Data Engineering Project

Colin Gallagher

Introduction

- The goal of this project was to create a data storage and processing pipeline that is designed in an efficient, modularized, and maintainable manner.
- Creation of a PGA Tour Statistics Dashboard web app
- Make a way for a user to visualize the data made available on the PGA Tour Stats website
- Useful for anyone interested in the game of golf who also wants to view the data in more visually pleasing way (not just numbers in a table)

Methodology

- Scraped statistics from the PGA Tour website using BeautifulSoup
- Chose 36 statistics of interest (all of them was too much for this project)
- With 36 base links, all other links to be scraped were generated
  - Each stat, year, and tournament had a code
  - For every stat, every year, every tournament in that year made a link
  - Total of 37,119 links
- A table was scraped from each link and saved as a dataframe
- This dataframe would continue to be concatenated until finished with a certain statistic, then saved as a table in a SQL database, repeat for every statistic

Methodology Continued

- SQL Database
  - 36 tables
  - 7,436,863 rows of data total
  - Each table is a statistic
  - Each row is a player’s stat, during a certain season/year, through a certain tournament
- Data Processing / Web App
  - Using Python, Pandas, SQL Alchemy, and Streamlit a web app was created for making visualizations
  - Data loaded in as Pandas data frame using SQL Alchemy, filters are applied using Pandas and Streamlit, and Streamlit is used for data visualizations and web app hosting

Results


Enter Golfer's Name
Rory McIlroy
Select a range of years
2009
2022
1950
2022
Years selected: 2009 to 2022
Scoring
Driving
Approach
Chipping
Putting
Money
PGA Tour Statistics Dashboard
Scoring
SG: Total
Scoring Average
Money
Career Earnings

Results

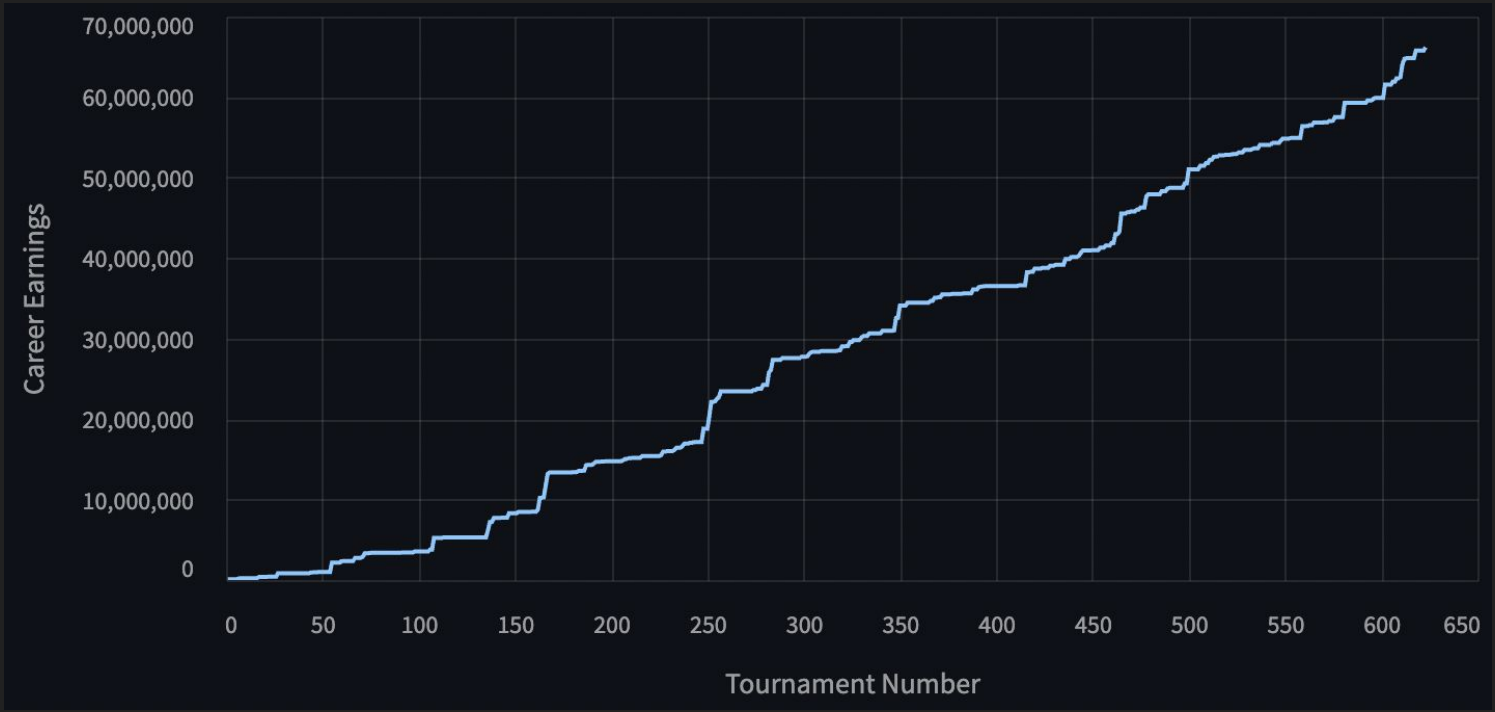
Money

☑ Career Earnings

Career Earnings

	Tour	Career Earnings	Year	Time Period	Tournament
0	1	90,222.0000	2009	Year-To-Date through	The Honda Classic
1	2	90,222.0000	2009	Year-To-Date through	Puerto Rico Open presented by Banco Popular
2	3	90,222.0000	2009	Year-To-Date through	World Golf Championships-CA Championship
3	4	90,222.0000	2009	Year-To-Date through	Transitions Championship
4	5	90,222.0000	2009	Year-To-Date through	Arnold Palmer Invitational presented by MasterCard
5	6	154,551.0000	2009	Year-To-Date through	Shell Houston Open
6	7	225,951.0000	2009	Year-To-Date through	Masters Tournament
7	8	238,548.0000	2009	Year-To-Date through	Verizon Heritage
8	9	238,548.0000	2009	Year-To-Date through	Zurich Classic of New Orleans
9	10	238,548.0000	2009	Year-To-Date through	Quail Hollow Championship

Results


70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
Career Earnings
0
50
100
150
200
250
300
350
400
450
500
550
600
650
Tournament Number

Conclusions

- A data storage and processing pipeline was created for PGA Tour statistics
- Data was scraped from PGA Tour website and a SQL Database was created
- The data from the database was used to create a web app to make visuals for the statistics using Python and Streamlit

Future Work

- Make more interesting visuals
  - Only have how stats change over time
  - Compare how player goes against the average
- Have access to huge database with lots of data, so a good start for making regression models for predictions for every tournament
- Figure out how to host entire database for Streamlit
  - Misunderstood how much could be uploaded to Github (25MB)
  - Streamlit can support up to 1 GB
  - Entire database is ~800 MB total
  - Only have one table (171,037 rows) on the Github hosted Streamlit app right now

Questions?