Week4 EDA

```r
# Load the 'college_expenses_and_enrollment.Rds' file into a data frame named 'dat'
dat <- import('college_expenses_and_enrollment.Rds')

# Display a variable summary table for the 'dat' data frame
vtable(dat)
```

```r
# Display the first few rows of the 'dat' data frame
head(dat,3)
```

```
  UNITID STABBR              type Total.Income Total.Expenses   Tuition
1 100654     AL     Public 4-Year    159374242      103277451  46521234
2 100663     AL     Public 4-Year   2958953209     2693408556 209220942
3 100690     AL Private NP 4-Year      8318086        3558638   6602757
    Federal     State   Local   Private      Sales  Research Public.Service
1  45236824  42153936       0   2281740   23180508  10150441       16175142
2 384940190 282821823 2350134 106376593 1973243527 297969696      172938412
3    904170         0       0    777723      33436         0              0
  Student.Services Instruction Academic.Support      Other
1         20550501    36797673          7735763   11867931
2         50272077   310526699        186799868 1674901804
3           771771     2358591           428276          0
                     Institution.Name Total.Enrollment Full.Time.Enrollment
1             Alabama A & M University             5859                 5040
2 University of Alabama at Birmingham            19535                12691
3                   Amridge University              597                  216
  Undergraduate.Enrollment
1                     4851
2                    12369
3                      294
```

Table 1: dat

Name	Class	Label	Values
UNITID	numeric	College ID	Num: 100654 to 491118
STABBR	factor	State	'AK' 'AL' 'AR' 'AS' 'AZ' and 54 m
type	factor	College Type (NP = Nonprofit)	'For-Profit 2-Year' 'For-Profit 4-Ye
Total.Income	numeric	Total Income	Num: 6067 to 9703300000
Total.Expenses	numeric	Total Expenses	Num: 4600 to 8629847000
Tuition	numeric	Tuition (Income)	Num: 0 to 1753736000
Federal	numeric	Federal Support (Income)	Num: 0 to 1071432208
State	numeric	State Support (Income)	Num: 0 to 909496000
Local	numeric	Local Support (Income)	Num: 0 to 386724142
Private	numeric	Private Support (Income)	Num: -246016 to 1568865000
Sales	numeric	Sales (Income)	Num: -1073664 to 8471746000
Research	numeric	Research (Expense)	Num: 0 to 2910863000
Public.Service	numeric	Public Service (Expense)	Num: -32 to 807399000
Student.Services	numeric	Student Services (Expense)	Num: 0 to 284363000
Instruction	numeric	Instruction (Expense)	Num: 0 to 2611938000
Academic.Support	numeric	Academic Support (Expense)	Num: 0 to 919577000
Other	numeric	Other Expenses	Num: -1 to 5543912000
Institution.Name	character	NULL	
Total.Enrollment	numeric	NULL	Num: 9 to 131629
Full.Time.Enrollment	numeric	NULL	Num: 0 to 131629
Undergraduate.Enrollment	numeric	NULL	Num: 0 to 103711

1. Initial analysis for finding Average total income and expenditure of Colleges grouped by state

In the initial analysis, the dataset was filtered to exclude rows where 'Total.Income' and 'Total.Expenses' were missing in the 'dat' data frame. Subsequently, a new data frame 'd1' was created, focusing exclusively on the filtered rows. Additional insights were derived by grouping the data based on the 'state' column, and the mean values for 'Total.Income' and 'Total.Expenses' were calculated for each state.

The resulting summary included the average total income and expenditure figures for each state. To narrow down the focus for further analysis, the top 10 states were selected based on the highest average total income. This step aims to identify states with notable financial performance, to perform a more detailed examination of these top-performing states in subsequent analyses.

```r
# Filter rows where 'Total.Income' and 'Total.Expenses' are
# not missing in the 'dat' data frame
d1 <- dat %>%
  filter(!is.na(Total.Income) & !is.na(Total.Expenses)) %>%

  # Create a new column 'state' with values from 'STABBR'
  mutate(state = STABBR) %>%

  # Group the data by 'state'
  group_by(state) %>%

  # Calculate the mean of 'Total.Income' and 'Total.Expenses' for each 'state'
  summarize(
    Average_Total_Income = mean(Total.Income),
    Average_Total_Expenditure = mean(Total.Expenses)
  ) %>%

  # Select the top 10 states based on the highest 'Average_Total_Income'
  top_n(10, Average_Total_Income)

head(d1)
```

```
# A tibble: 6 x 3
  state Average_Total_Income Average_Total_Expenditure
  <fct>                <dbl>                     <dbl>
1 CA              170930255.                137645484.
2 CT              245316413.                172946273.
3 DC              275871305.                216375249.
```

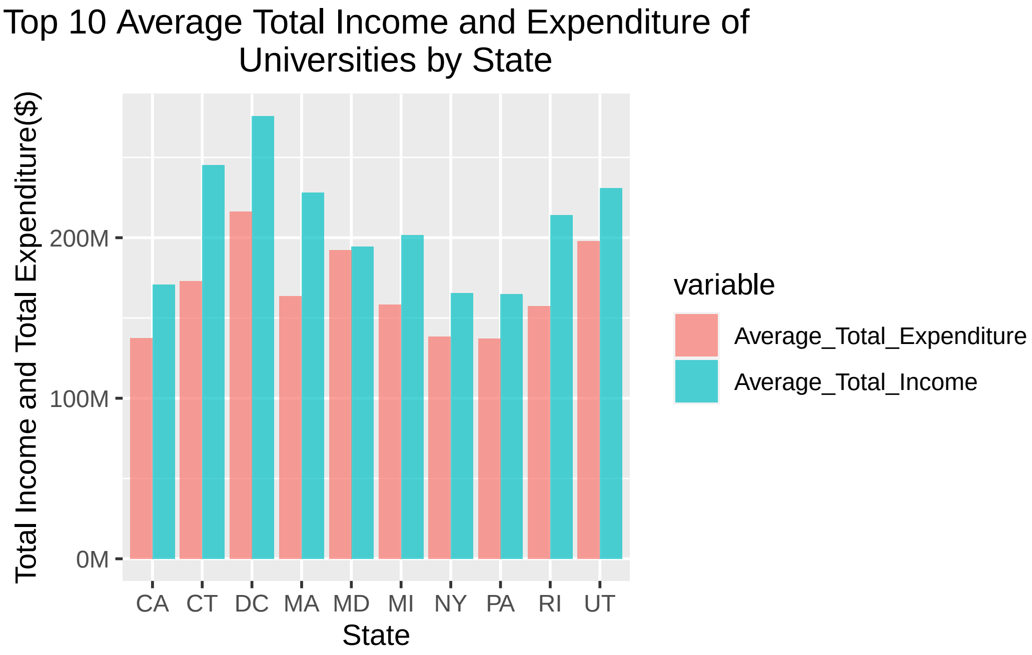
```
4 MA                  228205912.                      163532184.
5 MD                  194465962.                      192324739.
6 MI                  201618944.                      158325522.
```

Plotting a grouped bar graph for above analysis

A grouped bar plot was generated using ggplot, featuring the 'state' variable on the x-axis. The dataset 'df' was derived from 'd1' through the use of pivot_longer, transforming the columns 'Average_Total_Income' and 'Average_Total_Expenditure' into a long format with 'variable' representing the type of financial metric and 'value' capturing the corresponding values. The resulting plot visualized the average total income and expenditure of universities across the top 10 high income states, showcasing two distinct sets of bars differentiated by color.

The y-axis scale was adjusted to present values in millions, indicated by the 'M' suffix. The subsequent analysis will delve into the specific patterns and relationships revealed by this visualization, providing deeper insights into the financial dynamics of higher education institutions in these states.

```r
# Create a grouped bar plot using ggplot with 'state' on the x-axis
df <- d1 %>%
  pivot_longer(cols = c(Average_Total_Income, Average_Total_Expenditure),
  names_to = "variable", values_to = "value")

ggplot(data = df, aes(x = state, y = value, fill = variable)) +
  geom_col(position = "dodge", alpha = 0.7) +

  # Adjust the y-axis scale to display values in millions
  scale_y_continuous(labels = scales::label_number(scale = 1e-6, suffix = "M")) +

  # Set the main title and axis labels
  labs(
    title = "Top 10 Average Total Income and Expenditure of
    Universities by State",
    x = "State",
    y = "Total Income and Total Expenditure($)"
  ) +
theme(
    plot.title = element_text(hjust = 0.5)
)
```



Observation

It is observed that there are some states with less gap between the total income and expenditure of Colleges like MD, NY and PA while others like CT, DC and MA have a higher gap indicating that the expenditure is not in consistence with the income received.

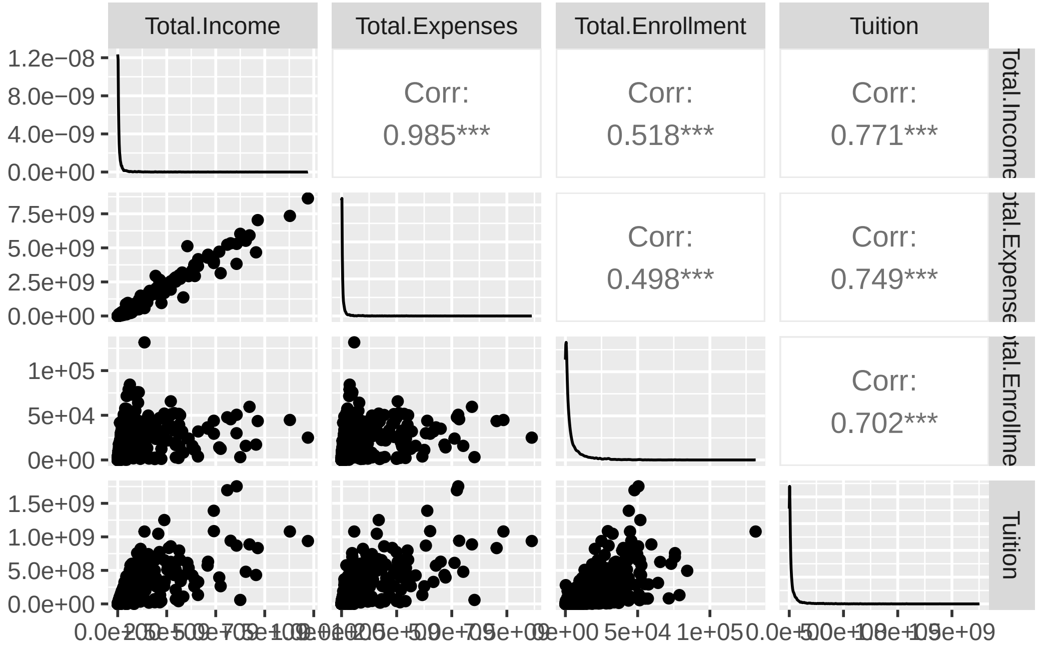
2. Generating a variable summary table for Total income, expenditure, enrollment and tuition

A variable summary table was generated using GGally::ggpairs, focusing on the selected columns (‘Total.Income,’ ‘Total.Expenses,’ ‘Total.Enrollment,’ and ‘Tuition’) within the ‘d3’ data frame. This analysis aimed to provide an insightful overview of the interrelationships and distributions among these key financial and enrollment metrics. Moving forward, the next analysis will delve into more specific patterns and correlations described by the summary table, offering a deeper understanding of the dynamics between total income, expenses, enrollment figures, and tuition costs within the dataset.

Plotting the correlation table

```
# Filter rows where 'Total.Income', 'Total.Expenses', 'Total.Enrollment',
# and 'Tuition' are not missing in the 'dat' data frame
```

```
d3 <- dat %>%
  filter(!is.na(Total.Income) & !is.na(Total.Expenses) &
           !is.na(Total.Enrollment) & !is.na(Tuition)) %>%

  # Select specific columns ('Total.Income', 'Total.Expenses',
  #'Total.Enrollment', 'Tuition') from the filtered data
  select('Total.Income', 'Total.Expenses', 'Total.Enrollment', 'Tuition')

# Create a variable summary table using GGally::ggpairs for the
# selected columns in the 'd3' data frame
GGally::ggpairs(data = d3)
```



Observation

There is a **very strong positive correlation** of 0.985 between total income and total expenditure. This means that as total income increases, total expenditure also tends to increase.

There is a **moderate positive correlation** of 0.518 between total income and total enrollment. This means that as total enrollment increases, total income also tends to increase, but the relationship is not as strong as the relationship between total income and total expenditure. Similarly, there is a moderate positive correlation of 0.498 between total expenditure and total enrollment.

There is a **strong positive correlation** of 0.771 between total income and tuition and 0.749 between total expenditure and tuition and 0.702 between total enrollment and tuition.

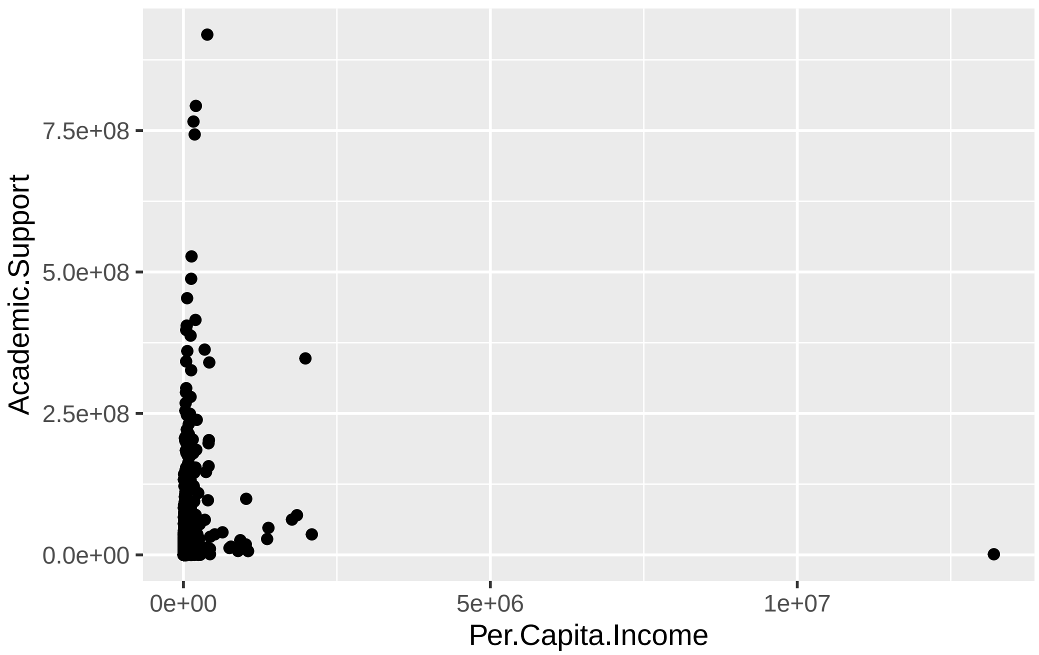
3. Plotting Percapita student income vs Academic support expenses

A new data frame ‘d4’ was created. It includes a derived column, ‘Per.Capita.Income,’ calculated by dividing ‘Total.Income’ by ‘Total.Enrollment.’ To gain further insights into the relationship between per capita income and academic support, a scatter plot was generated using ggplot.

```
# Filter rows where 'Total.Income' and 'Total.Enrollment' are not missing
# in the 'dat' data frame
d4 <- dat %>% filter(!is.na(Total.Income) & !is.na(Total.Enrollment)) %>%

  # Create a new column 'Per.Capita.Income' by dividing 'Total.Income'
  # by 'Total.Enrollment'
  mutate(Per.Capita.Income = Total.Income / Total.Enrollment)%>%
  select('Per.Capita.Income','Academic.Support')

# Display the first few rows of the resulting data frame 'd4'
head(d4)
```

```
  Per.Capita.Income Academic.Support
1          27201.61          7735763
2         151469.32        186799868
3          13933.14           428276
4          26639.64         13383035
5          25018.84         12078035
6          31241.59        102000404
```

```
# Create a scatter plot using ggplot for 'Per.Capita.Income' on the x-axis
# and 'Academic.Support' on the y-axis
ggplot(d4, aes(x = Per.Capita.Income, y = Academic.Support)) +

  # Add points to the scatter plot
  geom_point()
```



Observation

Per capita income of Colleges and Academic support expense have a weak positive correlation. Colleges with low percapita income tend to spend more on academic support while those with higher per capita tend to spend less on academic support.

4. Plotting Percapita student expenditure vs Instruction expenses

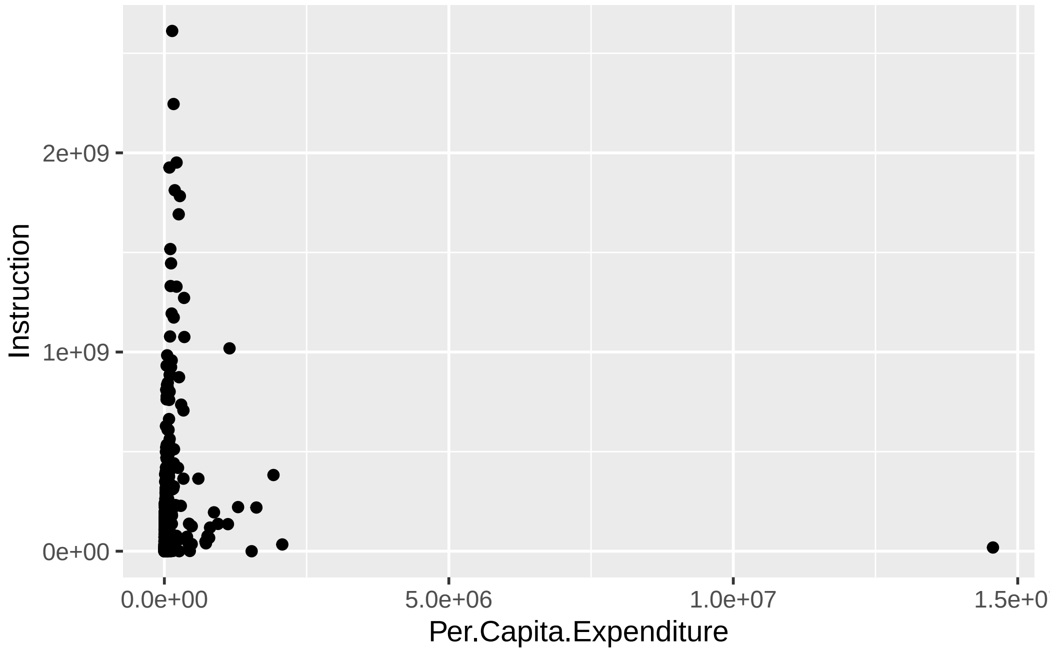
A new data frame 'd5' was created. It includes a derived column, 'Per.Capita.Expenditure,' calculated by dividing 'Total.Expenses' by 'Total.Enrollment.' To gain further insights into the relationship between per capita expenditure and instruction expenses, a scatter plot was generated using ggplot.

```
# Filter rows where 'Total.Expenses' and 'Total.Enrollment' are not
# missing in the 'dat' data frame
d5 <- dat %>% filter(!is.na(Total.Expenses) & !is.na(Total.Enrollment)) %>%
  # Create a new column 'Per.Capita.Expenditure' by dividing 'Total Expenses'
  #by 'Total.Enrollment'
  mutate(Per.Capita.Expenditure = Total.Expenses / Total.Enrollment)%>%
  select('Per.Capita.Expenditure','Instruction')

# Display the first few rows of the resulting data frame 'd5'
```

```
head(d5)
```

```
  Per.Capita.Expenditure Instruction
1              17627.146    36797673
2             137876.046   310526699
3               5960.868     2358591
4              22994.507    69698533
5              16819.052    40147376
6              24023.634   387462769
```

```
# Create a scatter plot using ggplot for 'Per.Capita.Expenditure' on
# the x-axis and 'Instruction' on the y-axis
ggplot(d5, aes(x = Per.Capita.Expenditure, y = Instruction)) +

  # Add points to the scatter plot
  geom_point()
```



Observation

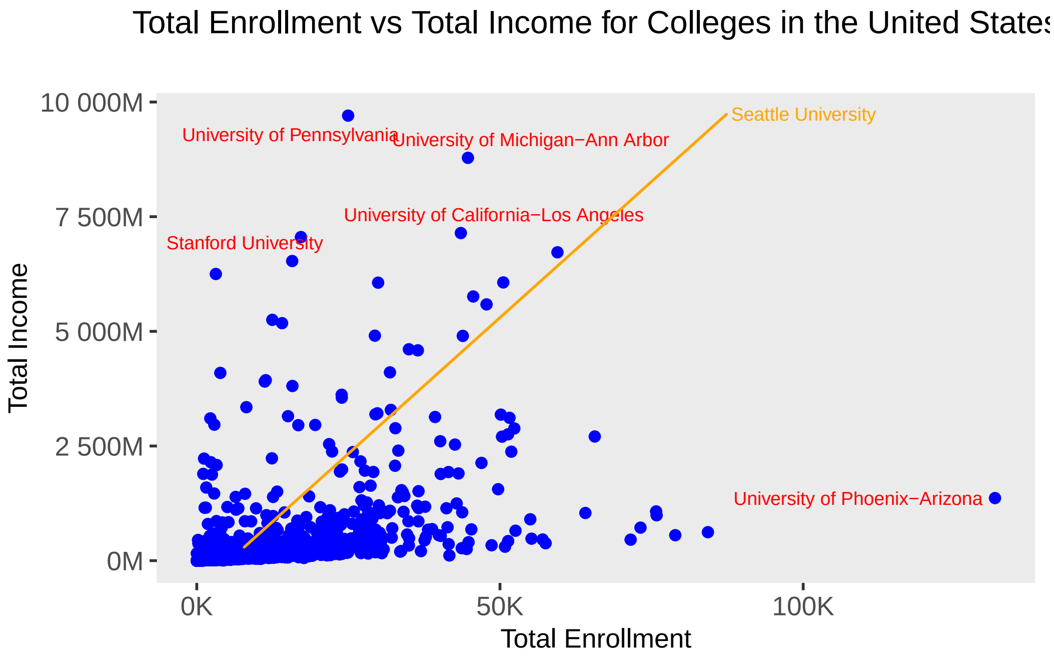
Per capita expenditure of Colleges and instruction expense have a weak positive correlation. Colleges with low percapita expenditure tend to spend more on instruction while those with higher per capita expenditure tend to spend less on instruction.

5. Plotting a scattered plot for Total income and Total enrollment

A scatter plot was generated using ggplot, depicting the relationship between 'Total.Enrollment' and 'Total.Income' for colleges in the United States. Points on the scatter plot were color-coded in blue, and red-colored text labels were added for data points representing institutions with 'Total.Income' greater than 7000000000, 'Total.Enrollment' greater than 100000, or belonging to "Seattle University."

The y-axis scale was adjusted to present values in millions ('M' suffix), and the x-axis scale was adjusted to display values in thousands ('K' suffix). This visual exploration aims to identify and highlight institutions with noteworthy enrollment and income characteristics. The subsequent analysis will delve into specific patterns and anomalies revealed in this scatter plot, providing deeper insights into the factors influencing the enrollment and income dynamics of these educational institutions.

```r
# Filter rows where 'Total.Income' and 'Total.Enrollment' are not missing
# in the 'dat' data frame
d2 <- dat %>% filter(!is.na(Total.Income) & !is.na(Total.Enrollment))

ggplot(d2, aes(x = Total.Enrollment, y = Total.Income)) +
  geom_point(color = 'blue') +

  geom_text_repel(
    aes(label = ifelse(Total.Income > 7000000000 | Total.Enrollment >
    100000, Institution.Name, "")),
    color = "red",
    size = 2.5,
    max.overlaps = Inf
  ) +

  geom_text_repel(
    aes(label = ifelse(Institution.Name == "Seattle University", Institution.Name, "")),
    color = "orange",
    size = 2.5,
    max.overlaps = Inf
  ) +

  # Adjust the y-axis scale to display values in millions
  scale_y_continuous(labels = scales::label_number(scale = 1e-6, suffix = "M")) +
  # Adjust the x-axis scale to display values in thousands
  scale_x_continuous(labels = scales::label_number(scale = 1e-3,
                                                   suffix = "K")) +

  # Set the main title and axis labels for the plot
  labs(title = "Total Enrollment vs Total Income for Colleges in the United States",
       x = "Total Enrollment", y = "Total Income") +
  theme(
    # Remove background checks
    panel.grid.major = element_blank(),  # Remove major grid lines
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    # Adjust axis text size
    axis.text = element_text(size = 10, margin = margin(b=20)),
    # Adjust label text size
    text = element_text(size = 10),
    # Adjust title placement
    plot.title = element_text(size = 11.5, hjust = 0.5,
                              margin = margin(b=20))

)
```



Observation

The scatter plot exhibits a positive correlation between total enrollment and total income. This means that colleges with higher enrollment tend to have higher total income. However, the correlation is not perfect, and there are some outliers. For example, the University of Phoenix-Arizona has a relatively high enrollment but a relatively low total income.

Impact of Enrollment on Income:

Universities with high enrollment and high income like University of Michigan and University of California, likely benefit from both tuition revenue and other income sources like federal and state support.

Universities with high Enrollment, Low Income could suggest lower tuition fees per student or limited external funding.

Universities with low enrollment but high income like Stanford University may rely more heavily on federal or state funding or other alternative income sources besides tuition.

Universities with low Enrollment, Low Income could indicate low tuition fees, limited external funding, and a smaller student body overall. A large number of colleges fall under this category according to the plot.