COGS9: Introduction to Data Science
Final Project
Due date: Friday 2020 December 18 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
Brandon	Born	A14346622
Nicholas	Chua	A16584026
Matthew	Yom	A15436904
Jonathan Li	Li	A16687285
Austin John	Felices	A15692230

Question

Clearly state the specific data science question you're interested in answering. This question can be the same as what you submitted for your project proposal. Alternatively, you can edit your original question or change your topic completely. (2 pts)

To what extent did the Astros Cheating scandal in 2017 impact their batting statistics and records?

Hypothesis

Write down your group's hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?). (2 pts)

The Astros cheating scandal impacted their records, statistics in an observable way. Looking at the batting averages for the Astros' home and away games in 2017, we immediately find discrepancies between the 2 sets of statistics. We will examine this discrepancy and find statistical significance that further demonstrates their cheating. Our hypothesis is that we will find that the Astros' batting percentages were extremely high compared to other teams and compared to themselves in seasons where they didn't cheat.

Background Information

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources are fine. (3 pts)

In baseball, the catcher puts down a number of fingers to communicate to the pitcher what type of pitch to throw, whether it be a fastball, changeup, slider, curveball, etc. The ball becomes much easier for the batter to hit if he knows what type of pitch the pitcher will throw, as he will have a better gauge of where the ball will be, when it will cross the plate, etc. Teams try to steal these signs to let the batter know what pitch will be thrown. However, there are rules and limitations to this sign stealing.
https://en.wikipedia.org/wiki/Houston_Astros_sign_stealing_scandal

In 2017, the Houston Astros won the MLB World Series. However, a whistleblower on that 2017 team revealed that they cheated during the 2017 season, up to and including the 2017 World Series. The Astros stole catcher signs by putting a camera in center field and streamed the opposing teams' catcher signs into their dugout during home games and would relay them to the batter at the plate by banging on a trash can. The act of stealing signs in itself is not against the rules of baseball, however live streaming of opposing catcher signs and relaying them from the dugout is against the rules.
https://bleacherreport.com/articles/2862408-mike-fiers-admits-astros-stole-signs-electronically-during-2017-mlb-season

It's also important to note that the MLB had determined that they had cheated and punishments were given to the team as a whole as well as a few of it's executives and team members.
https://mlb.nbcsports.com/2020/01/13/mlbs-punishment-for-astros-was-both-harsh-and-not-enough/

Data

Include a description of the perfect dataset you would need 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)

Firstly, the perfect dataset would have to be separated by the games that the Astros did cheat in, versus the games the Astros did not. The number of observations would be the number of games in the 2017 to 2020 seasons, each teams' home versus away statistics in each season, and each player's home versus away statistics in each season. Variables to be collected would be batting averages, on base percentages, slugging percentages of all players, and the home and away records of all teams over a few seasons. It also would separate the batting statistics mentioned by the games that the Astros cheated in, versus the games they did not.

An online dataset that's available that could be used to answer this question would be the data from baseball-reference.com. This website contains all the statistics from current and prior seasons. The number

of observations this contains is the same number of observations described above, as it also contains the number of games in the 2017 to 2020 seasons, each teams' home versus away statistics in each season, and each player's home versus away statistics in each season. It also contains all of the variables we need in the batting averages, on base percentages, and slugging percentages, as well as each team's home and away records. The thing that differs from this dataset to the ideal one is that there is no way for us to tell which games exactly the Astros cheated in, which is why we separated it by home versus away, because the Astros scandal was conducted primarily at home games. It also does not look like they have the statistics we need separated by Home versus Away, so we would have to transform the data from Baseball-Reference in order to get the data into the form that can be used to analyze our question. We also understand that we should web scrape this page for the website since there is no provided CSV that's in tidy format for us to use. This would be an extra step in our data collection process.

Ethical Considerations

Read about Deon's data science ethics checklist and consider the topics discussed in lecture. Then, discuss what ethical considerations must be made when answering your specific data science question. Brainstorm and explain how you would address these considerations for each of the following categories in your specific project: Data Collection, Data Storage, Analysis, Modeling, Deployment. Feel free to write about additional ethical considerations you would make that aren't included on the checklist. Note that data privacy is NOT the only ethical consideration for a data science project. It is a piece, but there is a lot more that has to be considered. (3 pts)

One ethical consideration our group would make sure to avoid is p-hacking since we are fairly certain that the Astros' cheating methods positively influenced their performance. Because we know that the Astros had cheated, we will naturally gravitate towards making decisions that will yield results demonstrating what we already know (falling for confirmation bias). The way we would address this is to make sure we choose a dataset to test before analyzing it for potential discrepancies between home and away stats such as record, batting average, on-base percentage, etc. We would also be sure to test one variable at a time and complete the test(s) before looking at the results in order to ensure p-hacking does not occur unknowingly.

Another ethical consideration we have to consider is obtaining the data. For this analysis we would have to carry out web scraping on certain sites that contain large amount of baseball data such as Baseball Reference, however these sites do not like when people obtain their data without consent or in general at all so we would have to contact some representatives of certain websites and describe to them what we would do with the data. Another way we could obtain our information is from official Major League Baseball sources released to the public, such as MLB Network and the League Office. In particular, we will use the information gathered in the Commissioner's report released by the League after their investigations and use that as a guide to see what games we should look at, what sort of data can be used, etc.

One more possible ethical consideration is how we ensure fairness across groups. In the discussion section, we mention how it's possible that other teams had cheated during games as well. This can cause a discrepancy in our analysis as if we were to compare the Astros to a team that also cheated, the Astros'

statistics might not look as if they benefited that much at all. To address this, we need to be wary of which teams we compare the Astros to and possibly remove teams that were also confirmed to be cheating from our data to correctly show the benefits that the Astros gained from their sign stealing method.

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)

For our Data Analysis, we will be using the Python coding environment to incorporate these five different Data Analysis methods: Data Collection, Data Wrangling, Statistical Analysis/Tests, Predictive Analysis, and Data Visualization.
The Python libraries or packages we would use include Beautiful Soup for Data Collection and Data Wrangling; Pandas for Data Collection and Data Wrangling; Matplotlib for Data Visualization; Numpy for Statistical Analysis/Tests and Predictive Analysis ; Statsmodels for Statistical Analysis/Tests; scikit-learn for Predictive Analysis; and SciPy for Predictive Analysis.

Data Collection: For our Analysis, we will be web scraping from our website linked in the Data Wrangling section holding the dataset of the Astros batting averages for Home and Away Batting Averages, and other

MLB team's respective Home vs. Away Batting Averages. This website does not have any possible way to download its data that we need, nor does it have an API to easily obtain the data. Thus, we will have to web scrape in order to get the data we need from the website. A lot of data that we need is not transformed so that we can look at home versus away statistics, so it means that we will have to scrape data from the box scores of every game and then separate them ourselves in order to summarize or categorize the data in order to do the analysis we want to. This also means that there later on, we will have to do a lot of cleaning in order to get the data into a form that's useful. If it were easily downloadable or accessible through the API, there would be more of a guarantee that we could obtain better structured data. However, because we have to web scrape, a lot more cleaning and transformation will be needed as described in the data wrangling section since the data will be unstructured.

Data Wrangling: Using the data obtained from Baseball-Reference (baseball-reference.com), we will look at the offensive stats of the Houston Astros during their home games vs their away games. Because of how the data is set up on the website, it will require wrangling this data in order to get it into a form that would be useful to perform the statistical analysis and tests described below. Because we would have to web scrape the data, we will have to do a lot of cleaning of the data. Once we have web scraped and cleaned it, we also will have to transform the data since the website does not explicitly separate the data by home and away like we need. So we will have to transform a lot of the data in order to get that home versus away form that we need in order to perform an analysis or test on it. These variables of home statistics and away statistics will need to be calculated after obtaining the data from the box scores of every game, separating the type of statistic into home versus away, and then calculate the statistic for the entire season. Throughout this process, it will be important to follow the tidy data practices so that we can take specific parts of our data to perform multiple types of analysis and tests with ease.

Statistical Analysis/Tests: We will utilize a two-sample t-test to compare the batting averages between the Astro's home and away games during the year of 2017 when the Astros were accused of cheating. This is not enough to fully analyze our question, as when teams play at home, they generally have better statistics than if they were playing away due to other factors such as the crowd or fans. We address this in the other tests below.

We will also conduct an ANOVA test to compare the batting averages between the Astros Home and Away games with all other MLB teams' batting averages for Home and Away games during the 2017 season. We have over 30 teams that played during the 2017 season, so this test can include all such batting statistics mentioned such as batting average, on base percentage, and slugging percentage.

Another comparison we will make is comparing the Astros' batting averages in 2017 (Home games) relative to the Astros' batting averages in the time after they were caught cheating (Home games). We will use a two sample t-test for this comparison. We can also do this for away games comparing the 2017 season to other seasons. This comparison will help to better understand how a batter's knowledge of the pitch that's coming will affect their performance without the other factors such as a home crowd cheering them on compared to the first t-test mentioned above.

Predictive Analysis: Through using linear regression, we can compare the offensive stats (batting averages, on-base percentage, slugging percentage) of the Houston Astros and compare them to the stats of the rest of the league. We initially considered including raw numbers of hits (doubles, home runs, etc.), however we ultimately decided against this because the 2020 season was a shortened season due to the Covid-19 pandemic, so it wouldn't be fair to compare the raw numbers of the Astros 2020 season vs the 2017 season. Baseball percentage statistics are proportional, however, so they are more comparable between 2020 and 2017.

We will also implement a machine learning model that allows a user to provide hypothetical Astros game stats (including batting averages, on-base, on base percentages, and slugging percentages). We will change different features being entered such as the batting averages or slugging percentages of a given game to establish our initial model. Using a random portion of our data, we will train the model, and see how accurately the model works on new, unseen data, making changes to the model's features and hyperparameters as a result. This will be a continuous process of seeing results and integrating changes until we are satisfied with a working model. Using how accurate our machine learning model is at predicting popularity, we will be able to analyze and conclude that finding an Astros game played either home/away in 2017 is predictable based on pure quantitative data.
One major aspect that we will pay attention to is adjusting our hyper parameter for tree depth to an optimal number. By creating a graph that denotes the accuracy of both test and training using mean absolute error in comparison to increasing tree depth, we will find a tree depth that produces the best balance of accuracy among both the training and test data.

Data Visualization: To effectively visualize and emphasize any potential disparities of home vs. away statistics in the 2017 Astros organization, we would use a grouped bar chart to display batting averages, on-base averages, and slugging percentages of players.

We will also use two scatter plots: one scatter plot shows a correlation between the batting averages vs number of games won for all the teams including the Astros and another scatter plot shows a correlation between the batting averages vs number of games lost. We will be looking to see if the Astros data point in both scatter plots is a possible outlier from the rest of the MLB team data points.

Another effective visualization also uses time-series plots to see the trend of batting averages for Home and Away games combined for the Astros 2017 season and the trend of batting averages for Home and Away games combined for other MLB teams that played in the 2017 season. We need to be able to see significant differences between Astros' batting averages for the Home and Away games combined during the regular season and postseason, and be able to see how the trend in the Astros' batting average time-series plot compares with the time-series plots of other MLB teams' respective batting averages for Home and Away games combined. That means looking at the major spikes or trends that were higher in batting averages during certain time periods of Home and Away games played for all MLB teams including the Astros. We should see that the Astros' time-series plots are abnormal from the rest of the time-series plots of other MLB teams that played the 2017 season.

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)

We would interpret the results of our analysis by determining how much the knowledge of what pitch is coming as a batter will affect a person's batting statistics. Our statistical analysis tests will be able to tell us this. Specifically the t-tests, if seeing that in a season like 2020 their batting statistics were much lower than in 2017 when they were confirmed to be cheating, we can see the difference in the statistics and see how much sign stealing really helped the batters if at all.

One potential pitfall of our investigation is that it is usually the case that teams perform better at home than they do in other stadiums, so in order to detect any disparities in the data we would have to determine some statistically significant difference in the Astros' home and away records from the records of other teams or even in the Astros' past years.

Another pitfall would be that some pitchers may have been able to thwart the Astros sign-stealing scheme by switching to more complex signs. For example, during the 2019 World Series, the Washington Nationals used an elaborate system of confusing signs to throw off the Astros and ultimately win the World Series. This would potentially affect the data from specific pitchers or teams we compare the Astros to and potentially render our analysis not significant.

Yet another pitfall that we would need to take into consideration would be that there is a very good chance that other teams have cheated as well. For example, a further investigation revealed that the 2018 Boston Red Sox also cheated by illegally stealing signs as well, albeit not to the same extent as the 2017 Astros, as they stole signs illegally but relayed them legally. However, there hasn't been any confirmation yet, so we will give the other teams the benefit of the doubt and assume that they played the game cleanly.

A potential limitation of our analysis is that we would want to analyze data from 2017 when the cheating scandal was at its height versus data from the 2020 season when the news had been broken about the scandal. This would reveal if the Astros had any significant decline in playing performance at home versus their potentially heightened performance at home. The difficulty we face here is that the 2020 season was shortened due to COVID-19 and would not be an entirely representative data set for analyzing team home versus away record, etc. because of the lack of data a shorter season yields. In addition, there were no crowds allowed at stadiums for most of the regular and postseason games during the 2020 season, so it is unclear if there is even a home-advantage during this season to examine. We would have to address this by looking at multiple years or seasons of data.

Other factors that we need to take into consideration is changes in roster and players leaving/being added. While the Astros' central core lineup largely remained intact, there were certain players that left or joined such as Maldonado. This is another factor that we have to take into consideration, so the best way to compare the 2017 Astros to the 2020 Astros is to compare only the players that were on both the 2017 and 2020 Astros teams.

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)

Brandon worked to elaborate on ethical considerations in the final project and helped brainstorm ideas for different types of statistical tests and graphs we could use for the Statistical inference and Data Visualization portions of the analysis section. Brandon also described a couple topics in the discussion portion of the final project and discussed ways we would go about avoiding or correcting these problems if they were to occur. Austin contributed by forming the group during the initial weeks by connecting with everyone via email and eventually creating a Facebook Messenger group for all of us to communicate our availability to work on this project. For the project specifically, Austin typed about using the ANOVA test to compare the batting averages between the Astros and all other teams that played during the 2017 MLB season, wrote about using scatter plots to see if the Astros are an outlier data point to the rest of the other MLB team's batting averages in comparison to the number of games won or lost, respectively, typed about using time-series plots to see the trend of the batting averages over time and watch for certain spikes or trends in batting averages for both the Astros and other MLB teams, and typed about how we would be using the Python coding environment and the respective packages or libraries we would be using for the five methods of Data Analysis we would be incorporating for our project. Nick revised a lot of what we had in the Data section from the first assignment, as that was one of our weaknesses looking at the feedback from assignment one. Nick also worked in the Analysis section about web scraping and data wrangling because of how the data is somewhat hard to obtain in a format that's useful for analysis and added parts to the statistical analysis section. Nick also helped to brainstorm ideas for other possible analysis methods and wrote some of the ethical considerations. Matthew came up with the idea of doing our project on the 2017 Astros cheating scandal. Matthew did most of the hypothesis and the background information, and I contributed towards the Discussion, Ethical Considerations, and Analysis Proposal sections, using my knowledge of the game of baseball to decide what stats to look at, what data we should analyze, what games we should compare, etc.. Finally, Jonathan came up with the majority of the overall project pipeline. From web scraping data from the baseball website to thinking of all the Python libraries that we would use. Jonathan also suggested using both linear regression and decision tree regression in the predictive analysis section to use as models for machine learning. For the rest of the project, Jonathan made suggestions that contributed to allowing the variety of sections to flow well together and made sure the logic was sound.