Homework 1

Collaboration Statement

The LaTeXtemplate I am using can be found on *overleaf.com*.

I also used the LaTeXintroduction website and chatted with Katherine Walton, Luis Gomez Flores and Jack Frumkes about the assignment.

I made extensive use of the Python and Pandas documentation while completing this assignment. Stack Overflow was also very helpful.

I've included a Jupyter notebook with my solution. You're welcome to run it in the same folder as *grades.csv*. It loads the file in a Pandas DataFrame and performs all the data manipulations for this assignment.

Task 1: Attribute Description

- Semester: currently *categorical*, represents the semester during which the student took this course. This probably should be *ordinal*, as there is a clear chronological order between semesters
- Student ID: *categorical*, identifies a student in the Emory system.
- Name: *categorical*, the first name of the student
- Section: *categorical*, the section of CS170 the student is in.
- Homework 1-5: *numerical ratio*, the grade for a given student for each homework. The maximum grade is 42 for each homework (40 + 2 bonus points for submitting early).
- Peer Evaluation: *numerical ratio*, the sum of the peer evaluation points earned by a student. The maximum score is 150
- Bonus: *numerical ratio*, most students do not have a value for this attribute. This represents the number of Bonus points a student earned.
- Quiz 1-12: *numerical ratio*, the grade for a given student for each quiz. The maximum grade for one quiz is 50.
- Quiz Adjustment: *numerical ratio*. This represents an adjustment to the quiz grade for exceptional purposes (I'm guessing a makeup or regrade)
- Drop Lowest Quiz 1-2: *numerical ratio*, opposites of the two lowest scores of the student, so that they cancel out those scores when added up.
- Final Exam: *numerical ratio*, final exam score, out of 150.

- Total Score: *numerical ratio*: sum of all previous numerical columns, represents the final score of a student out of 1000.
- Letter Grade: *ordinal*, equivalent letter grade according to the table in the syllabus.

Task 2: Missing Values

- **Homework 1-5**: It seems some people forgot to submit their homework... In most classes, the student would probably end up getting a zero for the given assignment (we fill the null value with a zero). If there are extraneous circumstances (illness, etc.) or we are very lenient, I suppose we could replace the null value with the average of the other homework assignments for that student. This however assumes that the student would perform as well on that missing homework as on the others, which is often not the case as assignments can vary greatly in breadth and depth throughout the semester. I'm leaning more towards filling with zeroes.
- **Peer evaluations**: The missing entries here could either be due to a student not completing ONE peer evaluation, or not completing ANY peer evaluations throughout the semester. If it's the latter, we can simply zero-fill the null values for the same reasons as **Homework 1-5**.

  After further looking at the data, it appears the vast majority of the people that had a null value in this field failed the course. Maybe these students withdrew from the course? In that case, it might be safer to delete their entries from the table, as these students are not representative of students that completed CS170. They should be analyzed separately (maybe this can allow us to discover some shortcomings in the course, or to find where students struggle the most).
- **Bonus**: 769 students do not have an entry for this field (this is much higher than for other attributes). This makes sense as the bonus is, by definition, optional. We can just fill null values with zeroes here. Other options do not really make sense.
- **Quiz 1-10**: The student most likely did not show to the quiz. We can fill this field with a zero. If the student made up the quiz, it will be reflected in the "Quiz Adjustment" Attribute.
- **Quiz Adjustment**: Most students don't need their quiz grade adjusted. The vast majority of students don't have an entry here. Filling in with zeroes makes most sense for this field.
- **Final Exam**: It appears some people did not take the final exam. It might be because they were taking the class pass-fail and could pass the class even if they got a zero on the final, or they dropped the course, or missed the exam. Filling in this field with zeroes makes most sense for the same reasons as explained previously.

Task 3: Re-encoding

- **Semester** can be separated in two attributes: *Year* and *Semester*, where

Semester
F18
S18
F17
F16

becomes

Year	Fall/Spring
18	F
18	S
17	F
16	F

  This way, we can easily compare all fall semesters against all spring semesters for CS170.

- **Section** is currently denoted by a number. Someone could assume that the section number is ordinal (1 is somehow ranked either higher or lower than 9. Furthermore, section 2 for the fall 2018 semester is not necessarily the same as section 2 for the spring 2016 semester. This can become important if we are trying to evaluate the performance of lab TAs for a semester. For this reason, I will concatenate the semester value with the section value

  For example:

  ```
  5
  ```

  Becomes:

  ```
  2018F-section5
  ```

Task 4: Scaling and z-scoring

1. Re-scaling is most appropriate to help readers of the data get an idea of what the grade represents. Most people are familiar with the 0-100 grading scale and can quickly estimate what letter grade a student got for an assignment.

2. Z-scoring using the mean of all semesters is only appropriate to do if the score distribution for each semester is homogeneous: if professors started grading harder in 2017, for example, the mean grade for each assignment would be skewed left. This would affect the z-scores of students who took the class before the grading change, and make them appear as better than their classmates from 2017 onwards.

3. This Z-Scoring method is most appropriate if there can be significant disparity in grading methods from one semester to the other. Overall, this should be the preferred method of calculating z-scores with this dataset.

Task 5: Summary Statistics

Summary statistics were computed for each attribute and included in *summary_statistics.csv*.

Task 7: Tools and Languages

Please find Task 6 at the bottom of the document.

- Python: General purpose programming language. I used this because I am familiar with it. It's easy and fast to write in. This is one of the most commonly used programming languages in data science.
- Pandas: A data manipulation library for Python. Pretty high learning curve compared to regular Python, but very powerful.
- matplotlib: A graphing library for python. Definitely more difficult to use than Excel graphing tools, but provides more rigid formatting.
- LaTeX: Typesetting language with which this document was made. High learning curve. Using a pre-made template made this tool easier to use.

Tools I considered using but did not use

- SQL: I am rather familiar with it, but did not have the time to set up a database with rules
- R: Similar to Python with pandas and other Data Science libraries. I'm more familiar with the Python syntax and did not have time to study up on R.

Part 8: LaTeX Document

NB: The code for this assignment is included in a separate Jupyter notebook for readability's sake.

Task 6: Charts

The following pages contain ten charts with analysis

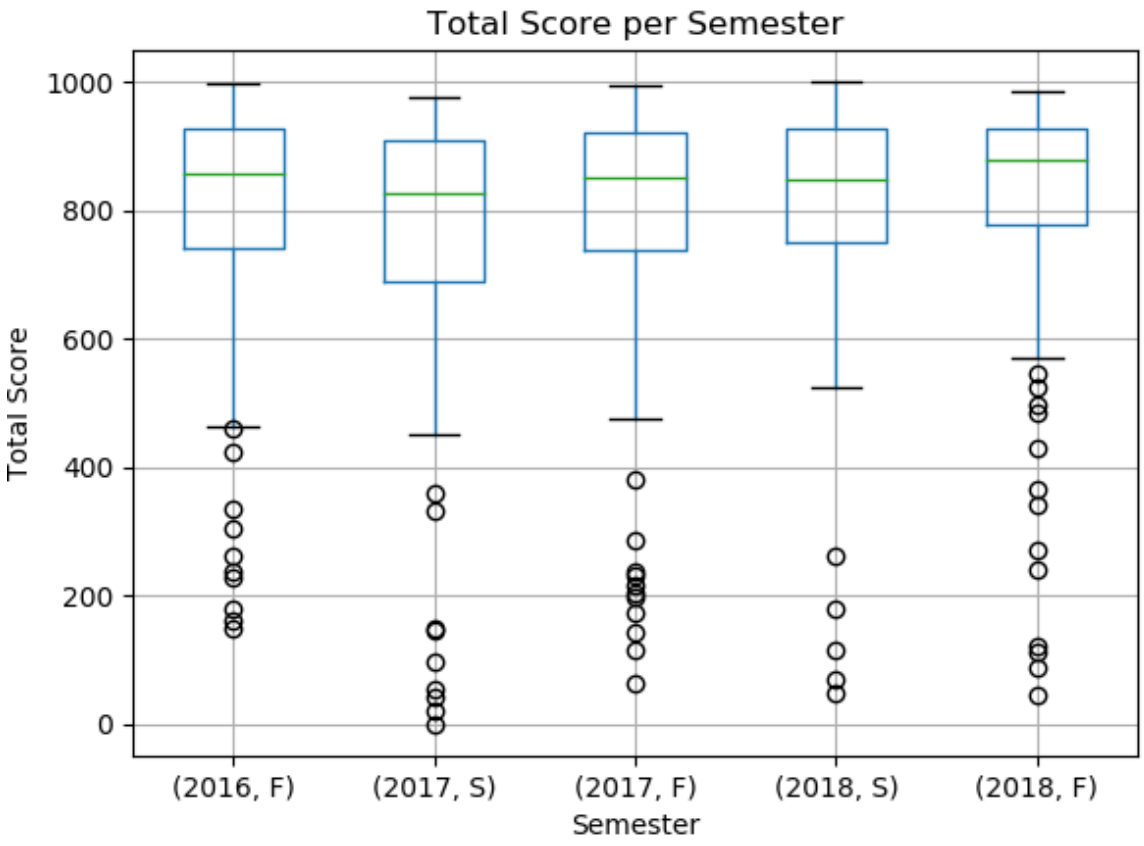

Figure 1: The Distribution of Total Scores for all sections for a given semester. Most semesters, the Total score distribution hovers around the low to mid 80s, which is expected from a college introduction course in computer science. The top 50% of students appears to have a lesser range of scores than the bottom 50%. Note the median total score appears to dip in the Spring of 2017, but then steadily seems to rise again. This could indicate some unforseen event that affected the students' ability to learn course material (Professor often missing that semester, Excessively difficult exams...) assuming student performance is relatively constant from one year to the next.

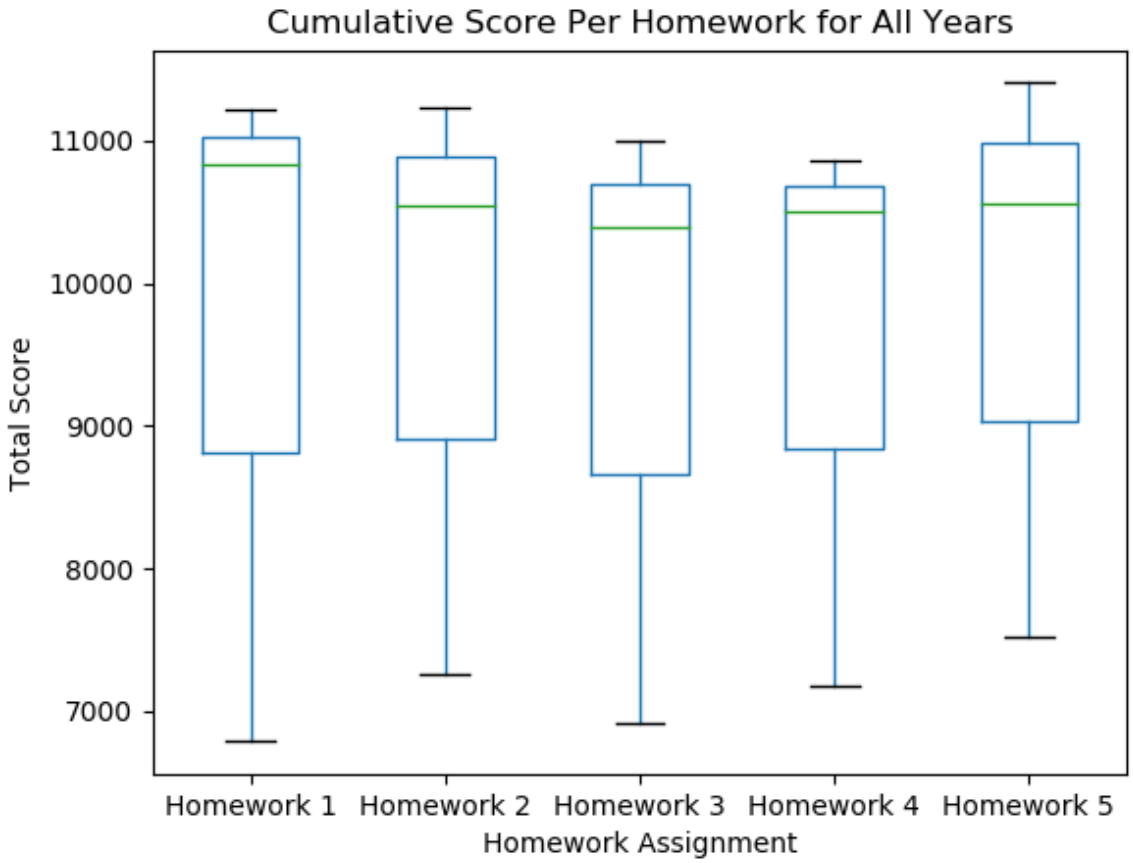

Figure 2: The sum of all student grades for a given homework assignment, per year. Note the grade range for the top 50% of students is extremely small compared to the bottom 50%. This suggests that there exists some significant disparity in student ability from year to year (This will be investigated further in another Figure). Median scores decrease until the middle of the semester (Homework 3), then increase again towards the end. This could suggest students are better assimilating the material as the class progresses. This could also indicate that students are unable to perform at their best during the middle of the semester. This would make sense considering this is when most students find themselves overwhelmed with the amount of homework assigned

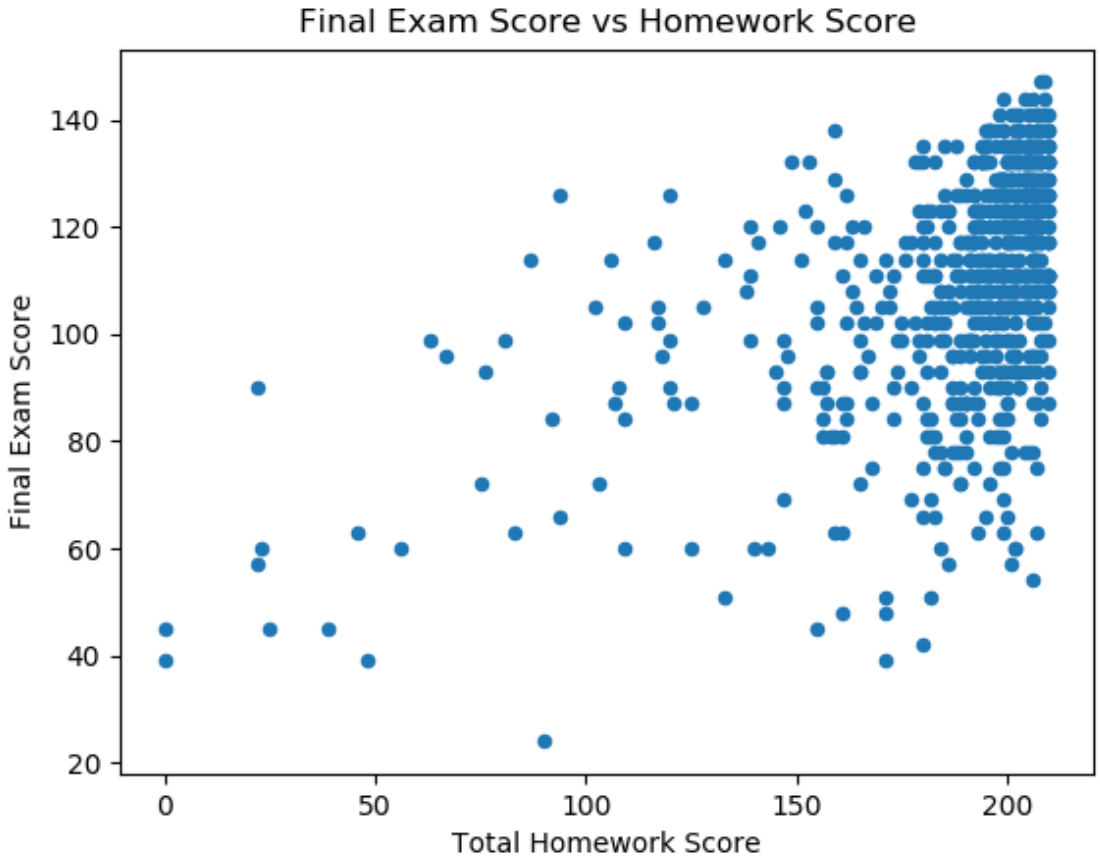

Figure 3: Final Exam Scores over Total Homework Scores per student that has taken CS170. The graph displays a moderately strong, positive correlation between Final Exam Score and Total Homework Score: Individuals that perform well on their homework assignments tend to do better on the Final Exam.

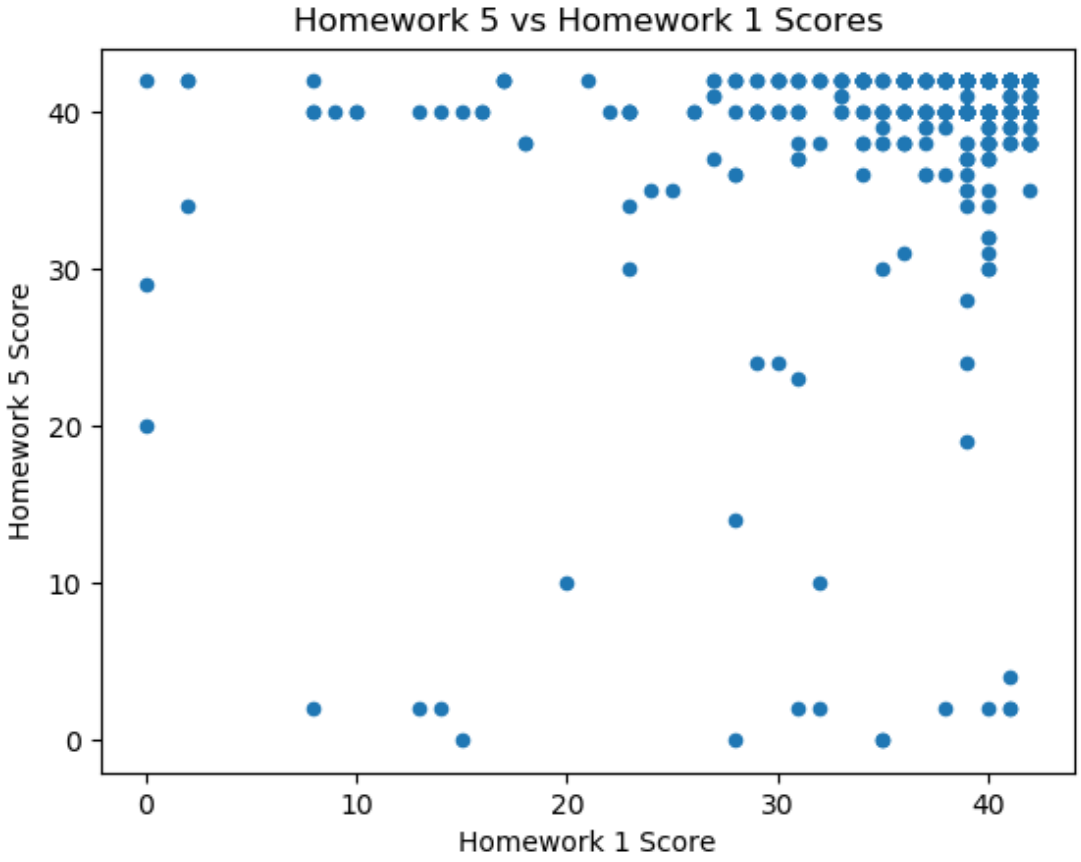

Figure 4: Comparison of Homework 1 and Homework 5 scores for every student. The objective of this graph is to observe any change in the performance of the student as the CS170 class progressed. It appears that students that performed well on the first Homework also performed well on the last Homework. A significant amount of students performed much better on their last homework than their first: this could be due to having better assimilated the class material after an entire semester.

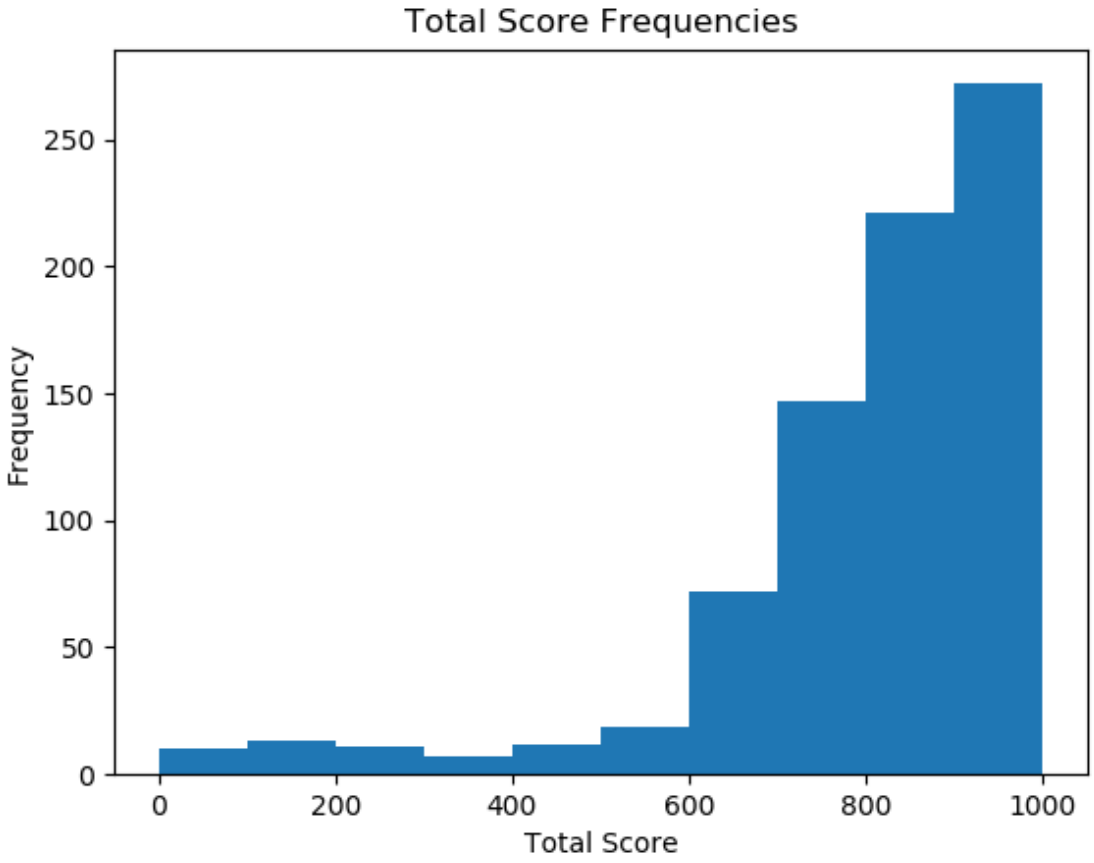

Figure 5: Frequencies of Total Scores for all years and sections. The distribution is highly skewed to the right, with most students scoring above 600 (passing the course). More students got A and A-s than any other letter grade in the course.

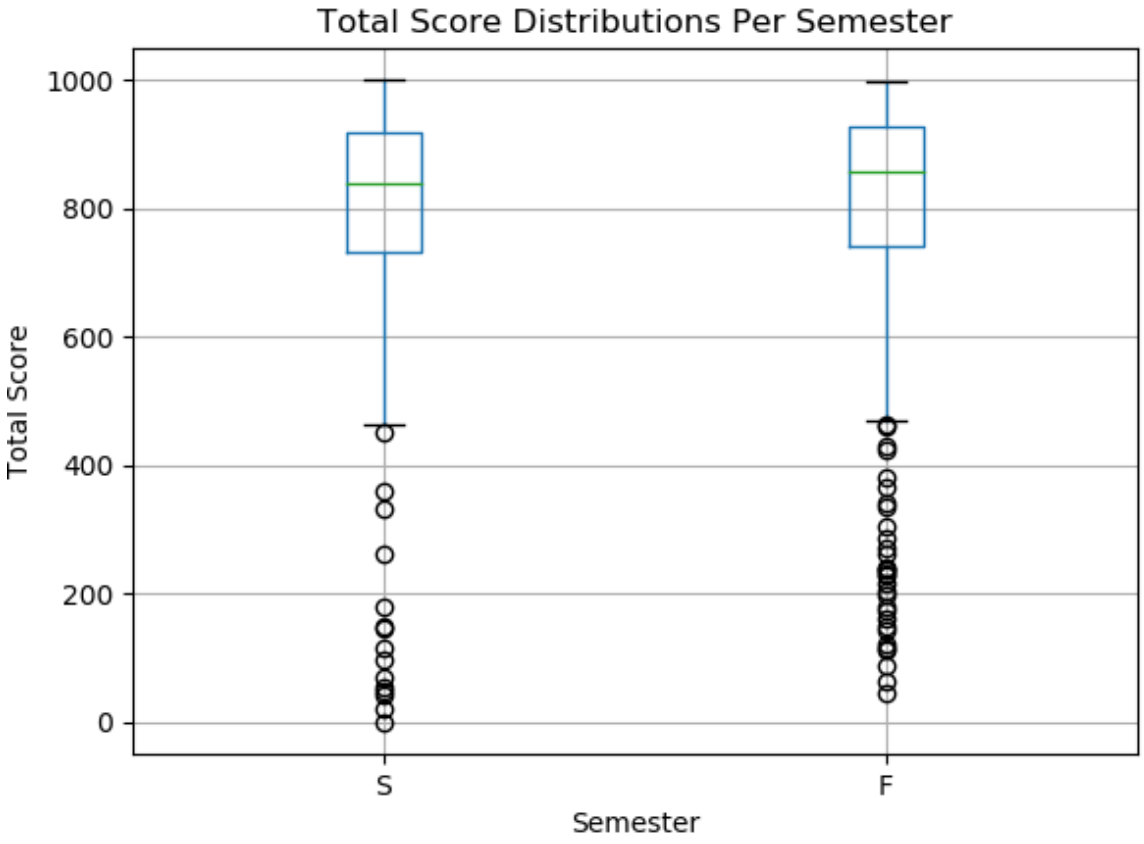

Figure 6: A comparison of Total Score Distribution for the Spring and the Fall Semesters. Both distributions display similarly shaped box plots. The median of the Spring distribution is slightly lower than that of the Fall distribution. This could be due to the fact that most Spring courses are somewhat faster paced than their Fall counterparts (professors dive faster into the material than in the Fall).

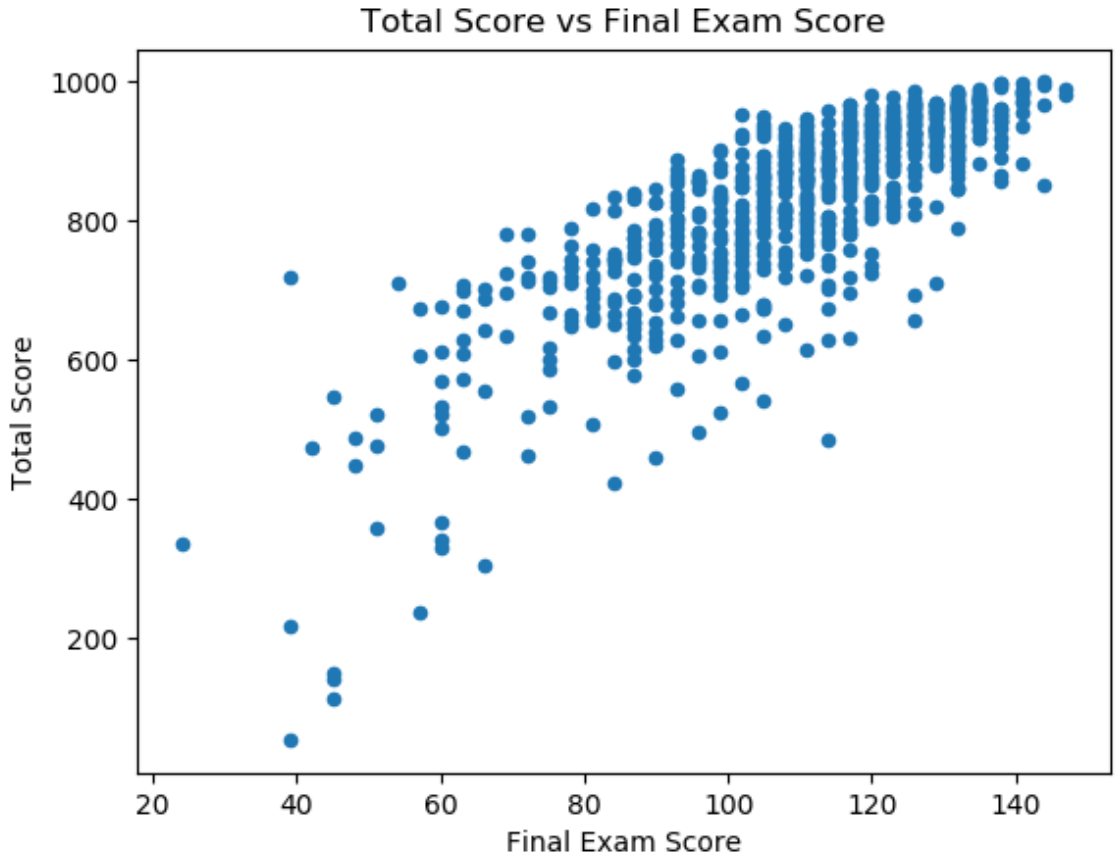

Figure 7: How Final Exam Score affect Total Score. Note that Final Exam Score only represents 15% of a student's total score for CS170. However, the correlation coefficient between the two is surprisingly high regardless (0.88), which suggests that Final Exam Score is a good predictor of a student's total score in the class.

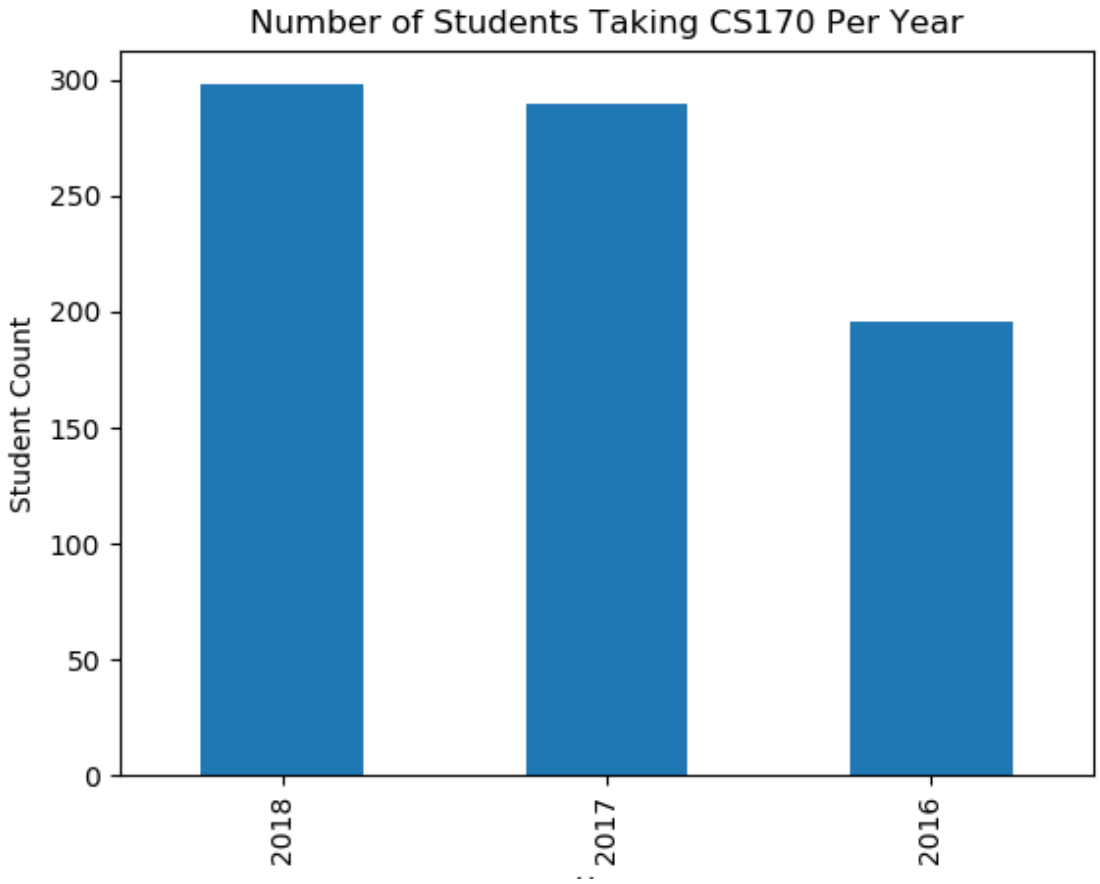

Figure 8: Student count per year for CS170. Note the number of students in CS170 has greatly increased over the years. This is most likely due to the increased popularity of Computer Science and its increased use in other fields (Mathematics, Physics, etc. but also Social Sciences). Around 300 people took CS170 in 2018.

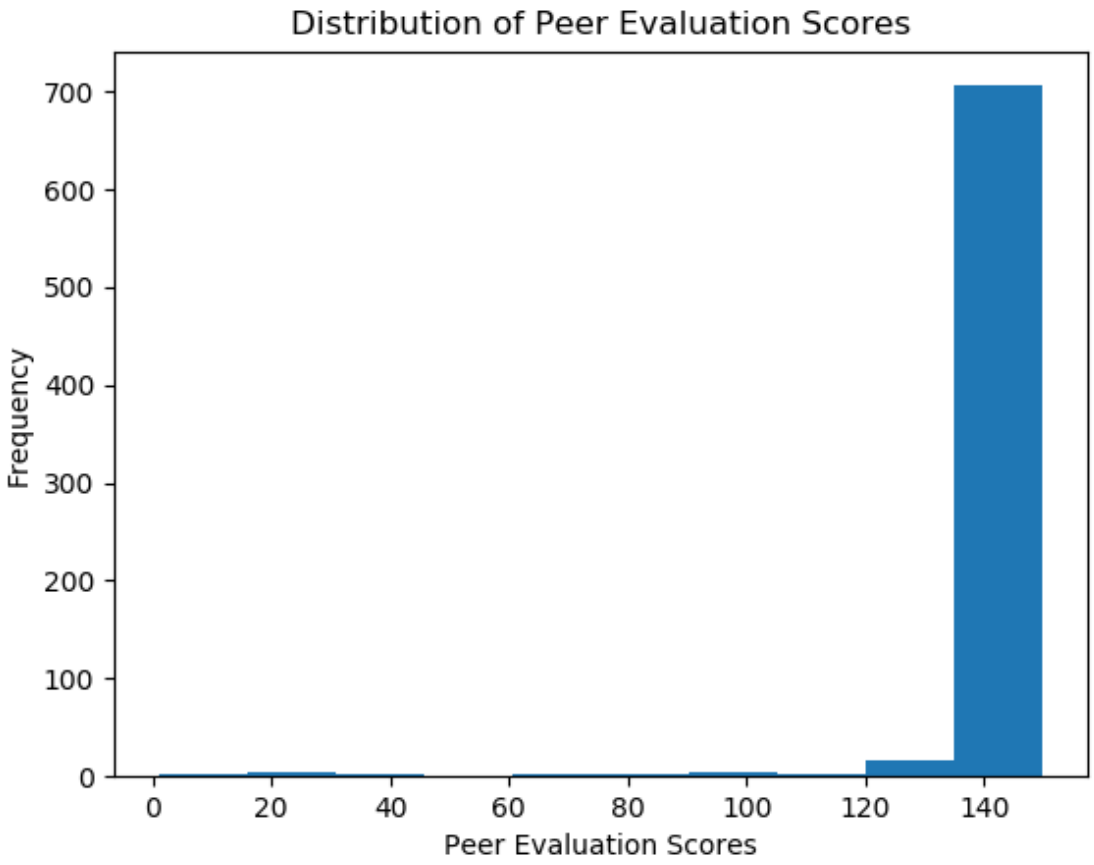

Figure 9: Peer Evaluation Scores for all years. The distribution is extremely skewed towards the right, suggesting the task is graded rather leniently

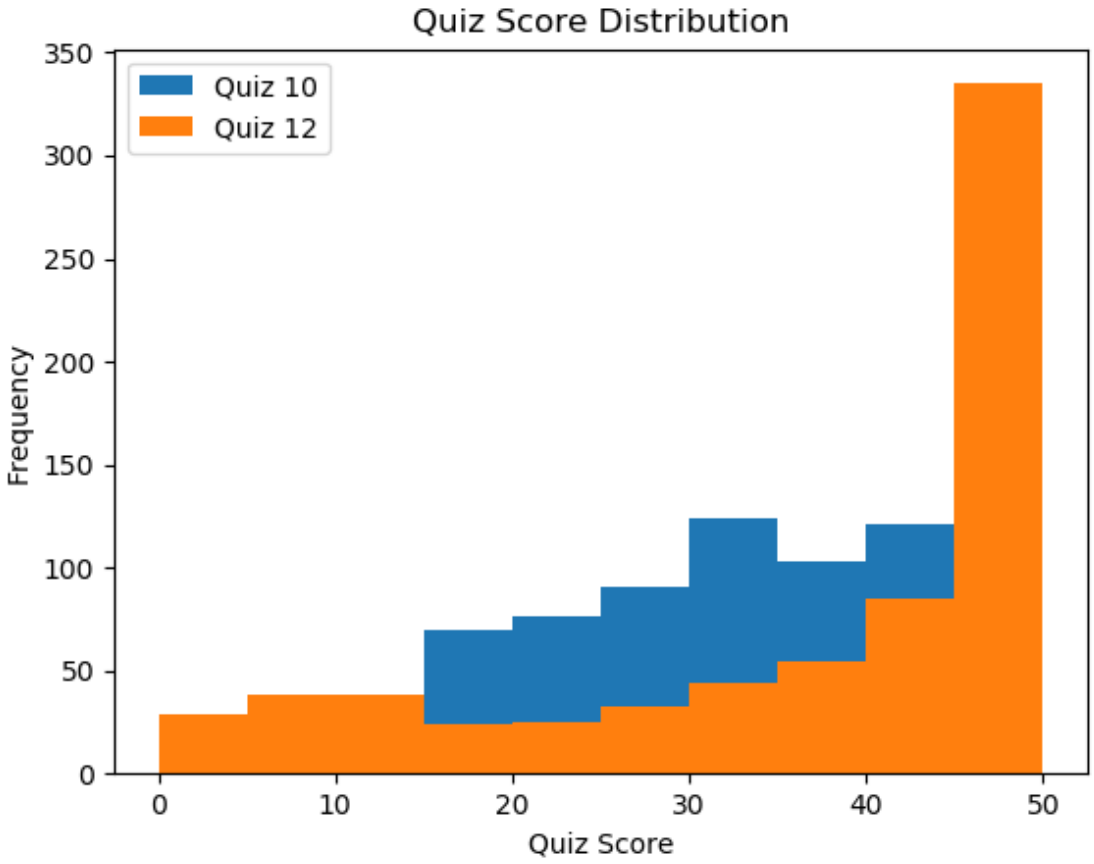

Figure 10: Comparison of Quiz 10 and Quiz 12 grades. Quiz 10, like many quiz distributions from the middle of the semester, has a very wide range of grades, with no grade being more common than another. Quiz 12, from the end of the semester, is extremely highly skewed towards the A range, which suggests this quiz was rather easy, maybe given as a grade booster to students before their final exam.