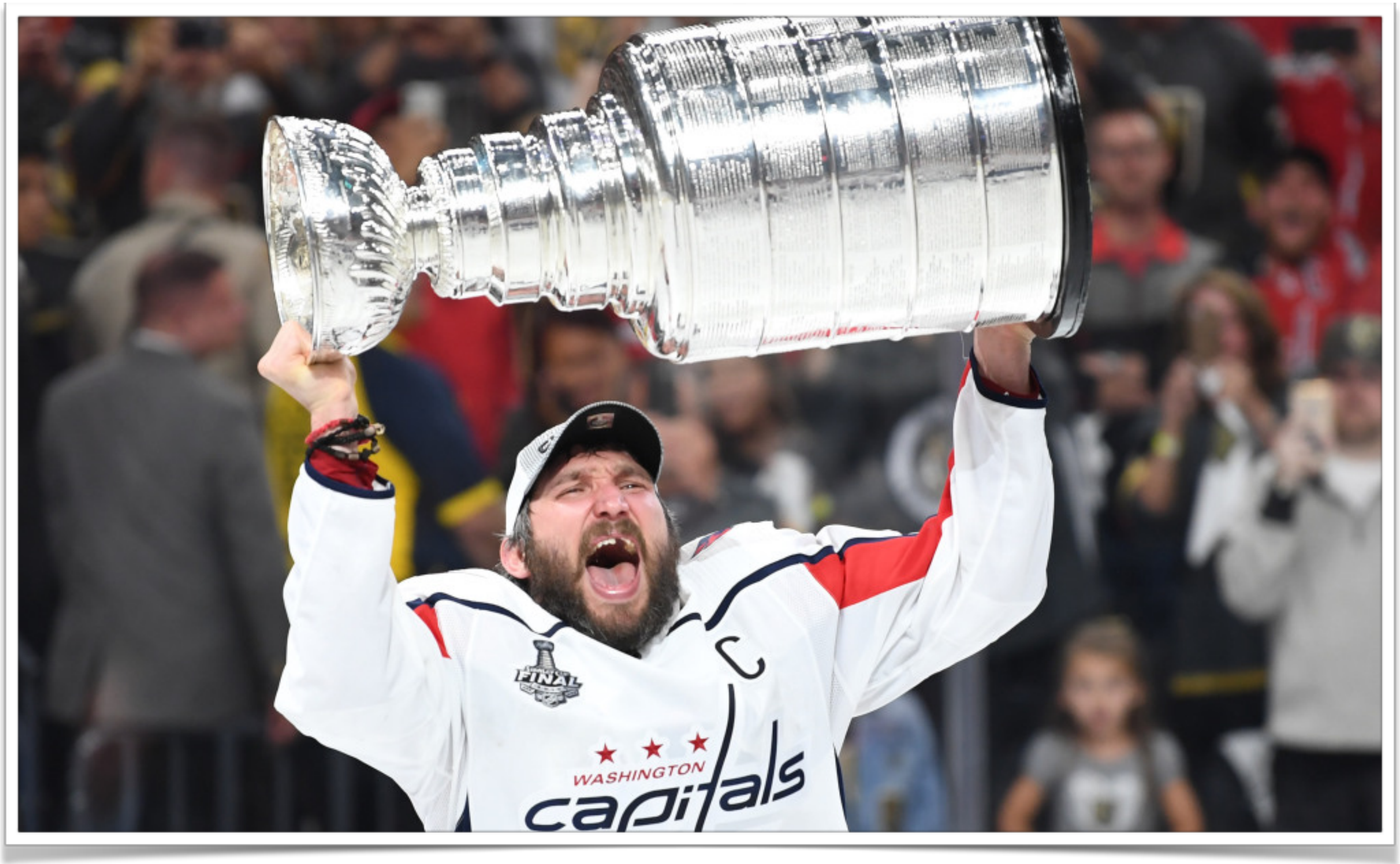

GW Data Science Capstone Pre-presentation

NHL Data API

What does it take to be a championship team?

Introduction

- The NHL publishes all of its robust data through an API on their website, the data includes play-by-play data dating back to 1917.
    - I intend to collect data back to the start of the Ovechkin era(2005-2006 season) in Washington, D.C.
    - The Washington Capitals have dominated the regular season for the majority of this time, yet fell short in the playoffs. The 2016 team, on paper, was the most talented team by far, yet they still failed to get past the second round of the playoffs. Then the 2017-2018 team filled with many new and young faces was written off as a rebuilding year team, but they won it all. My goal is to see if the data can tell us what made them different.

Problem Definition

- What variables are most important for a team to win the Stanley Cup?
  - Common data collected that could contribute:
    - Offensive - Goals, points(I'll define this but it is different than goals), shooting percentage.
    - Defensive - Hits, Blocked Shots, Goalie percentage
    - Win percentage(both home/away)
  - Data to look at that may contribute
    - Time on Ice for individual players
    - If there is performance changes due to player combinations
    - Penalty minutes per team/player

Methodology

- Using the NHL API I will harvest team, player and play by play data, and perform exploratory data analysis to identify predictor variables within the data.
- I will use a Generalized Linear Model/Logistic Regression to see if I can build a predictor model based on the identified variables.
- Additionally I plan to use a Bayesian GLM to see if I can see differences in the approach.
- I will attempt to wrap the results of my analysis in a dashboard for the presentation.
- Finally as close to the final presentation as possible I will harvest and clean the most up to date information for this season to run through both prediction models and report the results.(disclaimer - the hockey season will win through June, so I wont know until then if my predictions are accurate)

Challenges

- I have had some challenges harvesting the data
    - I have found several resources that have documented the api in detail, but it is still an undocumented resource and my inexperience with APIs in general and dealing with JSON data format has been an issue.
        - One of the resources I found is a python scraper package, that is very slow at pulling in the data, and a bit of a black box, the 2017-2018 data is good, but other data is not as robust - this could just be upgrades on the NHL side.
        - Data that I have pulled in from the API directly is in a heavily nested JSON structure, and I have had challenges unwrapping it, but I think I made some headway this weekend. Once I have a method worked out I will create an iterator to pull, clean, and store the data in .csv files
        - I am behind on my initial schedule, by about 2 weeks. I underestimated the work needed to get the data I want to use. I am confident once I have everything in and cleaned I will be able to make progress with the ML models.