**COGS9: Introduction to Data Science**
*Final Project*
**Due date:** Wednesday 2019 December 11 23:59:59
**Grading:** 10% of overall course grade. 40 points total.
*Completed as a group. One submission per group on Gradescope.*

**Group Member Information:**

Please read COGS 9 team policies to best understand how to approach group work and to understand what the expectations are of you in COGS 9.

| First Name | Last Name | PID |
|---|---|---|
| **Thanh** | **Luong** | **A15954790** |
| **Abigail** | **Han** | **A15392642** |
| **Saroop** | **Samra** | **A15383517** |
| **Amy** | **Nguyen** | **A15583168** |

**Question**

Is there an association between UCSD's change in relative college rankings in league tables and positive or negative sentiment from associated news stories to the rate of change in number of applicants applying to UCSD each year?

**Hypothesis**

Our hypothesis is that there is a correlation with the sentiment in news stories relating to college league tables in the year leading up to college applications that can be used to model the number of applicants that UCSD should expect will apply. News stories related to UCSD's recent increase in college ranking may result in an increase in applications; conversely, the opposite result, that negative new stories (or lack of positive stories) when college rankings fell, resulted in a reduction of college applicants to UCSD in the subsequent application period.

The justification for our hypothesis is based on the fact that members of this team read articles and publishings with college rankings and reviews before solidifying their decision to apply to UCSD. Our instinct is that the decision to apply to UCSD during the college applications process is influenced by the number of positive stories about UCSD; this might be a significant reason why UCSD has a recent record number of applicants in 2019.

If we can establish a model then we can enable college administrators at UCSD to better manage the resources needed for applicants. For example, ensuring there is an adequate server and database resources to cope with the number of applicants in the college portal. Conversely, if our data shows that there is little

to no correlation, then it can highlight that administrators should focus less of the marketing campaign (and budget) on the recent improvements in UCSD's college rankings.

**Background Information**

As reported by the San Diego Union Tribune (2019)[1], UCSD received a record of 118,359 applications in fall 2019. In the previous year, UCSD was ranked 2nd out of 50 colleges in Money magazine's annual list of the 50 Best Colleges in the United States[2]. We want to explore if a relationship exists between news stories that are reported in the year leading up to college admissions, to the impact on the number of applications. As the college rankings change frequently, both up and down, we want to see if this results in positive, negative or neutral sentiments in news stories, and establish if this drives an increase or decrease in college applications to UCSD.

Prior related work in this area includes Monks and Ehrenberg (1999), "The impact of US News World Report College Rankings and their impact on admissions outcomes at private colleges."[3] Their conclusion was that there is a positive association with rankings and a yield rate (applicants who end up accepting offers), and a positive association with rankings and average GPA of those who are given offers. Although this is not the specific question we wish to address, it is related to the impact of academic league tables. Furthermore, the results in this study would not have been significantly impacted by the Internet due to the limited number of websites in 1999. Our study will have much more data from news stories and college rankings that are published on the Internet.

The study from Sauder and Lancaster (2006), "Do Rankings Matter? The Effects of U.S. News & World Report Rankings on the Admissions Process of Law Schools"[4], shows college rankings "have significant effects on both the decisions of prospective students and the decisions schools make in the admissions process." Although this study was related to a small law school, we will attempt to *replicate* the study with UCSD, a large public university.

More recent work from Alter and Reback (2014), "True for Your School? How Changing Reputations Alter Demand for Selective U.S. Colleges,"[5] highlights "colleges receive fewer applications when peer universities earn high academic ratings." Their findings are based on two widely read authorities on college admissions, the Princeton Review and U.S. News and World Report's annual college rankings. Their research also indicates that the quality-of-life reputation scores by current students has an impact on the size of the applicant pool. In our study, our focus is narrower and will not include peer universities but the direct impact of UCSD's college rankings to UCSD's applicant pool.

---

[1] https://www.sandiegouniontribune.com/news/education/sd-me-ucsd-applications-20190129-story.html
[2] https://fox5sandiego.com/2018/08/13/ucsd-ranked-2nd-best-university-in-us
[3] https://www.researchgate.net/publication/37152774_The_Impact_of_US_News_World_Report_College_Rankings_on_Admissions_Outcomes_and_Pricing_Policies_at_Selective_Private_Institutions
[4] https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5893.2006.00261.x
[5] https://journals.sagepub.com/stoken/rbtfl/sWgSTO0MRTZKE/full

**Data**

Include a description of the perfect dataset you would need want to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the datasets limitations and how it differs from your ideal dataset. If you collected your own data, explain what information you collected, from whom you collected it, and a link to the data. (2 pts)

The ideal dataset would keep track of UCSD rankings in various categories, number of applicants over the years, and the rate of change of number of applicants between each year. It would have as many observations as the number of years UCSD has been established. The variables we would collect are the applicant counts and various category rankings ranging from academics to sports to food quality. Essentially, the variables would keep track of factors at UCSD that would change over the years and affect student perception of the school. Additionally, there would be another variable that would factor positive and negative news about UCSD during that year to determine whether the news has a role in determining student enrollment. Using this dataset, we can weigh each factor against the application trend. The factor with the biggest effect on the number of applicants will have the strongest correlation with the number of people applying to UCSD.

Ultimately, we could not find a single dataset that contained all the variables we wanted. To compensate, we decided to compile information from a variety of sources. Our first source is a table for UCSD's applicant headcount (https://www.universityofcalifornia.edu/infocenter/admissions-residency-and-ethnicity). The relevant variable we can extract from this is the number of total applicants every year. We would use a formula to compute the rate of change between each year. Our second source would be acquired by web-scraping a website that lists the U.S. News' national university rankings from 1983 to 2020 and also Times Higher Education (https://publicuniversityhonors.com/tag/u-s-news-historical-college-rankings/, https://publicuniversityhonors.com/2015/06/13/u-s-news-national-university-rankings-2008-present/, https://publicuniversityhonors.com/2016/09/18/average-u-s-news-rankings-for-126-universities-2010-1017) . Since the starting year for the ranking data is 1983, a limitation of this dataset is we would be missing data from when UCSD was first founded in 1960 to 1982. Our final source would be news articles on UCSD, which we can find by web-scraping all articles with UCSD as a keyword. Such articles would likely come from the UCSD Guardian, San Diego Tribune, and local news sites representing the San Diego area. To get the data from the news articles, we would web scrape a few pages for every year. With this dataset, we would only have observations from 1983 to 2020 because we do not have ranking data past 1983.

**Ethical Considerations**

1. Data Collection

The ethical considerations that must be made when answering our question during the Data Collection process would mostly apply to having the right to use the articles, documents, and sources that we find. The articles we will be using would consist of reports like annual US News & World Report college rankings or Forbes annual top colleges. Since we are not collecting personal data from applicants and solely comparing *numbers* of applicants from one year to another, we do not have to worry about each applicant's consent to their information (this would be a different story, however, if we were comparing the caliber of students who apply from one year to the next). We are only using them as a datapoint, collectively. Data Collection ethical considerations also include mitigating biases. Since personal information is not a consideration regarding the applicants of UCSD, we do not need to worry about protecting applicants' information. However, when we fact-check how and from where the ranking articles and sites are retrieving their ranks for each college, we need to be cautious that *we, as experimenters,* are not crossing boundaries and retrieving data like photos and names to be attached to our findings.

2. Data Storage
   For data storage, the ethical consideration would be whether the data needs to be deleted after use and whether the data holds sensitive information about specific, identifiable people. If the data holds sensitive information, we would have to monitor who has access to the dataset. Since this project uses generic data about a school's rankings, we would not have to worry about holding sensitive information unless we use people's personal information to determine ranking. However, we may have to worry about whether the data needs to be deleted since we would ideally be using data from the U.S. World News, which may have a policy on whether users can store their data obtained from web scraping.

3. Analysis
   For Data Analysis, the ethical considerations would revolve around how we associate/interpret the rankings and sources we use to the number of applicants. Although we will not state that the rankings *cause* the number of applicants to UCSD to rise or fall, our analysis might conclude that there is an association. In terms of missing perspectives, we run a risk of not obtaining data from articles and sources that may, in reality, influence the number of applicants to UCSD. We cannot, however, *ensure* that we focus on the "right" and most relevant rankings but can only decide based on popularity of the source, number of times a website was viewed, etc. In terms of dataset bias, this is the whole point of the rankings. Each source's rankings will probably be different, and the way that they decided their rankings was up to them -- they have their own bias, thus why one source is putting out a different list of rankings as the next source. Bias, in this sense, is actually what we are examining; we are examining the bias in the rankings as a source for applicant numbers rising and falling.

4. Modeling
   Ethical considerations include fair modeling of each source used in the data and displaying the data in such a way where users can understand each ranking, news article, etc. from an unbiased viewpoint. We must consider modeling our study in such a way that the rankings and articles about UCSD that we include are all equally represented and the inferences made are not biased towards a

more positive or negative perspective. Additionally, any algorithms that are made to predict the data must be clear for users to understand and tested to ensure accuracy.

5. Deployment
   In the event that our previous data must be restored for any reason (current data is inaccurate or lost), the previous data will be accessible by the year. Since league rankings, news articles, and the college application process varies every year, we must ensure that the model is updated to be relevant. Thus, we must update our model to the accurate predictions of student applications based on the most recent rankings and articles. To address unintended use, we must consider how users may abuse the results to negatively portray UCSD and how many students apply here. This can be identified through careful tracking of users who access the model or republish our information with other biased interpretations.

**Analysis Proposal**

Here, you will propose how you would use and analyze data to answer your question of interest but are not required to carry out the analysis to answer your question of interest. You will describe in detail what you would need to do to prepare your dataset for analysis (data wrangling) and what type of analysis you would do to answer your question of interest and explain how you would interpret the results from this analysis. We are looking for the correct conceptual understanding and application of ideas discussed in class, not specific and technical implementations. For example, if you are applying machine learning to some categorical data, it's important to specify whether you will be performing regression or classification. If you are unsure about the details of anything above, ask on Piazza, come to office hours, and/or do further research on your own (Stack Exchange, Google, Wikipedia, etc.).

Specifically, you are required to incorporate *at least four different methods*, exploring ideas from a combination of:

- Data Collection (web scraping, APIs, etc.)
- Data Wrangling
- Descriptive & Exploratory Data Analysis (summary stats, correlation, etc.)
- Data Visualization
- Statistical Analysis (Inference, A|B testing, etc.)
- Predictive Analysis (machine learning, classification, regression, etc.)
- Text Analysis (Sentiment Analysis, TF-IDF, etc.)
- Geospatial Analysis (choropleth maps, geospatial statistics, etc.)

(15 pts)

## Method 1: Data Collection

*How would we gather and prepare our data?*

The first step in analyzing the data is to make sure we have a sufficient amount of the data to do a reasonable analysis with. Too few data can result in an analysis that is easily skewed by outliers. However, as well as the quantity of the data, it is critical that the quality of the data is high otherwise we could get biased results. The fact is that the data collection and data wrangling constitute often over 80% of the time spent on typical studies. So we would spend a large portion of our time in this effort and avoid the temptation to rush to analysis without sufficient data and that is high quality with data wrangling techniques to have tidy data. The data collection we actually did perform using a selection of websites:
https://www.universityofcalifornia.edu/infocenter/admissions-residency-and-ethnicity
https://publicuniversityhonors.com/tag/u-s-news-historical-college-rankings/,
https://publicuniversityhonors.com/2015/06/13/u-s-news-national-university-rankings-2008-present/,
https://publicuniversityhonors.com/2016/09/18/average-u-s-news-rankings-for-126-universities-2010-1017

## Method 2: Data Wrangling

*Web scraping and Data Wrangling: University of California Admissions Statistics*
https://www.universityofcalifornia.edu/infocenter/admissions-residency-and-ethnicity
For the admissions data we scraped from the UC admissions and residency and ethnicity website. We were careful not to use ethnicity or gender in gathering our data. We gathered data from the earliest records (1996) to the latest data (2018). This required multiple queries as the website had restrictions on the number of years you could select. As the applications are consistently going up each year, we are interested in the rate of change of applications which does vary up and down for our analysis. The rate of change of applications was not immediately available on the website but we could create this data by gathering the number of applicants each year and then using a formula to calculate the yearly increase from year to year. Below is a sample of the data we collected:

| Year | Number Applicants<br><br>Extracted from Website | Yearly Application Increase %<br><br>Using Formulae |
|---|---|---|
| 1996 | 23687 | 0.00% |
| 1997 | 25098 | 5.96% |
| 1998 | 28090 | 11.92% |
| … | …. | …. |
| 2016 | 84206 | 7.87% |
| 2017 | 88448 | 5.04% |
| 2018 | 97898 | 10.68% |

*Web scraping and Data Wrangling: Academic League Tables*

For the academic league tables we scraped data from US News and also Timers Higher Education. However, for US News, the data required using a third party website that had the data but it was spread over multiple articles over multiple web pages. This required carefully copying of the data from each page and then putting it all together in one data set. There were cases where the same years were recorded in multiple web pages so we had to be careful to make sure the data was consistent. Finally, the Times Score did not have data from earlier than 2017, and because the number of observations was small any model using this date would likely be affected by outliers. Below is a sample of the data we collected:

| Year | US News Rankings | Times Score |
|---|---|---|
| 1996 | 43 | |
| 1997 | 34 | |
| 1998 | 33 | |
| ... | …. | …. |
| 2017 | 44 | 73.2 |
| 2018 | 42 | 78.7 |
| 2019 | 41 | 79.7 |
| 2020 | 37 | 78.8 |

The remaining data collection and wrangling would be for new stories that generated positive and negative sentiments, We would use this data for sentiment analysis. For this we would have used APIs to query for the data and cleanup the data (e.g. remove non essential words). Our approach, given time, would have been to develop a python script which we could use on websites to gather words used in the articles that had UCSD in the title of the article.

## Method 3: Descriptive Analysis

*What did we do after preparing our data? - Descriptive analysis*

Before moving on to detailed analysis we do descriptive and exploratory analysis. This involves finding basic statistics such as average, standard deviation and checking correlation of data variables. In fact we were able to do some descriptive analysis as we had collected the admissions and league table data. For example, we determined that the mean yearly annual increase was 6.48%, with a range of -4.42% to 15.58%. The summary of our descriptive analysis is below:

| | US News Rankings | Times Score | Number Applicants | Yearly Application Increase % |
|---|---|---|---|---|
| Mean | 36.04 | 74.00 | 51526.04 | 6.48% |
| Min | 31.00 | 67.40 | 25098.00 | -4.42% |

| | | | | |
|---|---|---|---|---|
| Max | 44.00 | 79.70 | 97898.00 | 15.85% |
| STD | 4.03 | 4.41 | 21185.96 | 5.69% |

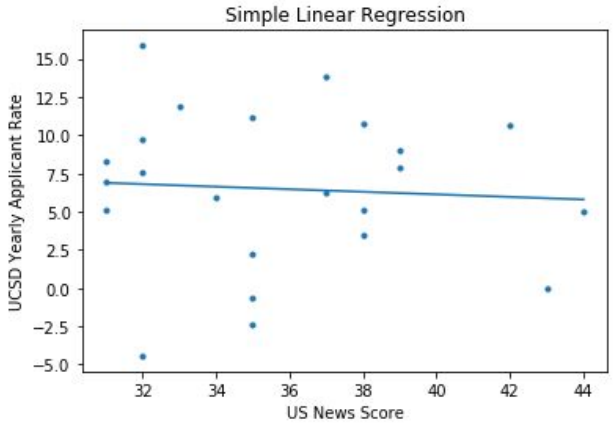
Our initial descriptive analysis we felt that there was sufficient justification that our analysis would produce meaningful results as the rankings showed a range of high and low data together with the yearly applications which had positive and negative values. This meant the data would have cases where the years had lower applicants and other years were the applicants were higher than average. Assuming this correlated (or not) with the rankings then we could use this to evaluate if our hypothesis was correct (or not).

## Method 4: Predictive Analysis

*What model would we use for analysis? Multiple Linear Regression*

We would develop a multiple linear regression model. That is specifically because we want our model to give us back a number not just a classification: the predictor variables would be the two variables, the US News league table and the Times Score. Together we would augment additional predictor variables from our sentiment analysis. The model would then allow us to predict the yearly application increase. The next step for analysis and visualization would require more sophisticated tools, in particular using python and numpy. We attempted in our extra credit using the data we collected, which didn't do full sentiment analysis but did have the league table predictor variables. Finally, we would also calculate the p-value using the null hypothesis that the results are just by random chance. We would want to verify if our p-value was sufficiently be categorized as statistically significant (<5%) or even highly statistically significant (<1%). We would be cognizant not to change our data once we do this p-value analysis, we do not want to fall into the trap of “fixing” our data to meet a particular p-value (p-hacking).

Below is a linear regression we would later perform using the actual data we collected.




## Method 5: Text Analysis (Sentiment Analysis, TF-IDF, etc.)

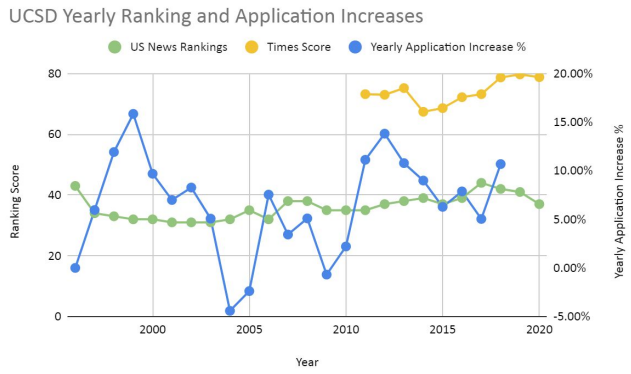
Sentiment analysis would also be beneficial in understanding articles and the writers’ tones for each university, specifically UCSD compared to the other Universities of California. Adapting the sentiment lexicon called SocialSent from a Stanford University project would be useful since this lexicon includes

domain-specific sentiment analysis. SocialSent's code is publicly available for use and is downloadable on GitHub. Thus, more context would be taken into consideration when analyzing the wording of the articles. To do this, tokenization and TF-IDF are useful to determine key phrases and unique words in the different passages. We can compare to see if certain words are unique words of more than one article. With this, we can analyze any additional rankings or sentiments on UCSD than the article explicitly reveals.

*How would we optimize our model?*
If the number of league tables predictor variables were too large then the cost of data storage and data computation would also be expensive. In this case we would do Principal Component Analysis (PCA) to reduce the number of predictor variables to the ones which had the most useful data. For example, if we had data from 20 different league table scores, we would do PCA analysis and see which variables had the most variation and select those and drop the other variables.

## Method 6: Data Visualization

*How would we visualize our data?*
After our analysis was complete, we were ready to start to visualize the data.The first step we did was to do a simple chart by putting the data into a spreadsheet. In particular we wanted to do a regression of the league table score as the predictor variables which we could use multiple regression to estimate the yearly application increase. So using a scatter plot made the most sense. Once we have the scatter plot then we could do a best fit line using multiple variables, we attempted this in our extra credit project where we used python and numpy together with a visualization package.




On our scatter plot we would also add the line of best fit using the linear regression. We would also want to highlight the R coefficient.

**Discussion**

How would you interpret the results of your proposed analysis? What are the limitations, pitfalls, and potential confounds of your methods, or biases in your data sources? (e.g., how does the selection of the sources of your crowds affect your outcomes?) How would you set out to address them? In addition, outline

how you would address any societal and/or ethical implications of your proposed project discussed in your Ethical Considerations section. (10 pts)

*How would you interpret the results of your proposed analysis?*
Our analysis will produce a multiple linear regression model. We would use that model to predict the rate of UCSD application rates based on the predictor variables of academic league standings and sentiment analysis from news stories relating to UCSD. We would want to validate out model by finding the pearson correlation factor "r" that we would expect to be much higher than zero (a zero would mean no correlation). If instead it was a negative value that would be surprising, negative league table would entice more applicants. A perfect 1.0 factor would also be surprising and would be "too good to be true". In these cases we would want to double check our data.

*What are the limitations, pitfalls, and potential confounds of your methods, or biases in your data sources? (e.g., how does the selection of the sources of your crowds affect your outcomes?)*
Biases in our data sources can be found depending on how we choose the websites to web-scrape. If we were to primarily web scrape articles on the UCSD website, we would get a more positive sentiment analysis. A potential confounder for the data collected from the rankings for Times Higher Education and US World News is the consideration of class size, which is a criteria for both their ranking methodology (https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2019-methodology, https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings). A limitation of sentiment analysis is words are taken out of context because each word is considered separately. This is an issue if someone wrote an article about UCSD being "not good" and the algorithm reads it as two terms: "not" and "good." If this kind of pairing is consistent throughout the article without some way of accounting for it, we would get an inaccurate sentiment analysis. Another limitation of sentiment analysis is its inability to assess sarcasm. Given a negatively written satirical piece, the algorithm may say the piece has positive sentiment. Sentiment analysis could also be a limitation in that a review does not necessarily include a ton of positively or negatively interpreted words but instead shows the ranking through using words like "large" or "few" (ex: large classrooms, few students per class) which more so explain and give a description of aspects of the school but do not necessarily have a sentiment towards them (and the sentiment analysis could easily misinterpret the sentiment in its context).

*How would you set out to address them?*
To address bias in our data sources, we would have to ensure that we are getting sources from a variety of websites. By doing so, we would have a variety of perspectives on UCSD. To address the potential confounder for the ranking data, we would have to create a formula to factor out the common criteria for one of the datasets. We do this to ensure that the common criteria is only counted once. To address the limitation of analyzing words individually for our sentiment analysis, we use a pre-existing dataset of positive, negative, and neutral sentiments on articles given certain keywords in conjunction with Naive Bayes to better classify articles as positive, negative, or neutral. For our limitation on detecting sarcasm, we could use a dataset that determined sarcasm based on keywords and likewise use Naive Bayes to classify whether articles are sarcastic. Sarcastic articles would have their sentiment flipped in order to reflect the true tone of the article.

*In addition, outline how you would address any societal and/or ethical implications of your proposed project discussed in your Ethical Considerations section.*

In terms of data wrangling, for our exploration into our question to be unbiased, we must trace how the college ranking is determined for each source that we find and where they retrieved the data from. For example, if a source we use is ranking solely based off of how high the GPA of incoming students are, we must look into how that information is being presented and fact-check if those statistics are actually true. For cross-checking purposes, we need to find databases and APIs that give us back only numerical values regarding publicly-found statistics like average accepted GPA from year to year, acceptance rate, money donated from alumni, etc., NOT information like applications, names, photos, or specifics that denote each individual; we want to look at the statistics as a whole. When it comes to analysis of data and our own biases, we must make sure that our own ideas of how *we* would rank the colleges does not play a factor in our analysis. Although we cannot decide a concrete number or ratio, in this stage of the experiment, we would have to look at the overall rankings from all of the sources and compare that to the number of applicants at UCSD (rising/falling). We will use statistics measures like Pearson correlation to determine if there is an association and verify that the p-value is statistically significant in order to draw a conclusion.

We also must be careful of web scraping and only use data that is accessible to us. Since our results would, in theory, go back to UCSD for feedback on how they should distribute their budget to get more applicants, we also have to be careful of not mixing up causation with correlation. If we obtain a statistically significant p-value after doing our analysis, this does not automatically mean that UCSD should spend more money on marketing campaigns and improving ranking. So, we have to be careful in communicating results.

**Group Participation**

Include one paragraph briefly outlining the contribution of each group member throughout the quarter while working on this project. Each of you must also fill out the survey (link provided toward the end of the quarter) about individual and group participation. (3 pts)

This was a group project and each person on the team made significant contributions. Saroop worked on the Hypothesis and Background Data in the original proposal, and in the final project did the data collection and data wrangling and wrote analysis section related to data collection, wrangling, and multiple linear regression. Amy worked on the Data and Data Storage under Ethical Considerations portions in the original proposal, and in the final project, on the first two questions of the Discussion portion and reviewing responses for continuity. Thanh contributed on Modeling and Deployment under Ethical Considerations for the first proposal and examined Text Analysis and Societal/Ethical Implications for the final. Abby

contributed on Data Collection and Analysis under Ethical Considerations for the first proposal and contributed to the Societal/Ethical Implications and Limitation of Bias for the final.

Finally, we learned a lot from this project including in particular group participation. We were all busy but the project was interesting and learning and applying the concepts we learned in lectures made this an enjoyable and valuable learning experience.