Analyzing Enrollment of Core Classes for CS/MATH/STA/ECON Majors.

```r
# Load necessary libraries
library(dplyr)
library(readr)
library(tidyr)
library(ggplot2)
library(stringr)
```

```r
# Load the data
data <- read_csv("data/data.csv")
```

```
New names:
Rows: 28 Columns: 28
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(1): ...1 dbl (27): ECON 101, ECON 104, ECON 201, ECON 204, ECON 205 (B.S.),
ECON 210,...
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 message.
* `` -> `...1`
```

```r
colnames(data)[1] <- "Semester"

# Function to identify academic years based on semesters
identify_academic_year <- function(semester) {
  year <- substr(semester, 2, 4)
  if(substr(semester, 1, 1) == "F") {
    return(paste0("AY", year, "-", as.numeric(year) + 1))
  } else {
    return(paste0("AY", as.numeric(year) - 1, "-", year))
  }
}

# Add academic year column
data$AcademicYear <- sapply(data$Semester, identify_academic_year)

# Group data by academic year, then sum each course's enrollment numbers
yearly_sum <- data %>%
  group_by(AcademicYear) %>%
  summarise(across(-Semester, sum, na.rm = TRUE))
```

```r
# Now, pivot the data to long format to identify active years easily
data_long <- yearly_sum %>%
  pivot_longer(-AcademicYear, names_to = "Course", values_to = "Enrollment")

# Identify years with any active enrollments (sum of enrollments across courses > 0)
active_years <- data_long %>%
  group_by(AcademicYear) %>%
  summarise(TotalEnrollment = sum(Enrollment, na.rm = TRUE)) %>%
  filter(TotalEnrollment > 0) %>%
  select(AcademicYear)

# Filter the original yearly_sum to include only active years
yearly_sum_active <- yearly_sum %>%
  filter(AcademicYear %in% active_years$AcademicYear)

# Print the yearly sum of enrollments for each course in active years
print(yearly_sum_active)
```

```
# A tibble: 6 x 28
  AcademicYear `ECON 101` `ECON 104` `ECON 201` `ECON 204` `ECON 205 (B.S.)`
  <chr>             <dbl>      <dbl>      <dbl>      <dbl>              <dbl>
1 AY18-19             295          0        329          0                179
2 AY19-20             502          0        419          0                274
3 AY20-21             422        308        374        250                223
4 AY21-22             447        215        365        207                157
5 AY22-23             515        335        244        197                107
6 AY23-24             668        314        537        228                141
# i 22 more variables: `ECON 210` <dbl>, `COMPSCI 101` <dbl>,
#   `COMPSCI 201` <dbl>, `COMPSCI 210` <dbl>, `COMPSCI 250` <dbl>,
#   `COMPSCI 230` <dbl>, `COMPSCI 330` <dbl>, `MATH 221` <dbl>,
#   `MATH 245` <dbl>, `MATH 401` <dbl>, `MATH 431` <dbl>, `MATH 501` <dbl>,
#   `MATH 531` <dbl>, `STA 199` <dbl>, `STA 210` <dbl>, `STA 211` <dbl>,
#   `STA 360` <dbl>, `STA 250` <dbl>, `STA 432` <dbl>, `STA 440` <dbl>,
#   `PHY 161` <dbl>, `PHY 162` <dbl>
```

```r
# Pivot the data to a long format for easier calculations
data_long <- yearly_sum %>%
  pivot_longer(-AcademicYear, names_to = "Course", values_to = "Enrollment") %>%
  filter(Enrollment > 0) # Ensure we only consider courses with enrollments
```

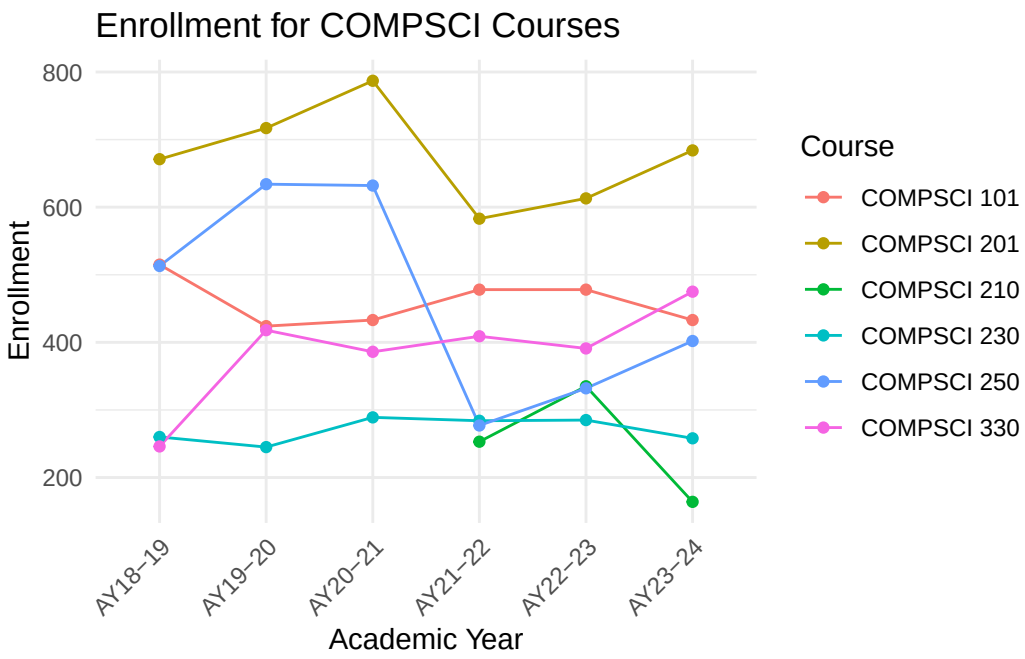
```r
# Fill in missing enrollments with 0 for courses not offered in some years
data_long <- data_long %>%
  group_by(Course, AcademicYear) %>%
  summarize(Enrollment = sum(Enrollment, na.rm = TRUE), .groups = 'drop')

# Calculate the percentage change
data_long <- data_long %>%
  arrange(Course, AcademicYear) %>%
  group_by(Course) %>%
  mutate(Percent_Change = if_else(Enrollment == 0, 0, (Enrollment - lag(Enrollment, defaul
  ungroup()

# Replace NA in Percent_Change with 0 for the first year a course is introduced
data_long$Percent_Change[is.na(data_long$Percent_Change)] <- 0

# View the result
print(data_long)
```

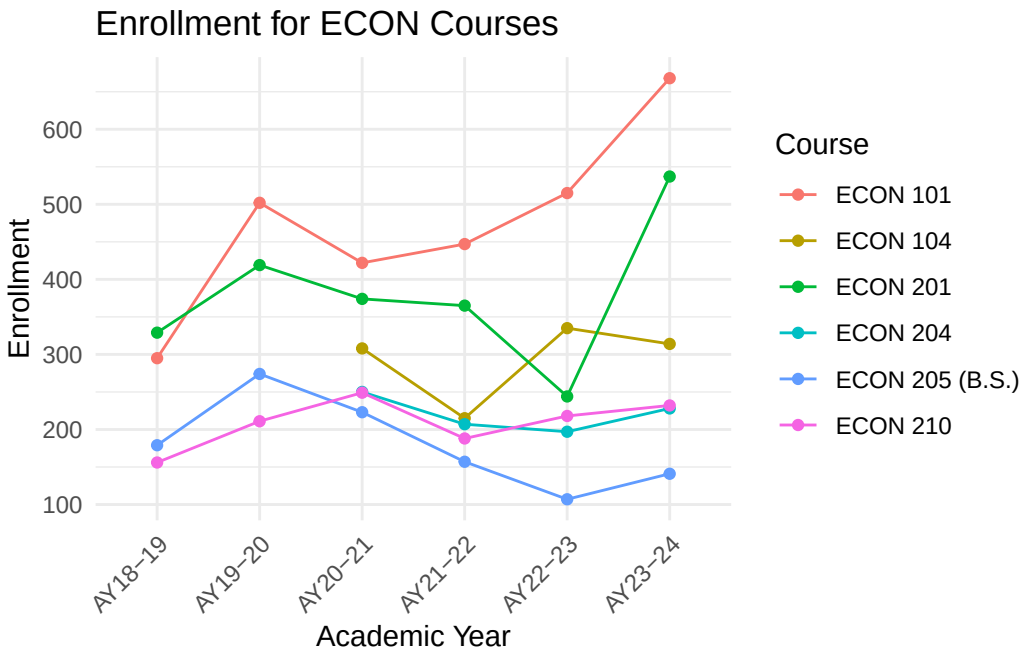
```
# A tibble: 144 x 4
   Course      AcademicYear Enrollment Percent_Change
   <chr>       <chr>             <dbl>          <dbl>
 1 COMPSCI 101 AY18-19             515           0
 2 COMPSCI 101 AY19-20             424         -17.7
 3 COMPSCI 101 AY20-21             433           2.12
 4 COMPSCI 101 AY21-22             478          10.4
 5 COMPSCI 101 AY22-23             478           0
 6 COMPSCI 101 AY23-24             433          -9.41
 7 COMPSCI 201 AY18-19             671           0
 8 COMPSCI 201 AY19-20             717           6.86
 9 COMPSCI 201 AY20-21             787           9.76
10 COMPSCI 201 AY21-22             583         -25.9
# i 134 more rows
```

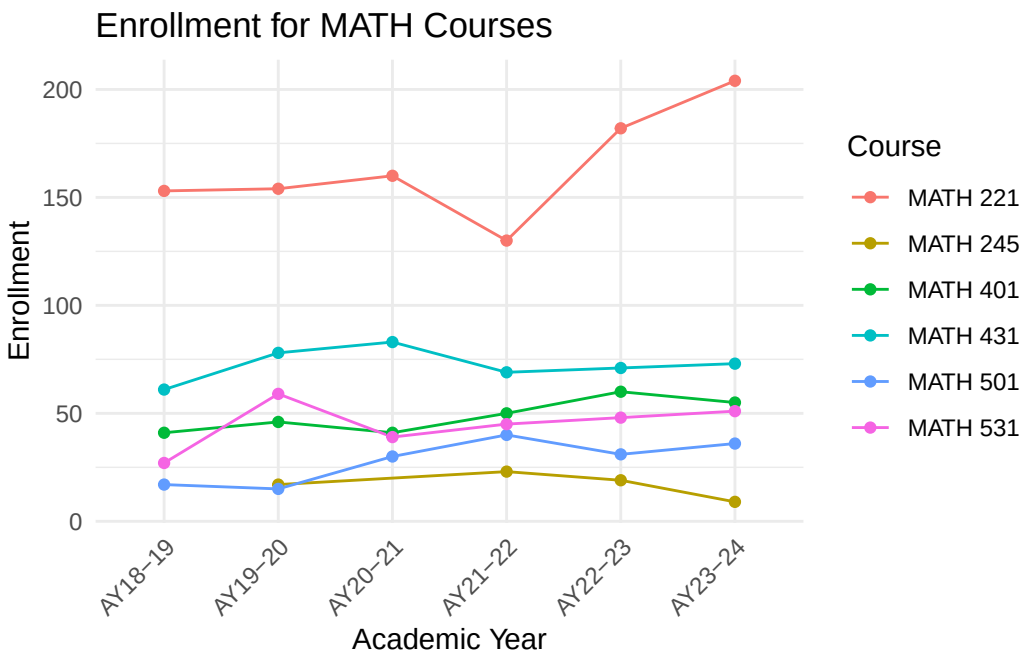
```r
# Identify unique course prefixes (assuming course codes are consistent in format)
course_prefixes <- unique(gsub("([A-Z]+).*", "\\1", data_long$Course))

# Print course prefixes for verification
print(course_prefixes)
```

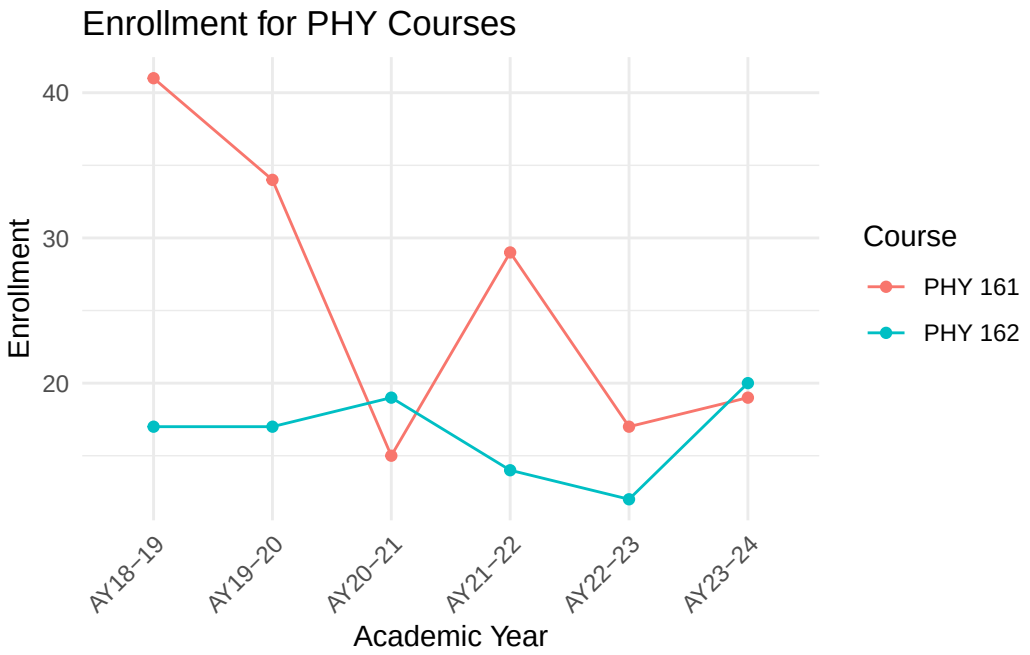
```
[1] "COMPSCI" "ECON"    "MATH"    "PHY"     "STA"
```

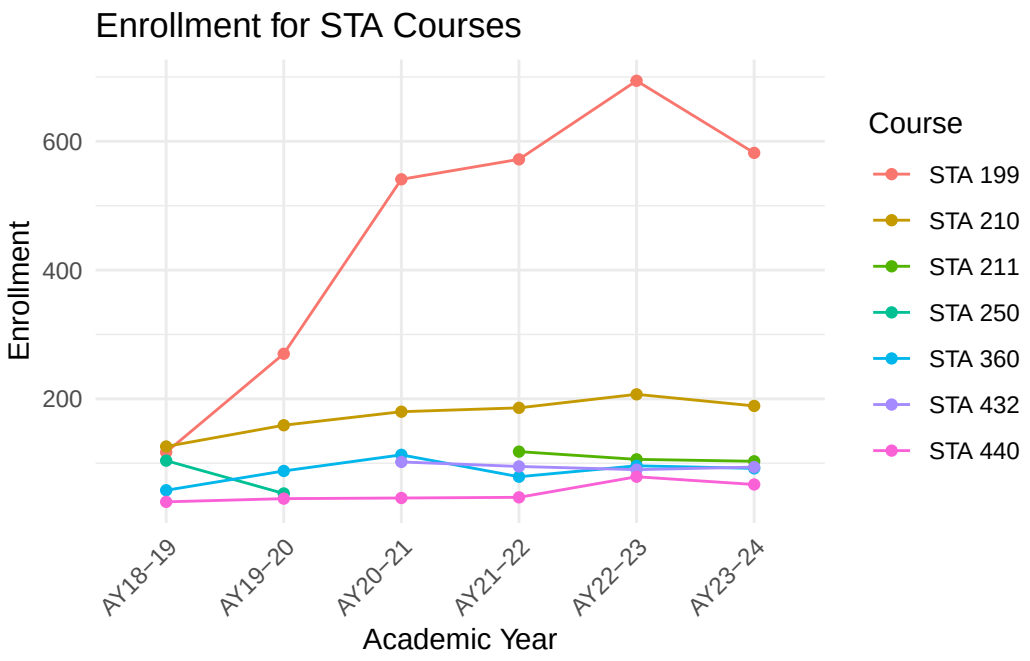
```r
# Function to plot data for each subject area
plot_subject_area <- function(data, metric, course_prefixes) {
  for(prefix in course_prefixes) {
    current_data <- data %>%
      filter(str_detect(Course, pattern = paste0("^", prefix))) %>%
      arrange(Course, AcademicYear)

    p <- ggplot(current_data, aes(x = AcademicYear, y = !!sym(metric), color = Course, gro
      geom_line() +
      geom_point() +
      theme_minimal() +
      labs(title = paste(metric, "for", prefix, "Courses"),
           x = "Academic Year", y = metric, color = "Course") +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))

    print(p)
  }
}

# Identify unique course prefixes
course_prefixes <- unique(str_extract(data_long$Course, "^[A-Z]+"))

# Plot total enrollments for each subject area
plot_subject_area(data_long, "Enrollment", course_prefixes)
```


Enrollment for COMPSCI Courses
800
600
400
200
Enrollment
AY18–19
AY19–20
AY20–21
AY21–22
AY22–23
AY23–24
Academic Year
Course
COMPSCI 101
COMPSCI 201
COMPSCI 210
COMPSCI 230
COMPSCI 250
COMPSCI 330

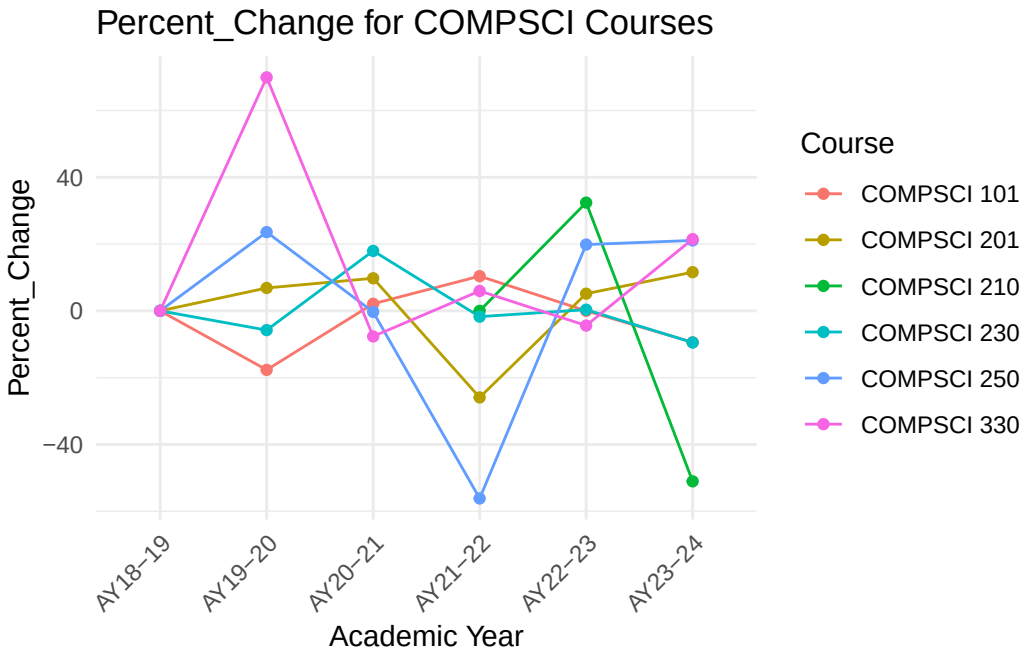

Enrollment for ECON Courses
600
500
400
300
200
100
Enrollment
AY18–19
AY19–20
AY20–21
AY21–22
AY22–23
AY23–24
Academic Year
Course
ECON 101
ECON 104
ECON 201
ECON 204
ECON 205 (B.S.)
ECON 210

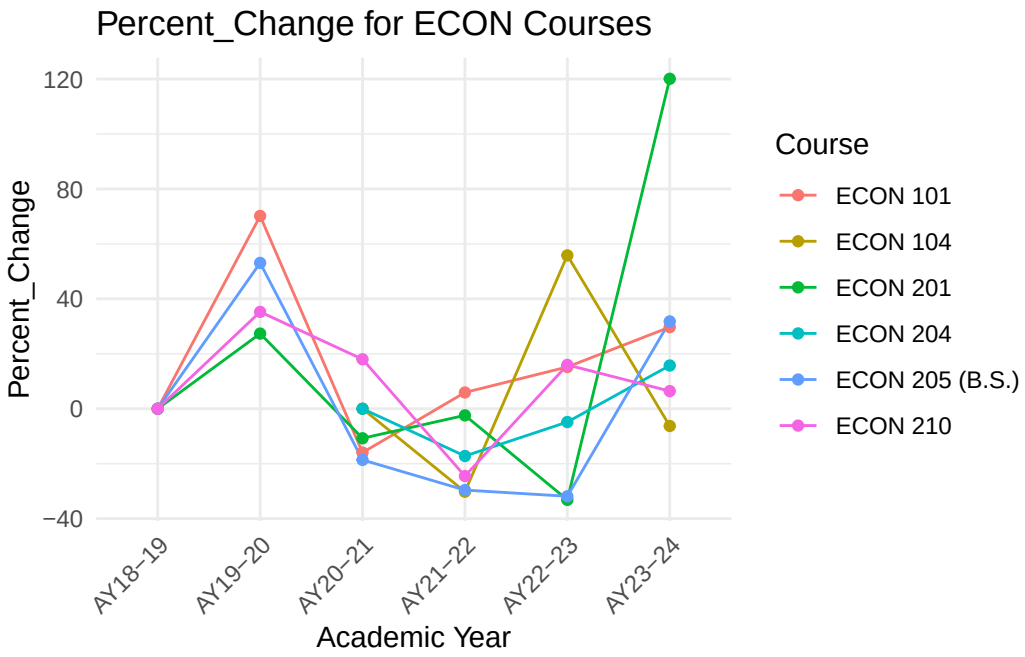

Enrollment for MATH Courses
200
150
100
50
0
Enrollment
AY18–19
AY19–20
AY20–21
AY21–22
AY22–23
AY23–24
Academic Year
Course
MATH 221
MATH 245
MATH 401
MATH 431
MATH 501
MATH 531

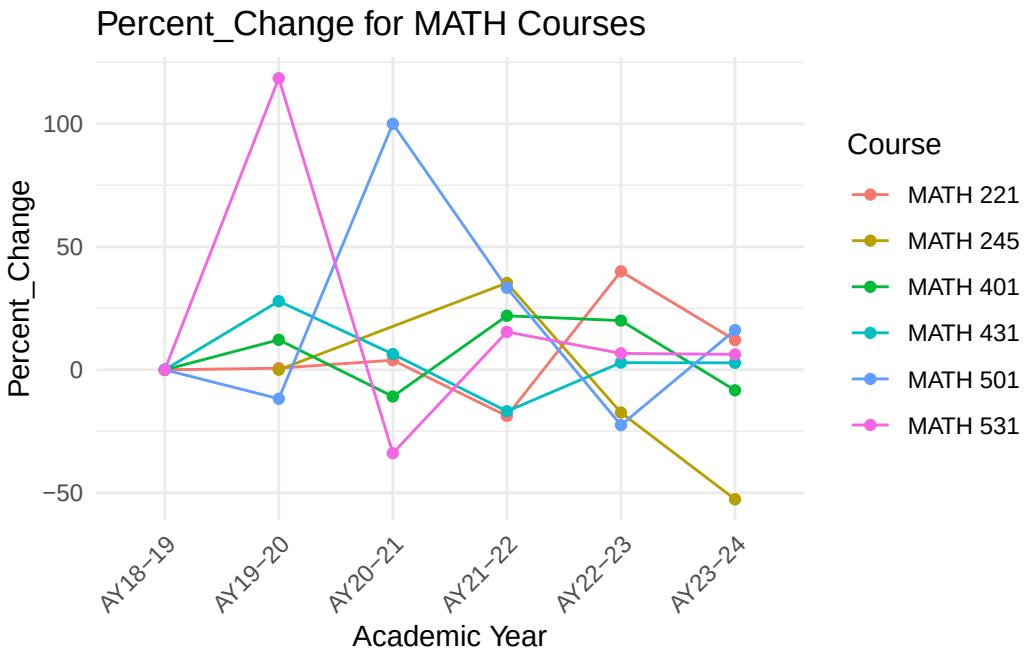

Enrollment for PHY Courses
40
30
20
Enrollment
AY18–19
AY19–20
AY20–21
AY21–22
AY22–23
AY23–24
Academic Year
Course
PHY 161
PHY 162

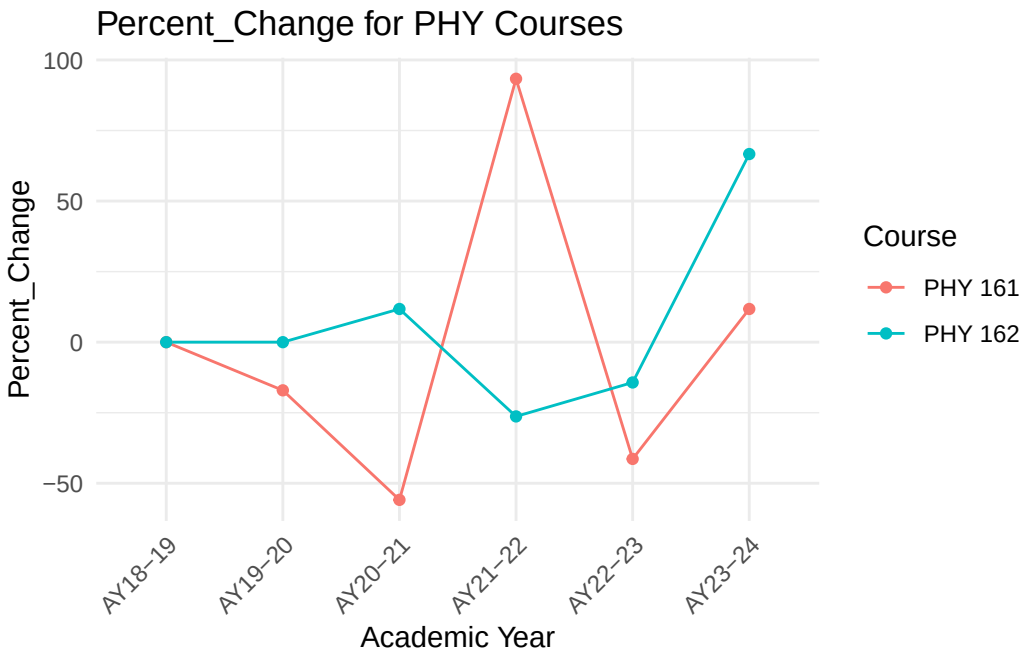


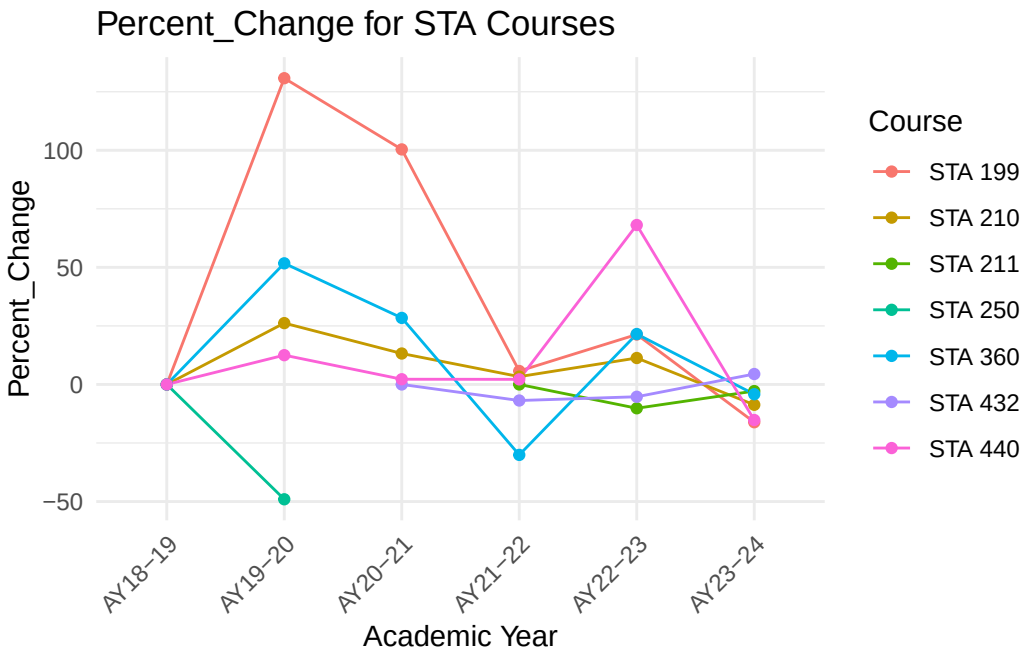

```
# Plot percentage change for each subject area
plot_subject_area(data_long, "Percent_Change", course_prefixes)
```




Percent_Change for ECON Courses
Percent_Change
120
80
40
0
−40
AY18−19
AY19−20
AY20−21
AY21−22
AY22−23
AY23−24
Academic Year
Course
ECON 101
ECON 104
ECON 201
ECON 204
ECON 205 (B.S.)
ECON 210


Percent_Change for MATH Courses
Percent_Change
100
50
0
−50
AY18−19
AY19−20
AY20−21
AY21−22
AY22−23
AY23−24
Academic Year
Course
MATH 221
MATH 245
MATH 401
MATH 431
MATH 501
MATH 531

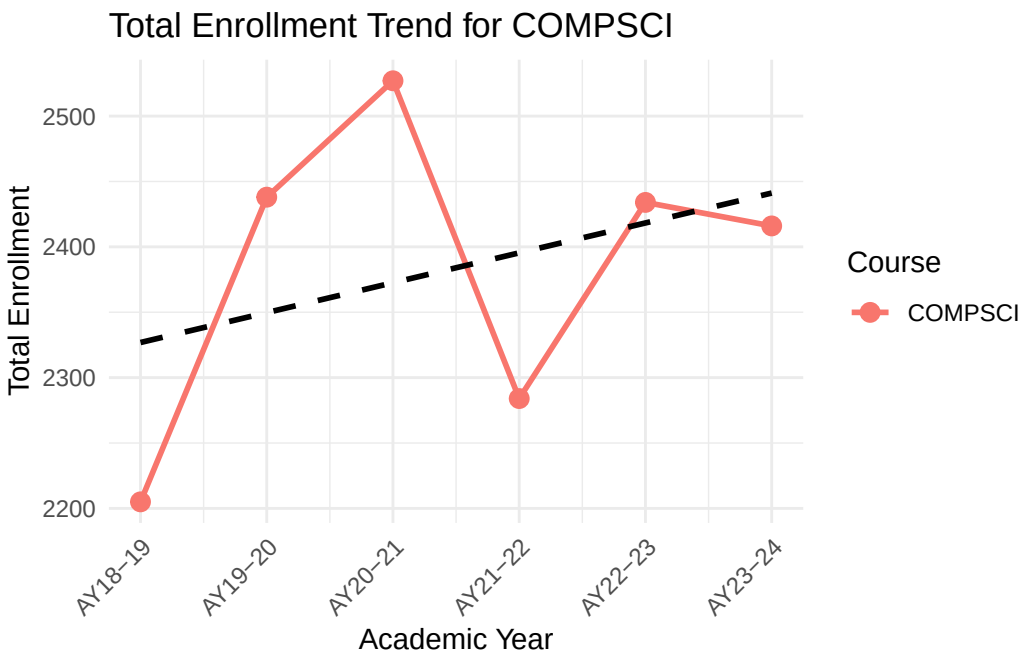

Percent_Change for PHY Courses
100
50
0
−50
Percent_Change
AY18−19
AY19−20
AY20−21
AY21−22
AY22−23
AY23−24
Academic Year
Course
PHY 161
PHY 162


Percent_Change for STA Courses
100
50
0
−50
Percent_Change
AY18−19
AY19−20
AY20−21
AY21−22
AY22−23
AY23−24
Academic Year
Course
STA 199
STA 210
STA 211
STA 250
STA 360
STA 432
STA 440

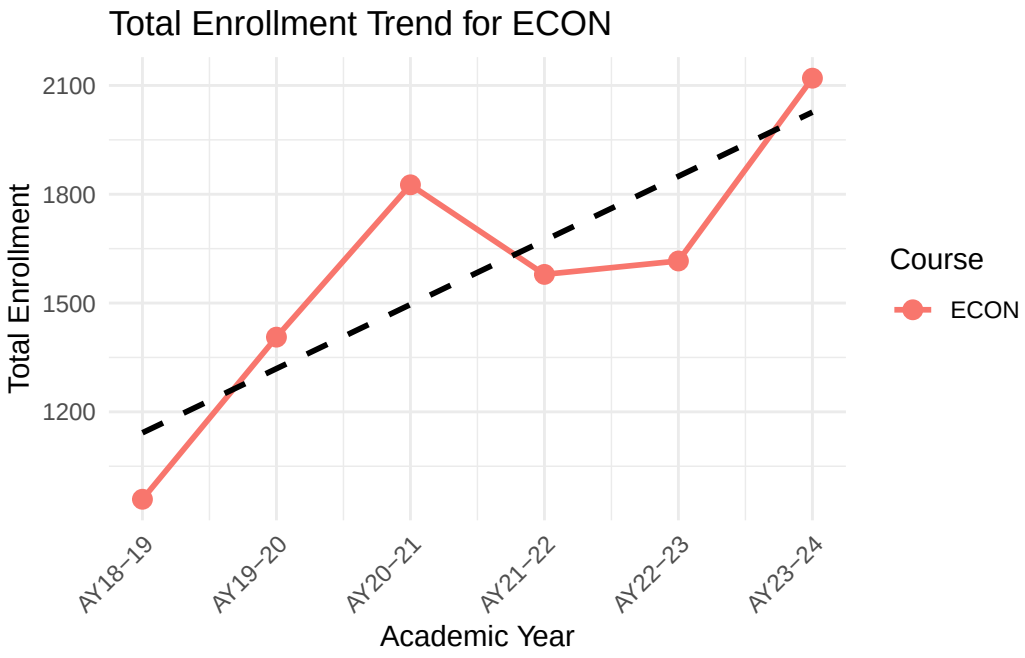
```r
# Loop through each course prefix to fit linear models
model_results <- list()
for(prefix in course_prefixes) {
  # Filter data for the current prefix
  department_subset <- data_long %>%
    filter(grepl(paste0("^", prefix), Course)) %>%
    group_by(AcademicYear) %>%
    summarize(TotalEnrollment = sum(Enrollment, na.rm = TRUE), .groups = 'drop') %>%
    mutate(AcademicYearNumeric = as.numeric(factor(AcademicYear)))

  # Fit the linear model
  model <- lm(TotalEnrollment ~ AcademicYearNumeric, data = department_subset)
  model_results[[prefix]] <- list("model" = model, "data" = department_subset)
}

# Optionally, print summary of models for each prefix
# lapply(model_results, function(x) summary(x$model))
```

```r
for(prefix in course_prefixes) {
  department_subset <- model_results[[prefix]]$data
  p <- ggplot(department_subset, aes(x = AcademicYearNumeric, y = TotalEnrollment)) +
    geom_line(aes(color = prefix), size = 1) +  # Line for each prefix
    geom_point(aes(color = prefix), size = 3) +  # Dots for each data point
    geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "black", size = 1)
    scale_x_continuous(breaks = department_subset$AcademicYearNumeric, labels = department
    labs(title = paste("Total Enrollment Trend for", prefix),
         x = "Academic Year", y = "Total Enrollment",
         color = "Course", linetype = "Trendline") +  # Adjust legend title for linetype
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(p)
}
```

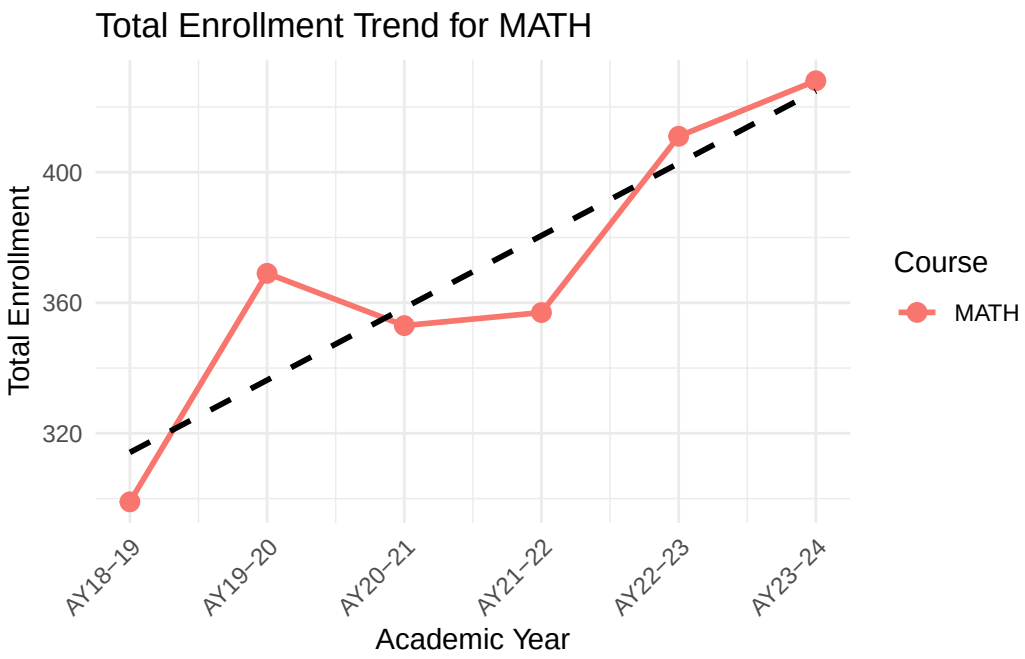
```
`geom_smooth()` using formula = 'y ~ x'
```



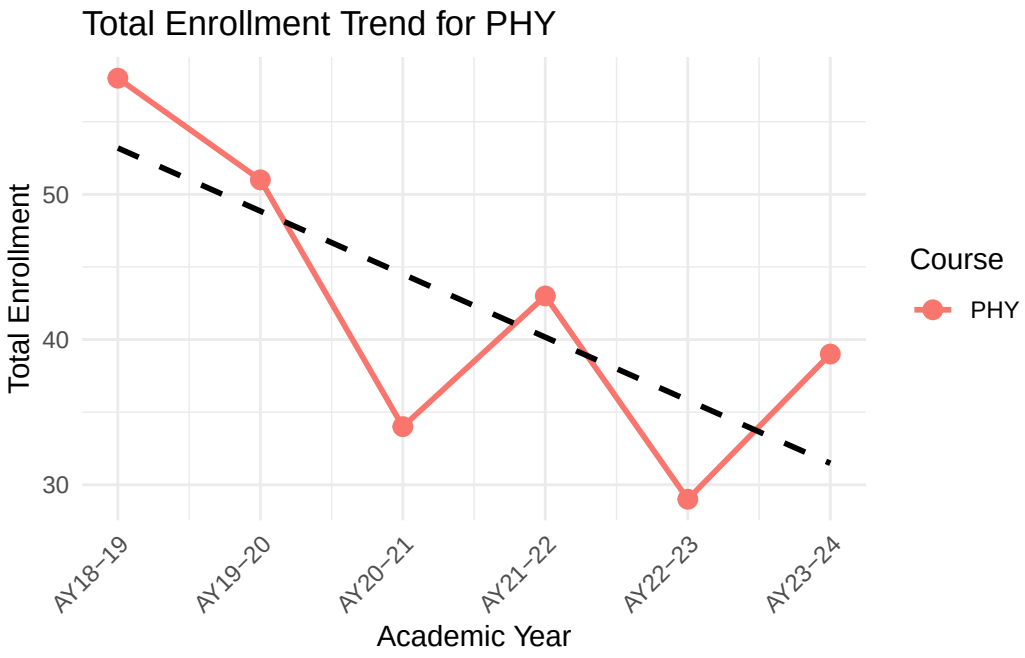

```
`geom_smooth()` using formula = 'y ~ x'
```



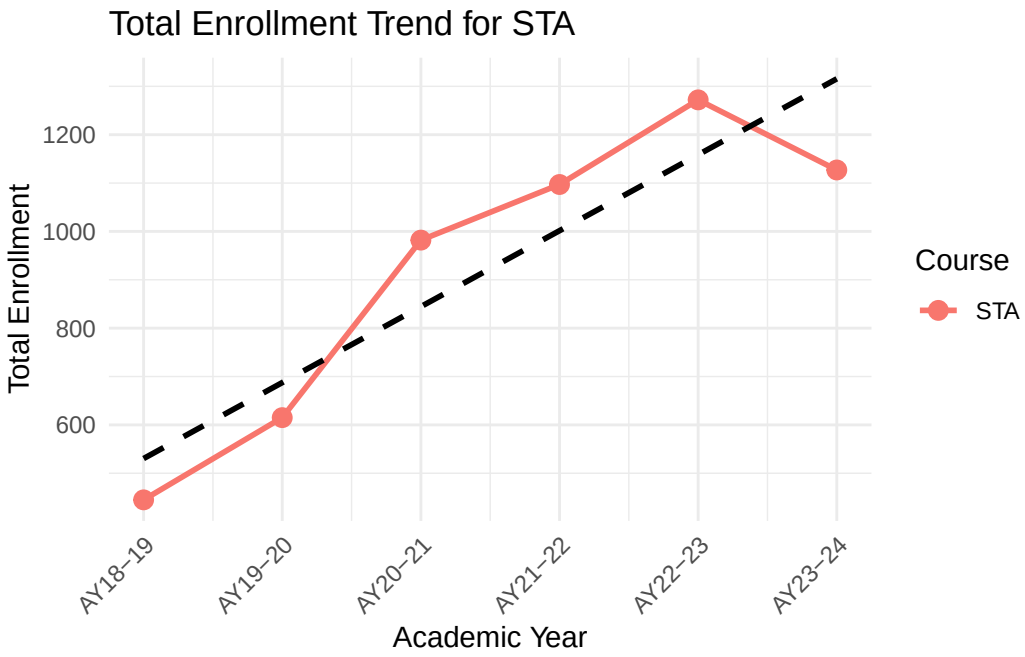

```
`geom_smooth()` using formula = 'y ~ x'
```




```
`geom_smooth()` using formula = 'y ~ x'
```




```
`geom_smooth()` using formula = 'y ~ x'
```


Total Enrollment Trend for STA
Total Enrollment
1200
1000
800
600
AY18–19
AY19–20
AY20–21
AY21–22
AY22–23
AY23–24
Academic Year
Course
STA