**IMPROVING ACADEMIC PERFORMANCE IN THE HISTORY DEPARTMENT**

**Purpose**:

* Improving academic performance of students at Wesleyan College within the History Department by viewing the factors that influence the final grade of the class.
* Subsequently make recommendations for improving the academic performance based on predictive and sentiment analysis, as well as survey results.
* Fostering the visibility of data in the non-STEM department
* Possible factors influencing the final grade based on data set:

Class level

Major/Minor

Year in which History class was taken

Semester

History Course Level

Papers

Exams

**Methodology**:

**Analytical tools used** : Python and R: allowed access to data, visualization of the raw data, predictive analytics and sentiment analysis

**Macro methodology:** Scientific Research Methodology/ Scientific method: research-discovery-learning analytics and for data exploration - research question, hypothesis, examination of research question and hypothesis, collecting data and testing hypothesis through linear regression, analysis and generalization.

**Micro methodology:** Group I (data exploration and discovery - descriptive analysis - hypothesis testing and linear regression) and Group III (models, solutions dependent on data - classification/sentiment analysis)

**Multiple Linear Regression & Hypothesis Testing - Final Grade Data**

Multiple linear regression is a commonly used type of predictive analysis that allows you to determine whether a set of predictor variables can predict the outcome of a dependent variable, as well as which variables are significant predictors of the outcome variable. Once the p - value is less than 0.05, it means that the variable is statistically significant, the null hypothesis is rejected and the alternative hypothesis is accepted. This is because there is less than a 5% probability the null is correct. However, once it is greater than 0.05, it means the null hypothesis is true:

**Null hypothesis (H0): There is no correlation between the selected independent variables and the final grade in the history department.**

**Alternate hypothesis (H1): There is evident correlation between the selected independent variables and the final grade in the history department.**

Multiple linear regression will be used in order to predict the relationship between the final grades, and the independent variables in the dataset. The p-value will be used to determine whether the null hypothesis should be rejected or not. This lm() function in R will be used to perform the linear regression.

**Data cleaning:** In order to prepare the data before the linear model was created, the dependent variable (final grade) had to be converted from character type to integer type in order to indicate that it was of an ordinal data type suitable for the regression model in the Excel sheet before it was imported into R. The independent variables remained the same as they were all suitable for regression analysis. No further cleansing was done. After this, data visualization was performed to see how the data was distributed across the dataset. The linear model was then created to view the correlation between the independent variables and the dependent variable.

**Sentiment Analysis & Classification - Survey Data**

This is the process of computationally identifying and categorizing opinions from pieces of text, and determining whether the writer’s attitude towards a product or topic in this case is positive, negative or neutral. Sentiment analysis is split into 5 steps under the TextBlob library in Python:

Step 1 - Tokenization: dividing the data or pieces of statements into separate parts

Step 2 - Cleaning the data (removing unnecessary characters and stopwords - does not add any substantial meaning to a sentence - words like he, the, she, they, etc.

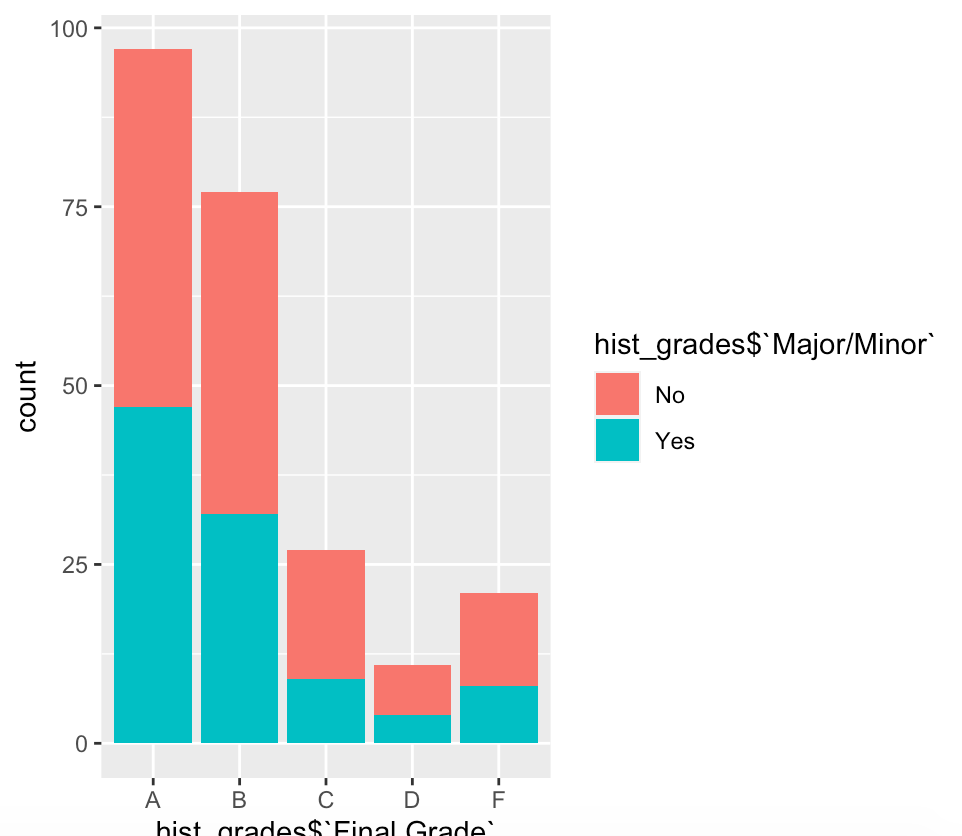
Step 3 - Classification - determining whether or not a word is positive, negative or neutral - positive words are +1, negative words are -1 and neutral are 0.

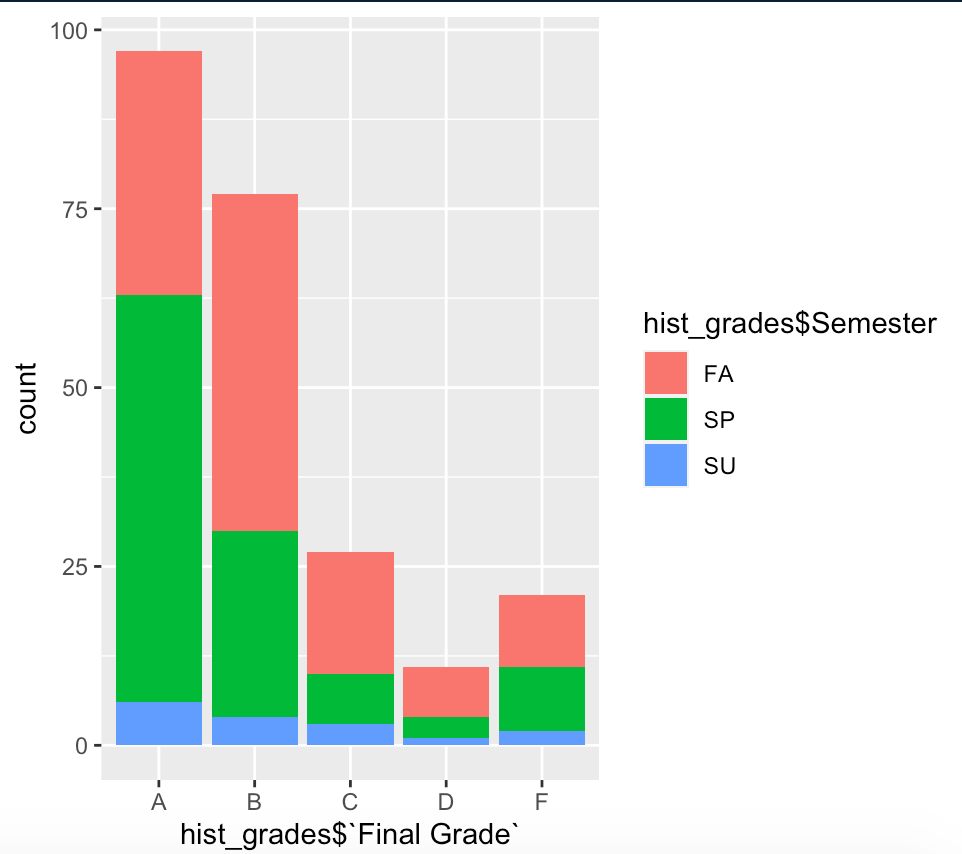
Step 4 - apply supervised learning algorithms; train the model with lexicons and test it on the analyzing statements.

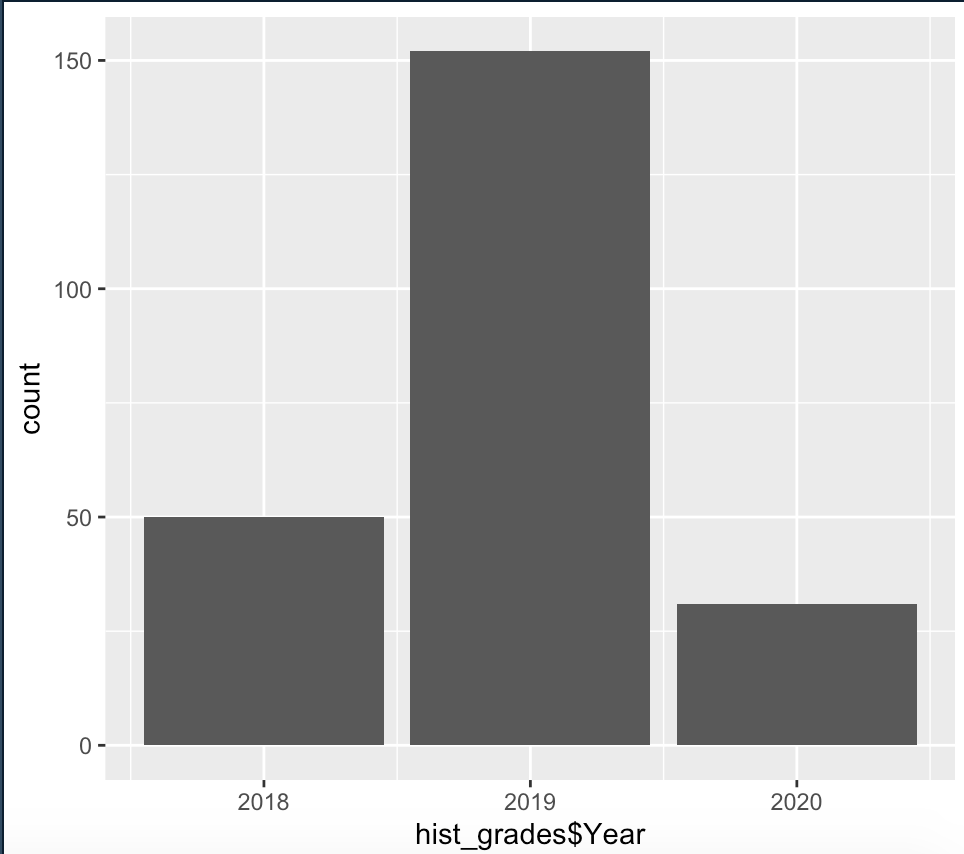
Step 5 - calculation of the final sentiment score

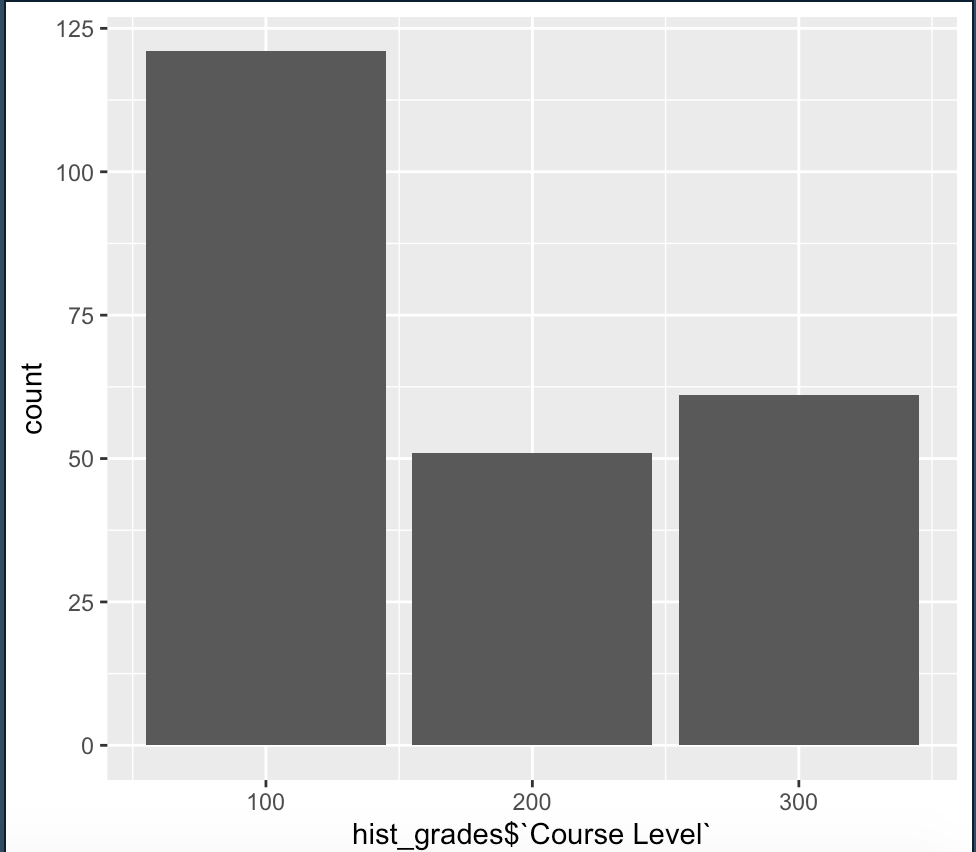
Data was also derived from a survey conveyed amongst 35 students who had either taken the History class in the past or were currently taking it. Their responses were downloaded from Google Forms into an Excel spreadsheet, which was further converted into a text file in order to perform sentiment analysis for the data. Data about how the course could be improved and how valuable they thought the course was was split into two different variables to decipher the sentiment analysis for the two different topics. Data cleaning was automated by the TextBlob library which got rid of the stopwords and unnecessary characters that might harm the sentiment analysis. The supervised learning algorithms under this library additionally automated the model training and calculation of the final sentiment score for the variable about improving the course and the other variable about how valuable the course was. A word cloud was then created using the WordCloud library in Python to also view the more prominent and frequent words about the variability of the History courses and to conduct text analysis.

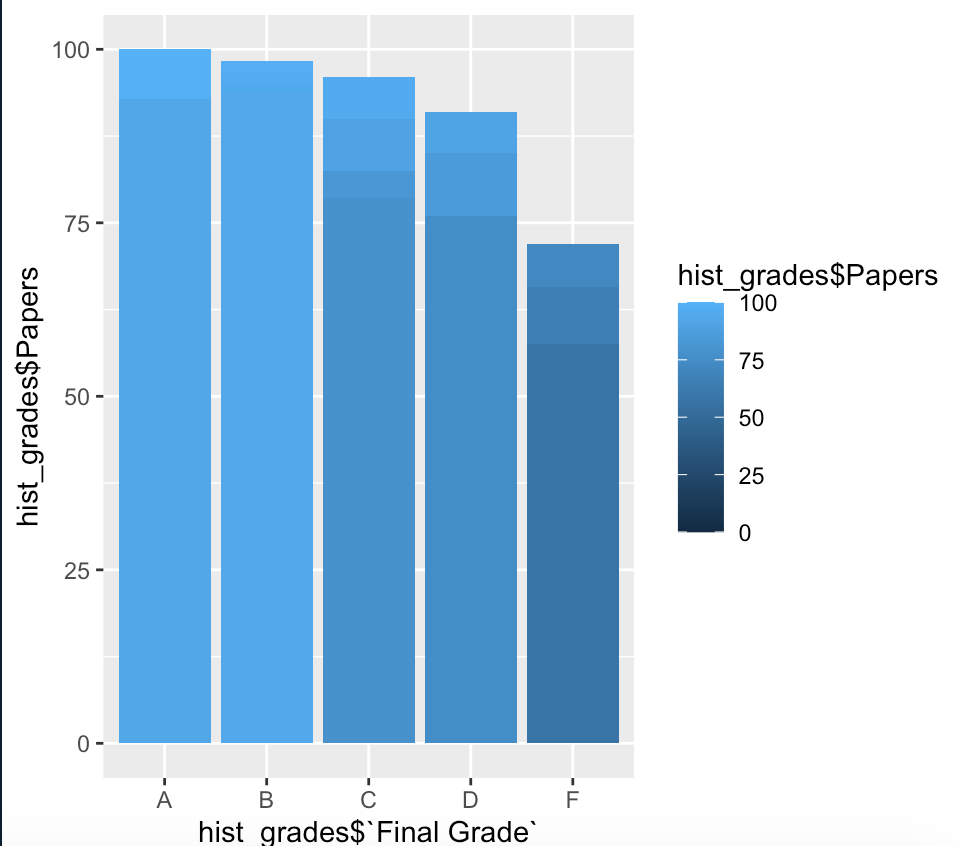
**Data (Final Grade)**:

Data consisting of both quantitative and qualitative data; nominal, ordinal and discrete data used

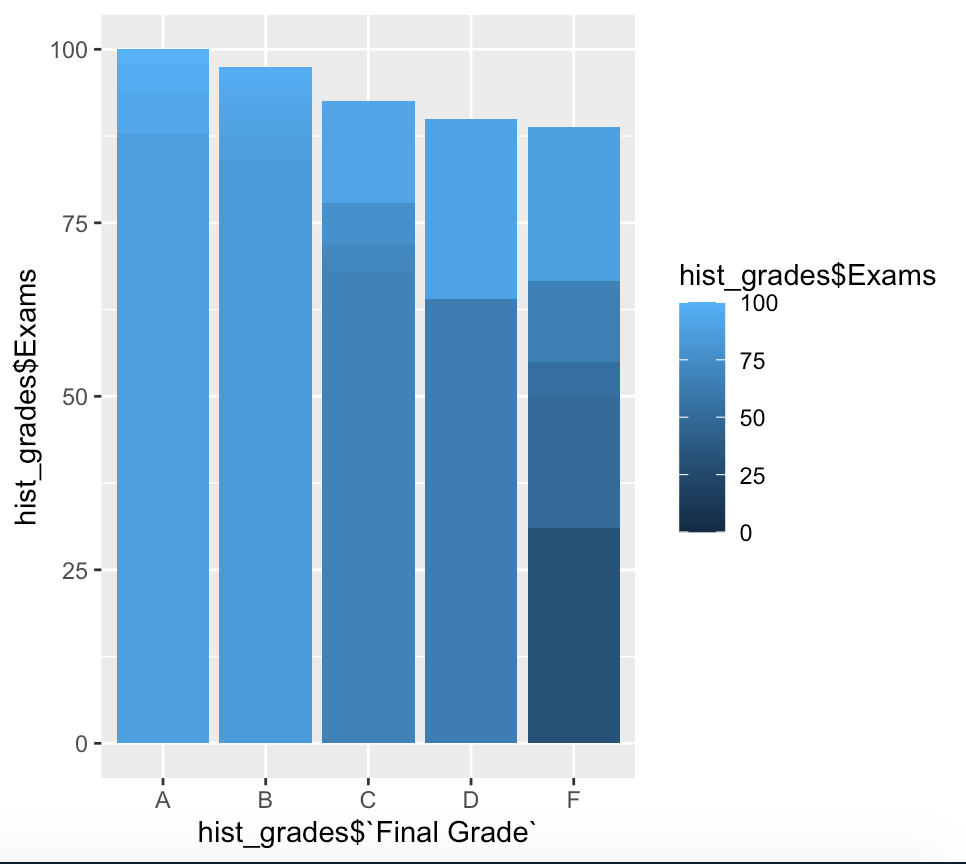


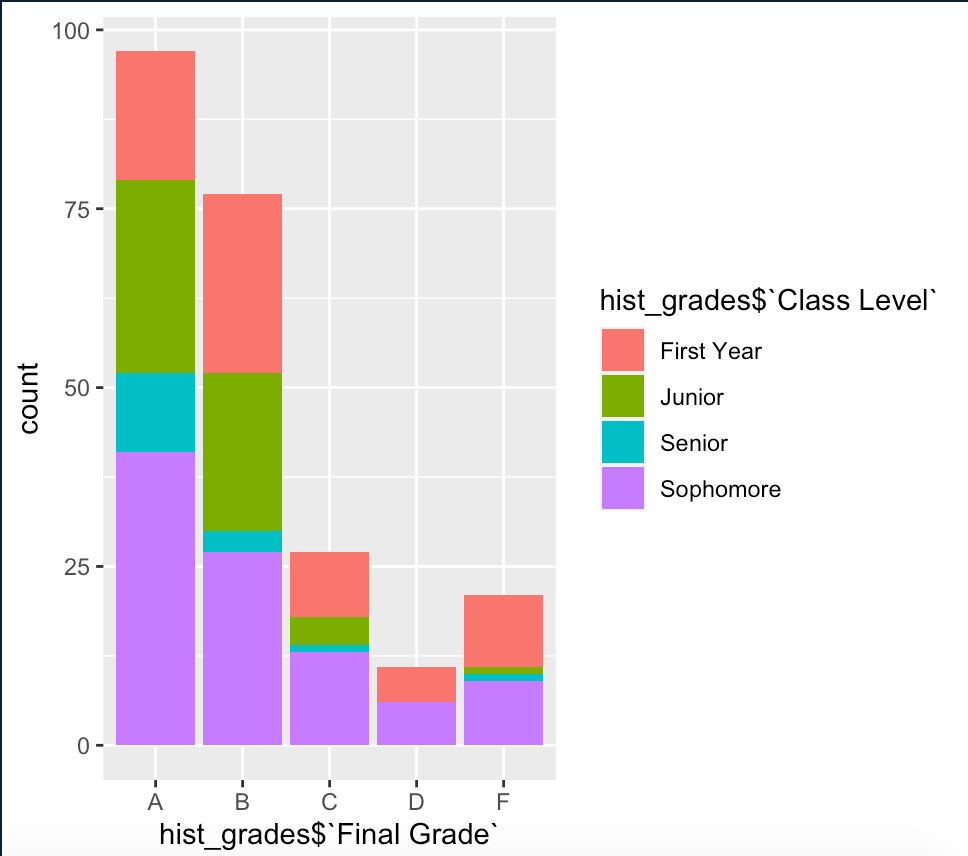




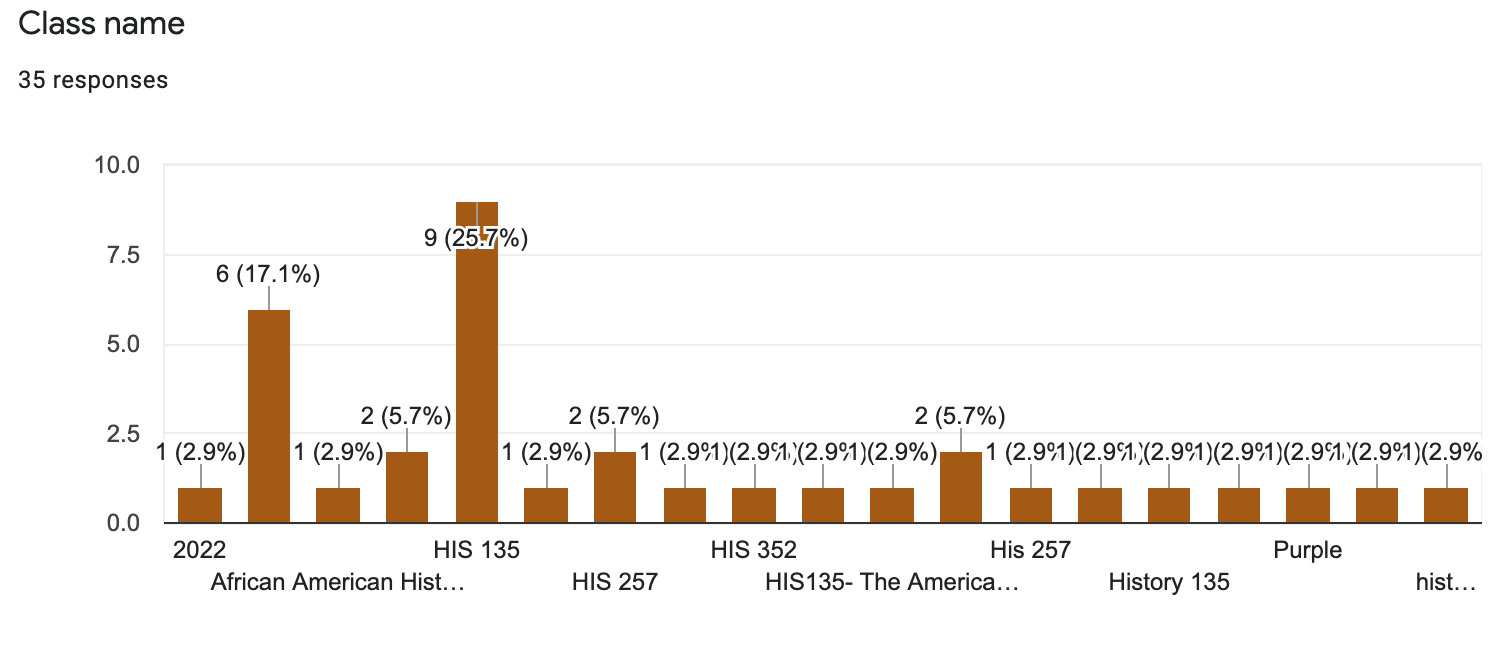


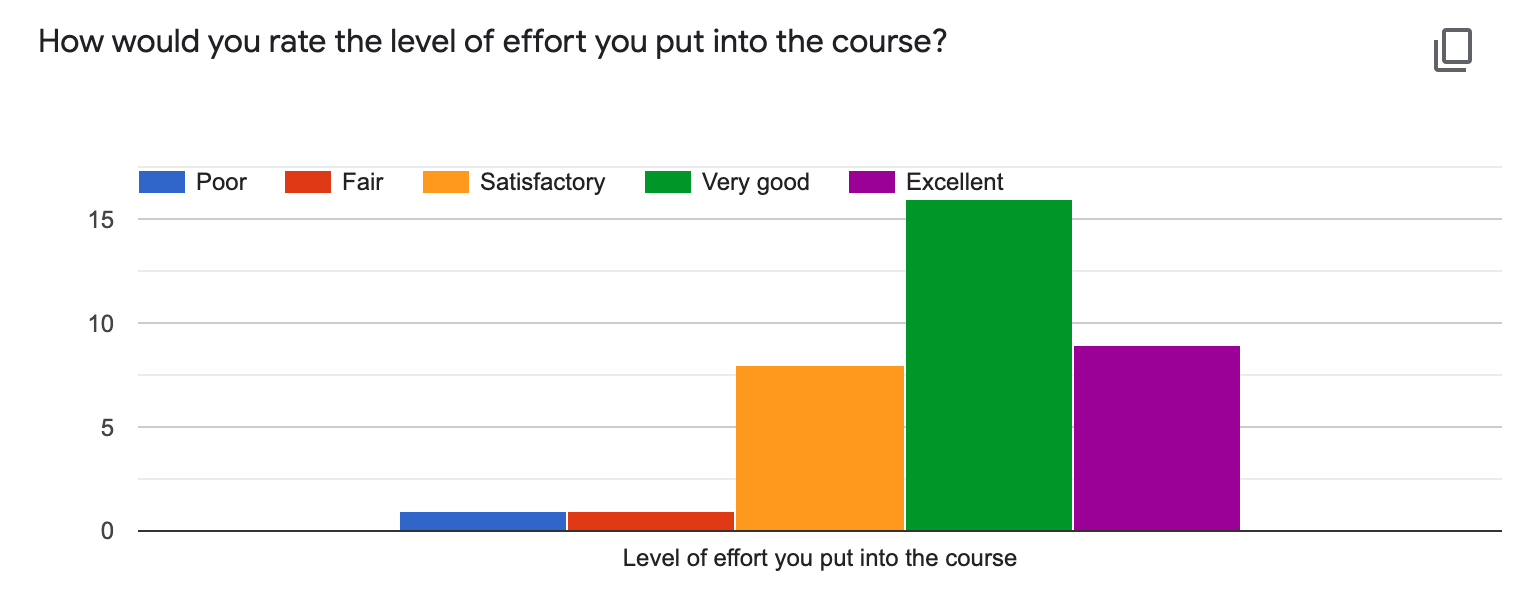




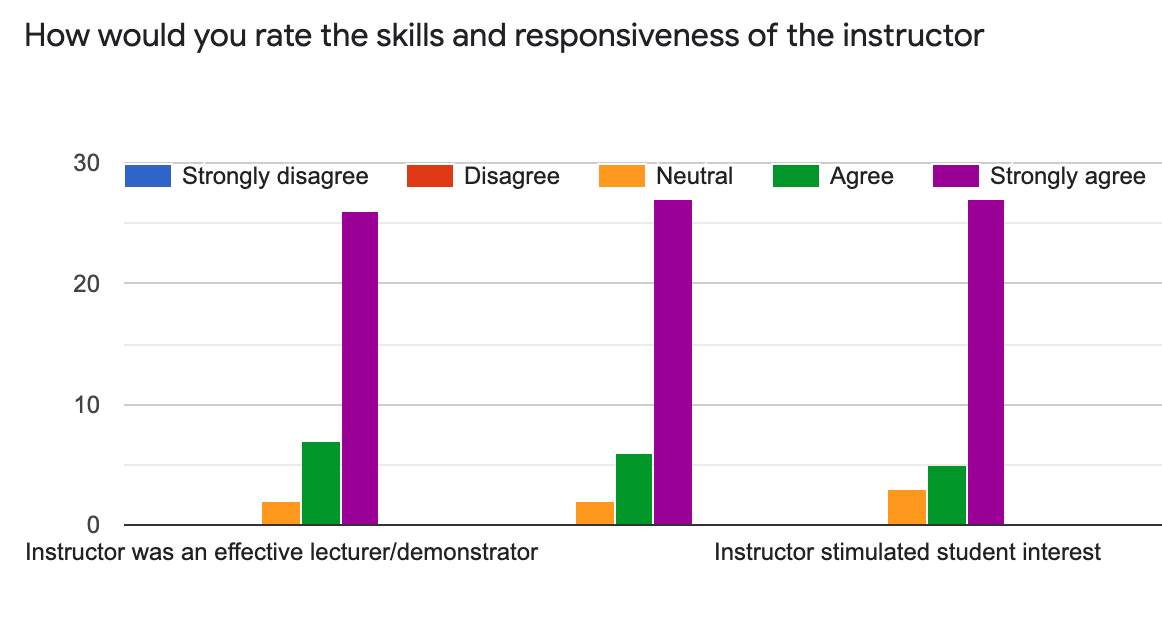


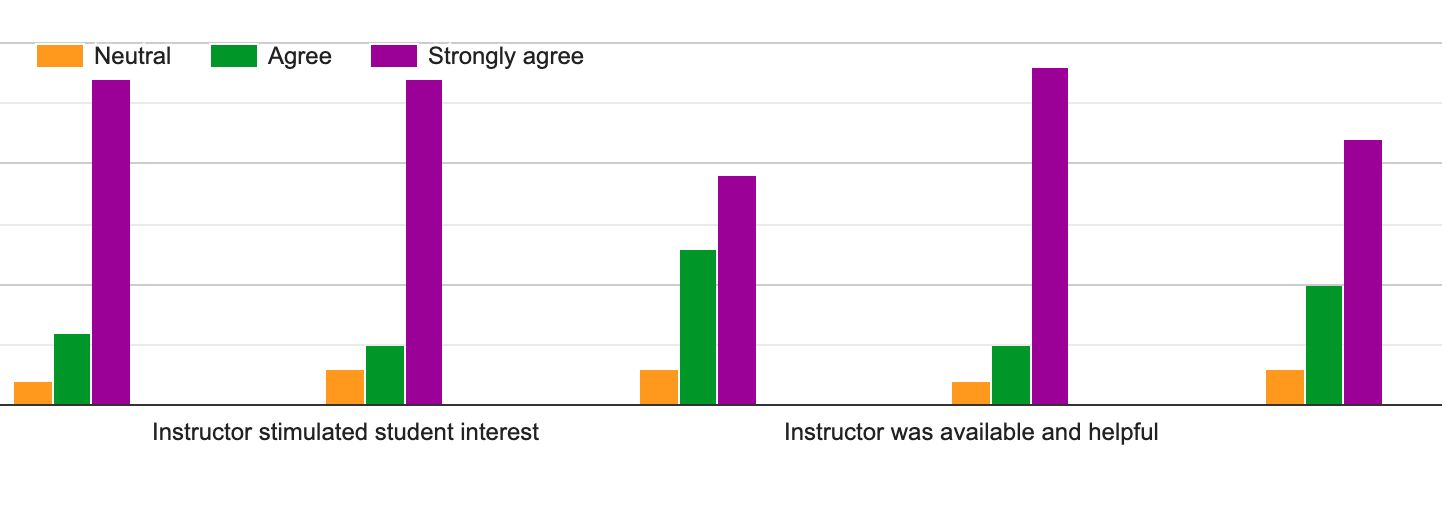


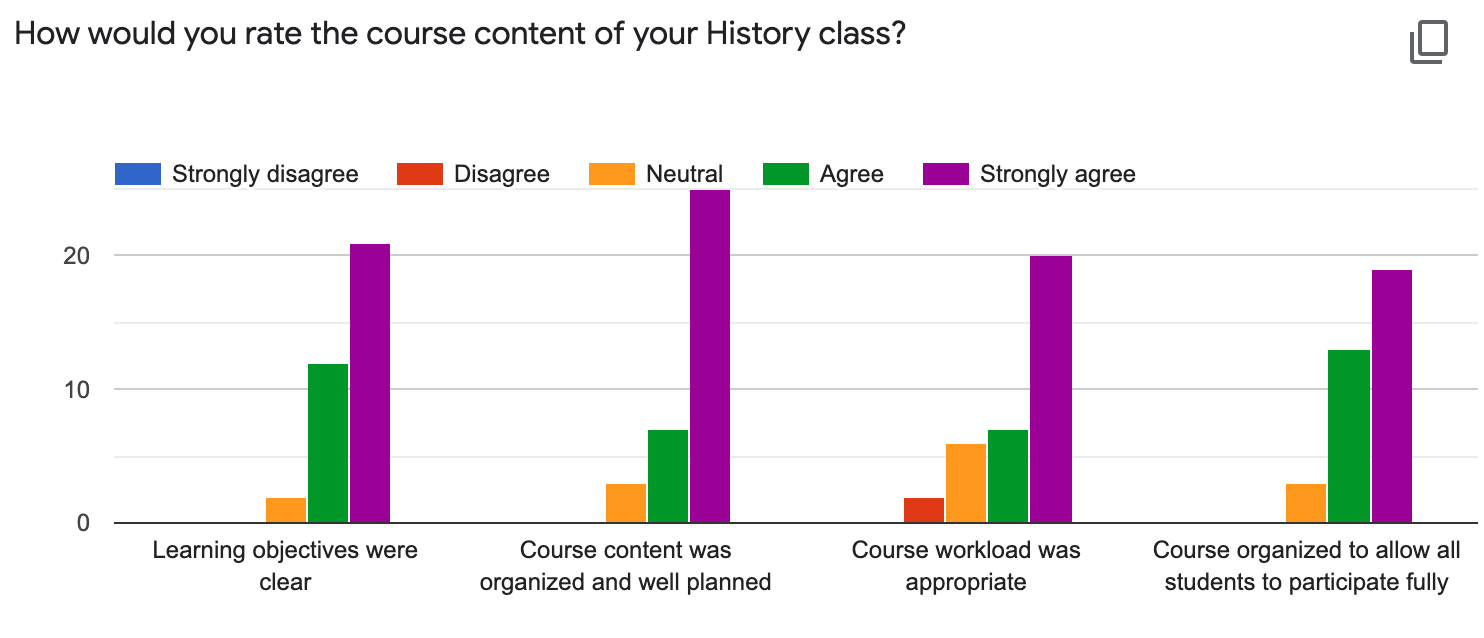


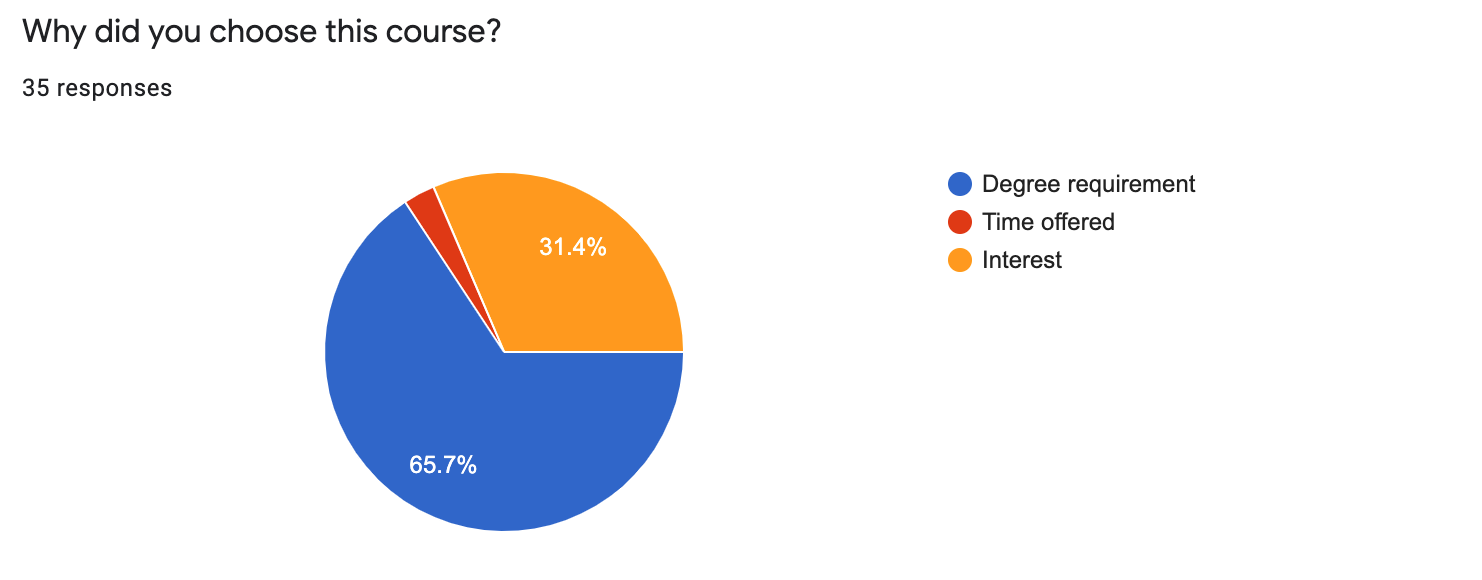


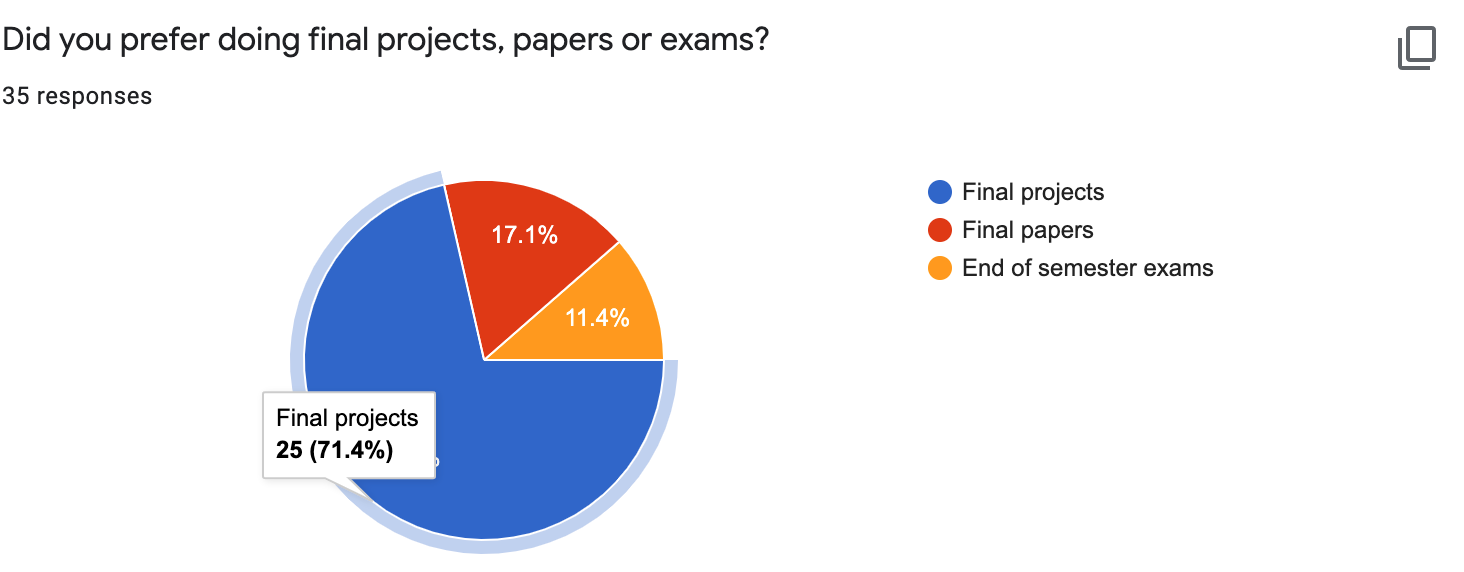




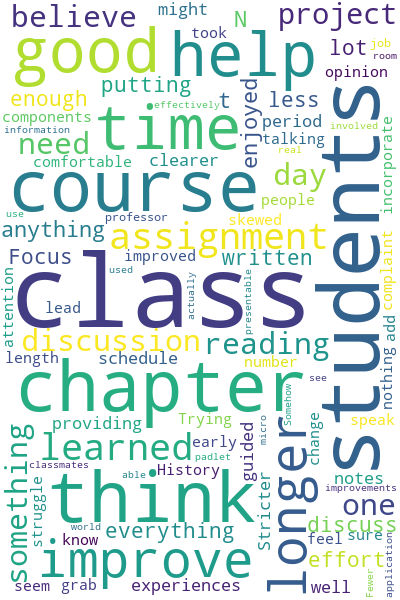


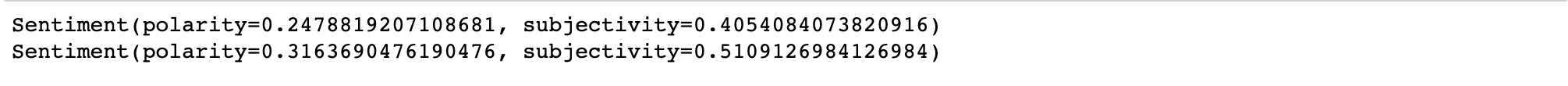
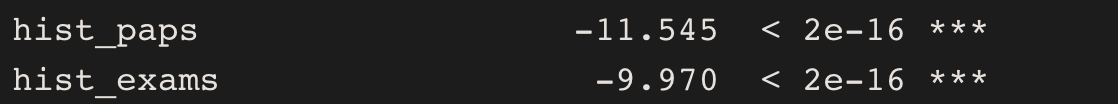
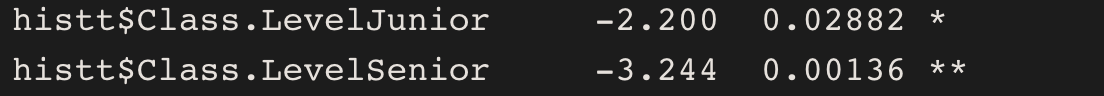






(Relevant question as this is usually a huge percentage of the final grade in the class)





**Analysis:**

* **Linear Regression:** Based on the results from the linear regression model, it is evident the following independent variables had high significance and can reject the null hypothesis:

- Class Level (whether or not they were a senior or junior)

- Semester (Spring semester)

- Papers

- Exams

It makes perfect sense thar the papers and exams had the highest significance, as they usually wield a huge influence on the final grade due to the percentage distribution (papers and exams are worth a lot more of the final grade compared to in class discussions, assignments and final projects). Because these variables were statistically significant, the null hypothesis can be rejected in their cases, meaning there was a correlation between these four independent variables and the final grade of the class.

* Sentiment Analysis:

From the analysis of how valuable the course was for students, polarity (meaning how positive or negative their feedback was), was almost neutral. Even though no inferences can be made from this, it was still a good sign that the overall feedback was not negative. The subjectivity (how opinionated the feedback was), also neared neutrality. In the case of how the course could be improved, polarity was only slightly higher, and neared neutrality which showed, again, that the overall feedback was not negative, the same thing occurred with subjectivity.

**Recommendations & Conclusion**:

* Based on the analysis conducted in this project, it would not be too presumptuous to state that more emphasis should be placed on the final exams and papers. As they influence the final grade the highest, the final grade could be improved by providing more clarity in terms of what is expected, providing study guides before the exam or final paper is due, and even through individual consultation sessions between the professor and students who believe they may be lagging behind in their classes.
* In addition to this, improvement could also be seen if there was a shift in the weight of the final project, which from the survey data, shows that more students prefer doing. Improvement of the final grade can be seen by accommodating to the students’ overall interests, as seen from the responses.

**Deployment**

It is clear that this model will require application development: because the model created applies to only the students involved in the dataset, it cannot be applied generally to all students that may come to Wesleyan College, as the responses may vary and the significant independent variables may also change. Thus, it is important to keep tracking the model quality from time to time - preferably yearly - to esnure that proper attention is paid to the necessary factors involved each year. It is also important to evaluate the departmental benefit of the model over time - i.e. whether or not this model works to the benefit of the professors in this department over time.

**Closing remarks**

All in all, this project was particularly important in not only finding out what was necessary to focus on for the improvement of grades, but also in finding out students’ opinions about History courses taken at Wesleyan college.