
##                 CV             0.58               1.23         1.00
0.62
##           Skewness             0.00               0.42         0.00
-0.03
##       SE.Skewness             0.03               0.03         0.03
0.03
##           Kurtosis            -1.19              -1.82        -2.00
-1.22
##            N.Valid          9000.00            9000.00      9000.00
9000.00
##          Pct.Valid           100.00             100.00       100.00
100.00
##
## Table: Table continues below
##
##
##
##                     NumberOfVideosWatched   QuizScores   TimeSpentOnCourse
UserID
## ----------------- ---------------------- ------------ ------------------
---------
##               Mean                  10.02        74.71              50.16
4498.89
##            Std.Dev                   6.03        14.38              28.49
2596.85
##                Min                   0.00        50.01               1.01
1.00
##                 Q1                   5.00        62.28              25.44
```

```
2251.50
##            Median                   10.00        74.74                49.82
4483.50
##                Q3                   15.00        87.02                75.07
6751.50
##               Max                   20.00        99.99                99.99
9000.00
##               MAD                    7.41        18.32                36.80
3335.85
##               IQR                   10.00        24.74                49.63
4499.50
##                CV                    0.60         0.19                 0.57
0.58
##          Skewness                    0.00         0.02                 0.02
0.00
##       SE.Skewness                    0.03         0.03                 0.03
0.03
##          Kurtosis                   -1.21        -1.18                -1.19
-1.21
##           N.Valid                 9000.00      9000.00              9000.00
9000.00
##         Pct.Valid                  100.00       100.00               100.00
100.00

# summary statistics for specific columns
for (col_name in colnames(data_cleaned)) {
  cat("\nSummary for column:", col_name, "\n")
  print(summary(data_cleaned[[col_name]]))
}

##
## Summary for column: UserID
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1    2252    4484    4499    6751    9000
##
## Summary for column: CourseCategory
##    Length     Class      Mode
##      9000 character character
##
## Summary for column: TimeSpentOnCourse
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.005  25.441  49.818  50.164  75.070  99.993
##
## Summary for column: NumberOfVideosWatched
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   10.00   10.02   15.00   20.00
##
## Summary for column: NumberOfQuizzesTaken
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   5.000   5.091   8.000  10.000
```

```
##
## Summary for column: QuizScores
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.01   62.28   74.74   74.71   87.02   99.99
##
## Summary for column: CompletionRate
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00933 25.65361 50.26412 50.34015 75.57249 99.97971
##
## Summary for column: DeviceType
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5007  1.0000  1.0000
##
## Summary for column: CourseCompletion
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3964  1.0000  1.0000
```

```
# install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
# Filter columns which only include numeric values for plotting
data_for_plotting <- data_cleaned %>% select(-c(1, 2,ncol(data_cleaned)-1, nc
ol(data_cleaned)))

# Ensure all selected columns are numeric
data_for_plotting <- data_for_plotting %>% select_if(is.numeric)

# Check the structure of the numeric data
str(data_for_plotting)
```

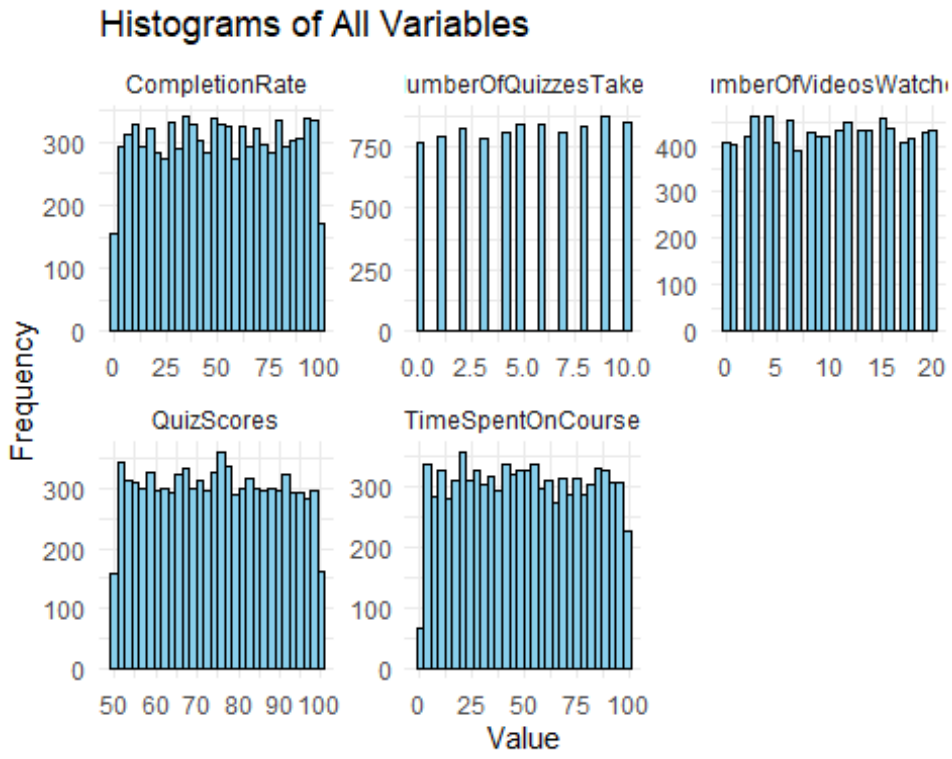
```
## tibble [9,000 × 5] (S3: tbl_df/tbl/data.frame)
##  $ TimeSpentOnCourse    : num [1:9000] 30 27.8 86.8 35 92.5 ...
##  $ NumberOfVideosWatched: num [1:9000] 17 1 14 17 16 12 10 16 8 15 ...
##  $ NumberOfQuizzesTaken : num [1:9000] 3 5 2 10 0 7 2 3 4 10 ...
##  $ QuizScores           : num [1:9000] 50.4 62.6 78.5 59.2 98.4 ...
##  $ CompletionRate       : num [1:9000] 20.9 65.6 63.8 95.4 18.1 ...
```

```
# Print the column names to ensure they are correct
print(colnames(data_for_plotting))
```

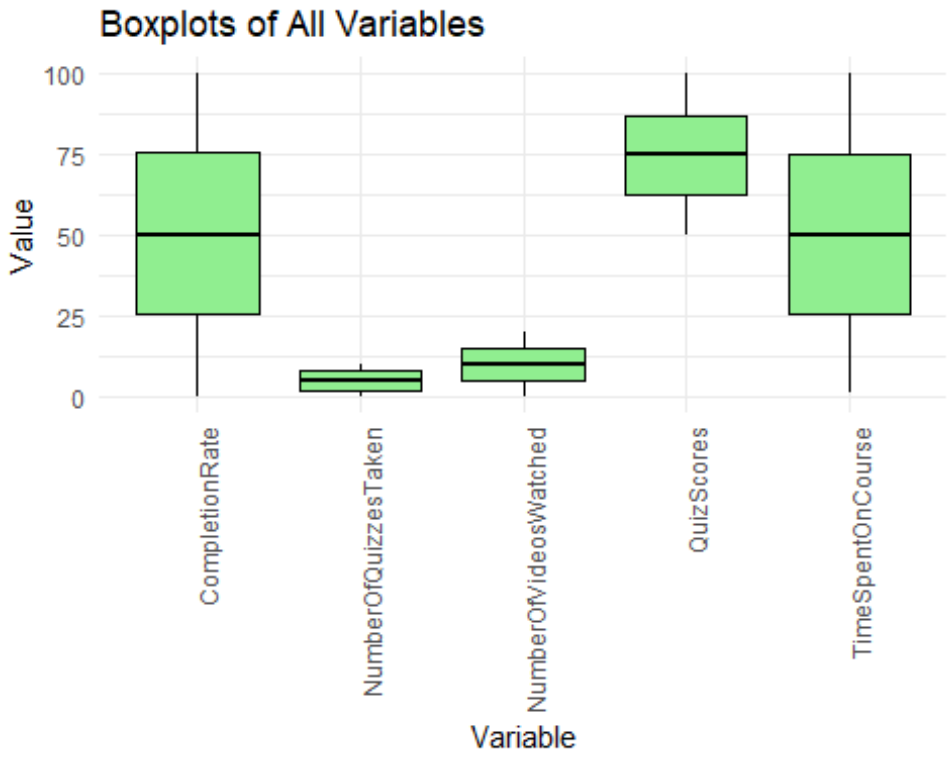
```
## [1] "TimeSpentOnCourse"     "NumberOfVideosWatched" "NumberOfQuizzesTaken"
## [4] "QuizScores"            "CompletionRate"
```

```
# Plotting histograms and boxplots
# Gather the numeric data into a long format for easier plotting with ggplot2
data_long <- data_for_plotting %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value
")
```

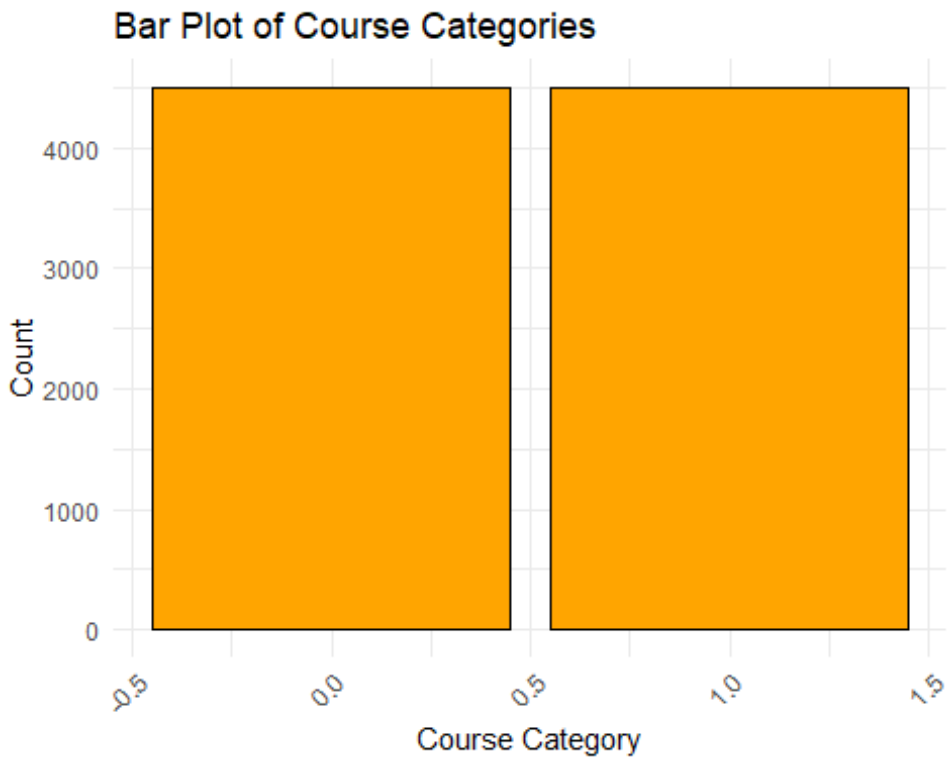
```r
# Combined histograms
ggplot(data_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  facet_wrap(~ variable, scales = "free") +
  theme_minimal() +
  labs(title = "Histograms of All Variables", x = "Value", y = "Frequency")
```



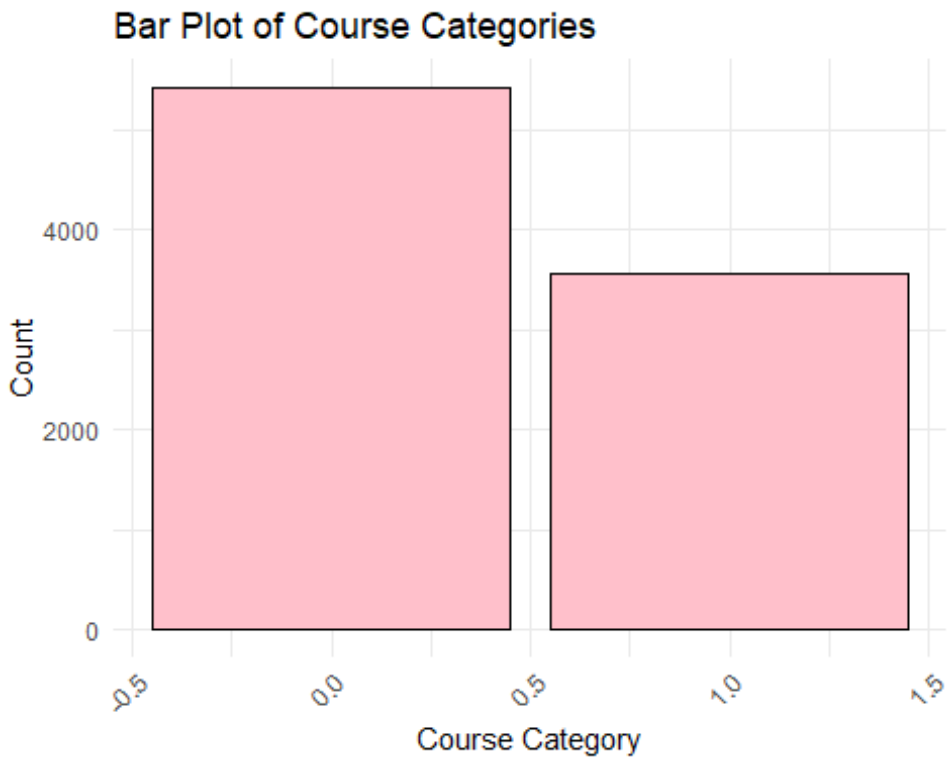

```r
# Combined boxplots
ggplot(data_long, aes(x = variable, y = value)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Boxplots of All Variables", x = "Variable", y = "Value")
```



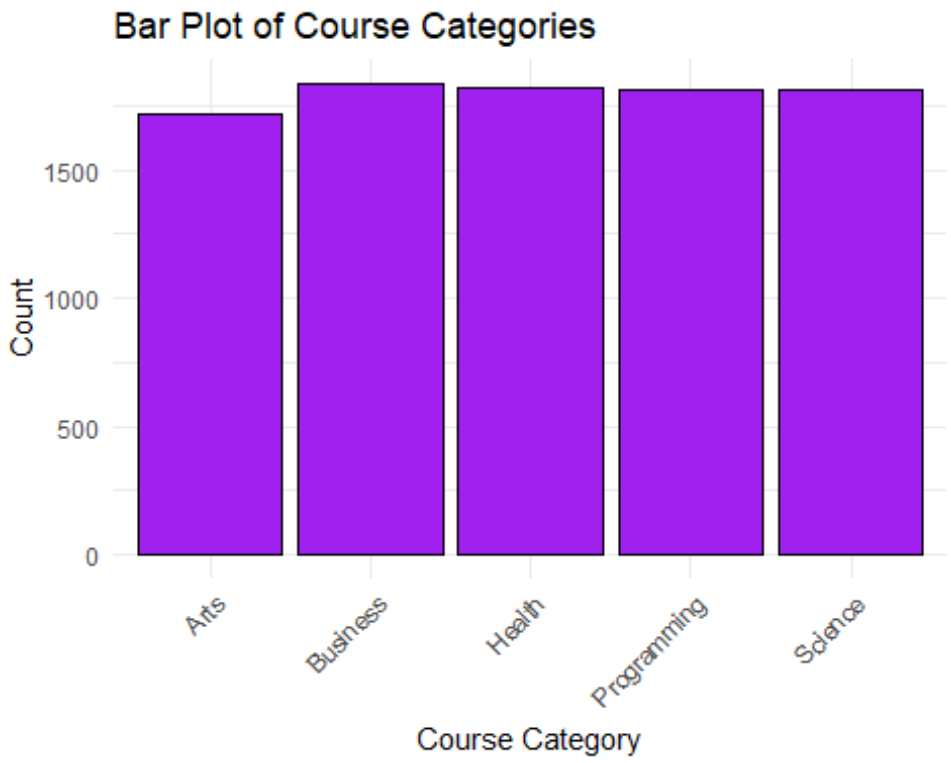

```r
# Bar Plot for DeviceType
ggplot(data_cleaned, aes(x = DeviceType)) +
  geom_bar(fill = "orange", color = "black") +
  theme_minimal() +
  labs(title = "Bar Plot of Course Categories", x = "Course Category", y = "C
ount") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Bar Plot for CourseCompletion
ggplot(data_cleaned, aes(x = CourseCompletion)) +
  geom_bar(fill = "pink", color = "black") +
  theme_minimal() +
  labs(title = "Bar Plot of Course Categories", x = "Course Category", y = "C
ount") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Bar Plot for CourseCategory
ggplot(data_cleaned, aes(x = CourseCategory)) +
  geom_bar(fill = "purple", color = "black") +
  theme_minimal() +
  labs(title = "Bar Plot of Course Categories", x = "Course Category", y = "C
ount") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Pie Chart for CourseCategory
data_cleaned %>%
  count(CourseCategory) %>%
  ggplot(aes(x = "", y = n, fill = CourseCategory)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  theme_minimal() +
  labs(title = "Pie Chart of Course Categories", x = "", y = "") +
  theme(axis.text.x = element_blank(), axis.ticks = element_blank(), panel.gr
id = element_blank())
```

Pie Chart of Course Categories