INFO 2950 PROJECT: Should I Let RateMyProfessor Choose My Classes?

Contributors: Carl Huang (ch976), Mu-Chieh (Jay) Huang (mh989), Evan Vu (ev239), Mith Patel (mpp59)

Table of Contents

1. INTRODUCTION

What is the context of the work?

Whether it's making a complex topic seem simple or turning a boring subject interesting, choosing a good professor can be one of the biggest determinants of how much a student enjoys a class. As a result, students often take RateMyProfessor data into account before enrolling in a certain class, and it serves as an index for selecting classes. Therefore, we would like to know if the ratings and difficulty of professors on RateMyProfessor are accurately reflective and representative. Through this project, we

want to explore the relationship between RateMyProfessor ratings, dificulty levels, number of reviews, and student performances in terms of grades.

With this project, we would like to see if there is a correlation between a professor’s difficulty and rating, and their average grade point average. Down the line, we could fit a multi-variable regression in order to predict what a professor’s median grade would be depending on what the difficulty and rating of that professor was.

The project focuses on exploring the relationship between professors' ratings and difficulty levels, obtained from RateMyProfessor, and the corresponding impact on the average grade point averages (GPAs) of classes taught by these professors at Cornell University. The aim is to determine whether students' perceptions of professors—as reflected in ratings and perceived difficulty—correlate with the actual academic outcomes (median grades) in their courses. The data analyzed includes:

1. Professors' ratings, difficulty levels, views from RateMyProfessor.
2. Median grades from classes taught by these professors – sourced from a dataset compiled by Cornell University students.

The project is set in an academic context, specifically looking at the dynamics of teacher evaluation from RateMyProfessor and reported student performance for around 500 classes within Cornell University. This analysis has broader implications for understanding how subjective evaluations of teaching staff might relate to objective educational outcomes like grades.

These are our research questions:

- Can a professor’s overall rating and difficulty level predict the instructor’s grade point average?
- How do a professor's ratings impact the grade distribution of the class?
- Do the ratings and difficulty level actually reflect how well students do?

What are your main findings? Include a brief summary of your results.

Hypothesis 1: Influence of Professor Ratings on Average Class Grades:

- Null Hypothesis (H0): No relationship exists between professor ratings and average class grades.
- Alternative Hypothesis (H1): There is a relationship between professor ratings and average class grades.
- Detailed Finding: The statistical analysis reveals a significant, albeit small, relationship between professors' ratings and the average grades of their classes. The low p-value indicates that the likelihood of this finding being due to chance is

minimal. However, the small coefficient (0.0343) and the low R-squared value (0.008) suggest that the strength of this relationship is weak. It implies that while students' perceptions of professors, as reflected in ratings, have some bearing on their grades, numerous other factors not accounted for in the model likely play a more substantial role in determining academic outcomes.

Hypothesis 2: Combined Impact of Ratings and Perceived Difficulty on Average Grades:

- Null Hypothesis for Ratings ($H0_1$) and Difficulty ($H0_2$): Ratings and perceived difficulty have no effect on average grades.
- Alternative Hypothesis for Ratings ($H1_1$) and Difficulty ($H1_2$): Ratings and perceived difficulty have an effect on average grades.
- Detailed Finding: The statistical analysis for ratings did not show a significant impact on average grades, leading to the retention of the null hypothesis for ratings. In contrast, the null hypothesis for perceived difficulty was rejected due to a significant p-value. The negative coefficient (-0.1670) for difficulty implies that classes perceived as more challenging are associated with lower average grades. The model's modest R-squared value of 0.114 suggests that about 11.4% of the variation in grades can be attributed to these factors, highlighting the influence of other unmeasured variables.

Hypothesis 3: Relationship Between Professor Ratings and Number of Reviews:

- Null Hypothesis (H0): Professor ratings do not influence the number of reviews.
- Alternative Hypothesis (H1): Professor ratings affect the number of reviews.
- Detailed Finding: The analysis indicates a significant inverse relationship between ratings and the number of reviews, as evidenced by the negative coefficient (-6.4802) and the p-value (0.034). This suggests that professors with higher ratings tend to receive fewer reviews. The low R-squared value (0.009), however, denotes that this factor alone does not substantially explain the variation in the number of reviews, implying that other unexplored factors may influence students' decisions to leave reviews.

The study sheds light on a number of important facets of the connection between academic achievement and student assessments at Cornell University. It highlights the poor predictive ability of professor ratings on student grades, implying that although a link exists, it is not significant enough to be the only measure of a teacher's performance or a student's progress. The debate over academic rigor and its effects on student performance gains important context from the noteworthy influence of perceived difficulty on grades. The inverse correlation between ratings and reviews initiates a conversation regarding the character of student feedback and the driving forces behind it. This result suggests that there may be a bias in the evaluation process, as students

may be more likely to provide feedback based on extreme experiences, whether they are extremely good or unfavorable. Overall, the research provides a nuanced perspective on the use of student evaluations in assessing academic performance. It highlights the necessity of considering a range of factors, beyond student perceptions, in evaluating teaching effectiveness and shaping educational strategies. Future research should aim to incorporate these additional variables to present a more comprehensive understanding of the determinants of academic success.

2. DATA DESCRIPTION

A) MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- This dataset was created to observe and analyze the relationship between Cornell University's professor and their median grades for their classes. It allows us to track a professor's difficulty, rating, and average median grades for their students across multiple classes that they have taught. We want to investigate how a professor's ratings impact the grade distribution of the class and how a professor's overall rating and level of difficulty predict the instructor's grade point average.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number?

- The raw dataset of median grades by class was created and updated by various students of Cornell University on behalf of Cornell University.
- The main source of funding for this dataset comes from the website RateMyProfessor and from a self-reported dataset on Google Sheets, which is updated and self-reported by students. We obtained the data from the website through an API and the Google Sheets from a Teaching Assistant. The raw data is sourced from the Cornell University student population.

B) COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

- The instances of the dataset represent the relationships between the ratings/difficulty of Cornell professors on RateMyProfessor and the average median of how well students actually do in terms of grades. The types of instances include names, ratings, difficulty, and number of reviews of professors and the average median grades students received.

How many instances are there in total (of each type, if appropriate)?

- There are five instances in total: professor name, professor rating, professor difficulty, number of reviews for each professor, and grade point average students received by professors.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

- The instance of professor names consists of the last names of professors in string.
- Professor ratings consist of the ratings of each professor on a scale from 1 to 5 in float.
- Difficulty of professors consist of the level of difficulty of each professor on a scale from 1 to 5 in float.
- The professors' number of ratings consist of the number of ratings each professor received in int.
- The average median grades of students consist of great point average students received by professors in float.

Is any information missing from individual instances? If so, please provide an explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

- The raw dataset of median grades by class has some missing data (NaN) in the columns of professor, grade, and number of students because they are unavailable in the get go. Moreover, the merged dataset had zeros in the column of "Num_reviews", which indicates that the corresponding professors do not have a rating on RateMyProfessor. We eventually eliminated these rows for a more polished version of our dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

- Professor Van Es appears twice in the "Professor" column because the cases of his name does not match. For example, one name is "van Es," and the other is "Van Es". This is the only minor issue we have with our dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external

resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- The merged dataset contains the raw data of median grades by class and the data retrieved from the RateMyProfessor API. It is an individual dataset that does not link to any external resources, but we did get our professor's ratings and the difficulty from https://www.ratemyprofessors.com/ using the API.

Since we use the website's data to complete our dataset, we cannot guarantee that the content will remain constant since anyone could give a new rating any day. Also, this dataset only accounted for semesters from the past 2-3 years, so there might even be an update on the raw dataset of median grades by class. We do not have official archival versions for our dataset since students create and update one while the ratings and difficulty are retrieved using an API. Our dataset has no restrictions like licenses or fees.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

- Our dataset contains students' median grade point average corresponding to the instructing professor, which may be sensitive for some students.

C) COLLECTION PROCESS

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?

- The columns of the data frame are as follows: Professor, Average_grade, Ratings, Difficulty, and Num_reviews. Data in the columns of Professor and Average_grade were acquired from the raw dataset of median grades by class. On the other hand, the data of Ratings, Difficulty, and Num_reviews were acquired from the RateMyProfessor API.

If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

- The data reported by subjects was not verified. We are unable to do this because the data is completely anonymous and we are not able to access files containing the grades of students for these classes and professors.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

- We used a Python API to web scrape RateMyProfessor from https://docs.google.com/spreadsheets/d/1817WfShAi7RHWWJR2c95YUgX37TkjUsLQM The performance of the API was verified through extensive testing by checking the information returned from the API with the website. The median grades for classes were obtained from the Google Sheets directly using the Pandas library.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

- Only students were involved in the collection process. They self-reported reviews and grades for the classes and professors. They were not compensated as everything was unpaid.

Over what timeframe was the data collected?

- The data was collected over the timeframe of Fall 2016 to Spring 2033.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

- The people that are involved in the data collection are unaware of the data collection for our specific use of the data. The data on RateMyProfessor are self-reported by students who submit with the intention that they are submitting a rating for a professor to report overall opinions and personal experiences that they had with a specific professor in a specific class. This means that the reviews submitted on RateMyProfessor are submitted with no knowledge of our data analysis.

D) PREPROCESSING/CLEANING/LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- One part of the process that influenced what data was not observed and recorded is how we merged the data points. Since the Google Sheets contained less professors than the RateMyProfessor data we are recording only the professors found in the Google Sheets. Another part of the process that influenced what data was not observed was when we had to drop na values in the data frame gathered from the Google Sheets since it contained missing values. This means that we had to remove rows of data where there was an instance of na from our observation. In addition, we also had to remove rows where the column values for Num_reviews was 0 which removed professors with no reviews on RateMyProfessor from our observations. We also took the average of median grades across all of the classes taught by the professor meaning that we only observe the average of median grades rather than individual grades.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

- The raw source data for the Google Sheets of self-reported median grades for Cornell University classes can be found at https://docs.google.com/spreadsheets/d/1817WfShAi7RHWWJR2c95YUgX37TkjUsLQM The raw source data for the API that web scrapes the RateMyProfessor website for rating, difficulty, and number of reviews can be found at https://github.com/tisuela/ratemyprof-api.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

- The API that we used to web scrape information about professors from RateMyProfessor can be found at https://github.com/tisuela/ratemyprof-api.

E) USES

What (other) tasks could the dataset be used for?

- The dataset could also be used for professor evaluation and their performance and overall Cornell University student performance on classes.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

- The dataset becomes more and more outdated as more semesters pass by since the dataset records only the data for the semesters from FA16 to SP23.

Are there tasks for which the dataset should not be used? If so, please provide a description

- The dataset should not be used for investigating the correlation between the opinions of a professor and the average median grades of each of the classes for universities other than Cornell University.

3. PREREGISTRATION STATEMENT

Hypothesis 1: Students' median grade point averages are higher when they take classes taught by professors with higher overall ratings.

Analysis 1: We run a linear regression where we input grades (average median GPA) and output of professors' ratings (the average of all ratings). Because our measurement will not have signed variables, we will test whether $\beta_{rating} > 0$

Hypothesis 2: Professors' overall rating and difficulty can significantly predict the median grade in their class. Higher professor ratings and lower perceived difficulty levels are associated with higher median grades.

Analysis 2: We run a multi-variable linear regression where we input the professors' overall ratings and difficulty level as independent variables, and the median grade as a dependent variable.

Hypothesis 3: There is an inverse relationship between the number of reviews a professor receives and their rating. It is hypothesized that professors with higher ratings will receive a lower number of reviews, suggesting that students are less inclined to leave feedback when their experience is overwhelmingly positive.

Analysis 3: To test this hypothesis, a regression analysis will be conducted where the dependent variable is the number of reviews each professor receives, as a proxy for student engagement or the need to provide feedback. The independent variable in this analysis is the professor's overall rating. The aim of this analysis is to explore if there's a correlation between the professor's rating and the number of reviews they receive, as indicated by the β coefficient in the regression model. A negative β coefficient would imply that higher-rated professors tend to receive fewer reviews, potentially indicating a general satisfaction with their teaching that diminishes the perceived need for feedback.

4. DATA ANALYSIS

A) ANALYSIS OVERVIEW

Here is an overview of our data analysis:

- Average and standard deviation grade of Cornell in terms of GPA
- Bar plot of grade ranges in terms of GPA
- Scatterplot of ratings and difficulty
- Scatterplot of rating and GPA
- Scatterplot of difficulty and GPA
- Covariance matrix of ratings, difficulty, and GPA
- Distribution of average GPA by rating
- Distribution of average GPA by difficulty
- Multivariable regression model
- Correlation matrix of of Difficulty, Ratings, and Average_grade
- Heatmap of Difficulty, Ratings, and Average_grade

In [4]:
```
#importing libraries
import pandas as pd
import numpy as np
import ratemyprofessor
import seaborn as sns
import matplotlib.pyplot as plt
import duckdb
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
```

In [5]:
```
#loading csv file
df = pd.read_csv('final_df.csv')
df.head(50)
```

Out[5]:

	Professor	Average_grade	Ratings	Difficulty	Num_reviews
0	Chang	4.000000	4.5	2.0	29
1	Moghimi	4.000000	5.0	2.0	6
2	van Es	4.000000	3.9	2.8	82
3	Riley	3.325000	5.0	3.0	12
4	Van Es	4.000000	3.9	2.8	82
5	Alexander	3.575000	3.9	3.3	31
6	MacMahon	3.300000	4.5	3.4	44
7	Kourkoutis	3.300000	4.8	3.5	7
8	Wickham	3.300000	3.2	3.6	20
9	Fuchs	3.750000	4.7	2.1	60
10	Wise	3.300000	4.4	4.0	10
11	Kriner	3.300000	3.9	3.4	18
12	Stacey	3.300000	4.0	3.6	13
13	Nicholson	3.680000	4.8	2.8	27
14	Goldfarb	3.300000	1.3	4.9	10
15	March	3.650000	2.7	3.3	10
16	Howarth	3.150000	3.9	2.4	12
17	Campbell	3.300000	4.1	4.3	31
18	Huffaker	3.300000	3.5	3.9	36
19	Tumbar	3.300000	4.4	3.2	13
20	Fromme	3.300000	4.5	3.1	9
21	Liu	3.650000	4.3	3.1	12
22	Blake	3.150000	4.4	3.3	6
23	Blankenship	3.650000	4.1	3.6	44
24	Winans	3.300000	4.0	3.1	20
25	Scheinberg	3.300000	3.3	3.4	8
26	Doerschuk	3.300000	2.7	3.5	15
27	Saikia	3.300000	3.9	3.4	8
28	Entner	3.500000	2.4	2.8	51
29	Giles	3.300000	3.5	4.7	11
30	Lorey	3.225000	5.0	3.9	9
31	Davis	3.614286	4.0	3.2	43

	Professor	Average_grade	Ratings	Difficulty	Num_reviews
32	Kinsland	3.075000	3.1	4.0	27
33	Lin	3.300000	4.3	2.5	15
34	Milner	3.300000	3.5	3.8	62
35	Ezra	3.300000	3.0	4.8	9
36	Wilson	3.300000	4.1	3.6	9
37	Varner	4.000000	4.1	2.6	9
38	Duncan	3.000000	4.4	4.0	56
39	Schuldt	3.300000	4.5	3.0	6
40	Schmidt	3.300000	2.7	3.4	47
41	Sen	3.566667	2.8	3.1	25
42	White	3.575000	4.6	3.3	117
43	Schalekamp	3.400000	5.0	2.8	16
44	van Zuylen	3.240000	4.0	4.3	16
45	Clarkson	3.375000	5.0	4.0	33
46	Foster	3.433333	4.5	4.0	20
47	Hsu	3.300000	3.3	3.4	52
48	Bracy	3.300000	3.5	4.0	58
49	Alvisi	3.150000	2.5	4.5	29

B) BASIC SUMMARY OF DATASET

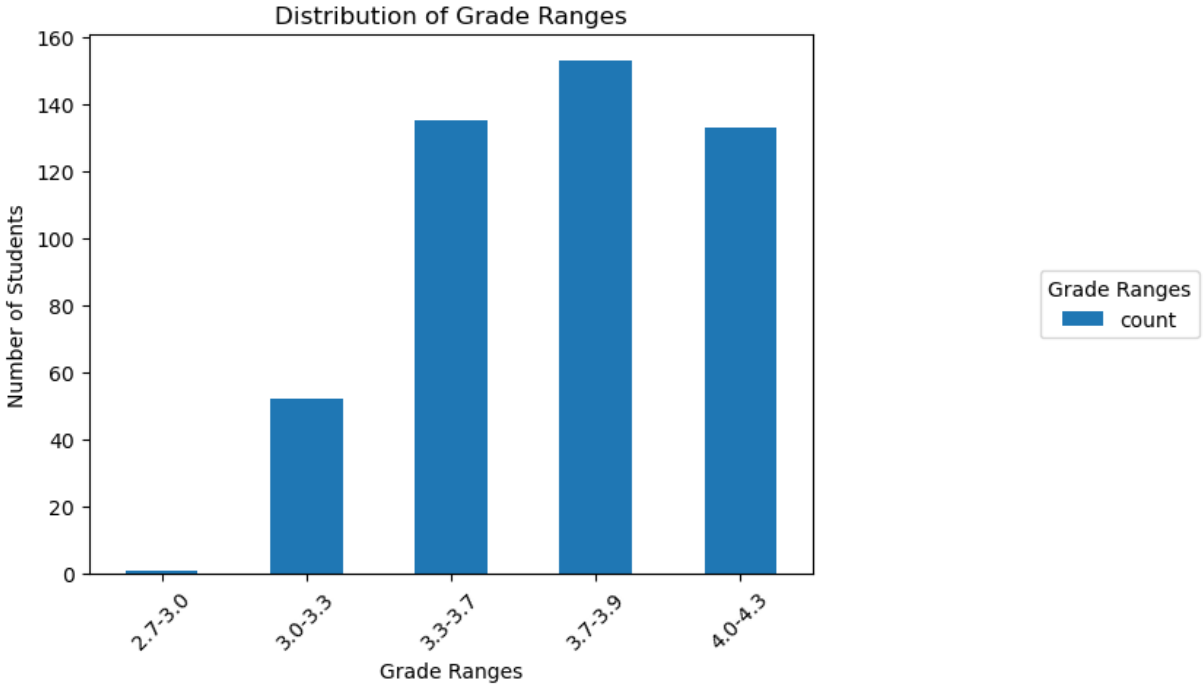
In [ ]:
```
#summary of dataset: mean and standard deviation
avg_grade = df['Average_grade'].mean()
std_grade = df['Average_grade'].std()
print(f'Average grade of Cornell: {avg_grade}')
print(f'Standard Deviation of grade of Cornell: {std_grade}')
```

```
Average grade of Cornell: 3.65549798985052
Standard Deviation of grade of Cornell: 0.3376116011279313
```

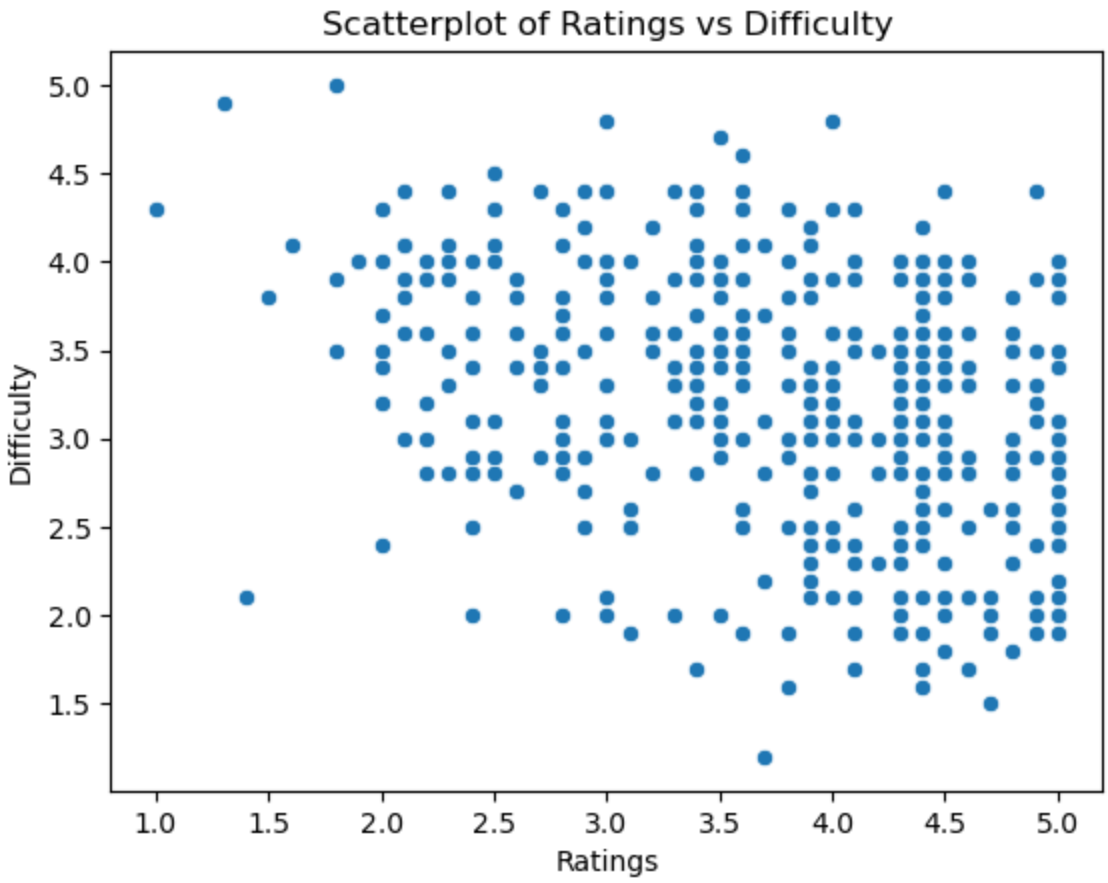
C) VISUAL DATA REPRESENTATIONS

In [7]:
```
#Pie chart of grades, ex. A+, A, A-, etc.

bins = [2.7, 3.0, 3.3, 3.7, 3.9, 4.3]

#Bin the grades using pandas.cut
#Round values to 2 decimal places
df['Average_grade_rounded'] = df['Average_grade'].round(2)
df['Grade_Ranges'] = pd.cut(df['Average_grade_rounded'], bins,\
```

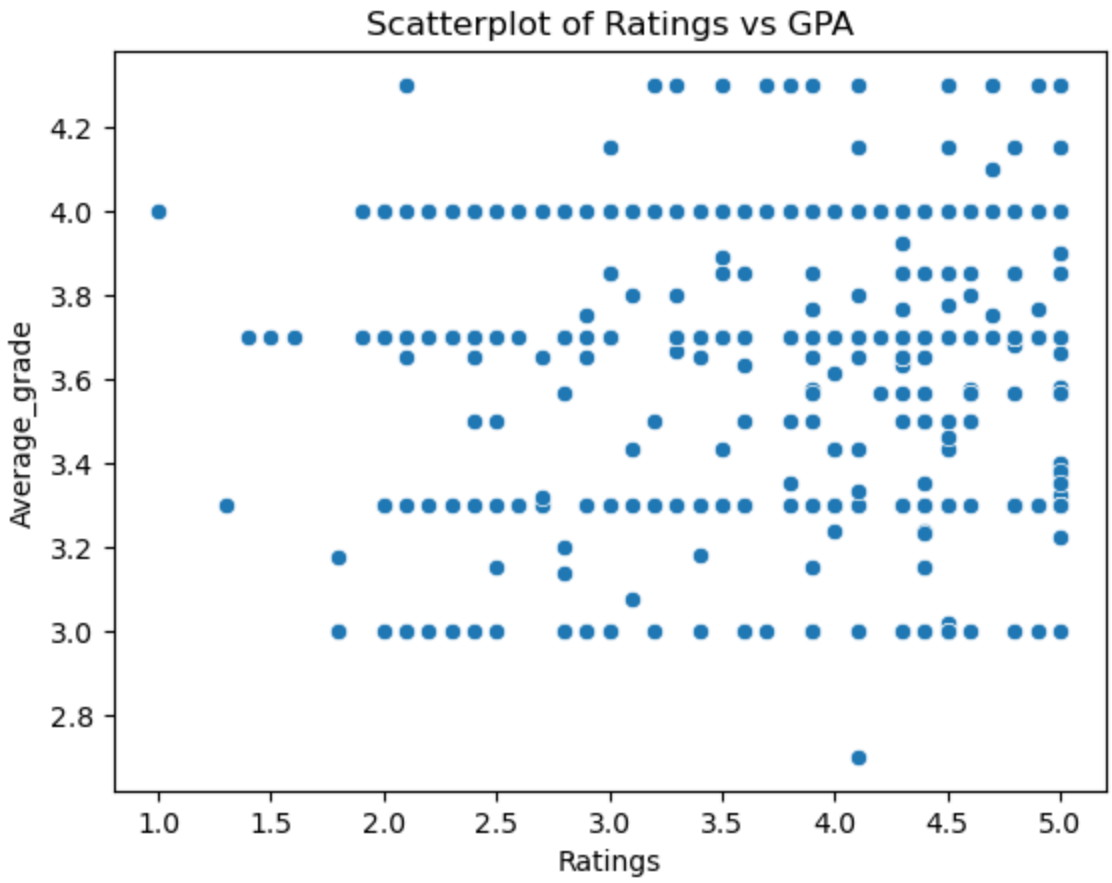
```
        right=False,
        labels=['2.7-3.0', \
                '3.0-3.3', '3.3-3.7', '3.7-3.9', '4.0-4.3'])

# Get the value counts of the binned grades
grades_count = df['Grade_Ranges'].value_counts().sort_index()

# Plot
fig, ax = plt.subplots()
grades_count.plot(kind='bar', title='Distribution of Grade Ranges', \
        ax=ax)

# Add a legend on the right side
ax.legend(title="Grade Ranges", loc="center left", \
        bbox_to_anchor=(1.3, 0.5))

ax.set_xlabel('Grade Ranges')
ax.set_ylabel('Number of Students')
ax.set_xticklabels(grades_count.index, rotation=45)

plt.show()
```



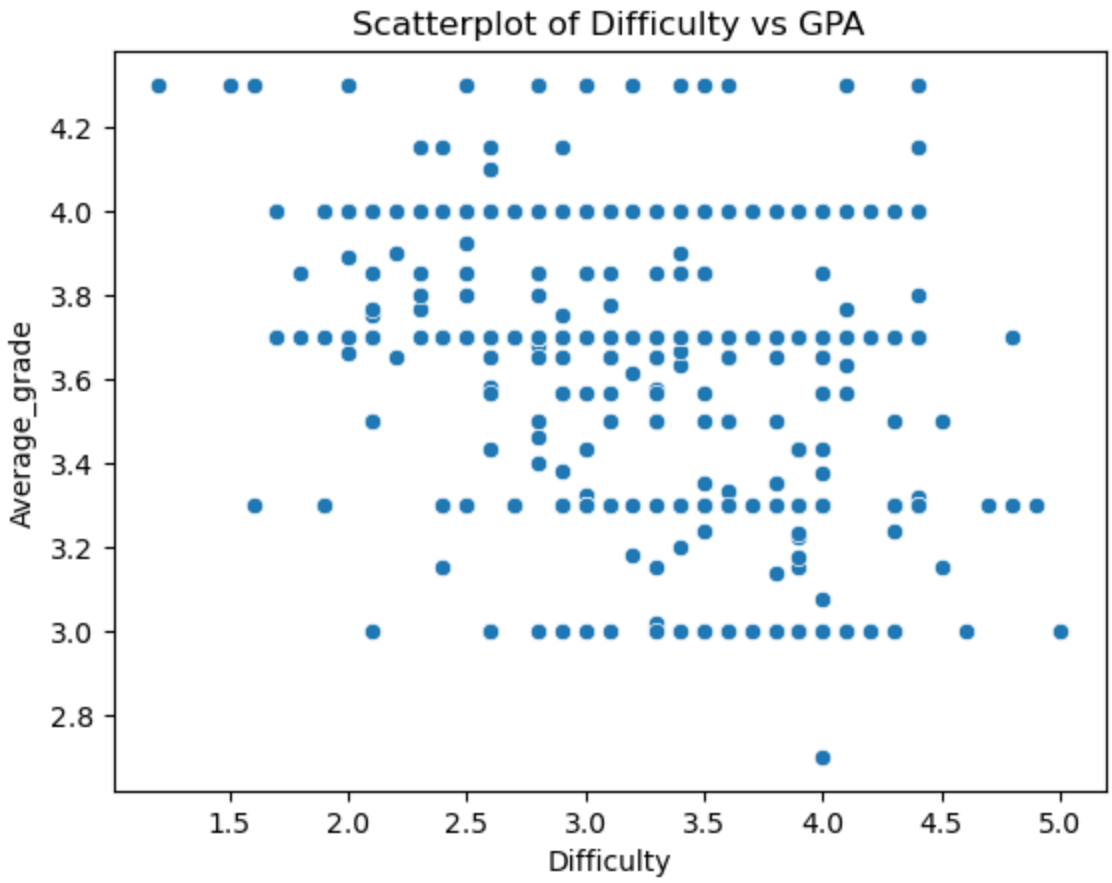
In [8]:
```
# 3. Scatterplot of ratings and difficulty
sns.scatterplot(x=df['Ratings'], y=df['Difficulty'])
plt.title('Scatterplot of Ratings vs Difficulty')
plt.show()
```



In [9]:
```
# 4. Scatterplot of rating and GPA
sns.scatterplot(x=df['Ratings'], y=df['Average_grade'])
plt.title('Scatterplot of Ratings vs GPA')
plt.show()
```



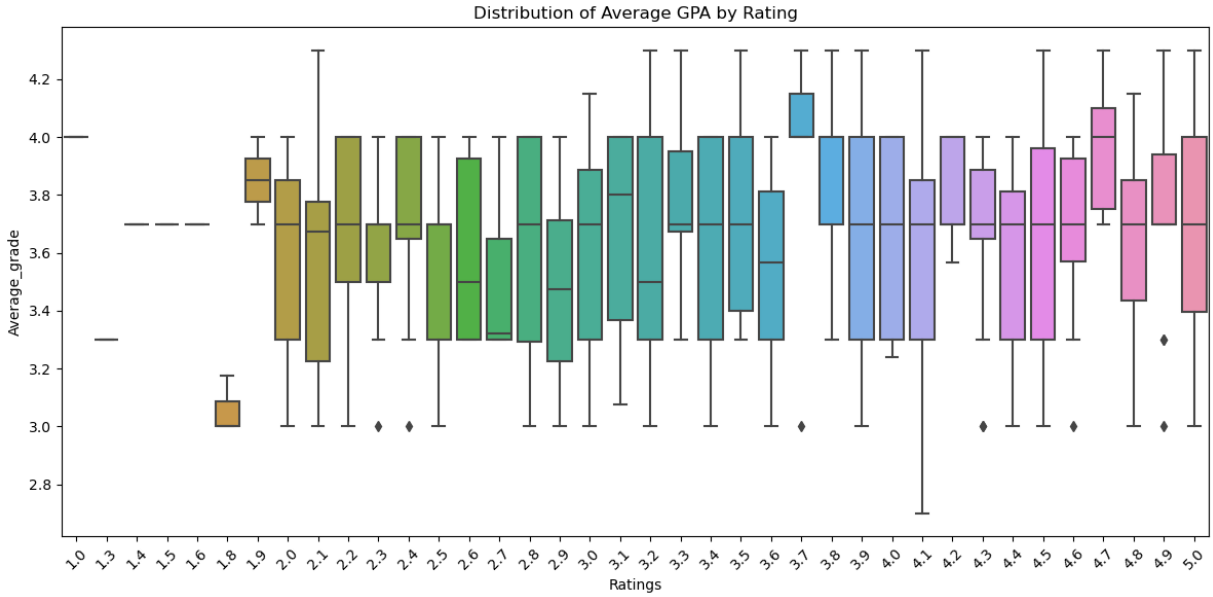
In [10]:
```
# 5. Scatterplot of difficulty and GPA
sns.scatterplot(x=df['Difficulty'], y=df['Average_grade'])
plt.title('Scatterplot of Difficulty vs GPA')
plt.show()
```



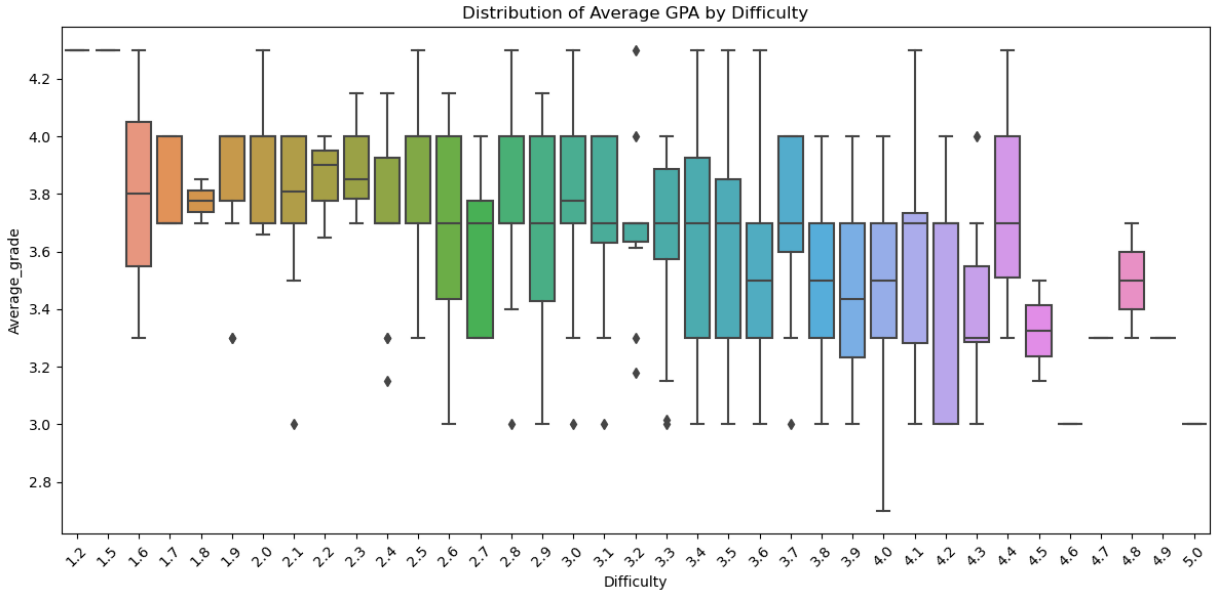
In [11]:
```
# 6. Covariance matrix of ratings, difficulty, and GPA
cov_matrix = df[['Ratings', 'Difficulty', 'Average_grade']].cov()
print(f'Covariance matrix:\n{cov_matrix}')
```

```
Covariance matrix:
                Ratings  Difficulty  Average_grade
Ratings        0.785927   -0.205476       0.026929
Difficulty    -0.205476    0.486677      -0.079335
Average_grade  0.026929   -0.079335       0.113982
```

In [12]:
```
# 7. Distribution of average GPA by rating

#Create the box plot
#Adjust the figure size for better visibility
plt.figure(figsize=(12, 6))
sns.boxplot(x=df['Ratings'], y=df['Average_grade'])

# Set the title and rotate the x-axis labels
plt.title('Distribution of Average GPA by Rating')
plt.xticks(rotation=45)  # Rotate the x-axis labels

plt.tight_layout()  # Ensure everything fits without overlap
plt.show()
```



In [13]:
```
# 8. Distribution of average GPA by difficulty

# Create the box plot
# Adjust the figure size for better visibility
plt.figure(figsize=(12, 6))
sns.boxplot(x=df['Difficulty'], y=df['Average_grade'])

# Set the title and rotate the x-axis labels
plt.title('Distribution of Average GPA by Difficulty')
plt.xticks(rotation=45)  # Rotate the x-axis labels

plt.tight_layout()  # Ensure everything fits without overlap
plt.show()
```

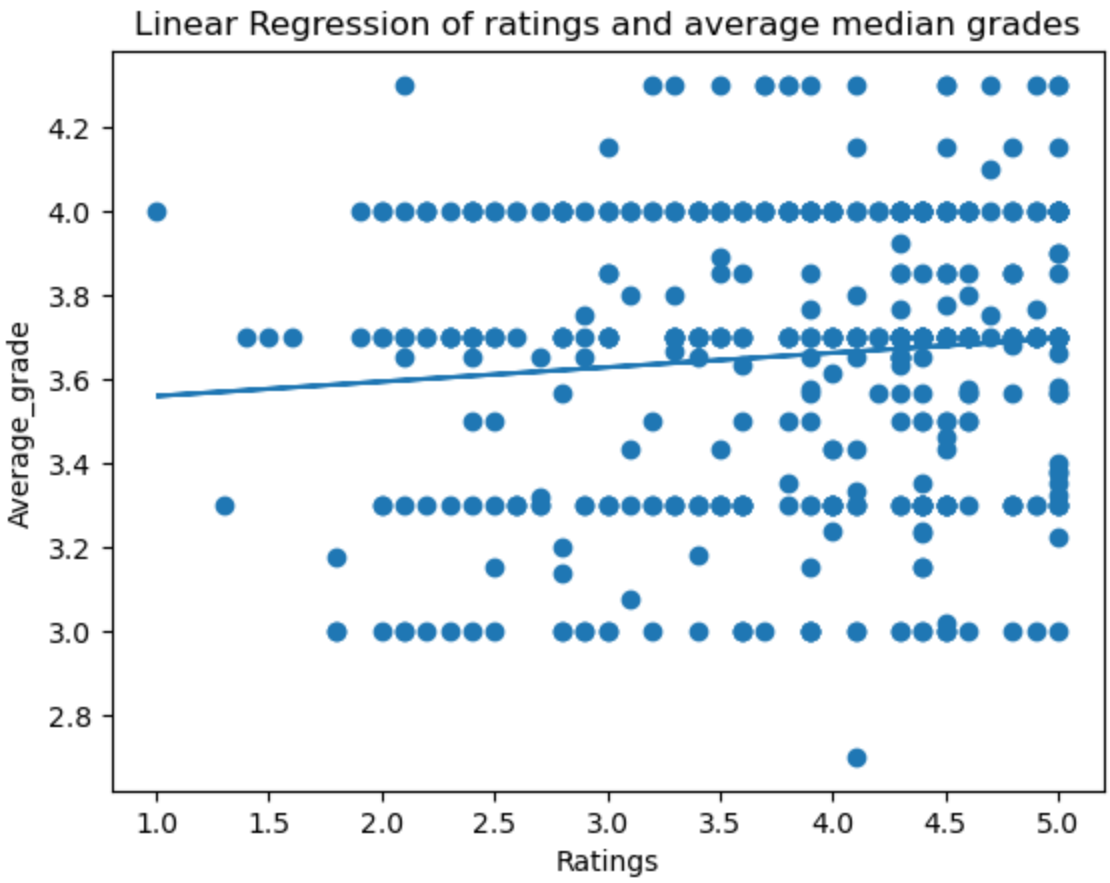


In [24]:
```
#9 Linear Regression of ratings and average median grades,
#using ratings to predict average grades
x = df['Ratings'].values.reshape(-1, 1)
y = df['Average_grade']
```

```
reg = LinearRegression().fit(x, y)
plt.xlabel("Ratings")
plt.ylabel("Average_grade")
plt.title('Linear Regression of ratings and \
    average median grades')
print(f'Regression Coefficients: {reg.coef_}')
print(f'Intercept: {reg.intercept_}')
plt.scatter(x, y)
plt.plot(x, reg.predict(x))
```

```
Regression Coefficients: [0.0342644]
Intercept: 3.524824759060496
```

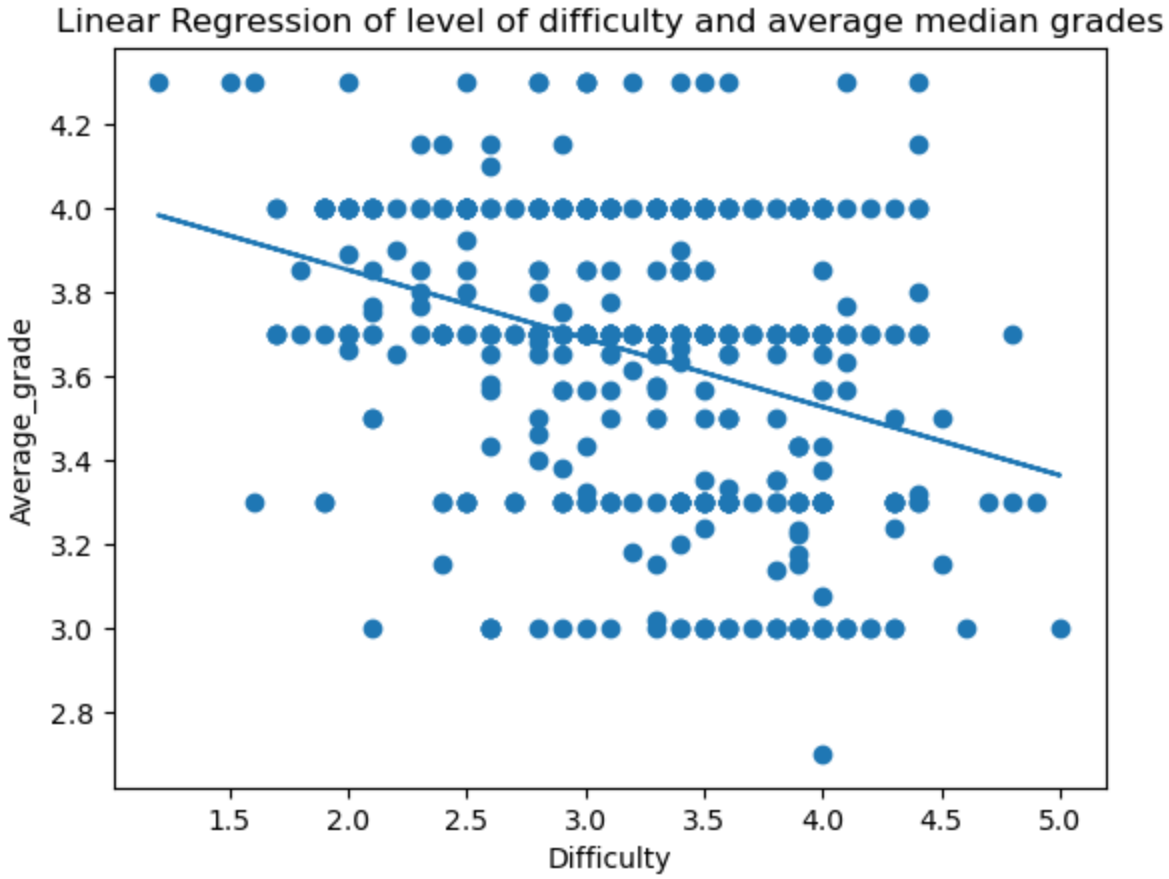
Out[24]: [<matplotlib.lines.Line2D at 0x293c80310>]



In [25]:
```
#10 Linear Regression of level of difficulty and average median grades,
#using difficulty to predict average grades
x = df['Difficulty'].values.reshape(-1, 1)
y = df['Average_grade']
plt.xlabel("Difficulty")
plt.ylabel("Average_grade")
plt.title('Linear Regression of level of \
    difficulty and average median grades')
reg = LinearRegression().fit(x, y)
print(f'Regression Coefficients: {reg.coef_}')
print(f'Intercept: {reg.intercept_}')
plt.scatter(x, y)
plt.plot(x, reg.predict(x))
```

```
Regression Coefficients: [-0.16301271]
Intercept: 4.178303028360556
```

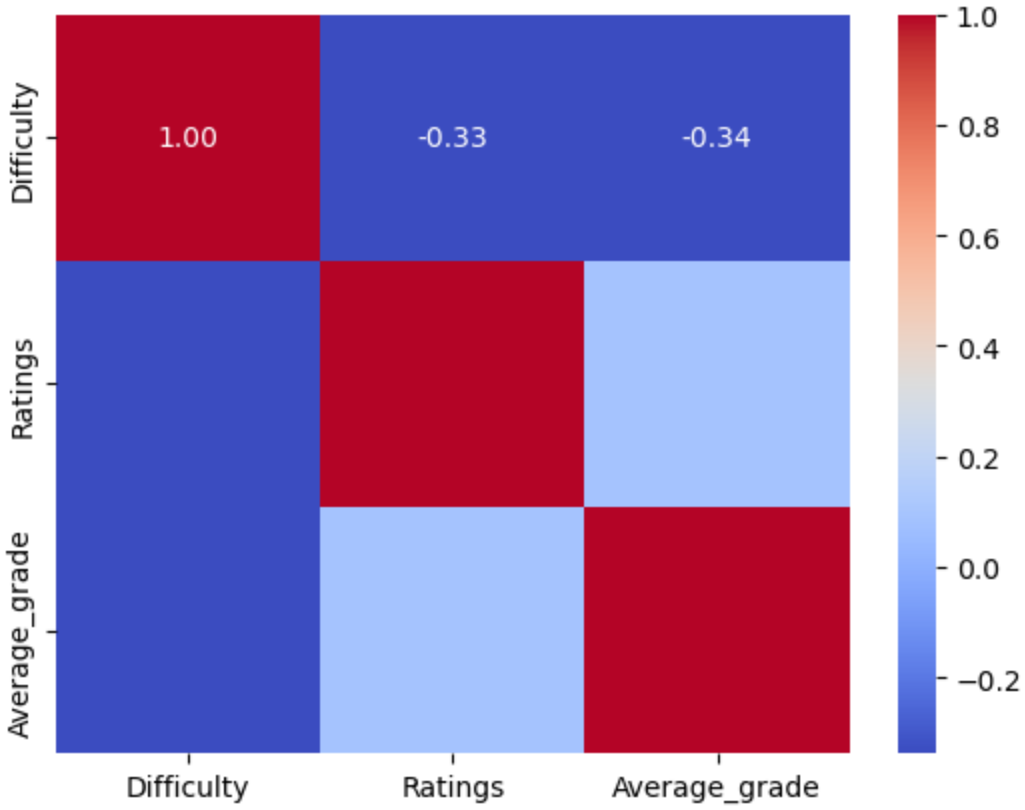
Out[25]: [<matplotlib.lines.Line2D at 0x29386d810>]



In [16]:
```
#11 Correlation matrix and heatmap
#correlation matrix and heatmap of \
    # "Difficulty", "Ratings", and "Average_grade"
corr_matrix = df[['Difficulty', 'Ratings', 'Average_grade']].corr()
print(corr_matrix)
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm')
```

```
              Difficulty   Ratings  Average_grade
Difficulty      1.000000 -0.332239      -0.336841
Ratings        -0.332239  1.000000       0.089974
Average_grade  -0.336841  0.089974       1.000000
```

Out[16]: <Axes: >



5. EVALUATION OF SIGNIFICANCE

Hypothesis 1: Students' median grade point averages are higher when they take classes taught by professors with higher overall ratings.

Analysis 1: We run a linear regression where we input grades (average median GPA) and output of professors' ratings (the average of all ratings). Because our measurement will not have signed variables, we will test whether βrating> 0

In [17]:
```
# creating regression model
X = sm.add_constant(df['Ratings'])
Y = df['Average_grade']
#training
model = sm.OLS(Y, X)
results = model.fit()
#sumamrize
summary = results.summary()
print(summary)
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:          Average_grade   R-squared:                       0.0
08
Model:                            OLS   Adj. R-squared:                  0.0
06
Method:                 Least Squares   F-statistic:                     3.9
83
Date:                Mon, 04 Dec 2023   Prob (F-statistic):             0.04
65
Time:                        21:07:40   Log-Likelihood:                -160.
72
No. Observations:                 490   AIC:                             32
5.4
Df Residuals:                     488   BIC:                             33
3.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const          3.5248      0.067     52.437      0.000       3.393       3.6
57
Ratings        0.0343      0.017      1.996      0.047       0.001       0.0
68
==============================================================================
==
Omnibus:                       29.474   Durbin-Watson:                   0.5
14
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.5
78
Skew:                          -0.341   Prob(JB):                     9.24e-
05
Kurtosis:                       2.334   Cond. No.                        1
8.4
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

The OLS regression results indicate that the model has an R-squared of 0.008, with an adjusted R-squared of 0.006. The F-statistic for the model is 3.983 with a p-value of 0.047. There are 490 observations included in the analysis. The condition number is 18.4. The coefficient for the intercept (const) is 3.5248 with a standard error of 0.067 and is statistically significant with a p-value of 0.000. The coefficient for "Ratings" is 0.0343 with a standard error of 0.017, and its p-value is 0.047, which is on the threshold of statistical significance at the 5% level.

In [27]:
```
#making regression plot
sns.regplot(data = df, x = df['Ratings'], y = df['Average_grade'])\
    .set_title("Regression plot of Ratings vs. Average_grade")
```

Out[27]: Text(0.5, 1.0, 'Regression plot of Ratings vs. Average_grade')

Hypothesis 2: Professors' overall rating and difficulty can significantly predict the median grade in their class. Higher professor ratings and lower perceived difficulty levels are associated with higher median grades.

Analysis 2: We run a multi-variable linear regression where we input the professors' overall ratings and difficulty level as independent variables, and the median grade as a dependent variable.

In [19]:
```
#input variables
X = df[['Ratings', 'Difficulty']]
X = sm.add_constant(X)
Y = df['Average_grade']
#training
model = sm.OLS(Y, X)
results = model.fit()
#Get results
summary = results.summary()
print(summary)
```

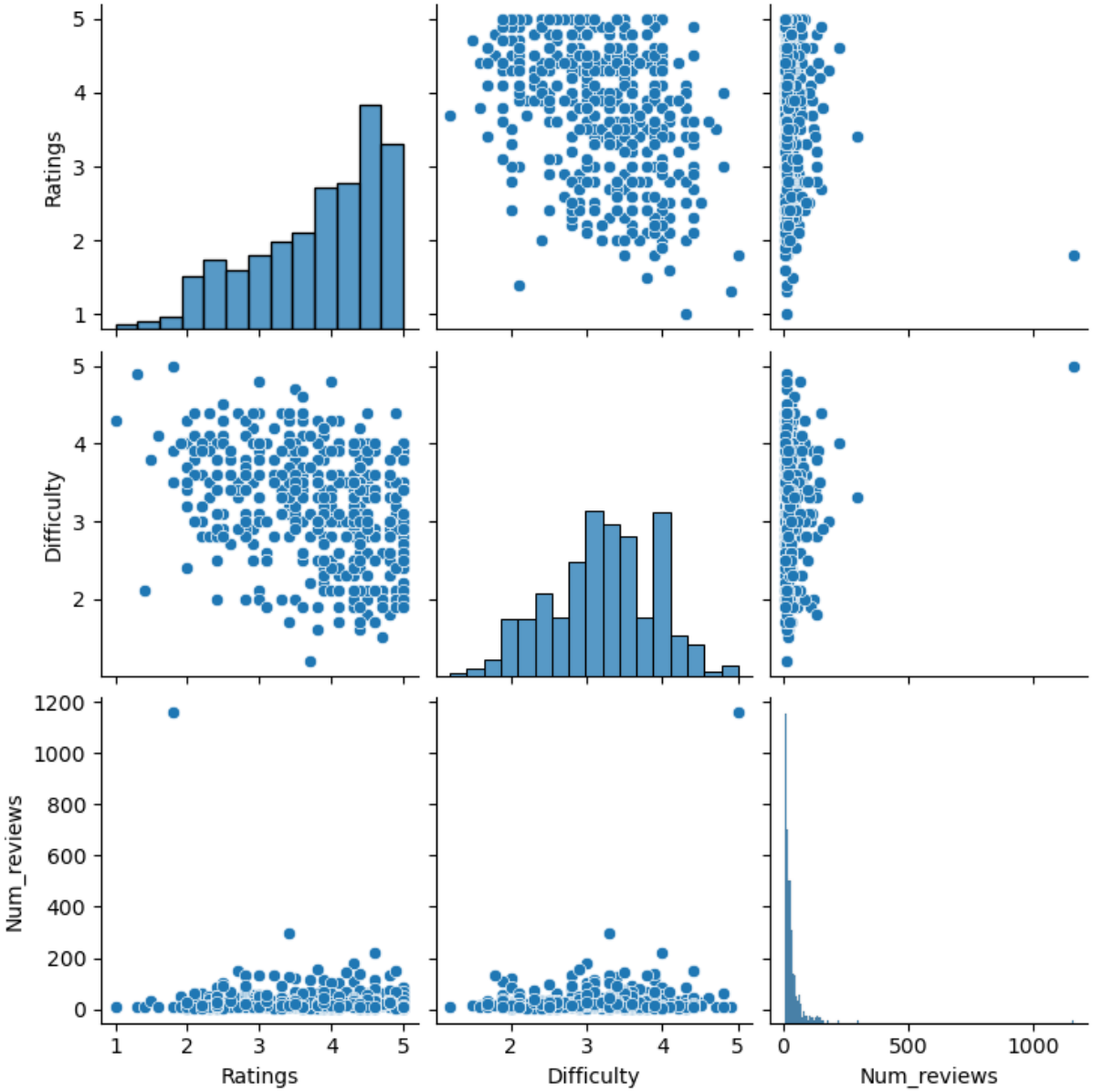
```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:          Average_grade   R-squared:                       0.1
14
Model:                            OLS   Adj. R-squared:                  0.1
10
Method:                 Least Squares   F-statistic:                     31.
33
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           1.58e-
13
Time:                        21:07:40   Log-Likelihood:                -133.
05
No. Observations:                 490   AIC:                             27
2.1
Df Residuals:                     487   BIC:                             28
4.7
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const          4.2268      0.112     37.791      0.000       4.007       4.4
47
Ratings       -0.0094      0.017     -0.545      0.586      -0.043       0.0
24
Difficulty    -0.1670      0.022     -7.630      0.000      -0.210      -0.1
24
==============================================================================
==
Omnibus:                       10.227   Durbin-Watson:                   0.6
76
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                9.0
41
Skew:                          -0.268   Prob(JB):                       0.01
09
Kurtosis:                       2.607   Cond. No.                         4
0.7
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

The OLS regression results show an R-squared of 0.114 and an adjusted R-squared of 0.110. The model's F-statistic is 31.33, with a p-value effectively at zero (1.58e-13), indicating that the model is statistically significant. The number of observations in the model is 490. The condition number is 40.7. The regression includes two independent variables: "Ratings" and "Difficulty". The coefficient for "Ratings" is -0.0094, with a standard error of 0.017, t-statistic of -0.545, and a p-value of 0.586, suggesting it is not

statistically significant. The coefficient for "Difficulty" is -0.1670, with a standard error of 0.022, t-statistic of -7.630, and a p-value of 0.000, indicating statistical significance.

In [20]:
```
#making pairplot
sns.pairplot(df[['Ratings', 'Difficulty', 'Num_reviews']])
```

Out[20]:
```
<seaborn.axisgrid.PairGrid at 0x2940b2b50>
```



Hypothesis 3: There is an inverse relationship between the number of reviews a professor receives and their rating. It is hypothesized that professors with higher ratings will receive a lower number of reviews, suggesting that students are less inclined to leave feedback when their experience is overwhelmingly positive.

Analysis 3: To test this hypothesis, a regression analysis will be conducted where the dependent variable is the number of reviews each professor receives, as a proxy for student engagement or the need to provide feedback. The independent variable in this analysis is the professor's overall rating. The aim of this analysis is to explore if there's a

correlation between the professor's rating and the number of reviews they receive, as indicated by the β coefficient in the regression model. A negative β coefficient would imply that higher-rated professors tend to receive fewer reviews, potentially indicating a general satisfaction with their teaching that diminishes the perceived need for feedback.

In [21]:
```
#input variables
X = sm.add_constant(df['Ratings'])
Y = df['Num_reviews']
#training
model = sm.OLS(Y, X)
results = model.fit()
#summarize
summary = results.summary()
print(summary)
```

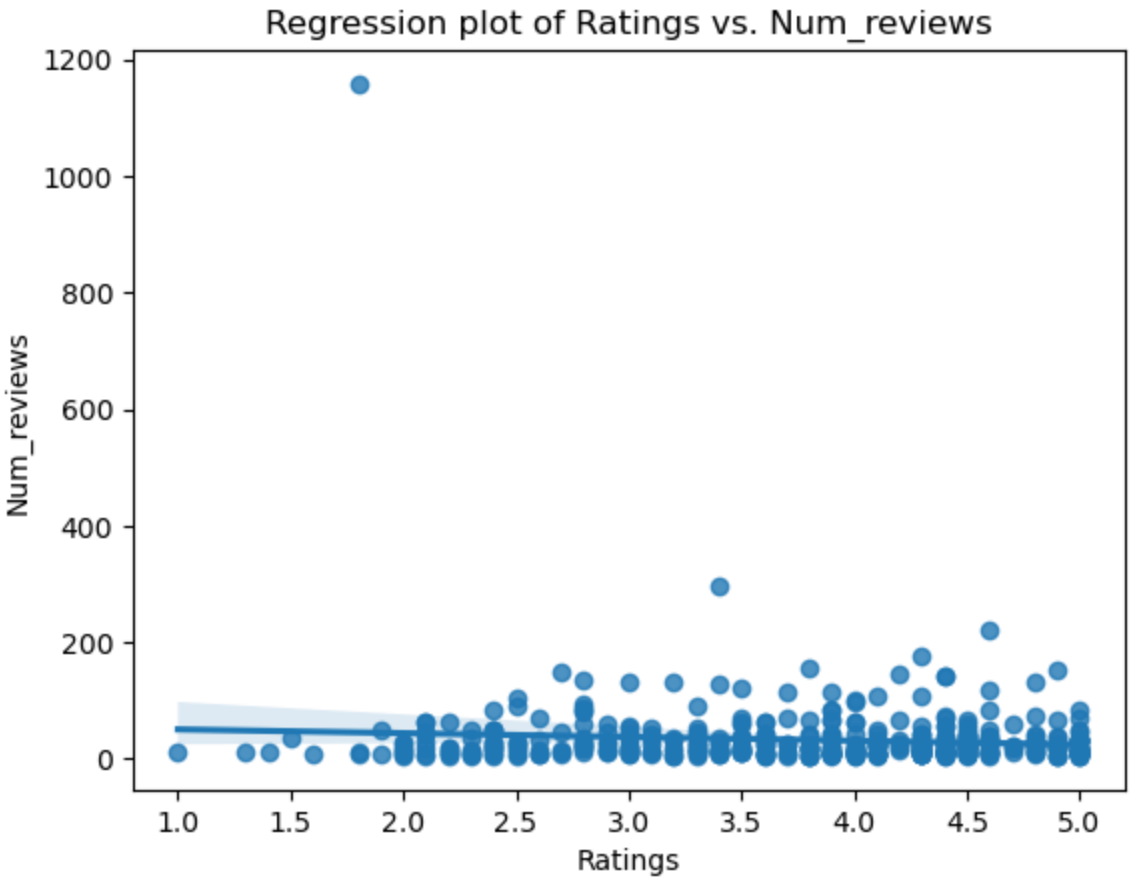
```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            Num_reviews   R-squared:                       0.0
09
Model:                            OLS   Adj. R-squared:                  0.0
07
Method:                 Least Squares   F-statistic:                     4.5
07
Date:                Mon, 04 Dec 2023   Prob (F-statistic):             0.03
43
Time:                        21:07:41   Log-Likelihood:                -269
9.2
No. Observations:                 490   AIC:                             540
2.
Df Residuals:                     488   BIC:                             541
1.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const         56.9033     11.950      4.762      0.000      33.423      80.3
84
Ratings       -6.4802      3.052     -2.123      0.034     -12.478      -0.4
83
==============================================================================
==
Omnibus:                      971.163   Durbin-Watson:                   1.7
70
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1235392.0
70
Skew:                          13.652   Prob(JB):                         0.
00
Kurtosis:                     247.466   Cond. No.                         1
8.4
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

The OLS regression output indicates that the model has an R-squared of 0.009 and an adjusted R-squared of 0.007. The dependent variable for the model is "Num_reviews". The F-statistic is 4.507 with an associated p-value of 0.0343, suggesting the model fit is statistically significant. The total number of observations in the model is 490, with 488 degrees of freedom for the residuals and one degree of freedom for the model. The condition number is 18.4. The regression coefficients include an intercept ("const") of 56.9033 with a standard error of 11.950, and a coefficient for "Ratings" of -6.4802 with

a standard error of 3.052. The "Ratings" variable also appears to be statistically significant with a p-value of 0.034.

In [28]:
```
#making regression plot
sns.regplot(data = df, x = df['Ratings'], y = df['Num_reviews'])\
    .set_title("Regression plot of Ratings vs. Num_reviews")
```

Out[28]: Text(0.5, 1.0, 'Regression plot of Ratings vs. Num_reviews')



6. INTERPRETATION AND CONCLUSION

Hypothesis 1:

Null Hypothesis (H0): There is no relationship between the professor's rating and the average grade of their class. Alternative Hypothesis (H1): There is a relationship between the professor's rating and the average grade of their class. Given that the p-value (0.0343) is less than 0.05, you would reject the null hypothesis in favor of the alternative hypothesis, concluding that there is a statistically significant relationship between the professor's rating and the average grade of their class, albeit a small one.

The coefficient of 0.0343 for the professor's ratings indicates a modest impact on the average grade. Since the ratings can range from 0 to 5, the full range of potential ratings could increase the average grade by a maximum of 0.0343 × 5= 0.1715, assuming a linear relationship holds across the entire range of ratings. This is a relatively small

increase, considering the full scale of typical grading systems (which often go from 0 to 100). The p-value is used to determine the statistical significance of each coefficient. It's calculated based on the t-statistic derived from the estimated coefficient divided by its standard error. The p-value (0.047), in this case, slightly below the 0.05 threshold for significance, suggests caution in interpreting the impact of professor ratings. Given the proximity to the cutoff and the low R-squared value (0.008), while there is evidence to suggest that ratings are associated with average grades, the strength of this association is weak. The low R-squared value (0.008) indicates that the effect size of professor ratings on average grades is small. This implies that other factors not included in the model are responsible for most of the variation in average grades. In an educational context, this could mean that while professor ratings have a detectable relationship with average grades, they are not a strong predictor on their own. Student abilities, teaching methods, course content, and assessment styles are likely also influential. The F-statistic in our regression analysis is derived by comparing the variance explained by the model (SSR) to the variance unexplained (SSE). It's calculated as the ratio of the mean regression sum of squares (MSR) to the mean error sum of squares (MSE), where MSR is SSR divided by the model's degrees of freedom and MSE is SSE divided by the error's degrees of freedom. In our study, a high F-statistic would indicate that variables like professor ratings significantly influence class GPA, rejecting the null hypothesis of no effect. The F-statistic for this hypothesis was 3.983, indicating that the model is statistically significant, suggesting that professor ratings significantly impact average class grades. The condition number in our regression analysis, which is a measure of multicollinearity or the degree of intercorrelation among the independent variables, is derived from the eigenvalues of the matrix X′X, where X is the matrix of our independent variables. It's calculated as the square root of the ratio of the largest eigenvalue to the smallest. In the context of our study, a high condition number would suggest a strong multicollinearity issue. Our condition number was 18.4, which was lower than 30. A condition number above 30 often indicates multicollinearity issues, which may affect the stability and reliability of the coefficient estimates; this means that we have no multicollinearity issues.

Hypothesis 2:

Null Hypothesis for Ratings ($H0_1$): The professor's rating has no effect on the average grade of their class. Alternative Hypothesis for Ratings ($H1_1$): The professor's rating has an effect on the average grade of their class. Null Hypothesis for Difficulty ($H0_2$): The perceived difficulty level of a class has no effect on the average class grade. Alternative Hypothesis for Difficulty ($H1_2$): The perceived difficulty level of a class has an effect on the average grade of the class. Given the p-value of 0.586 for Ratings, we would fail to reject the null hypothesis $H0_1$, indicating that there is no statistically significant evidence that professor ratings affect the average grade. Conversely, with a p-value of 0.000 for Difficulty, we would reject the null hypothesis $H0_2$, concluding that there is statistically

significant evidence that the perceived difficulty of a class negatively affects the average grade.

The coefficient for the professor's ratings is -0.0094, which is not statistically significant (p-value = 0.586). This suggests that this dataset has no clear relationship between professor ratings and average class grades. The coefficient for difficulty is -0.1670 and is statistically significant (p-value = 0.000). This implies that classes perceived to be more difficult are associated with lower average grades. Specifically, for each one-point increase in difficulty (on an unspecified scale), the average grade is expected to decrease by 0.1670 points. The R-squared value is 0.114, indicating that the model explains approximately 11.4% of the variability in the average grades, which includes both professor ratings and class difficulty. This is a modest amount, suggesting that while the model captures some of the factors influencing average grades, other variables are not included that also affect this outcome. The F-statistic is 31.33 with a very low p-value, which shows that the model as a whole is statistically significant. Despite the high p-value for professor ratings (0.586), indicating their individual insignificance in affecting average class grades, the overall model demonstrates significant predictive power, as evidenced by a high F-statistic of 31.33. This suggests that while professor ratings alone do not significantly impact average grades, their combined effect with other factors like perceived difficulty is substantial. Essentially, the ratings, when considered in isolation, do not show a strong correlation with grade averages. However, the significance of the overall model points to the importance of a multifactorial approach in understanding class grades. It underscores that factors beyond professor ratings, possibly including the perceived difficulty, play a crucial role in this dynamic. Therefore, although professor ratings don't emerge as a standalone predictor of grades, their contribution, in conjunction with other variables in our model, cannot be discounted in explaining the variability in class grades. The condition number we received was a value of 40.7. Since this is a value greater than 30, it suggests that there were multicollinearity issues. One possibility for the higher value of the condition number is the inherent relationship between the independent variables in our model, specifically professor ratings and perceived difficulty. For instance, professors with higher ratings might generally be perceived as less difficult, or vice versa. This interdependence could inflate the condition number, indicating that these variables are not as distinct as we would ideally want for a clear-cut analysis. Moreover, multicollinearity does not invalidate our model but warrants a cautious interpretation of the coefficients. It suggests that while we can conclude that these variables have a combined effect on class grades, pinpointing the exact contribution of each variable becomes more complex.

Hypothesis 3:

Null Hypothesis (H0): The professor's rating does not have an effect on the number of reviews they receive. Alternative Hypothesis (H1): The professor's rating does have an

effect on the number of reviews they receive. Given the p-value of 0.034, which is less than the common alpha level of 0.05, we would reject the null hypothesis in favor of the alternative hypothesis. This suggests there is a statistically significant effect of the professor's rating on the number of reviews they receive, with higher ratings associated with fewer reviews.

The coefficient for the professor's ratings is -6.4802, with a p-value of 0.034. This is statistically significant at the 5% level and supports Hypothesis 3, suggesting an inverse relationship between professor ratings and the number of reviews. For every one-unit increase in ratings (on a 0-5 scale), the number of reviews decreases by approximately 6.48, on average. The R-squared value is 0.009, indicating that the ratings explain only 0.9% of the variability in the number of reviews. This suggests that while the relationship is statistically significant, the overall effect of ratings on the number of reviews is quite small. The model's F-statistic is 4.507 with a p-value of 0.0343, which shows that the model is statistically significant. This means that the relationship observed between ratings and the number of reviews is unlikely to be due to random chance. The analysis confirms the proposed inverse relationship: higher-rated professors tend to receive fewer reviews. This result aligns with the idea that students may be less inclined to leave feedback when they have a positive experience. However, the low R-squared value indicates that ratings are not a strong predictor of the number of reviews, and there are likely other factors influencing the decision to leave feedback that are not included in the model. Our condition number was 18.4 once again, which was lower than 30. A condition number above 30 often indicates multicollinearity issues, which may affect the stability and reliability of the coefficient estimates; this means that we have no multicollinearity issues.

7. LIMITATIONS

DATA LIMITATIONS

- The biases that can be found in the datasets of the RateMyProfessor ratings and the median grades of each class taught by a professor.
- RateMyProfessors is a website based on self-reporting data by students of Cornell which means that the ratings posted tend to be posted by students who are on the extreme ends of opinions on a professor. Students who really like a professor or really hate a professor are more likely to submit a rating and skew the data more towards extreme ends.
- The median grades of each class is similar in that it is also a self-reported data set (dataset created and updated by students) which means that students who did better on classes are more likely to submit their grades making the data more biased towards the upper end of grades in the class.

- There were more ratings on professors in the RateMyProfessor than there were for the dataset for the median grade which means not all of the professors on RateMyProfessor can be used as analysis for the comparison between professor rating and the grade distribution of their class.
- The dataset contains median grades and professor reviews from the COVID-19 pandemic years where opinions and grades of professors and classes are affected by quarantine and online learning. This can skew the results for certain classes for certain years and make certain comparisons of data points inaccurate.
- The median grades are in letters, so the grades are not as accurate as we desired.
- These datasets are constantly being updated daily. This means that our datasets will become more outdated by time and will need to be re fetched and recleaned for our data and analysis to be accurate.
- Certain professors and classes are more popular than others meaning that certain classes would have more data than others. This could lead to discrepancies in our analysis of data across different classes and professors.
- The datasets used are only for Cornell University classes and professors meaning that the results and conclusions from the analysis cannot be accurately translated to professors and classes of other universities.
- Our datasets depend heavily on RateMyProfessor and the self-reported Google Sheets for median grades of Cornell University classes. This means that if any classes or professors are missing from either of the dataset we are unable to include their data in our analysis.
- The data for both datasets are self-reported meaning that there can be false data points being reported and we cannot find out which ones are untrue and which ones are true.
- There is human error contained within the Google Sheets. This means that some of the data has missing column values which forces us to exclude the data point from the final data frame. Some of the professor names were spelt incorrectly and some of the professor names were listed in an incorrect and ambiguous format. One example was how “various professors” was the value of the name of the professor column. This means that we will inevitably lose data points in the process.

8. SOURCE CODE

GitHub link

9. ACKNOWLEDGEMENTS

SOURCES:

INSPIRATIONS:

- RateMyProfessor/Reddit/CUReview data correlation/accuracy to grade distribution (course reviews).
- https://www.cureviews.org/
- https://www.ratemyprofessors.com/school/298
- https://www.reddit.com/r/Cornell/

GRADE MEDIANS:

- https://docs.google.com/spreadsheets/d/1817WfShAi7RHWWJR2c95YUgX37TkjUsLQM

SCRAPING RMP:

- https://github.com/tisuela/ratemyprof-api
- https://pypi.org/project/RateMyProfessorAPI/

10. APPENDIX

In our GitHub Repository, please see...

- phase_2.ipynb - Data Cleaning section
- final_df.csv - final clean data
- classes.csv - raw dataset of median grades and professors