Math 5071-Final Project Presentation

Adrian Cao

Department of Statistics and Data Science
Washington University in St.Louis

2023-12-04

Central Question

What Factors Influence the Enrollment of Math courses?

What is Enrollment?

Data Introduction

First, we identify that the data set contains 11 variable columns:

1. X: Appears to be an index or identifier for each record.
2. StdtId: Encrypted Student ID.
3. **Sem: Semester of enrollment, including both year and semester code (05 for Fall, 02 for Spring).**
4. **Crs: Course number.**
5. Sec: Section number of the course.
6. SecType: Section type, where 'S' indicates a standard lecture section.
7. Units: The number of units for the course.
8. DeanCd: Title of the course.
9. GradeOpt: Grade option selected by the student.
10. **PrimeDiv: Primary department of the student.**
11. YRLevel: Year level of the student.

Our analysis will focus on 100- and 200-level courses and exclude discussion sections. For this, we need to filter the data based on the Crs and SecType columns.

Quick Recap of the data

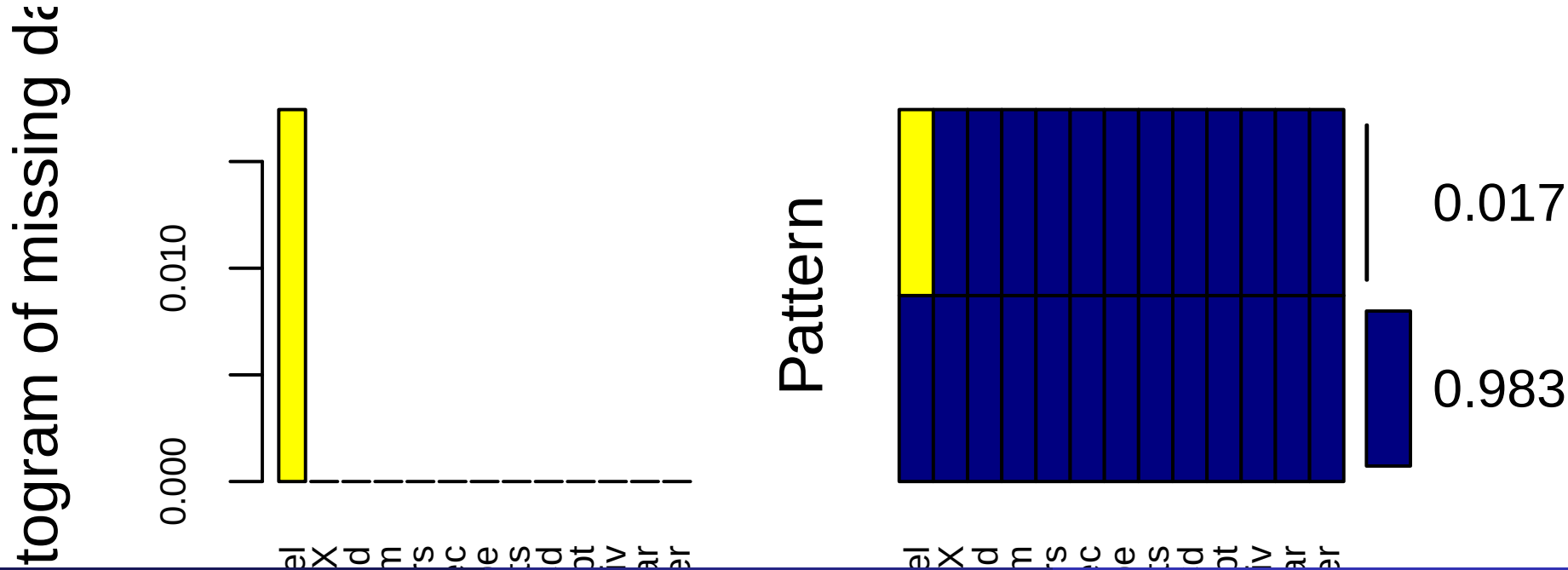
```
##   X StdtId    Sem Crs Sec SecType Units                    I
## 1 1 903629 201705 100  01       S     3 FOUNDATIONS FOR CAI
## 2 2 906787 201705 100  01       S     3 FOUNDATIONS FOR CAI
## 3 3 902106 201705 100  01       S     3 FOUNDATIONS FOR CAI
## 4 4 903791 201705 100  01       S     3 FOUNDATIONS FOR CAI
## 5 5 903781 201705 100  01       S     3 FOUNDATIONS FOR CAI
## 6 6 903684 201705 100  01       S     3 FOUNDATIONS FOR CAI
##   PrimeDiv YRLevel
## 1       LA       1
## 2       LA       1
## 3       LA       1
## 4       LA       1
## 5       LA       1
## 6       LA       1
```

Data Filter and Preparation

Here, I used 'dplyr' package to do the filter, following these steps:

- Filter the dataset for 100- and 200-level courses.
- Exclude discussion sections (SecType != 'S').
- Analyze enrollment trends over the years and during the COVID period.

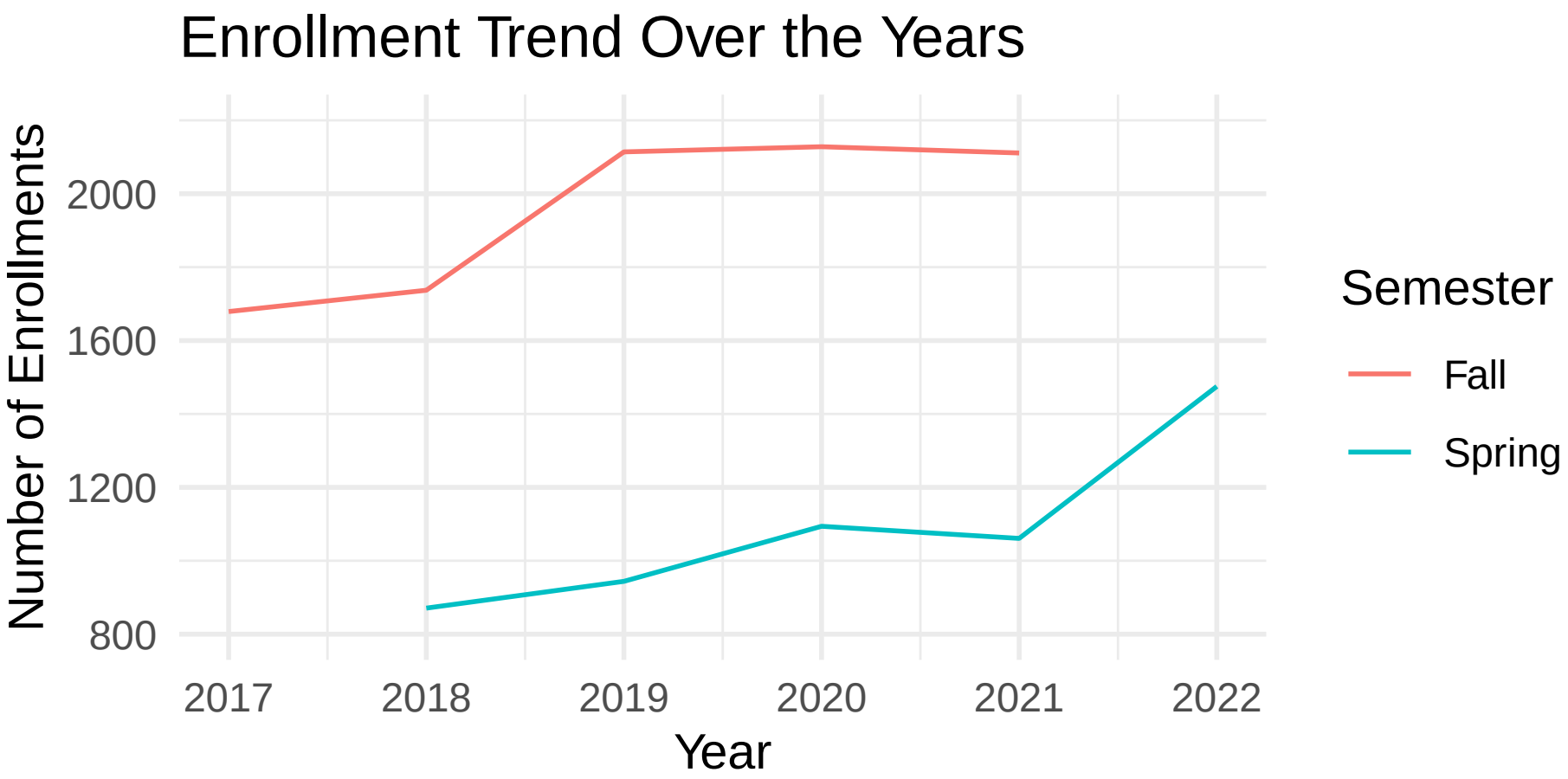
For the missingness, as there is only very few of missing values for YRlevel, so I just delete the missing ones.



Exploratory Data Analysis

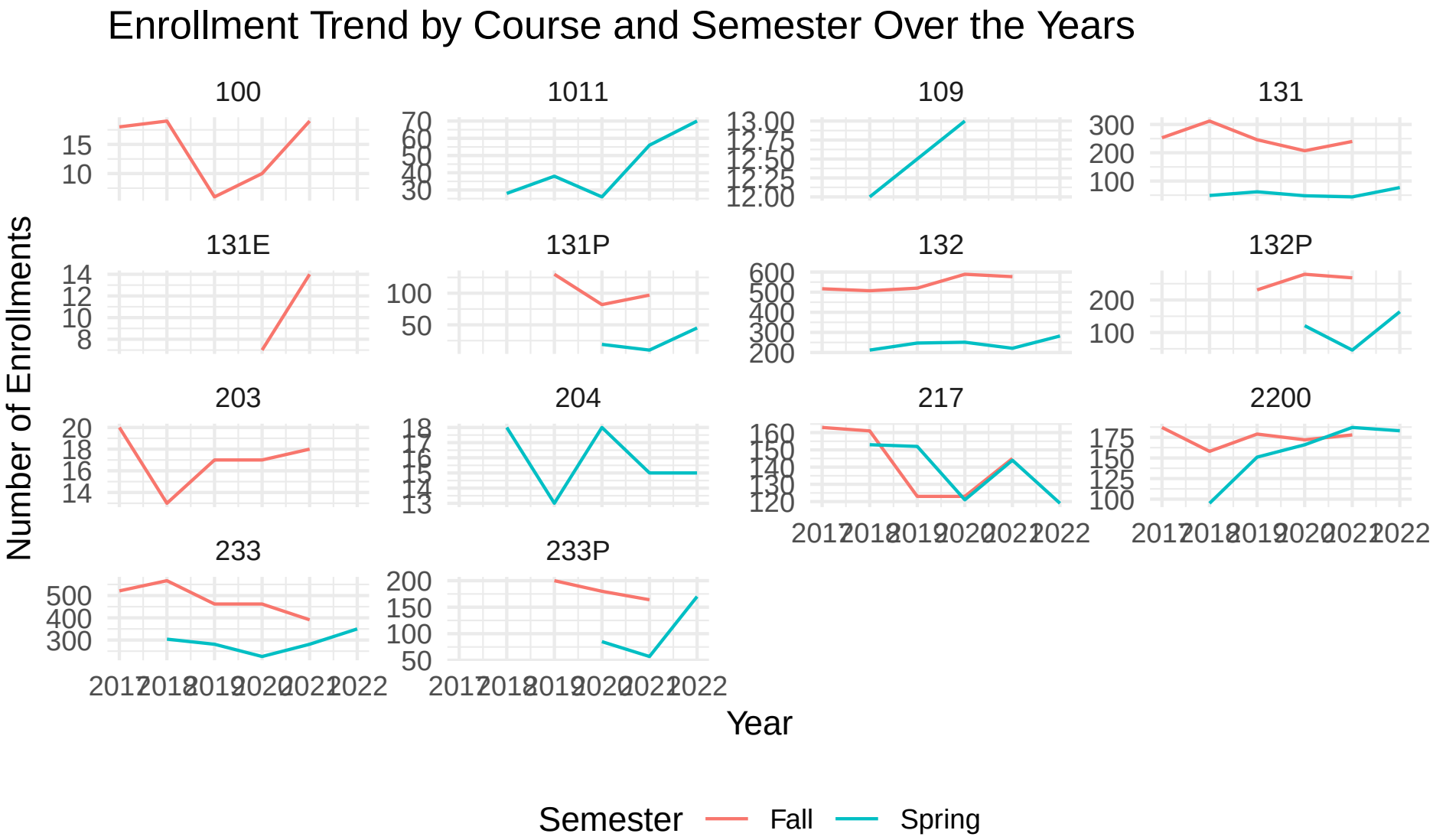
Before diving deep into the model building, we start with some exploratory data analysis to better understand the distribution and charateristics of the data, including trends over time, enrollment patterns by course, department, year levels, and COVID period.

Enrollment Distribution by Years

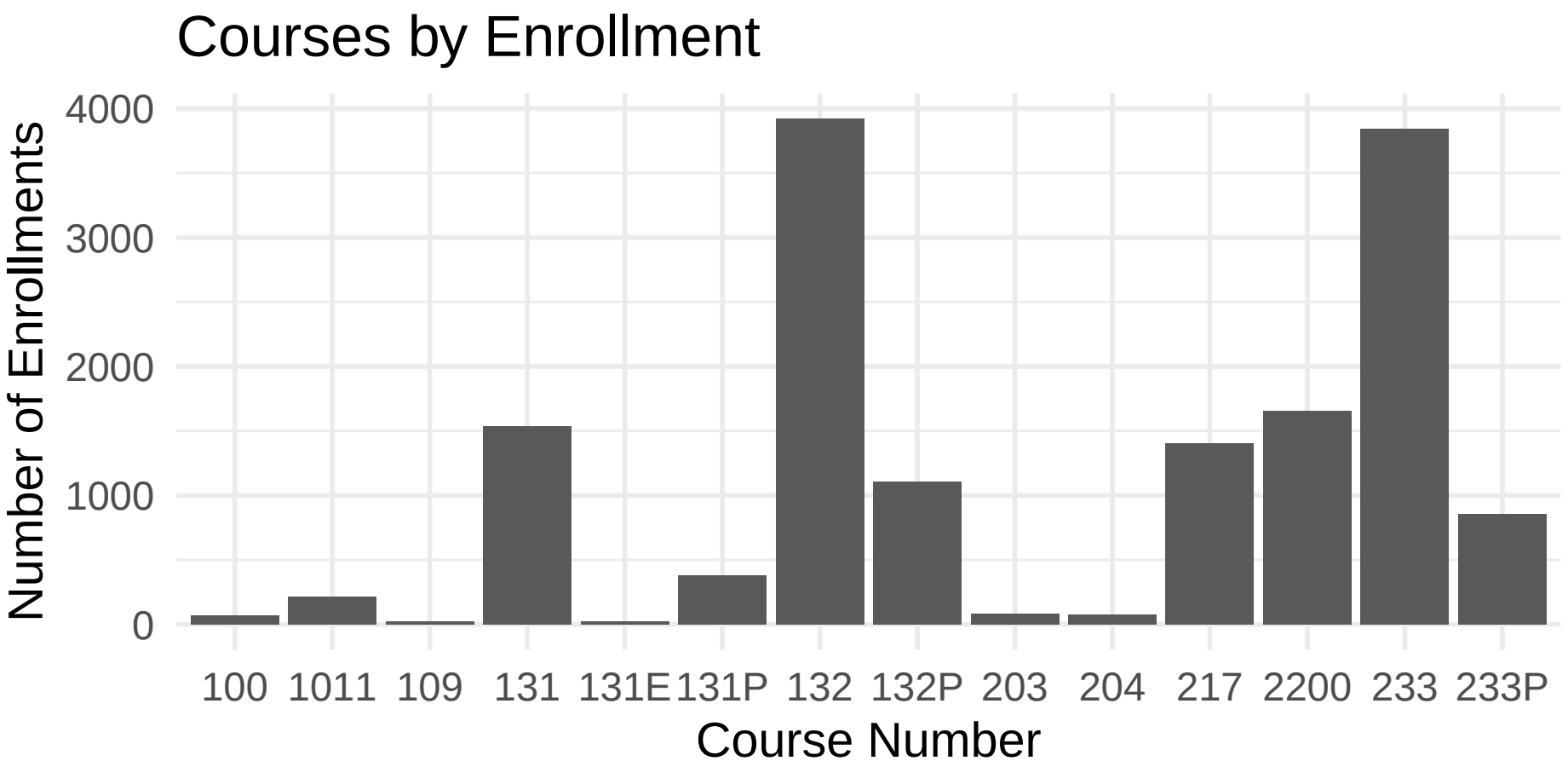


This graph provides an overview of how enrollment numbers have changed over time. We could tell from the graph that there would be more student enrolling in elementary math courses in fall compared to in spring semester.

Enrollment over Courses

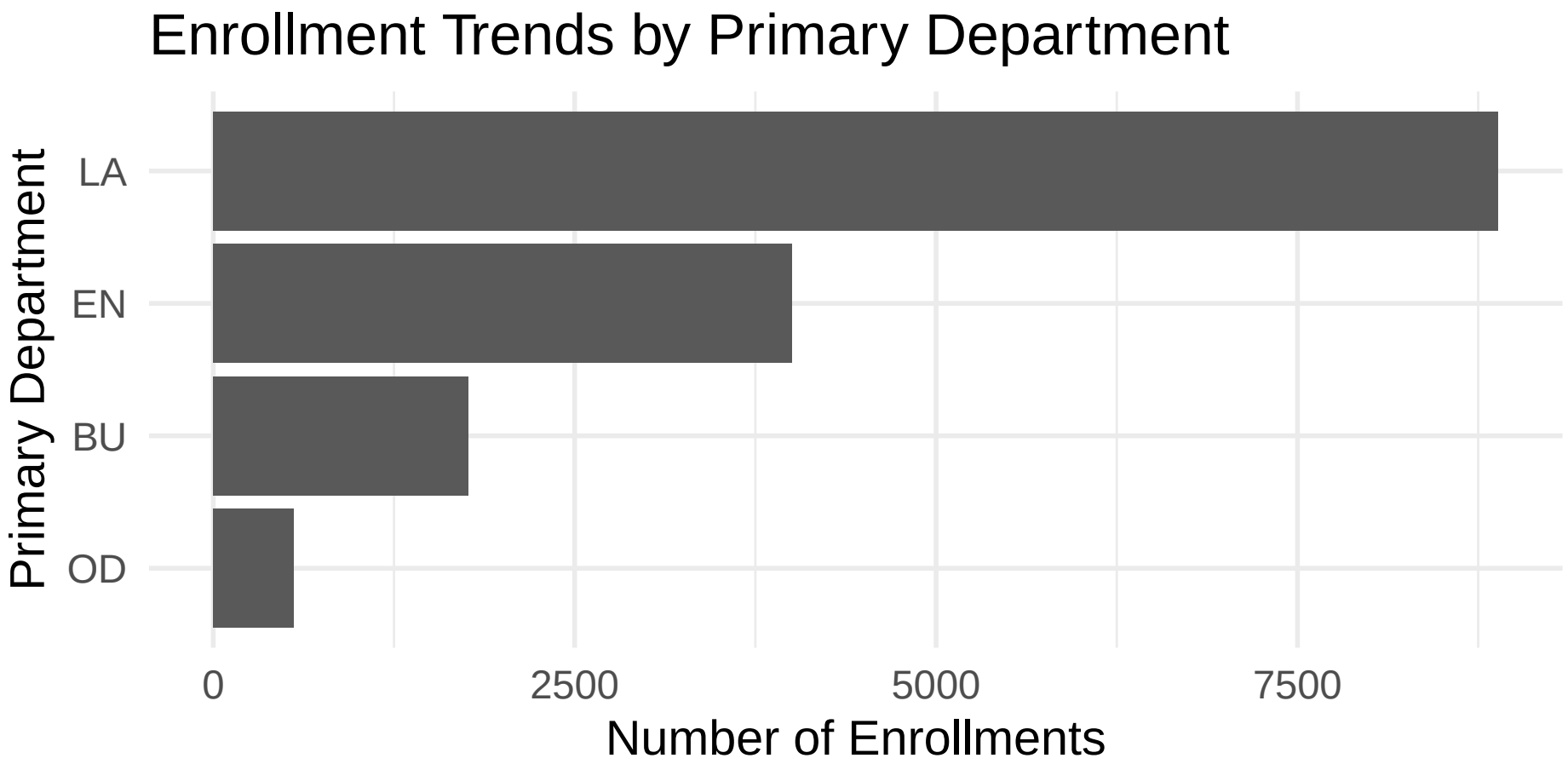


Enrollment Distribution Across Different Courses



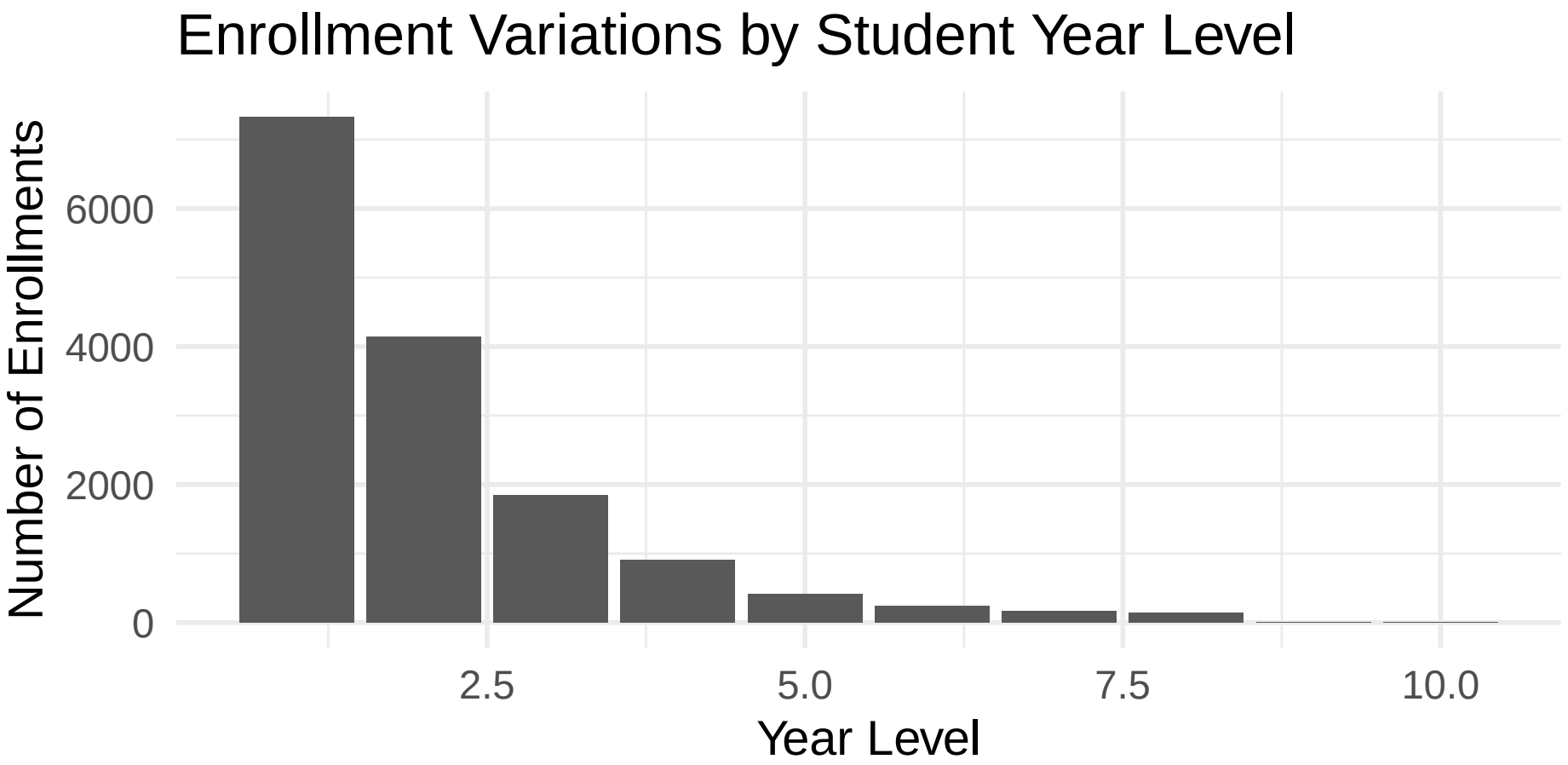
This graph displays the courses with the highest number of enrollments, identifying the most popular or required courses within the 100- and 200-level range would be calculus related courses.

Enrollment Distribution Across Departments



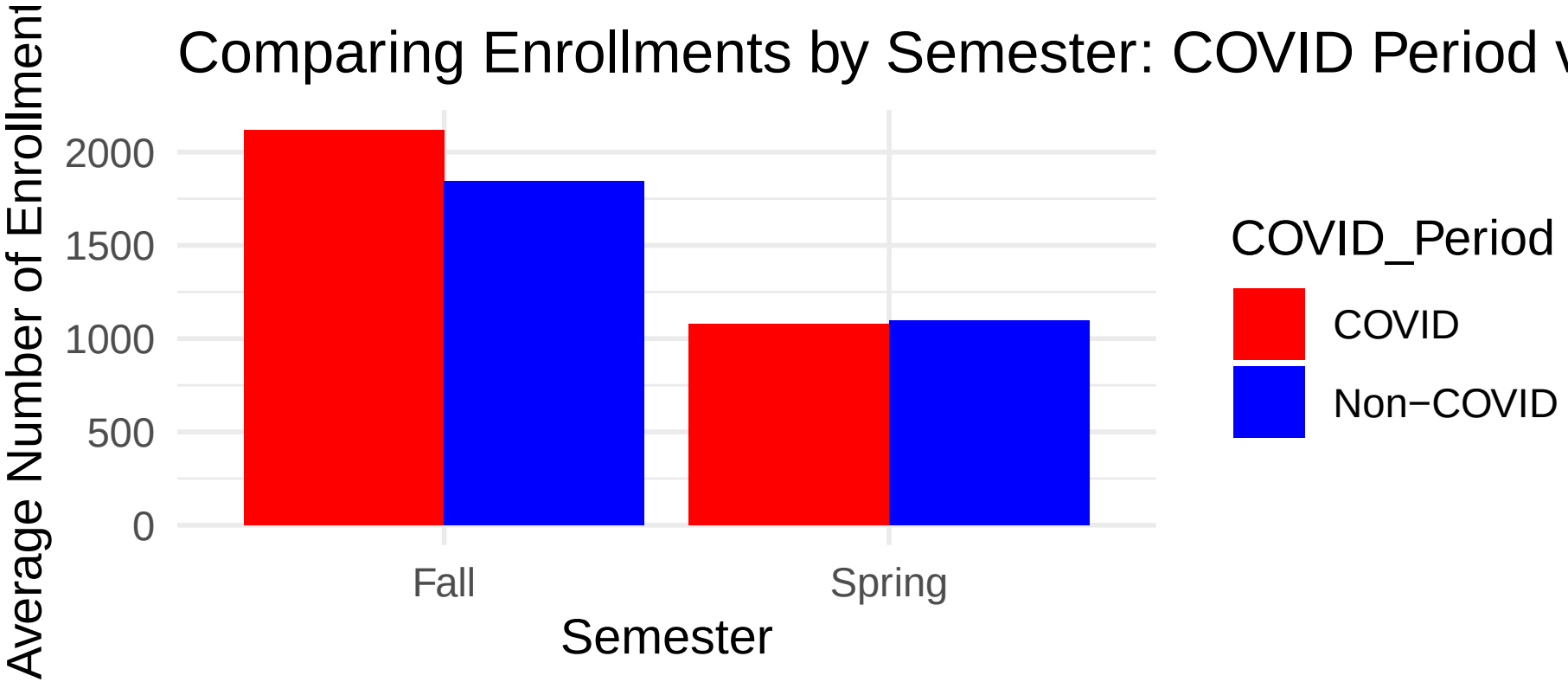
This chart illustrates the distribution of enrollments across different primary departments, indicating college of arts and sciences have the highest representation in these math courses. However, I found this a little uninformative as the LA would be the largest departments.

Enrollment Variations by Student Year Level



There are more student from year levels 1 or 2 to enroll in these courses as it is introductory courses. However, based on the data collection is based on the time it is collected, it would be fairly difficult to use this to determine student's standing.

COVID period versus non-COVID periods



This comparison provide insights into the impact of the pandemic on course enrollments. However, though it may seems virtually significant (or not), we may need more statistical testing to consolidate the result. Also, it could be due to the reason for increasing size of the university instead of just because of the pandemic.

Statistical Modeling

The first thing we do is to find the appropriate response variables. Given that the data is at the individual student enrollment level for each course, there are a couple of ways to construct the response variable:

- Binary Response (Logistic Regression): You could create a binary variable indicating whether a student is enrolled in a course or not (1 for enrolled, 0 for not enrolled). However, this approach might not be suitable since data seems to include only enrolled students.
- Count Response (Poisson or Negative Binomial Regression): A more appropriate approach might be to aggregate the data at a course level for each semester and use the total number of enrollments in each course as the response variable.

Hence, here I choose to aggregate data by semester to see how it is depends on the covariates.

Modeling

So the first model I build is a fairly simple model to see how average enrollment in intro-level math courses would be different by semester.

$$Enrollment_t = \beta_0 + \beta_1 * Year + \beta_2 * Semester + \epsilon_t$$

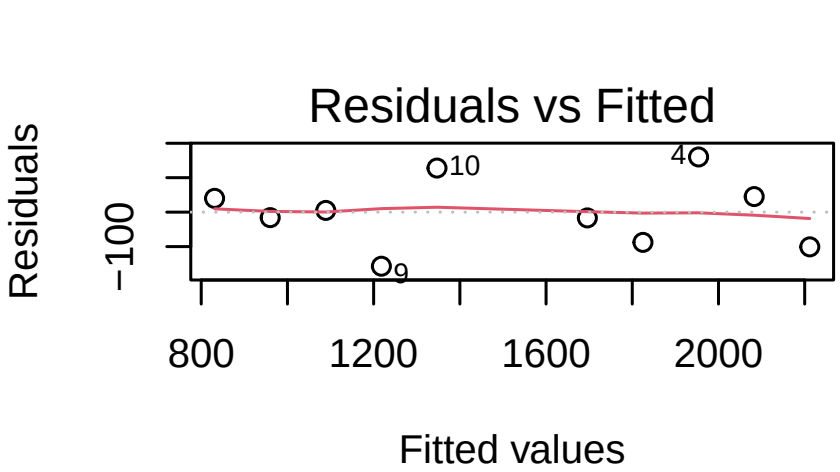
Table 1: Model Summary Statistics

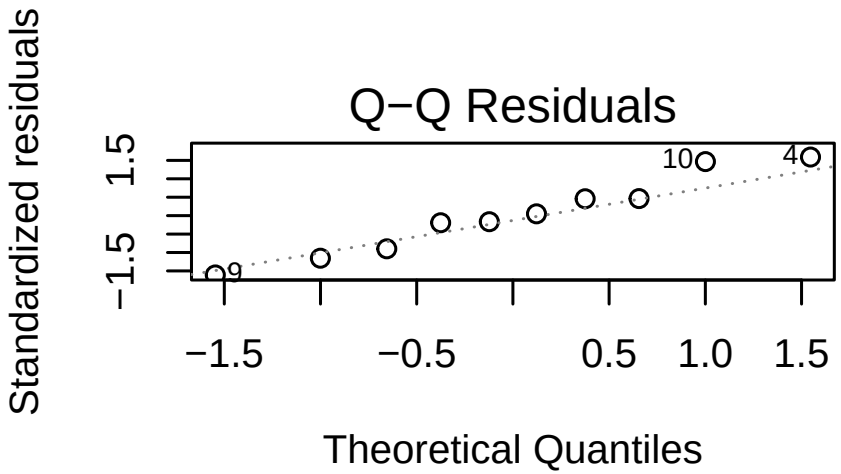
	Estimate	StdError	tValue	Pr
(Intercept)	1695.8	71.22311	23.809687	0.0000001
Year_Since_2017	129.0	25.18117	5.122875	0.0013644
SemesterSpring	-993.8	75.54352	-13.155332	0.0000034

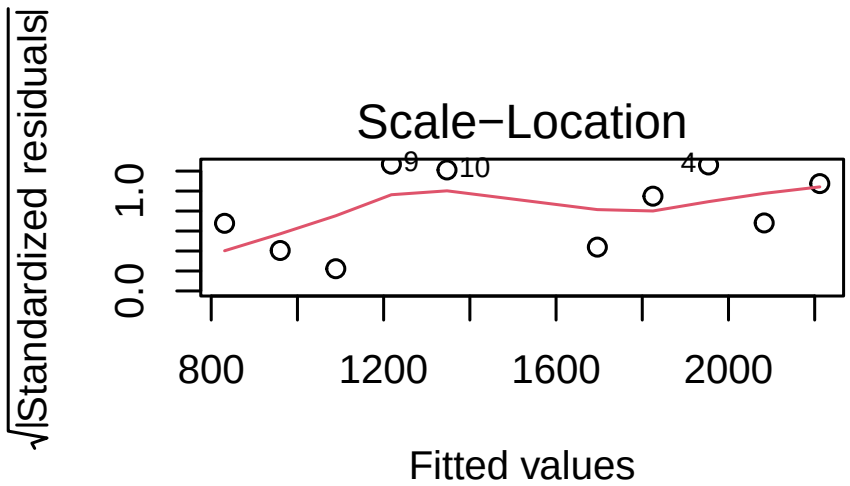
Table 2: Model Metrics

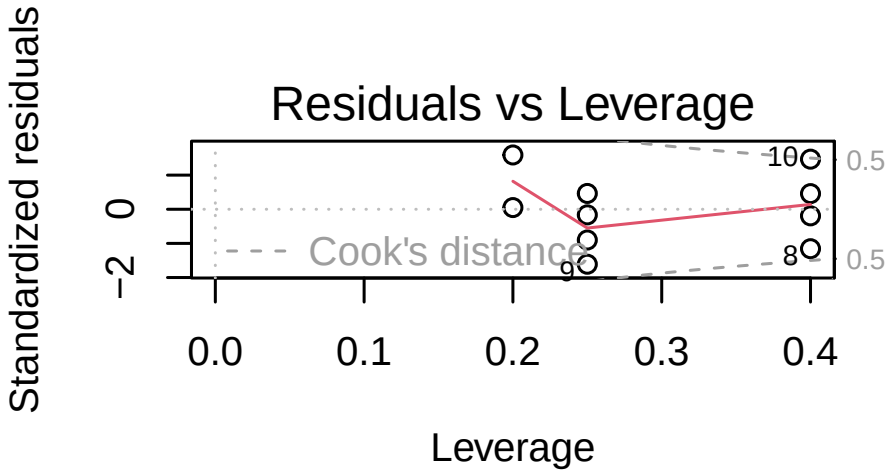
Metric	Value
R-squared	0.9612564
Adjusted R-squared	0.9501868

Include Diagnostic Plots:



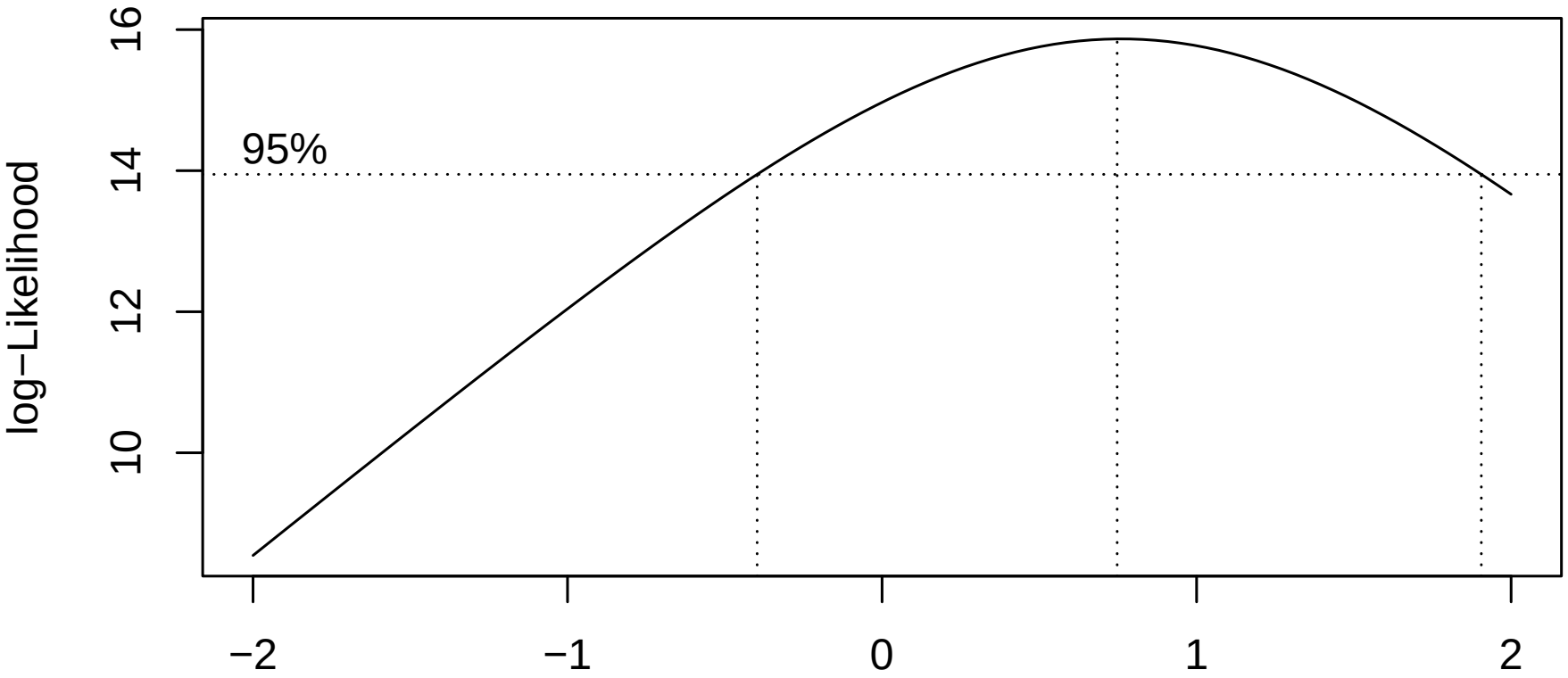






Box-Cox Transformation

Though the model diagotics looks so what satisfactory, we could still see if Box-Cox could make it better.



Adding Courses?

The model provides a clear indication of temporal trends in enrollment, but it's essential to consider external factors that could influence these trends. Future models could explore additional predictors, like specific course attributes, to gain more comprehensive insights.

Here, the reference group for Year, though continuous, is 2017. For semester, it's Fall, For courses, it's 132, calc 2.

Summary Statistics

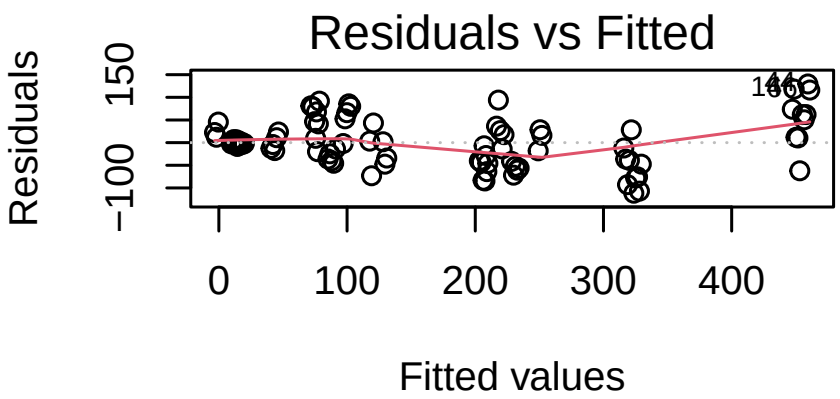
Table 3: Model Summary Statistics

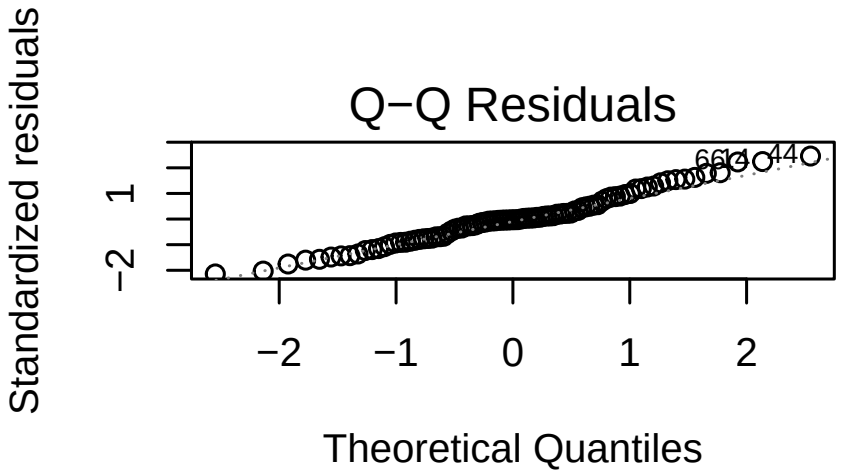
	Estimate	StdError	tValue	Pr
(Intercept)	455.086789	21.067800	21.6010588	0.0000000
Year_Since_2017	1.459547	4.520315	0.3228861	0.7476682
SemesterSpring	-132.871312	14.357416	-9.2545419	0.0000000
Crs100	-443.605882	31.519939	-14.0738179	0.0000000
Crs1011	-282.994118	31.519939	-8.9782572	0.0000000
Crs109	-312.634571	44.283415	-7.0598569	0.0000000
Crs131	-238.500000	25.127422	-9.4916223	0.0000000
Crs131E	-449.695203	44.570869	-10.0894421	0.0000000
Crs131P	-329.926214	29.364656	-11.2354870	0.0000000
Crs132P	-208.926214	29.364656	-7.1148871	0.0000000
Crs203	-441.005882	31.519939	-13.9913305	0.0000000
Crs204	-310.794118	31.519939	-9.8602386	0.0000000
Crs217	-251.900000	25.127422	-10.0249042	0.0000000
Crs2200	-226.700000	25.127422	-9.0220158	0.0000000
Crs233	-7.800000	25.127422	-0.3104178	0.7570929
Crs233P	-251.092880	29.364656	-8.5508537	0.0000000

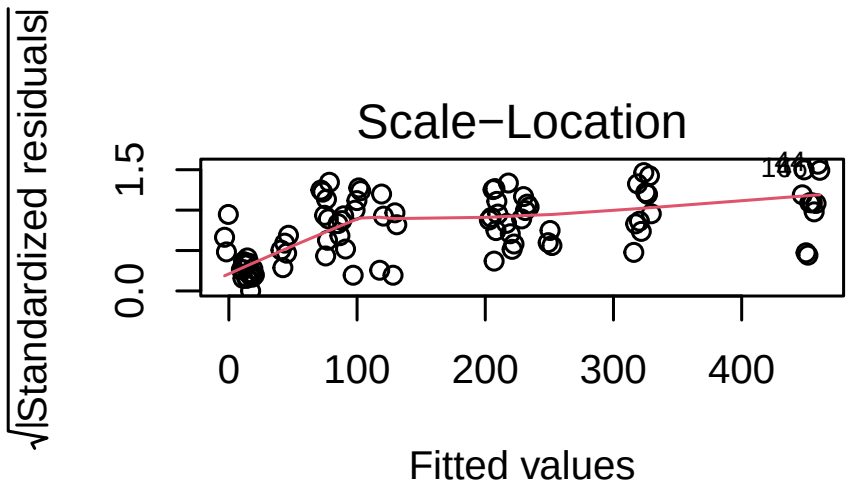
Table 4: Model Metrics

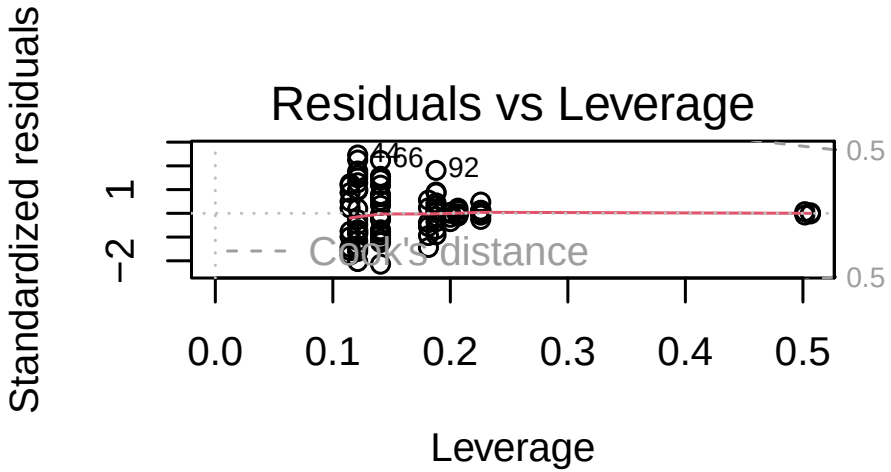
Metric	Value
R-squared	0.8862102
Adjusted R-squared	0.8637516

Model Diagotics









Not Year

The lack of significance for the year could be due to multicollinearity, especially if certain courses are only offered in specific years or if there are other time-related variables that are not included in the model.

Table 5: Model Summary Statistics

	Estimate	StdError	tValue	Pr
(Intercept)	458.0059	18.91842	24.2095192	0.0000000
SemesterSpring	-131.4118	13.54776	-9.6998867	0.0000000
Crs100	-443.6059	31.33607	-14.1563998	0.0000000
Crs1011	-282.9941	31.33607	-9.0309395	0.0000000
Crs109	-314.0941	43.79512	-7.1718975	0.0000000
Crs131	-238.5000	24.98084	-9.5473169	0.0000000
Crs131E	-447.5059	43.79512	-10.2181675	0.0000000
Crs131P	-328.4667	28.84539	-11.3871461	0.0000000
Crs132P	-207.4667	28.84539	-7.1923683	0.0000000
Crs203	-441.0059	31.33607	-14.0734283	0.0000000
Crs204	-310.7941	31.33607	-9.9180962	0.0000000
Crs217	-251.9000	24.98084	-10.0837280	0.0000000
Crs2200	-226.7000	24.98084	-9.0749549	0.0000000
Crs233	-7.8000	24.98084	-0.3122393	0.7557027
Crs233P	-249.6333	28.84539	-8.6541848	0.0000000

Table 6: Model Metrics

Metric	Value
R-squared	0.8860541
Adjusted R-squared	0.8653366

Interactions?

Table 7: Model Summary Statistics

	Estimate	StdError	tValue	Pr
(Intercept)	472.886538	22.933886	20.6195553	0.0000000
Year_Since_2017	-6.408926	6.200556	-1.0336050	0.3046423
SemesterSpring	-175.885334	27.503873	-6.3949297	0.0000000
Crs100	-445.668686	31.069190	-14.3443935	0.0000000
Crs1011	-280.931314	31.069190	-9.0421190	0.0000000
Crs109	-302.854611	43.949685	-6.8909393	0.0000000
Crs131	-238.500000	24.751665	-9.6357153	0.0000000
Crs131E	-439.955297	44.228080	-9.9474203	0.0000000
Crs131P	-329.850555	28.925566	-11.4034263	0.0000000
Crs132P	-208.850555	28.925566	-7.2202756	0.0000000
Crs203	-443.068686	31.069190	-14.2607093	0.0000000
Crs204	-308.731314	31.069190	-9.9368961	0.0000000
Crs217	-251.900000	24.751665	-10.1770930	0.0000000
Crs2200	-226.700000	24.751665	-9.1589797	0.0000000
Crs233	-7.800000	24.751665	-0.3151303	0.7535378
Crs233P	-251.017222	28.925566	-8.6780403	0.0000000
Year_Since_2017:SemesterSpring	15.585629	8.547243	1.8234686	0.0722168

Table 8: Model Metrics

Metric	Value
R-squared	0.8910407
Adjusted R-squared	0.8677961

Model Comparison

- **Small R-squared and Adjusted R-squared**: they all explain a similar amount of variance in enrollments and the number of predictors used is appropriate.
- **Year and Interaction not significant**: The inclusion of Year in the model does not add significant explanatory power. The interaction between Year and Semester is also not significant, suggesting that the effect of the semester on enrollments does not change over the years.
- **Semeter Coefficient Significant**: This indicates the seasonal effect is still strong in the model.
- **Course Coefficient Significant**: Enrollment numbers vary significantly by course

Final model that may be perferable:

$$Enrollment_{time,course} = \beta_0 + \beta_1 \times Semester + \beta_2 \times Crs + \epsilon$$

Cross-Validation

However, we may worry it could be some over-fitting problem given the number of courses included. Hence, we use cross-validation to check the stability of the model.

```
## Warning in predict.lm(modelFit, newdata): prediction from r
## attr(*, "non-estim") has doubtful cases

## Linear Regression
##
## 92 samples
##  2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 82, 83, 83, 83, 81, 82, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
```

Nonlinearity?

Also, as we are using count for our response variable, we also think about some nonlinear model such as Poisson Regression for some analysis. But this model holds similiar performance as our linear ones. So we may just perfer our previous one.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Enrollments ~ s(Year, k = 6) + Semester + s(Crs, bs = "re")
##
## Parametric coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       192.45      37.55   5.125 2.22e-06 ***
## SemesterSpring   -132.54      14.14  -9.371 2.77e-14 ***
## ---
```

Add Primary Departments?

Now, it could be primary department is also one important factor that incluences one's enrollment in math courses.

Table 9: Model Summary Statistics

	Estimate	StdError	tValue	Pr
(Intercept)	-1180.1492275	3703.432710	-0.3186636	0.7501983
Year	0.6140182	1.834343	0.3347347	0.7380538
SemesterSpring	-34.7136159	5.627289	-6.1687988	0.0000000
Crs1011	40.6477003	16.831532	2.4149733	0.0163206
Crs109	20.4535122	22.973739	0.8902997	0.3740022
Crs131	53.7785746	13.287265	4.0473773	0.0000656
Crs131E	2.6971986	20.952916	0.1287266	0.8976582
Crs131P	27.7338054	14.845847	1.8681188	0.0626980
Crs132	113.4035746	13.287265	8.5347567	0.0000000
Crs132P	60.9228897	14.528379	4.1933714	0.0000360
Crs203	-15.8059612	16.742415	-0.9440670	0.3458778
Crs204	17.7936365	17.571098	1.0126650	0.3120176
Crs217	51.4195578	13.601725	3.7803704	0.0001881
Crs2200	57.2727149	13.380344	4.2803619	0.0000250
Crs233	111.4535746	13.287265	8.3879996	0.0000000
Crs233P	50.3812231	14.528379	3.4677800	0.0005997
PrimeDivBU	-82.7900653	6.728374	-12.3046174	0.0000000
PrimeDivEN	-52.3140054	6.507087	-8.0395434	0.0000000

Table 10: Model Metrics

Metric	Value
R-squared	0.6021963
Adjusted R-squared	0.5788723

Cross-Validation

As we are adding more factors, it is important to keep on check for the overfitting problem. With only adding the PrimeDiv, it is still stable.

```
## Linear Regression
##
## 326 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 293, 293, 292, 293, 294, 293, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   43.47112  0.5695077  35.35742
##
## Tuning parameter 'intercept' was held constant at a value c
```

Year Delete and Interaction Check

The year variable is still not significant for our previous analysis. Hence, I am considering delete the year variable and add interaction terms with PrimeDiv and Crs to see if it would enhance the model performance.

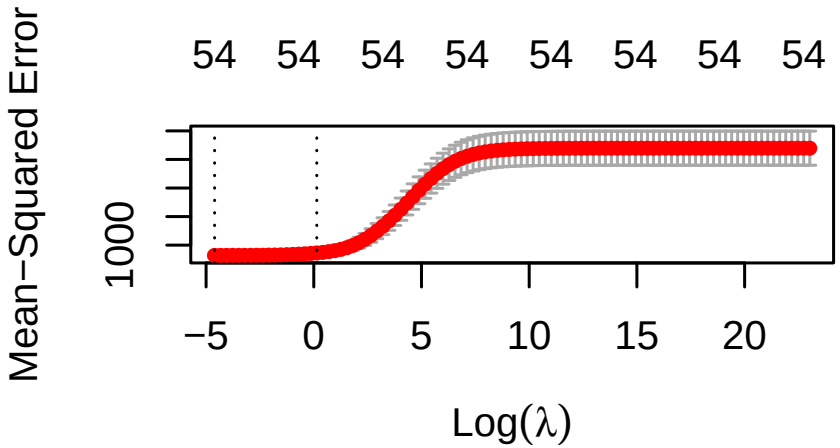
For the reasons that it would be a fairly long summary statistics, I am not showing the result here.

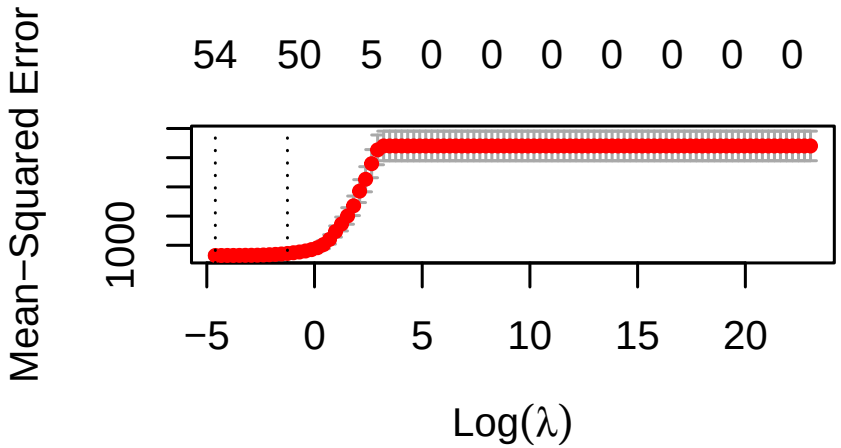
However, the result is that:

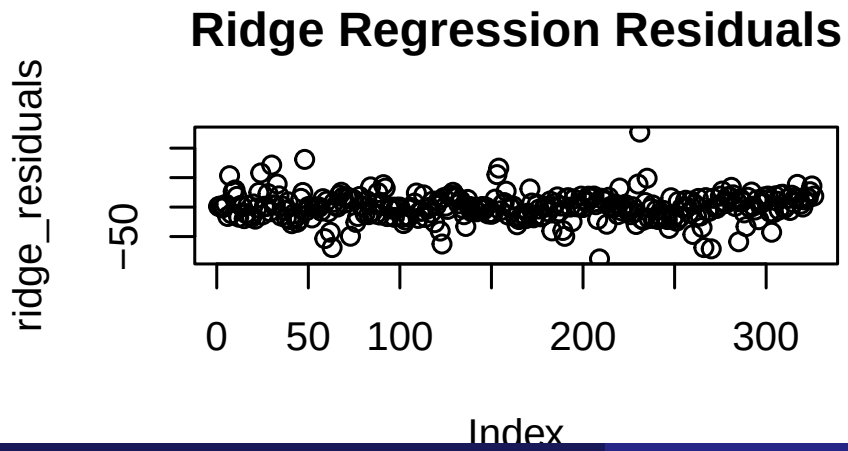
- Deleting year: This would not significant influence the performance of our model for that it is not changing R-squared, Adjusted R-squared and F-statistic significantly.
- Adding Interaction: This interaction could be a interesting perspective to check as we could see how different schools hold different performance for different courses. But it faces some over-fitting problems, also collinearity problem.

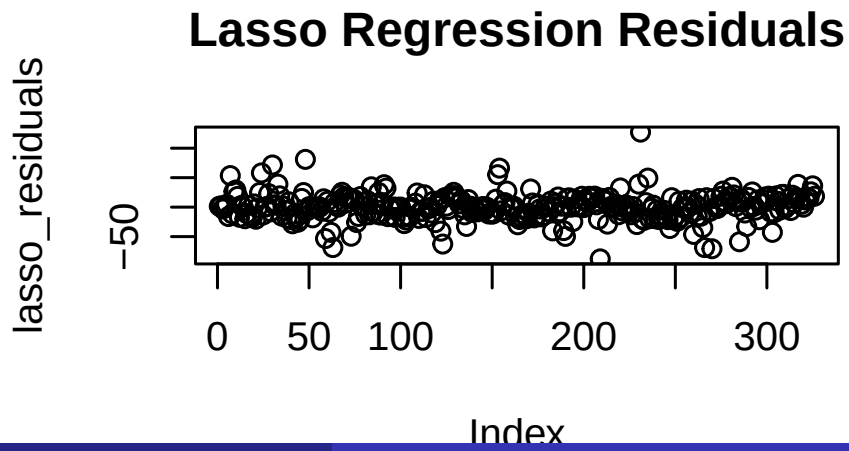
Trying to solve this Collinearity Problem

With this noticed, I tried to solve this problem through some penalized regression, such as LASSO and Ridge Regression we discussed in class.









Result for Panelized Regression

```
##
## Call:  cv.glmnet(x = predictors, y = response, lambda = 10
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure     SE Nonzero
## min   0.01   100   640.2  87.38      54
## 1se   1.15    83   725.8 105.15      54

##
## Call:  cv.glmnet(x = predictors, y = response, lambda = 10
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min 0.0100   100   651.1 82.22      54
## 1se 0.2848    88   720.3 87.33      52
```

Conclusion: Analyzing Factors Influencing Math Course Enrollments

Time-Based Trends (Model 1):

- Findings: A clear trend of increasing enrollments over the years, with higher enrollments in Fall compared to Spring.
- Implication: Indicates growing interest in math courses and potential seasonal influences on course selection.

Impact of Course Selection (Model 2):

- Findings: 'Year' becomes insignificant, likely due to collinearity. Significant variations in enrollments across different courses.
- Implication: Course selection plays a crucial role in enrollment numbers, overshadowing the year-over-year trend.

Influence of Primary Department (Model 3):

- Findings: Primary department is a significant factor, but adding it reduces the model's explanatory power of course numbers. Potential issues of collinearity and overfitting noted.
- Implication: The choice of major significantly influences course enrollment, but the relationship is complex and may interact with course selection.

Key Takeaways

- Enrollment trends in math courses are influenced by a combination of time (semester), course selection, and student's primary department.
- Course selection is a stronger predictor of enrollment numbers than time-based trends.
- The complex interplay between course choice and student's primary department suggests the need for nuanced curriculum planning and advisement.

Thank You and Questions

Thank You!

Questions?