DATA ANALYSIS OF THE 2021-22 @DOWNGOESBROWN NHL PREDICTION CONTEST

WILLIAM J TOWNSEND

BS Data Management & Data Analytics Capstone for

Western Governors University

Part A: Project Overview

A-1: Research Question/Organizational Need

This project will automate several portions of scraping, parsing, standardizing, and scoring nearly 1600 contest entries predicting various outcomes and events of the 2021-22 National Hockey League season. These entries all exist as unstructured qualitative data in the comments of an announcement article which would ordinarily have to be manually scored in a very tedious process. This process allows all entries to be evaluated quickly to determine the contest winner, as well as generating a variety of descriptive statistics on the dataset to be used for follow-up articles and engagement via social media.

A-2: Context & Background

Sean McIndoe, also known as Down Goes Brown, is a national NHL writer for The Athletic who has previously worked in similar capacities at Sportsnet, ESPN's Grantland, Vice Sports, and other outlets. McIndoe holds an annual "easy" prediction contest which requires readers to enter in the comments of his article a series of specific predictions for the forthcoming 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, worth increasing numbers of points (1, 3, 6, 10, 15 points for 1, 2, 3, 4, 5 correct answers) across 9 questions, with a 10th question that may receive only a single answer. If any provided answer for a question is incorrect, the reader receives 0 points for that question, which creates a mechanic in which readers may aggressively pursue more points in exchange for the increased risk of not receiving any points at all.

The annual contest started with the pandemic-shortened 2020-21 NHL season, and it was more popular than anticipated. Over 800 entries were made in the article's comments, each of which McIndoe had to read and manually enter in a spreadsheet over the course of the season. At the end of the season, these entries then had to be manually scored as well. Because of the sheer volume data and its unstructured and unstandardized nature, this was a labor-intensive and tedious process, and the manual nature of the data entry creates opportunities for error. This year's contest for the 2021-22 NHL season grew to nearly 1600 entries. A programmatic solution to gather, clean, and standardize the data would save significant time and reduce potential errors. This would also potentially allow the contest to scale even larger in future years.

Calculation of a contest winner is the primary 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 ways. This data will provide a fun avenue for McIndoe to engage with his readership and other NHL fans in a way that is consistent with his brand, as well as a unique approach for commentary on NHL news throughout the season, especially surprising NHL news or results. These results will also provide useful prompts for written and podcast content for McIndoe.

A-3: Summary of Published Works

On 27 Apr 2021, Sean McIndoe wrote an article for The Athletic titled, "Down Goes Brown: Remember that prediction contest that was supposed to be easy? So far, it hasn't been". This was McIndoe's first article following up on the 2020-21 prediction contest, and before discussing the actual status of the contest, he discussed the overwhelming response to it:

If you missed it the first time, or could use a refresher on who you picked, the original post is here. I thought it might be fun. The readers apparently agreed, because over 800 of you entered the contest. That was, uh, a bit of a problem, since I hadn't really thought through how I'd do the actual work of scoring this thing. Still haven't, in case you were wondering. Any students out there looking for an offseason job?

With that many entries, and only one first-place prize up for grabs, it quickly became apparent that playing it safe wouldn't win the day. Sure, you could have given one sure-thing answer for each question and banked points on all eight, but it wouldn't be enough to win. Somebody was going to have to run the table with eight perfect questions, each with multiple answers…

In fact, despite the over 800 entries, there's a decent chance that we won't get many perfect entries after all. It's possible we won't even get one.

This was, of course, the whole point. The "easy" contest wasn't ever supposed to be easy in the first place. Instead, it was a way to remind us that an NHL season always seems predictable until it starts to play out, at which point things go off the rails quickly. In fact, I originally assumed nobody would get a perfect score back when I figured I'd get 100 entries; when the number kept ballooning, I started to have my doubts.

I still do, and we'll see how it plays out. But for now, let's run through all eight questions and see where things are at.

This was the first article McIndoe published about the contest in the first year of its existence, and he acknowledges that the prediction contest was orders of magnitude more popular than he had anticipated. Where McIndoe expected to receive something like 100 entries,

he instead received over 800. McIndoe also notes that he was wholly unprepared for handling this number of entries, even joking about soliciting help with handling the unexpected deluge of contest entries that he had to process.

The article then went on to discuss each of the questions from that year's contest, going into detail on the results to that point in the NHL season and how they related to the contest. This primarily focused on the surprises from the NHL season that many entries got wrong, including how many entries took a 0 on a particular question, or that a number of entries correctly predicted. This article also commented on a number of metatextual concepts such as the strategies that different readers appeared to be trying to use within the contest, McIndoe's own expectations of certain questions, etc.

The next article that was written about the prediction contest was by McIndoe on 3 Aug 2021, titled "Down Goes Brown prediction contest results: Did anyone have a perfect entry? Would that be enough to win?". This article was the 2020-21 prediction contest wrap up and was released in the midst of the NHL offseason, a time when many NHL writers regularly struggle to generate interesting or engaging content. This piece included another acknowledgement about the unexpected popularity of the contest:

> *I'll admit, I didn't think the concept all the way through, especially when it came to tallying up the results. I'd have to do that by hand, which could get tricky. But I figured I'd get a few dozen entries, maybe even over 100 if people really seemed to like the idea, and that would be manageable.*
>
> *Then over 800 of you entered. Whoops.*

> *Ah well, what's one more addition to my "overly complicated spreadsheet" folder. After digging through the entries, I think I've found the highest scores, including our winner. Did anyone manage to pull off a perfect entry, with points on all eight questions? We'll get there, but first, let's walk through the questions and how you all did.*

This article also went into detail on each of the contest questions, looking at that year's performances and transactions that did (or more commonly, did not) make the cut. In doing so, McIndoe cited how many entries took a zero because of these performances. For example, on Question 5, requiring readers to "Name up to 5 goaltenders who will definitely start at least 60 percent of their team's regular-season games this year", McIndoe said:

> *There were also more than a few "obvious" picks that didn't make the cut, and this was the first question where we saw some real carnage. Sergei Bobrovsky was responsible for way more zeroes in this contest (about 140) than he was on opponent's scoreboards. Matt Murray took out over 100 entries, Carey Price was on the hook for over 200, and Frederik Andersen topped that with 250.*
>
> *But the big name here was Carter Hart, who showed up on over 400 entries. He was pretty much the perfect illustration of what this contest was trying to remind us. Coming into the season, it seemed inconceivable that a healthy Hart wouldn't be the Flyers' go-to guy. He was 22 and had two very good seasons under his belt, and was backed up by creaky veteran Brian Elliott. If anything, you'd have bet on Hart to be a Vezina candidate before you'd think he'd lose his full-time starter's job. But he did, posting awful numbers, and was going to miss our cutoff even before he was shut down due to injury at the end of the season. Sometimes, the surest of sure things don't go the way we're all expecting.*

This analysis continued for each of the eight questions in last year's contest. After a disclaimer acknowledging the difficulty of tabulating the scores ("It turned out to be a huge pain in the neck to sort through all these entries, and I did the best I could. I'm pretty sure I came out with the right winners, but…"), the article covered the top scoring entries, including a perfect entry:

> *Garret basically figured out a near-ideal strategy in terms of dodging wrong answers… Add it all up, and there's not a single miss to be found here. Brilliant work.*
>
> *Except for one thing. Garret didn't win.*
>
> *While it's perfect in terms of avoiding mistakes, his entry ends up being too conservative. With only 14 points on the back half, he leaves the door open for a more aggressive entry to pass him at the finish line.*

The "perfect" entry with zero wrong answers earned 74 points, several earned 75, and another earned 76. All were passed by an entry that earned 91 points to win the contest. For reference, a truly perfect entry (full points on every question) would've earned 120 points.

In addition to his writing for The Athletic, McIndoe also appears weekly on The Athletic Hockey Show, a podcast that he co-hosts with another NHL writer, Ian Mendes. Shortly after the publication of the previous contest wrap-up article, on the 5 Aug 2021 episode ("Nathan MacKinnon's eating habits, John Tortorella joins ESPN, the taboo of gambling allegations, and more") of the show at 20:59, Mendes started a conversation about the contest by asking about the process of figuring out the results:

Mendes: I want to know, how long did it take you (laughs) to calculate or tabulate the results of this thing, because it is insane to me that you would've gone through all of these submissions from readers and declared a winner.

McIndoe: Yeah, uh, I didn't think this through, is what happened. When I came up with the idea for the contest, I thought 'I'll get a hundred entries, that'll be alright, that'll be successful' and it ended up being over 800.

The two spoke for a few minutes about the contest, covering much of the same ground as the above articles. McIndoe explained that his process for handling that many entries was to just sit down and work on it for a little bit at a time, throughout the entire NHL season. This process was aided by eliminating the majority of entries without scoring them in full, once they were recognized to have earned too few points to be able to contend for first place. The two also joked about how this story, like many of the obscure or complex topics that McIndoe writes about, could justify a "Down Goes Brown Internship Program" because McIndoe is aware that some of these situations could be easily addressed through technology, but lacking that expertise, McIndoe has to research them "the hard way". The discussion wrapped up at 25:56, with this:

McIndoe: There were a lot of people that have offered help, and that's very cool of them. We'll see, I may come up with a better plan next year, or I may just wait until the day before the season and throw it out there without much thought and not learn from my mistakes.

Mendes: I love how when I first asked you about this, I think your answer was something like, 'Well, I thought it was a good idea', and I feel like that's the tagline for like 80% of your pieces, is "I thought it was a good idea. It seemed like a good idea at the time".

> *McIndoe: Yeah. 'It seemed like a good idea, it doesn't feel like that anymore, but I'm already too far in, so we're gonna do it.' Yeah, that's pretty much, I mean, that's gonna be, pretty much everything I write in August is going to fall into that category, so be ready.*

Each of these published works indicates that, while McIndoe was able to make his manual methodology work, an improved data analysis solution for this contest would be extremely welcome. This is supported by unpublished work as well, consisting of emails between myself and McIndoe starting on 14 Oct 2021. The 2021-22 prediction contest announcement was posted on 10 Oct 2021 to collect entries until 1930 on 12 Oct, when the NHL season started. When I reached out on 14 Oct to see if McIndoe would be interested in the scraped and standardized contest data, McIndoe agreed to take this, under the condition that I would accept compensation for the work because of the significant amount of time this would be saving him. This indicates that McIndoe sees a clear business case for himself to purchase this data analysis from me as a contractor, rather than generating it himself in a laborious manual process.

A-4: Summary of Data Analysis Solution

My solution to this labor-intensive process is to create a bespoke Python script to automate the gathering of contest entries, standardize them, and then score all of the entries. This script will scrape all 1600+ comments from a saved copy of the article on The Athletic announcing the contest, determining which comments are contest entries. It will then enter all of

those entries, including the entrant's name, into a dataframe. Answers will be standardized by using dictionaries, from entries such as "Det", "Detroit", "Detroit Red Wings", "Red Wings", "Detriot" to a single consistent entry form such as "DET".

A series of descriptive statistics will be generated to detail the answers for each question. This data will be compiled into a slide deck to be provided to McIndoe. The script will then score each of the entries, comparing each entrant's answers to a provided list of the correct answers for each question. Scores will be calculated for each entry to determine the contest winner, and these scores will be added to the dataframe containing the standardized entries. The dataframe will then be exported to a CSV, where some minor formatting will be performed to increase readability. This spreadsheet will then be provided to McIndoe as well.

Performing this data analysis will then provide McIndoe with the needed information to declare a contest winner and issue their prize, as well as supplemental data allowing him to generate written media, podcast content, or social media engagement with his readership about the contest, consistent with the previous year's usage.

A-5: Benefits of Data Analysis

This software-based process is anticipated to save significant time over the manual process that McIndoe used for the previous year's contest. While McIndoe introduced the bonus question as a way to "zero out" a large number of entries easily in an attempt to reduce the amount of time that would need to be spent on scoring, this would still leave him with needing to manually score a number of questions similar to last year's contest. This remains a large time investment, which could be improved with an automated process.

This data analysis solution also generates additional prompts for content, which is a large part of the point of the contest itself. While the prediction contest itself provides an additional topic for engagement or a unique approach for examining stories in the NHL season, the data within the contest is also of interest and ripe for discussion, as demonstrated in the published works about last year's contest. The ability to utilize this data for generating content is, however, limited by the ability to wrangle it.

Even prior to implementing the scoring of the contest entries, the project's data wrangling elements and subsequent descriptive statistics regarding the predictions made by readers is a great source of prompts for content creation or audience engagement. Scoring of all entries in the contest, rather than depending on eliminating a number of them, also allows for additional engagement in the context of allowing all readers to easily check their scores, in what position they finished, or even who may have come close to winning if not for a single mistake, in addition to generating descriptive statistics on the final outcomes of the contest as well. In total, this project intends to create more prompts for engagement or content, in greater detail and more quickly than if all of the same work was done by hand.

Part B: Data Analytics Project Plan

B-1: Project Goals, Objectives, & Deliverables

The goal of this project is to use Python scripting to automate the handling of all prediction contest entries through the process of gathering, standardizing, and scoring all of the contest entries, while providing descriptive statistical summaries of the contest answers. Accomplishing this goal will produce the following deliverables for McIndoe:

- A spreadsheet containing all contest entries in their entirety, standardized and scored.
- A slide deck containing various descriptive statistics for each question, as well as any other observations or trends observed in the course of the handling this data.

B-2: Project Scope

This project's scope is a standalone Python script to scrape contest entries, standardize contest entries, generate descriptive statistics, and score the contest entries, and then export the complete dataset to a spreadsheet. The spreadsheet may be modified slightly to enhance usability by non-technical users, such as freezing headers, alternating row backgrounds, and other such minor changes of a cosmetic nature. A slide deck will also be generated containing the descriptive statistics that are generated by the script.

Descriptive statistics to be included in the slide deck:

- Number of contest entries, number of possible answers, and answers actually provided
- Number of entries requiring specific programmatic intervention in order to correctly parse
- Top scores without question #10
- Top scores overall
- For each question:
    - Number of possible answers, and answers actually provided
    - Number of unique answers provided
    - For each unique answer, the frequency of answer selection
    - Distribution of number of correct

- Distribution of number of incorrect answers

The slide deck will also include any observations that strike me as being interesting or worth highlighting to McIndoe. This may include unusual standardization issues, verifying unexpected answers, acknowledging interpretation of vague entries, highlighting trends within the data, etc.

Descriptive statistics not included in the list defined above are outside of the scope of this project. Support of past or future prediction contests besides the 2021-22 NHL prediction contest is outside the scope of this project. If McIndoe chooses, he may share the results of this project at his discretion, but the hosting of the results or social media engagement (answering reader questions, explaining results, etc.) with his readership remains outside the scope of this project.

B-3: Project Methodology

This project will use a modified Agile methodology. While a waterfall methodology was considered, it was deemed to be inappropriate because of the regimented nature of its phases and their contents and goals, where the development of this project was much more uncertain and fluid in its development and eventual fruition. There will be no standups or other such collaborative elements of an Agile methodology because the project involves only one developer. Progress will be made through a series of short sprints, each dedicated to completing a new element of the automation script:

- Scraping of contest entries from comment section of contest announcement
- Fixing of scraped comments as needed to allow parsing
- Generation of dataframe and import of entries, including granular answers

- Fixing of dataframe contents as needed to allow standardization
- Standardization of answers for all 10 questions
- Automated entry scoring system

Each of these elements of the project requires the preceding step to be completed before that step may be initiated. Two additional steps exist, which do not have this specific requirement: development of the slide deck, and export of the data to a CSV. Each of these steps will be completed in a single sprint, in which they will be the sole development focus. Quality control will be maintained throughout the process, including diligent documentation of decision points within the software or non-obvious coding solutions, and expected outputs will be compared to actual outputs to verify that the task is successfully completed in its entirety, before moving on to the next task/sprint.

B-4: Project Timeline & Milestones

The constituent elements of this project are projected to take approximately 80 hours. Following is a schedule of expected hours necessary for each project task:

Project Task/Component	Start Date	End Date	Projected Time Need	Milestone?
Scraper	5/31/2022	6/1/2022	4 hours	
Comment Parsing/Fixing	6/1/2022	6/2/2022	12 hours	
Generating of Dataframe	6/3/2022	6/3/2022	2 hours	**X**
Dataframe Fixer	6/6/2022	6/10/2022	16 hours	**X**
Standardization Operations	6/13/2022	6/17/2022	24 hours	**X**
Automated Scoring System	6/20/2022	6/24/2022	16 hours	**X**
Slide Deck	6/27/2022	6/30/2022	6 hours	
Export Operations	6/27/2022	6/30/2022	2 hours	

The primary milestones in this project are 1) the successful scraping and parsing of comments to successfully generate a dataframe, 2) completion of all dataframe fixing operations and allowance for standardization operations to begin, 3) completion of all standardization operations, and then 4) completion of the automatic scoring system. Each of these four milestones represents the completion of a large chunk of the project (16-24 hours) and allows for the following task to start.

B-5: Project Resources & Costs

The resources required for this project are:

- Approximately 80 hours of work time ($0)
- Python development environment via Anaconda ($0)
- Subscription to The Athletic ($72/2 years) to access the contest announcement page

The approximate total cost of this project is $72, the amount which I paid for my current 2-year subscription to The Athletic from Jul 2020 – Jul 2022. Other than the subscription cost to The Athletic, there is no other costs for this project.

B-6: Criteria for Project Success

For this project to be considered fully successful, the following must be accomplished:

- All contest entries must be scored correctly
- A CSV containing all contest data, including standardized entries and cosmetic changes made to enhance human readability, must be created and delivered to McIndoe

- Slide Deck containing descriptive statistics as defined within the project scope must be generated and delivered to McIndoe
- T-testing of the software-based scoring process against the manual scoring process must indicate that the software-based solution is faster

Part C: Data Analytics Solution Design

C-1: Project Hypothesis

Among other things, one of the goals of this project is to programmatically perform analysis of the contest data in a more efficient manner than a manual approach could have done, from collecting and standardizing the contest entries to scoring them and providing descriptive statistics about the entries and the results.

C-2: Analytical Method & Justification

This project will use a qualitative content analysis to provide various descriptive statistics about the qualitative data contained within the contest entries. This will be primarily done through in the slide decks provided to McIndoe, breaking down the nature of the contest entries (most/least common answers or selections per question, etc.), as well as the outcomes of the contest entries (distribution of points earn for each question, etc.).

Inferential statistics will be used to determine if this programmatic solution is faster than the original solution of manually scoring contest entries. Specifically, T-testing will be used to

compare the mean of one group (manual processing of contest entries) against another group (programmatic processing of contest entries).

C-3: Solution Environment & Tools

This project will use the following data analysis tools:

- Python & Jupyter Notebook via the Anaconda environment

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 and about the contest as a whole. 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, pandas, NumPy, and MatPlotLib libraries for Python

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.

- Microsoft Excel/Google Sheets

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.

C-4: Solution Statistical Significance

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.

$$H_0: \text{time}_{(\text{program})} - \text{time}_{(\text{manual})} \geq 0$$

$$H_1: \text{time}_{(\text{program})} - \text{time}_{(\text{manual})} < 0$$

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 submitted to the prediction contest. The time required for manual handling of contest entries can then be compared to the run time of the Python script to handle those entries programmatically.

To prove statistical significance of my results, I will use a standard significance level (α) of 0.01 and calculate a z-score to determine the number of standard deviations the observed datapoint is from the established mean. A p-value will then be determined for that observed datapoint. If the p-value is less than α (0.01), I will reject the null hypothesis in favor of the alternative hypothesis, that the programmatic data analysis solution is the more efficient solution.

The significance level (α) of 0.01 may be higher than needed (0.05 is a common standard), but using this more stringent level means that if the outcome is considered to be statistically significant, it is very strong evidence that this is actually the case. Given the nature of this project as a unique one-time study of this data, it seems reasonable to push for very strong evidence if I am to reject the null hypothesis. This method of attempting to disprove the null hypothesis by comparing manual analysis vs automated analysis of the data at hand will clearly demonstrate if the difference in time is a result of random sampling or if the new automated process is truly an improvement.

C-5: Solution Practical Significance

The practical significance of this project is to reduce the amount of time that McIndoe would otherwise have spent on analyzing the contest entries. Manual scoring is, by McIndoe's admission, a more laborious process than he had anticipated. The reduction in time needed to score the contest entries frees up time to be spent on other obligations. This is further evidenced by McIndoe's usage of the bonus question in this year's contest, as it was specifically intended to be an easy way to "void" a large number of contest entries, though it would not be able to be used until near the end of the season when the answers to that question were clear. Additionally,

the ability to standardize and obtain descriptive statistics regarding the submitted answers provides the opportunity to create content about the contest throughout the season.

For the previous year's contest, McIndoe was only able to write about the contest entries at the end of the regular season, because of how long it took to get those entries entered into a spreadsheet, even prior to the final scoring. Provision of this data and the associated analysis allows McIndoe to instead do multiple updates throughout the NHL season and to regularly engage with his audience about the contest via social media. Essentially, the data analysis provided by this project and the speed with which it is completed provides a tremendous amount of prompts for content creation throughout the 2021-22 NHL season.

C-6: Solution Visualizations

A normal distribution will be graphed of the times taken to manually score all contest entries, communicating both the mean time needed and the standard deviation. On this visualization, the time taken by the programmatic solution will also be graphed, demonstrating its position relative to the overall distribution. This will help demonstrate the statistical significance of the programmatic solution in a clear visualization.

Part D: Description of Dataset

D-1: Data Source

The source of the data used in this dataset is the 1600+ comments on McIndoe's 2021-22 prediction contest article, found at 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/. Contest entries are entered into the comments of this article by DGB readers, along with other non-entrant comments. This project will scrape the contest entries from a locally saved copy of this website, created on 12 Oct 2021 at 1930 ET when the contest closed to new entries, to avoid consuming excessive bandwidth or other resources from The Athletic's website.

To score the contest entries, other data will need to be collected to verify happenings from the 2021-22 National Hockey League season, including playoff qualification, coach & general manager changes, player statistics, awards voting, and player transactions. This data is published by the NHL and its member teams and collected in several locations. This project will primarily use https://www.capfriendly.com/ for transaction data, Wikipedia for hiring date information on coaches and general managers, and Professional Hockey Writers Association releases for awards voting data. The remaining statistical/team data will be gathered from https://www.hockey-reference.com/.

D-2: Suitability of Dataset

This dataset is the only collection of the contest entries. There is no other viable dataset that could be used to handle the analysis of the 2021-22 prediction contest and its entries.

D-3: Data Collection Methodology

A copy of the contest announcement webpage was made using Google Chrome (Right Click > Save As) at approximately 1930 ET on 12 Oct 2021, the time of the contest's closing to

new entries. A Python script will be created which uses the BeautifulSoup library to scrape entries from this saved copy of the webpage. BeautifulSoup will parse the contents of the HTML page, pulling out the contents of the tags containing the id “parent-comment-container”, which contains all elements of each comment made on the web page. The author of each comment will be extracted from the tag “comment-author-text” and placed into a list of authors. The contents of each comment will be extracted from the tag “comment-text-container” and placed into a list of comments.

Once scraped, a check will be performed to verify the list of authors and comments are of equal length. As all contest entries are at least 8 lines long, comments will be checked for the number of new line characters they contain (“\n”). Any comment containing less than 8 lines will have its index noted in a list. The noted indexes will then be used to remove both the comment and the associated author, in descending order to avoid moving the list’s indexes as it is being processed. A check will be performed for each author’s name for uniqueness. If an author’s name is not unique within the dataset, the index of that author’s position in the list of authors will be appended to the name.

For each comment, the comment will be split by lines and have numerous standardizations or checks performed against each line to verify the line does contain contest answers. Once the line is identified as having answers, it will be split into a list of those constituent answers. The associated comment author will be the first entry in a temporary list, to which each of the answers will be appended. Once the comment is done being processed, this temporary list will be appended to a dataframe generated from the pandas Python library, and a new comment will be handled by the parser function.

The majority of contest entries will be able to be programmatically handled, but some number of entries will need to be individually addressed to fix formatting issues which prevent programmatic handling. These entries will be modified programmatically, and the changes made will be recorded and documented to maintain accountability. The answers in the dataframe will then be standardized using Python dictionaries. Once standardized, the dataset will be complete.

Once complete, each entry may be scored, and the descriptive statistics may be generated. Because each question actually has three possibilities as it pertains to scoring (correct, incorrect, and did not participate), contest entries will account for this by using NaN (not-a-number) to represent answers which were not completed by the entrant. During the scoring process, while incorrect answers are penalized, NaN or unattempted answers will not be penalized and thus not counted as either being correct or incorrect.

D-4: Data Quality & Completeness

The raw data set from the contest announcement is fully complete, in that it is the only location where contest entries can be created, and entries do not have to be gathered from any other source. The unstructured nature of the data and its being wholly qualitative in nature does lead to it being of very poor quality. Addressing this is anticipated to be by far the most difficult and time consuming part of the project. Entries are at least formatted by having one set of answers on each line, per the contest announcement rules, but beyond that, there exists almost no standardization whatsoever. Issues with the raw data include, but are not limited to:

- Some entrants included extra line breaks within their answers (blank lines)

- Some entrants alternated lines, between writing the question and then the answers (alternating lines of value with lines of no value)
- Some entrants broke their answers up across multiple lines
- Some entrants submitted their answers in the form of "DET GM" rather than "Yzerman"
- Most entrants prefixed a question answer with data such as "1)", "Q1:", "Question 1:", etc., but this prefixing is inconsistent and not all entries use such prefixes
- Some entrants used non-English characters in player names, such as "è", "é", "or "ê", or similarly, various punctuation symbols such as a simple apostrophe, the "fancy" apostrophe, the hyphen, or even the tick mark (`)
- Delimiters between answers are inconsistent, often even within the same entry, as most entrants used commas, but slashes, dashes, periods, ampersands, semicolons, the word "and", hard tabs, or even no delimiter at all are all common cases as well
- Some entrants did not use periods as a delimiter, but abbreviated player names with periods, such as "C. McDavid, L. Draisaitl"
- Some entrants used city names for teams, while others used team names, or both, or abbreviations
- Some entrants skipped questions entirely, providing 0 answers to a question
- Spelling is extremely inconsistent, such as 86 different spellings of Tampa Bay Lightning goaltender Andrei Vasilevskiy requiring standardization to a single uniquely identifiable "Vasilevskiy" as the worst example, though answers using punctuation in a name (such as Rod Brind'Amour) are similarly problematic

- Most entrants included any commentary or additional information on their entries at the end of their entry, but others prefixed their entry with such, while a handful interspersed their entries with such commentary
- Several comments did not include entries at all, but instead were discussion or commentary on other user's entries
- Some entrants mistakenly included more than the maximum 5 answers per question
- Some entries were unclear, such as listing only a last name (for example, "Smith") when multiple players or staff members may have that name
- Some entrants were made on mobile devices, leading to autocorrect "fixing" names such as "Bednar" to "bed are" or "Bernard"
- Some entrants mixed up questions, such as providing answers to Q4 (general managers) in Q3, and answers to Q3 (coaches) in Q4

The biggest portions of this project involve addressing these issues, both to allow for proper parsing of the entries through issues with delimiters or line breaks into a dataframe and for effective automatic scoring of the entries by authoritatively standardizing each answer. In all cases, this script will refuse to reject a contest entry, as this authority is out of my scope and lies with McIndoe. As such, all attempted entries will be handled and included in the project.

Where interpretation of an entrant's intent is required, the script will generate a notice of this interpretation and documentation of the rationale for the outcome will be made. Interpretations of an entrant's unclear answers will always be made to be consistent with the most popular choices in the contest results. For example, the unclear "Smith" will be interpreted as Reilly Smith, a top line forward for the Vegas Golden Knights, rather than Brendan Smith, a

fringe NHL defenseman for the Carolina Hurricanes, because Reilly Smith has been more commonly selected in other contest entries as an answer to various questions.

The final output of this project will have thoroughly addressed these data quality issues. As a result, the resulting spreadsheet will be completely standardized and all data will be of high quality such that it could easily be used in any other analysis which this project may inspire.

D-5: Data Governance Concerns & Precautions

The contest entries themselves are all made in the form of comments made to the bottom of a web page on a subscription sports site, The Athletic. As a result, this data has all been placed out into the public by the contest entrants, which significantly mitigates data governance concerns for this project. The Athletic's commenting system partially anonymizes comment authors, by appending only the comment author's first name and last initial to a comment (for example, my comments are made as Joe T). The information provided back to McIndoe, such as contest entry scores, descriptive statistics regarding the dataset, and interesting observations I make about the dataset are thus entirely based upon publicly available information. In fact, these communications are actually intended for further publication by McIndoe himself, whether through written articles, podcast discussion, social media engagement, or any other means of content generation.

With no sensitive information or personal data (beyond first name and last initial) used in the project, and the personal data that does exist is consensually published by the contest entrants, there is no need for any special precautions to be taken with this data. To put it simply – anyone who paid for a subscription to The Athletic could go and get any of the information that

this project has organized, standardized, and analyzed. This is a positive in my view because it creates transparency and accountability by allowing for my work to be thoroughly checked by anyone with an interest to do so.

Part E: Sources

(Note: Because The Athletic is a subscription site and content is behind a paywall, I have generated PDFs of the necessary pages which are sources for this project and attached those PDFs to this submission.)

McIndoe, S. (2021a, April 27). *Down Goes Brown: Remember that prediction contest that was supposed to be easy? So far, it hasn't been*. The Athletic. https://theathletic.com/2543980/2021/04/27/down-goes-brown-remember-that-prediction-contest-that-was-supposed-to-be-easy-so-far-it-hasnt-been/ (Attached as 202021 Contest 02 due to paywall)

McIndoe, S. (2021b, August 3). *Down Goes Brown prediction contest results: Did anyone have a perfect entry? Would that be enough to win?* The Athletic. https://theathletic.com/2747031/2021/08/03/down-goes-brown-prediction-contest-results-did-anyone-have-a-perfect-entry-would-that-be-enough-to-win/ (Attached as 202021 Contest 03 due to paywall)

McIndoe, S. (2021c, October 10). *Down Goes Brown: Predict the NHL season with the return of the contest that's so easy it's almost impossible*. The Athletic.

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/

(Attached as 202122 Contest 01 due to paywall)

The Athletic Hockey Show. (2021, August 5). *Nathan MacKinnon's eating habits, John Tortorella joins ESPN, the taboo of gambling allegations and more* [Podcast]. Apple Podcasts. https://podcasts.apple.com/ca/podcast/nathan-mackinnons-eating-habits-john-tortorella-joins/id1546282862?i=1000531118508

Asdf

Asdf

Asdf

Asdf