Data Management and Data Analytics Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Student Name: William J Townsend

Student ID: 003397146

Capstone Project Name: Calculating the Results of the 2021-22 @DownGoesBrown NHL Prediction Contest

Project Topic: This project will programmatically scrape, parse, standardize, and score nearly 1600 contest entries in an "easy" annual prediction contest held by Sean McIndoe, also known as @DownGoesBrown, to predict various results and happenings within the 2021-22 National Hockey League (NHL) season. Using the Python programming language and libraries such as Beautiful Soup, pandas, and NumPy, this project will automatically score all entries in the 2021-22 prediction contest, doing so in a far faster fashion than the labor-intensive process of manually scoring each entry. In the process of doing so, it will also generate a number of other descriptive statistics to be used by McIndoe in articles or other engagement with readers about the status of the contest and related NHL news.

Research Question: Analysis of the nearly 1600 entrants into the 2021-22 @DownGoesBrown NHL prediction contest is a laborious process, owing to both the volume of entries and the unstructured nature of those entries. This project aims to create a Python script to automate this analysis to be more efficient while still providing a depth of descriptive analytics intended to facilitate their use in social media, written, or podcast pieces by McIndoe.

Hypothesis: In order to judge the efficiency of the Python script's work in scraping contest entries, standardizing them, and scoring them, the run time of the script can be compared to the amount of time that would be needed to manually enter and score those same entries. The null hypothesis will be that the programmatic solution is not faster than the manual process. The alternative hypothesis will be that the programmatic solution is faster than the manual process.

$H_0 : time_{(program)} - time_{(manual)} \geq 0$
$H_1 : time_{(program)} - time_{(manual)} < 0$

As the process of scoring all entries manually would take dozens of hours, a sample of the entries will be manually scored and then prorated out to the number of entries actually made in the prediction contest. A T-Test will then be used to compare the mean of a group of programmatic attempts to score the contest entries against the mean of a similarly sized group of manual attempts to score the contest entries.

Context: Sean McIndoe holds an "easy" annual prediction context for the forthcoming NHL season. This contest on The Athletic (a subscription-based sports page for which McIndoe is a senior NHL writer) requires readers to enter in the comments certain predictions for that NHL season, such as teams that will definitely make the playoffs or players to be voted into the top 10 for a major award. Readers provide up to 5 answers for the question, for increasing numbers of points (1, 3, 6, 10, 15 points for 1 – 5 correct answers), across 9 questions, with a 10th question that may receive only a single answer. If any answer for a question is incorrect, the reader receives 0 points for that question. This creates a fun contest mechanic in which readers may aggressively pursue more points in exchange for an increased risk of not receiving any points for that question.

Last year's contest was more popular than anticipated, with over 800 entries made in the article's comments, each of which McIndoe had to read and manually enter into a spreadsheet over the course of the season, scoring them at the end of the season. This was a labor-intensive and painstaking process, rife with opportunities for error because of the volume and unstructured nature of the data. This year's contest grew to nearly 1600 entries. A programmatic solution to gather, clean, and standardize the data would save significant time and reduce potential errors, as well as potentially allowing the contest to scale even further in future years. Similar contests have been created by other writers at The Athletic as well, which could potentially make use of this sort of scripting.

The calculation of a contest winner is one goal of this project, but the generation of various descriptive statistics regarding the contest entries provides additional content and context which may be used by McIndoe in other articles. McIndoe has previously used this data as a fun avenue to engage with his readership and other NHL fans in a way that is consistent with his particular brand of humor, cynicism, and deep knowledge of the game. Such data has also provided a unique approach for commentary on NHL news throughout the season, particularly from the perspective of how surprising various results or news stories are within the NHL. The ability to use this data was previously limited by the time and effort required to process it, but a programmatic solution and the associated statistical analysis would provide a streamlined process that allows for the data to be used more effectively, more quickly, and more broadly than it was able to be used previously because of the limitations of McIndoe's original process.

Data: Entries to Sean McIndoe's annual prediction contest are made in the comments of the article on The Athletic which announced this year's contest. Each comment consists of a first name and last initial, a date that the comment was made, and the comment contents. 1665 comments were added to this article, approximately 95% of which were contest entries. Each contest entry consists of at least 10 lines (if the commenter followed directions) of unstructured qualitative data lacking any sort of standardization, providing upwards of 46 different answers to prediction contest questions.

The entries include other extraneous information such as commentary on one's own entry, unique identifiers to facilitate finding one's entry via CTRL+F in the pool of 1665 comments, or other irrelevancies to the issue at hand. Additionally, approximately 5% of comments need to be ignored, as they consist of commentary by readers on their own or each other's contest entries. This data will be scraped from a saved copy of the webpage which was made at the time that the contest closed to new entries, at 1930 ET on 12 Oct 2021. This dataset is completely original.

Regarding the question of ownership of the data, to my knowledge, there is no particular

settled case law on the issue of who specifically "owns" comments made upon a website which allows such, and responsibility for such content (such as occurs when users violate the law in a comments section) is shared across multiple entities. The original commenter has created their particular contest entry, and to my thinking, would retain ultimate ownership of the ideas espoused within their particular entry. However, in posting their entry for inclusion in the contest, the entrant has placed this information into a domain that is publicly accessible to me as a subscriber of The Athletic to read, comment on, or even save. Additionally, in entering the contest, the entrant has at least consented to Sean McIndoe collecting and evaluating the data, and I was given permission by McIndoe in an email exchange on 14 Oct 2021 to do this on his behalf in exchange for future financial compensation.

In order to score the contest entries, data regarding how the 2021-22 NHL season has unfolded will also need to be gathered. This will largely come from the hockey-reference website, which is publically available and contains nearly any information that could be needed pertaining to the current and historical NHL seasons, including statistical accomplishments, transactions, standings, and awards information. I anticipate additional information may be gathered from the CapFriendly website regarding specific transaction & salary cap details. Any other data needed will be gathered from similarly public-facing sources, including press releases, websites, or reporters covering the NHL.

Data Gathering: The data needed for this project will be scraped from a saved copy of the website, using the Beautiful Soup 4.4.0 library for Python 3. This library allows users to scrape the XML and HTML of a website to find, among other things, content stored within particular tags. This library allows me to automatically scrape the web page containing the contest announcement and entries, to pull each entry out from the parent-comment-container tag, including the name of the entrant and the entirety of their comment, which will then be parsed to generate up to 46 unique answers to the 10 contest questions. In order to avoid repeatedly calling the website and potentially consuming unreasonable bandwidth belonging to The Athletic, a saved copy of the website will be used. This will allow an iterative approach to getting the scraper working, without using bandwidth or computing resources besides my own.

Once the comments are scraped, I will use Python to write code to process the scraped comments to uniquely identify each author and their associated comment. Contest entries will be identified from non-entry comments and removed from the pool of comments. Any contest entries which are formatted in such a way that precludes programmatic handling of the entry to distinguish answers from each other will be fixed to facilitate such. A parser will then "read" each comment and generate a dataframe using the pandas library, consisting of that entrant's uniquely identified name and up to 46 answers to the contest questions. Python dictionaries will be used to standardize the answers for each question and facilitate the scoring function, which will also be developed in Python.

Data Analytics Tools and Techniques:
This project will use the following data analysis techniques:

- Qualitative content analysis
- Thematic analysis
- T-testing
- Data wrangling
    - Data gathering (via web scraping)
    - Data assessment
    - Data cleaning

- Exploratory data analysis
- Data visualization & reporting
- Qualitative content analysis
- Thematic analysis
- T-Tests

This project will use the following data analysis tools:

- Jupyter Notebook
- Microsoft Excel/Google Sheets
- Python programming language in the Anaconda environment
- The following Python libraries:
    - Beautiful Soup
    - pandas
    - NumPy
    - MatPlotLib

Justification of Tools/Techniques: The core of this project is qualitative content analysis. The dataset consists entirely of unstructured and unstandardized qualitative data in the form of NHL teams, coaches, general managers, and player names. Qualitative content analysis allows us to evaluate patterns within a piece of content. In this case, that is definitively identifying the elements of the dataset through standardization operations, evaluating which elements are objectively right or wrong, and identifying the frequency of elements within the dataset. This will lead directly into performing thematic analysis, as the statistics generated and the observations made from those statistics will speak to views and opinions of Sean McIndoe's readership, which becomes content for future articles, podcasts, or other media that he may create.

T-testing will allow me to test my hypothesis, comparing the mean of one group (programmatic processing of contest entries) against another group (manual processing of contest entries).

Python provides a useful and robust programming language to perform this data analysis, and the Anaconda environment provides an all-in-one solution that allows this to be done in one convenient package. Jupyter Notebooks, as a part of that environment, allows for a "show-your-work" approach that easily moves back and forth between Python script and narrative output all in one place. Jupyter Notebooks also allow for the easy generation of a slide deck to be provided to McIndoe with summary descriptive statistics about each question. This allows me to do all of my work in a transparent fashion that can be verified by McIndoe if desired, which is an important consideration given that he has an established reputation and readership.

Beautiful Soup is a Python library which facilitates web scraping from HTML or XML tags. This will allow me to gather both comment contents and comment authors automatically, by allowing me to pull the contents of the appropriate tags within the published comment section on The Athletic.

The pandas library will allow me to place all of the data into a dataframe, which is essentially a Python table that is capable of sophisticated operations and manipulation upon the data within it. The NumPy library allows for placing Not-a-Number (NaN) values into the dataframe, as well as performing some other mathematical functions. The MatPlotLib library allows for the generation of highly customizable data visualizations.

At the end of the scoring process, the complete dataframe will be exported to a .csv file which is accessible with Microsoft Excel or Google Sheets. This .csv will be imported in either Excel or Sheets and formatted slightly to enhance readability. The completed sheet will then be provided to McIndoe.

Application Type, if applicable (select one):

☐ Mobile
☐ Web
☐ Stand-alone

Programming/Development Language(s), if applicable: Python 3.9

Operating System(s)/Platform(s), if applicable: The scripting is performed within a Jupyter Notebook in Windows 10 which generates output information for Sean McIndoe. A slide deck providing some descriptive data to McIndoe for use in writing about the contest will be provided, as will a spreadsheet containing the final contest results and scoring.

Database Management System, if applicable: No DBMS is used, as the data is placed into and manipulated within a pandas dataframe and then exported to a spreadsheet. This facilitates Sean McIndoe's use of the results and makes it easily sharable to readers, at his discretion.

Project Outcomes: A sortable spreadsheet will be created and provided to Sean McIndoe which scores out all of the entries to the prediction contest, including final scores for each entrant. This will demonstrate the winner of the contest, while retaining the answers so that they may be verified before a prize is awarded.

Additionally, a slide deck will be generated and provided to McIndoe with a variety of statistics generated from the script, including:
- Total number of entries
- Number of entries which required manual intervention in order to handle programmatically
- For each question, the total number of possible answers, and the number of answers provided
- For each question, a count of how many times each value was selected as an answer by an entrant. For example, how many people selected Colorado to make the playoffs, how many selected Tampa Bay to make the playoffs, etc. for all answer values for a question
The slide deck will also highlight any interesting observations that become apparent during the course of generating these statistics. The slide deck will be written in a fashion to be approachable to a non-technical user.

Projected Project End Date: 6/30/2022

Sources:
Sean McIndoe's 2021-22 Prediction Contest Announcement:
https://theathletic.com/2869497/2021/10/10/down-goes-brown-predict-the-nhl-season-with-the-return-of-the-contest-thats-so-easy-its-almost-impossible/
NOTE: The Athletic is a subscription-only site, and thus its content is behind a paywall. A PDF of the full article (4 pages of the original article, 607 pages of contest entries) will be uploaded alongside this form.

Sean McIndoe engagement/commentary using analysis from this project:
https://twitter.com/DownGoesBrown/status/1523674768439922688
https://twitter.com/DownGoesBrown/status/1523812031748612098

Beautiful Soup – Documentation:
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Human Subjects or Proprietary Information
Does your project involve the potential use of human subjects? (Y/N): N
Does your project involve the potential use of proprietary company information? (Y/N): N

STUDENT SIGNATURE

_____William J Townsend___1 Jun 2022_____

By signing and submitting this form, you acknowledge that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

TO BE FILLED BY A COURSE INSTRUCTOR

The capstone topic is approved by a course instructor.

Wednesday, June 1, 2022n

Project Compliance with IRB (Y/N): Y