Nathaniel Jones
Final Project – Data Mining
12/6/2022

Table of Contents

0. Executive Summary

The two-year withdrawal rate rapidly increased during the 17 school years observed in this data (Figure 0). Discovering the mechanisms that may have led to the cause of this shift could aid schools in reducing their two-year withdrawal rate. This project sought to investigate what role student sentiment, observed in Rate My Professor comments, played. A longitudinal model was built using a two-levels where the level two unit of observation is schools, and the level one unit of observation is school year. Using the withdrawal rate as the dependent variable, this project found that student sentiment may not be useful in predicting the two-year withdrawal rate. The following multi-level equation was found to correspond to the best model:

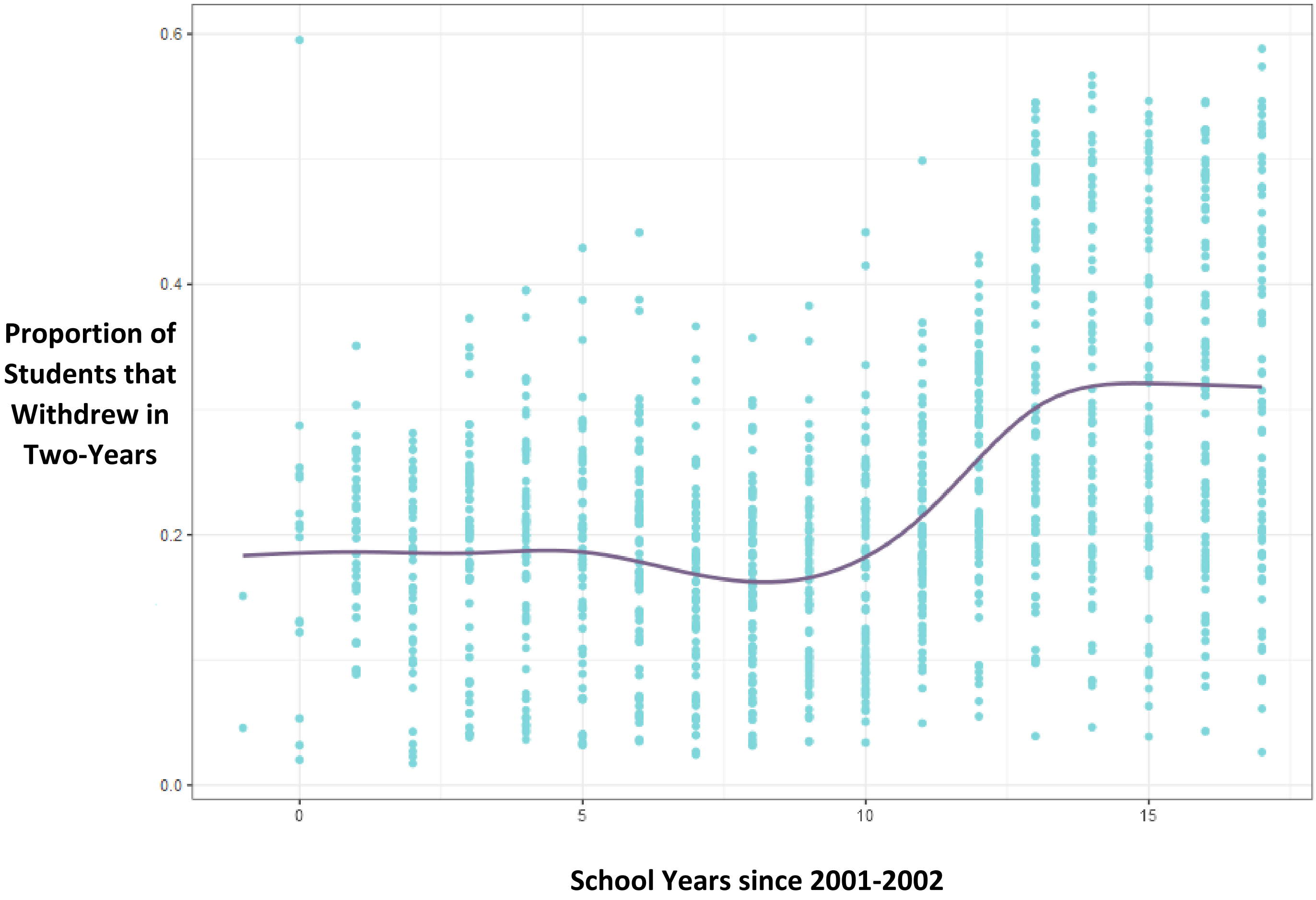
$$Two\ Year\ Withdrawal\ Rate = \ \beta_{0i} + \beta_{1i}(School\ Year)_{ti} + \beta_{2i}(School\ Year)_{ti}^{2} + error_{ti}$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(Student\ Sentiment) + U_{0i}$$

$$\beta_{1i} = \gamma_{10}$$

$$\beta_{2i} = \gamma_{20}$$

Figure 0: Two-Year Withdrawal Rate Over Time



1. Introduction

Often students polarized by either good or bad teachers navigate to Rate My Professor to voice their opinions and experiences on the courses they took over the previous semester. Utilizing the sentiment left in the comments reviewers give could potentially provide valuable insight into why some schools have high withdrawal rates. The goal of this research is to study the pattern of change in the two-year withdrawal rate at post-secondary schools over time, while accounting for the average student sentiment to see if a relationship between a school's average student sentiment and two-year withdrawal rate exists. Data was scraped from Rate My Professors and merged with the CollegeScorecard dataset to determine which post-secondary schools have a positive or negative average student sentiment over time. Afterwards, a longitudinal model was used to model the two-year withdrawal rate over time.

CollegeScorecard:
The CollegeScorecard dataset is released by the Department of Education through the Integrated Post-Secondary Education Data System (IPEDS) which collects nearly 3,000 parameters for 6,654 schools across the U.S per school year. In total, the CollegeScorecard dataset contains 176,688 observations across 25 school years (1996-1997 to 2020-2021). For this research project, four parameters were used:

- INSTNM: The name of the school.
- SCHOOL_YEAR: The school year (EX: 2020-2021).
- WDRAW_ORIG_YR2_RT: The withdrawal rate of students that withdrew from the original school they began at in two-years.

Rate My Professor:
Since 1999, Rate My Professors has collected more than 15 million reviews for over 1.3 million professors each with up to 19 parameters. For this project, six parameters were used:

- REVIEW_IDs: Three variables identify distinct schools (SID), professors (TID), and review instances (RID).
- DATE: The Month-Day-Year of the review.
- OVERALL: A overall rating (1-5) given by the reviewer.
- TAGS: Buttonized text strings reviewers can leave in addition to the comments.
- COMMENTS: The text reviews left by the student.

2. Methods

Data Collection and Wrangle:
The data collection process utilized a GitHub repository under the name 'ratemyprof-api' that returns all professor review data for a given school's SID. Since a school's SID is assigned by Rate My Professors and is not in the CollegeScorecard dataset, the SID's must be collected for each school. Python packages 'Requests' and 'BeautifulSoup' were used to search Rate My Professors for the school's URL directory and collect 3,988 SID's. Then each SID was used to collect all reviews corresponding to each school. Due to the number of reviews Rate My Professors has for each school, this project will focus on schools in Georgia where the highest degree offered is either a bachelor's or a graduate degree, meaning 97 of 3,988 SID's were used to collect 348,083 reviews. Some of these reviews were missing comments or contained only non-alphanumeric text and were dropped from the data, which left a total of 303,482 reviews. Some schools only have reviews for one or two school years. Since a longitudinal model will be constructed from this data, these schools were dropped from the dataset. Lastly, the school years prior to 2001-2002 contain little to no data (Figure 1) and were not used in this project. After filtering the data in this way, 286,288 reviews remained.

Text Preprocess:
Prior to performing the sentiment analysis, the review comments were made lowercase and cleaned of non-alphanumeric characters, html tags, URL links, and punctuation. In addition, the words within each review were removed if the word was considered a 'stopword' ('a', 'do', 'not'), and put through a process of lemmatization to determine and replace the word with the canonical form. Removing stopwords resulted in the loss of several thousand reviews containing valuable data (such as 'Take Them' or 'Do Not Take Them'). To mitigate this loss some stopwords, such as 'not,' 'very,' 'do,' and 'did,' were not removed.

Sentiment Analysis:
Using the cleaned review comments and a pre-trained language model ('siebert/sentiment-roberta-large-English'), a sentiment analysis was conducted on each of the comments to systematically identify the positive or negative emotional polarity of a given text. In result, each review comment received a decimal value between 0 (Negative) and 1 (POSITIVE). These scores were then aggregated for each professor and school over time to create the average student sentiment for each school.

Longitudinal Analysis:
A longitudinal model was used to model the pattern of change in the two-year withdrawal rate observed at post-secondary schools over time. This research project used a 2-level model where the level two unit of observation is schools, and the level one unit of observation is school year. Using the withdrawal rate as the dependent variable, time-only models were initially created and compared before including the average student sentiment. The best model was decided using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Each model created used the Restricted Maximum Likelihood (REML) estimation. In total, six different model structures were constructed for comparison.

3. Data Exploration

Figure 1 displays a large increase in reviews over time. Prior to the 2009-10 school year, the number of reviews is less than 10,000 each year. After the 2009-10 school year, the number of reviews increases above 20,000 reviews. The peak number of reviews was observed for the 2016-17 school year, with 27,870 reviews. As discussed in Section 2, this plot shows school years prior to 2001-2002 contain little to no data. In addition, these school years are low in frequency of schools (Figure 2). Since a longitudinal model will be built using the comments of these reviews, each school must be observed with at least three occasions, and each occasion needs at least three observations. School years prior to 2001-02 do not meet this requirement and will be dropped before modelling the dependent variable. Figure 3 below displays the average number of reviews per school over time. On the horizontal axis, 0 represents the 2001-02 school year and every unit increase on this axis represents an increase in the number of school years since 2001-02. This plot displays that the average number of reviews increase over time. Like Figure 1, this plot displays a large increase in the average number of reviews after the 2009-10 school year.

Figure 1: Frequency of Reviews over Time

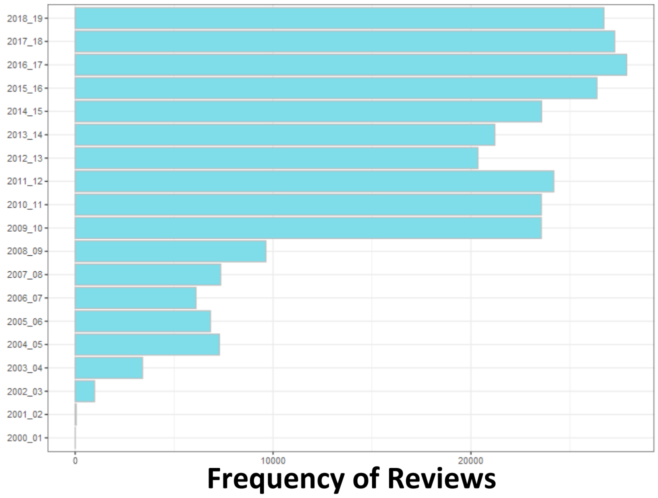


Figure 2: Frequency of Schools over Time

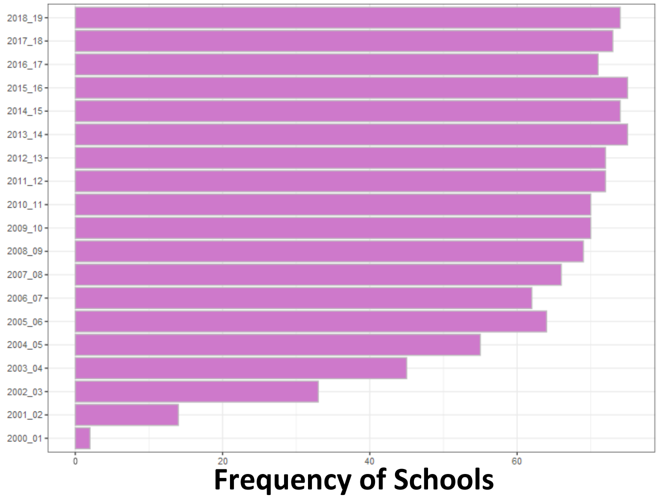


Figure 3: Average Number of Reviews per School over Time

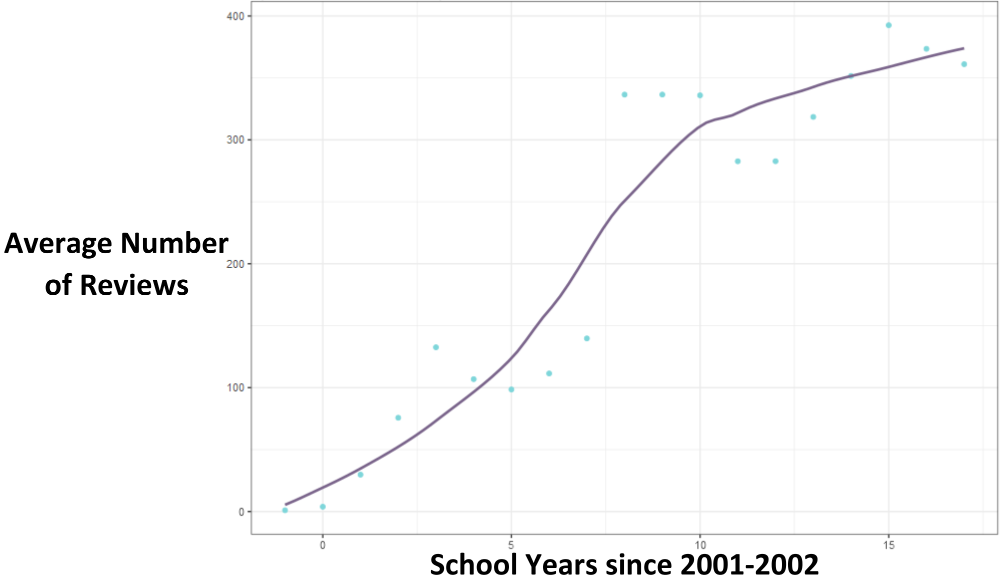
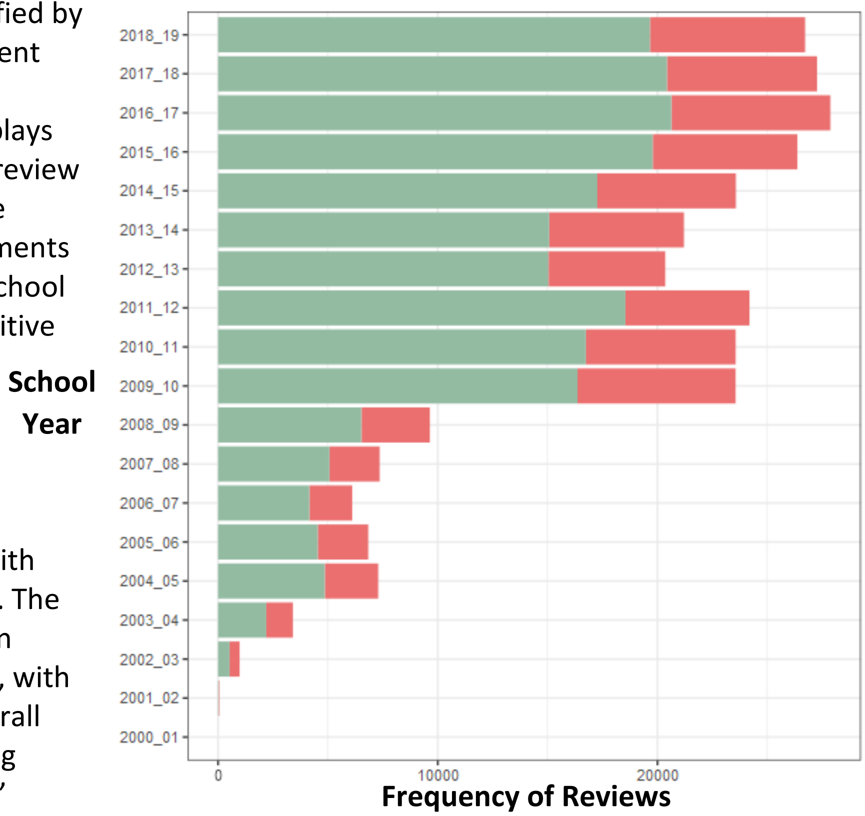


Figure 4 displays the number of reviews over time, stratified by the results of the sentiment analysis on the review comments. This plot displays the number of negative review comments in red and the number of positive comments in green. Regardless of school year, this plot shows positive sentiment is the most frequently observed sentiment in this data.

Figure 4: Frequency of Sentiment Analysis over Time



To verify the results of the sentiment analysis, a comparison was made with the overall review rating. The overall review rating is an integer between 1 and 5, with 1 representing 'poor overall rating' and 5 representing 'awesome overall rating.' Figure 5 displays the review frequency of each overall rating stratified by the review comment sentiment. As can be observed from this plot, reviews given a 1 (poor) typically correspond to a comment with negative sentiment while reviews given a 5 (awesome) typically correspond to a comment with positive sentiment.

Figure 5: Frequency of the Overall Review Ratings

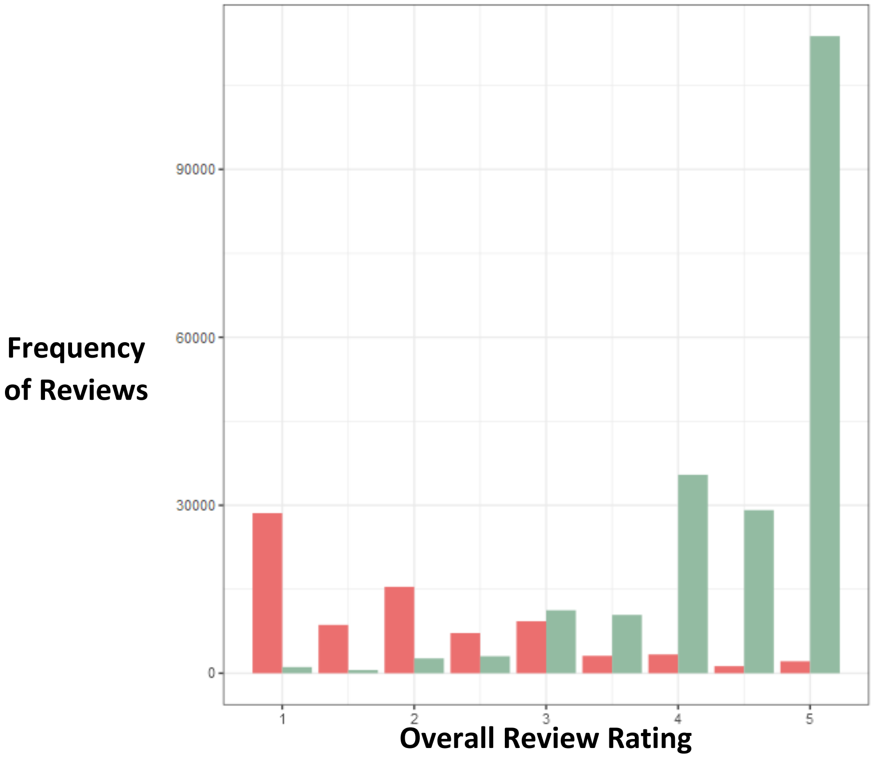
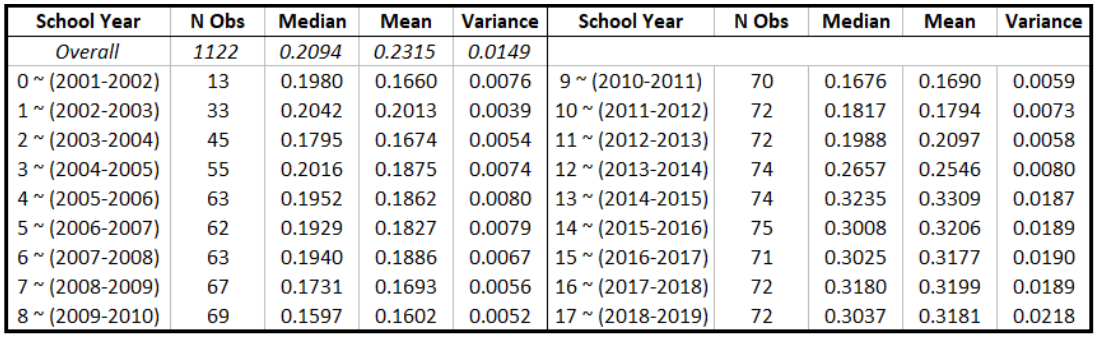


Table 1: Mean of Withdrawal Rate by School Year



School Year	N Obs	Median	Mean	Variance	School Year	N Obs	Median	Mean	Variance
Overall	*1122*	*0.2094*	*0.2315*	*0.0149*					
0 ~ (2001-2002)	13	0.1980	0.1660	0.0076	9 ~ (2010-2011)	70	0.1676	0.1690	0.0059
1 ~ (2002-2003)	33	0.2042	0.2013	0.0039	10 ~ (2011-2012)	72	0.1817	0.1794	0.0073
2 ~ (2003-2004)	45	0.1795	0.1674	0.0054	11 ~ (2012-2013)	72	0.1988	0.2097	0.0058
3 ~ (2004-2005)	55	0.2016	0.1875	0.0074	12 ~ (2013-2014)	74	0.2657	0.2546	0.0080
4 ~ (2005-2006)	63	0.1952	0.1862	0.0080	13 ~ (2014-2015)	74	0.3235	0.3309	0.0187
5 ~ (2006-2007)	62	0.1929	0.1827	0.0079	14 ~ (2015-2016)	75	0.3008	0.3206	0.0189
6 ~ (2007-2008)	63	0.1940	0.1886	0.0067	15 ~ (2016-2017)	71	0.3025	0.3177	0.0190
7 ~ (2008-2009)	67	0.1731	0.1693	0.0056	16 ~ (2017-2018)	72	0.3180	0.3199	0.0189
8 ~ (2009-2010)	69	0.1597	0.1602	0.0052	17 ~ (2018-2019)	72	0.3037	0.3181	0.0218

The grand mean of two-year withdrawal rate is 23%. This means that on average for the 17 school years observed, 23% of students withdrew within two-years from the original institution they began at. This value is not very different from the two-year withdrawal rate observed for the 2001-02 school year (17%) but is very different from the most recent school year (32%). Looking over time, the two-year withdrawal rate fluctuates between 16-20% for the first eight school years after 2001-02. Roughly starting in the 2010-11 school year, the two-year withdrawal rate begins increasing. The peak withdrawal rate (33%) was observed in the 2014-15 school year and has remained above 30% each year. Notably, the variances of these years (2014-15 to 2018-19) are significantly larger than the other observed school years. For example, the two-year withdrawal rate for the 2013-14 school year had a variance of 0.008 while in the next school year (2014-15) the variance more than doubled in value (0.0187). Figure 6 below displays a spaghetti plot of two-year withdrawal rate over time. The horizontal axis represents the school years since 2001-02 and each line on the plot shows the withdrawal rate of a different school. The most notable feature of this plot happens around 10 years after the 2001-02 school year (2011-12) where the withdrawal rate increases steeply for many schools.

Figure 6: Spaghetti Plot of the Two-Year Withdrawal Rate

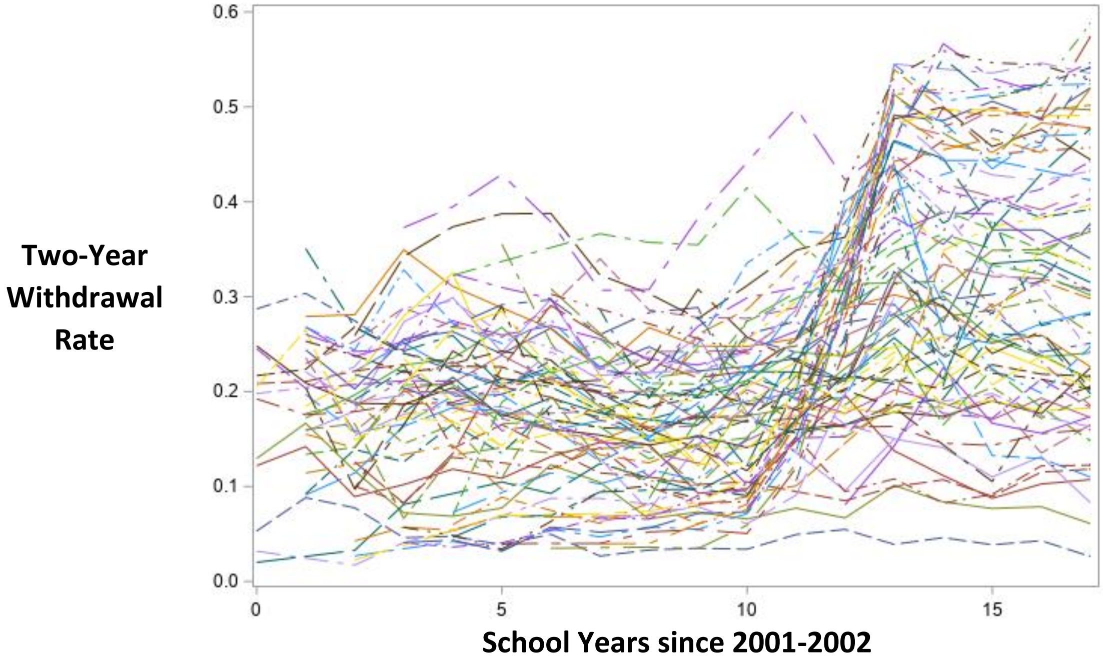
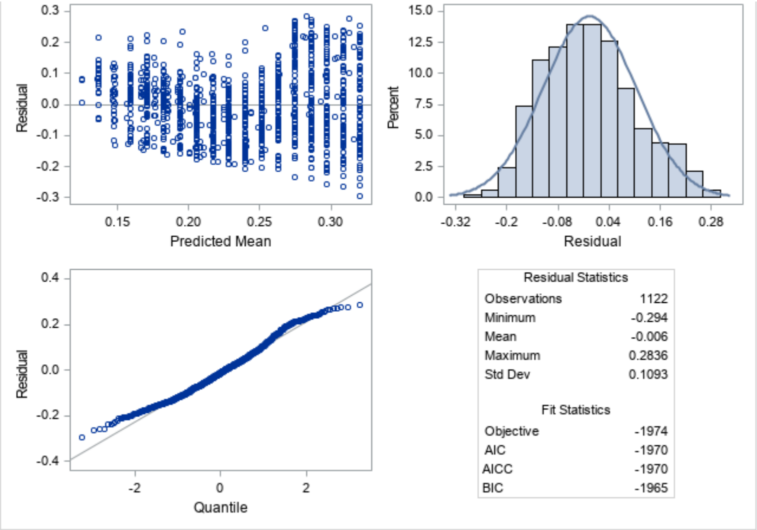


Figure 8: Test of Normally Distributed Residuals for the Dependent Variable

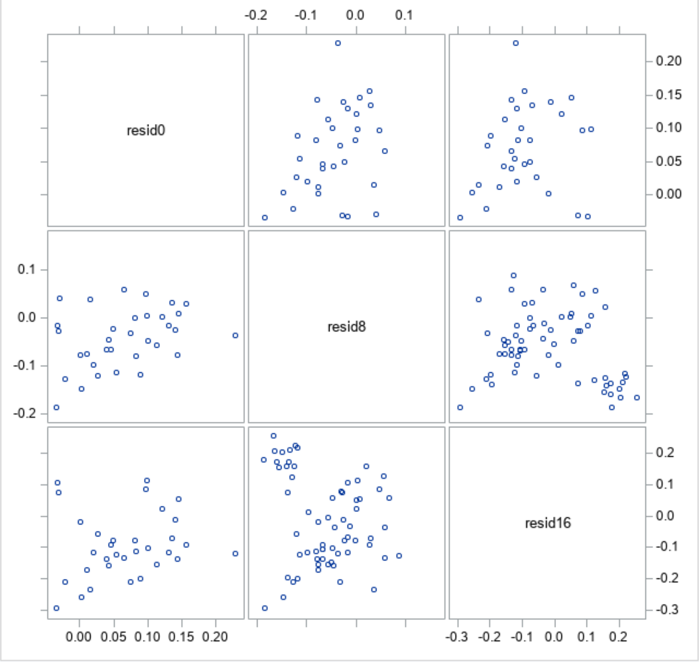


Residual Statistics	
Observations	1122
Minimum	-0.294
Mean	-0.006
Maximum	0.2836
Std Dev	0.1093

Fit Statistics	
Objective	-1974
AIC	-1970
AICC	-1970
BIC	-1965

Prior to modelling, both the dependent variable and the residuals of a general model using the dependent variable were tested for normality. Although the numeric testing values displayed violations to the assumption of normality in the Shapiro-Wilks and the Kolmogorov-Smirnov tests, these tests are sensitive to sample size. Using evidence from the Q-Q plot, it was determined that the dependent variable (Two-Year Withdrawal Rate) may be normally distributed. In addition, a residual scatterplot matrix was created to look at the correlation between different timepoint combinations. Figure 8 below displays combinations of the 2001-02, 2009-10, and 2018-19 school years. For the 2001-02 school year in the first column, the residuals display a cloud-like structure. This structure is observed again for the school year in the second column (2009-10), but for the 2018-19 school year a nonlinear pattern is observed between the 2009-10 and 2018-19 school years.

Figure 9: Residual Scatterplot Matrix



4. Results

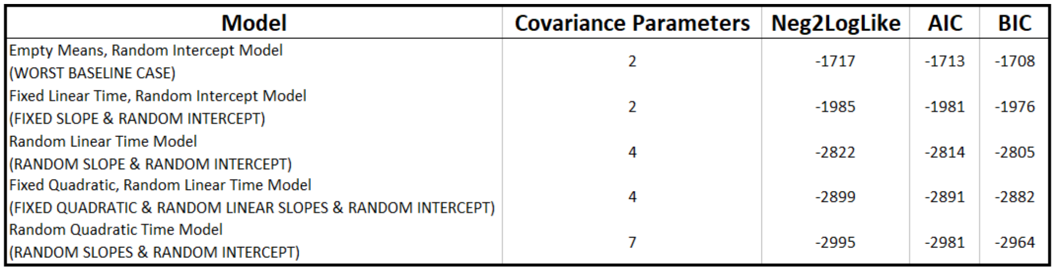
The withdrawal rate for each of the 1,122 occasions across the 97 Georgia schools over 18 school years were collected for this longitudinal analysis. The dependent variable, withdrawal rate, is a continuous variable with a value between 0 and 1. Since the earliest timepoint observed was the 2001-2002 school year, this analysis will center the data around this school year. Prior to modelling the normality of the dependent variable was assessed using a residual scatterplot, Q-Q plot, and Shapiro-Wilks test.

The first model created was an Unstructured, Saturated Means model but this model did not converge. The next model created was an Empty Means, Random Intercept Model which uses a Compound Symmetry covariance structure to partition the variance. This type of structure uses two parameters, the between-school variance and the residual within-school variance, and allows for the retrieval of a baseline value for the Intraclass Correlation Value (ICC) in order to assess the proportion of Level 2, between-person variance. Next, fixed and random effects (both linear and quadratic) were added to the model in an iterative process to determine if each effect contributed to the model fit by utilizing Wald test statistics and their corresponding p-values for the fixed effects, or a likelihood ratio test using -2 log likelihood (-2LL) for the random effects. Model fit for both random and fixed effects was also assessed by utilizing Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC).

Likelihood Ratio Test: The Likelihood Ratio Test compares the Random Intercept model to an error-only model to test whether the ICC is significantly greater than 0 (for p-values < 0.05). For this project, the Empty Means, Random Intercept model produced an ICC of 0.2862 which suggests that 28.62% of the total variance is due to between-school differences. The Likelihood Ratio Test for this model produced a Chi-Square statistic of 692 and a p-value less than 0.0001, which suggests that the ICC value (0.2862) is significantly greater than zero. Since 71.38% of the total variance is due to within-school differences, moving forward with a longitudinal model is appropriate.

The worst model was the empty means, Random Intercept model which produced an AIC of -1,713 and a BIC of -1,708 (Table 2). Since the baseline ICC value displayed 72% of the total variation of this data is due to time-variant differences, the empty means, random intercept model was expected to be the worse because this model does not account for the change over time. Adding a fixed linear time slope to the model slightly improves the AIC (-1,981) and BIC (-1,976) while adding a random linear time slope greatly improves the AIC (-2,814) and BIC (-2,805). As discussed earlier, the spaghetti plot (Figure 5) displayed a nonlinear trend ten years after the earliest observed School Year. For this reason, a quadratic time parameter was added to the model to form a Random Quadratic Time Model which produced the best (lowest) AIC and BIC values (AIC: -2,981, BIC: -2964, Table 2).

Table 2: Model Comparison of the Time-Only Models



Model	Covariance Parameters	Neg2LogLike	AIC	BIC
Empty Means, Random Intercept Model (WORST BASELINE CASE)	2	-1717	-1713	-1708
Fixed Linear Time, Random Intercept Model (FIXED SLOPE & RANDOM INTERCEPT)	2	-1985	-1981	-1976
Random Linear Time Model (RANDOM SLOPE & RANDOM INTERCEPT)	4	-2822	-2814	-2805
Fixed Quadratic, Random Linear Time Model (FIXED QUADRATIC & RANDOM LINEAR SLOPES & RANDOM INTERCEPT)	4	-2899	-2891	-2882
Random Quadratic Time Model (RANDOM SLOPES & RANDOM INTERCEPT)	7	-2995	-2981	-2964

After selecting the best time-only model, the average student sentiment will be added to the model as a level-2 predictor. The Random Quadratic Time Model was observed to have the lowest AIC (-2,981) and BIC (-2,964) and was chosen as the best time-only model. After adding the predictor variable, average student sentiment, the AIC and BIC expectedly increased to -2,974 (AIC) and -2,957 (BIC). For both the time-only model and the predictor model, the p-value for the intercept and quadratic slope were both significant. The linear slope produced a relatively small p-value (0.04) in the time-only model but the p-value of the linear slope in predictor model was much larger (0.10). In addition, the predictor variable, average student sentiment, did not produce a small enough p-value to be significant to the model (0.3259). The main effect of student sentiment is the difference in the expected withdrawal rate at time zero. Since this estimate (-0.007) was negative, the change in withdrawal rate decreases for a unit increase in the average student sentiment. Similarly, the effect of withdrawal rate on linear time is the difference in the expected linear rate of change at time zero per school year. Since the estimate is negative (-0.003), the difference in the linear rate of change at time zero per school year is decreasing for a unital increase in the average student sentiment. Lastly, the effect of withdrawal rate on quadratic time is the difference in half the rate of acceleration/deceleration of linear rate of change per school year. The estimate (0.00099) found by this model is positive which suggests that the difference in half the rate of acceleration/deceleration of linear rate of change per school year is increasing with a unital increase in the average student sentiment. Since these values are very close to zero, the rate of change in withdrawal rate over time is very subtle. The equation that describes the model for predicting the change in the two-year withdrawal rate by the average student sentiment over time is the following:

$$Two\ Year\ Withdrawal\ Rate = \beta_{0i} + \beta_{1i}(School\ Year)_{ti} + \beta_{2i}(School\ Year)_{ti}^2 + error_{ti}$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(Student\ Sentiment) + U_{0i}$$

$$\beta_{1i} = \gamma_{10}$$

$$\beta_{2i} = \gamma_{20}$$

Table 3: Model Comparison for the Best Models

* Bold Values are p < 0.05	**Best Time-Only Model** Random Quadratic Time Model			**Best Predictor Model** Random Quadratic Time Model			
Name	**Estimate**	**Std Error**	**p-value**	**Estimate**	**Std Error**	**p-value**	**Label**
Model for the Means							
Intercept	0.161	0.01252	**< 0.0001**	0.16490	0.013160	**< 0.0001**	γ_{00}
Linear Slope	-0.00328	0.00156	**0.04**	-0.00273	0.001654	0.1035	γ_{10}
Quadratic Slope	0.00101	0.00013	**< 0.0001**	0.00099	0.000135	**< 0.0001**	γ_{20}
Predictor				-0.00719	0.007311	0.3259	γ_{01}
Model for the Variance							
Random Intercept Variance	0.009679	0.002173	**< 0.0001**	0.009739	0.002182	**<.0001**	$\tau_{U_0}^2$
Intercept-Linear Slope Covariance	-0.00012	0.000203	0.5569	-0.00012	0.000203	0.5550	
Linear Slope Variance	0.00005	0.000028	**0.0344**	0.000049	0.000027	**0.0363**	$\tau_{U_1}^2$
Intercept-Quad Slope Covariance	-0.00007	0.000015	**< 0.0001**	-0.00007	0.000015	**<.0001**	
Linear-Quad Slope Covariance	-0.00000179	0.000002	0.4130	-0.00000173	0.000002	0.4280	
Quadratic Slope Variance	0.00000092	0	.	0.00000092	0	.	$\tau_{U_{01}}^2$
Residual Variance	0.002458	0.000116	**< 0.0001**	0.002459	0.000116	**<.0001**	σ_e^2
MultiLevel Model Fit							
Number of Parameters	7			7			
-2LL	-2994			-2988			
AIC	-2981			-2974			
BIC	-2964			-2957			

5. Conclusions

In conclusion, the two-year withdrawal rate rapidly increased during the 17 school years observed in this data. Discovering the mechanisms that may have led to the cause of this shift could aid schools in reducing their two-year withdrawal rate. This project sought to investigate what role student sentiment, observed in Rate My Professor comments, played. It was found that the average student sentiment may not be very useful in the prediction task. This may be caused by the aggregation step of this project where the student sentiment value was averaged for each school and school year. Instead, it may be better to add the sentiment of the individual reviews as a time-varying level-1 predictor. This will be the plan for a future research project where the total population of Rate My Professor reviews will be used.

6. References

- Language Model Used to perform the Sentiment Analysis:
    - ‘SiEBERT/sentiment-roberta-large-english’
    - https://huggingface.co/siebert/sentiment-roberta-large-english

- GitHub Library Used to Scrape Rate My Professor:
    - https://github.com/tisuela/ratemyprof-api

- The code used for this project can be found here:
    - https://github.com/njones738/Analytics_Day_Fall2022