**Assignment 2: Data Collection and Curation**

**Name: Tyler Conger Email: Tyler-Conger@Tyler-Conger.com**

Statement:

For this assignment's preparation, the author(s) have utilized Chat GPT-3.5, a language model created by OpenAI. Within this assignment, Chat GPT was used for purposes such as brainstorming, asking specific code error questions, and help with proof-reading”

**Objective**: In this assignment, students will hone their skills in data collection, data analysis, and visualization. They'll use web scraping to gather data and then apply statistical methods to derive insights.

**Question 1: Amazon Best-Seller Rank Analysis**

**Objective**: Analyze the Amazon best-seller ranks for books written by Steven S. Skiena and recommend a book for purchase.

*Tasks*:

1. **Web Scraping**: Write a program to scrape the Amazon best-seller rank for each of Skiena's books. (Note: Ensure you adhere to Amazon’s terms of service and robots.txt when scraping their website.)
   * Tools recommended: Python, Beautiful Soup.
   * For each book, gather information on its name, ISBN, and best-seller rank.

In the web scraping phase, I manually curated a list of Skiena's books on Amazon using ISBN numbers from a national database. This meticulous curation ensured dataset accuracy and reliability for subsequent analysis. I used only resource to determine a list of ISBN numbers published by Skiena. This was cross referenced with the list of books under his Amazon seller list.

Refer to the accompanying image below for a visual overview of the web scraping process. This graphic illustrates the key steps, emphasizing the interplay between manual curation, ISBN validation, and systematic data extraction. Once the book was in the list I created the algorithm to scrape it gathering information on ISBN, best-seller rank and title. This data was all stored in an array for later plot.

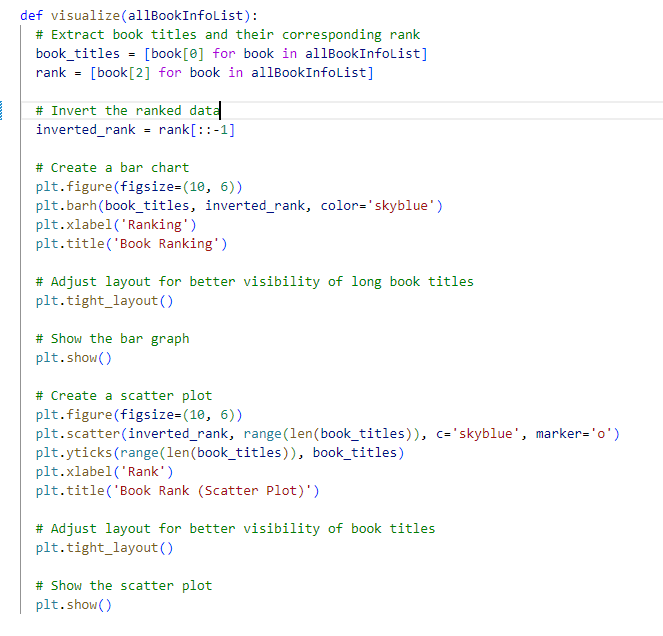
A white background with red and green text

Description automatically generated

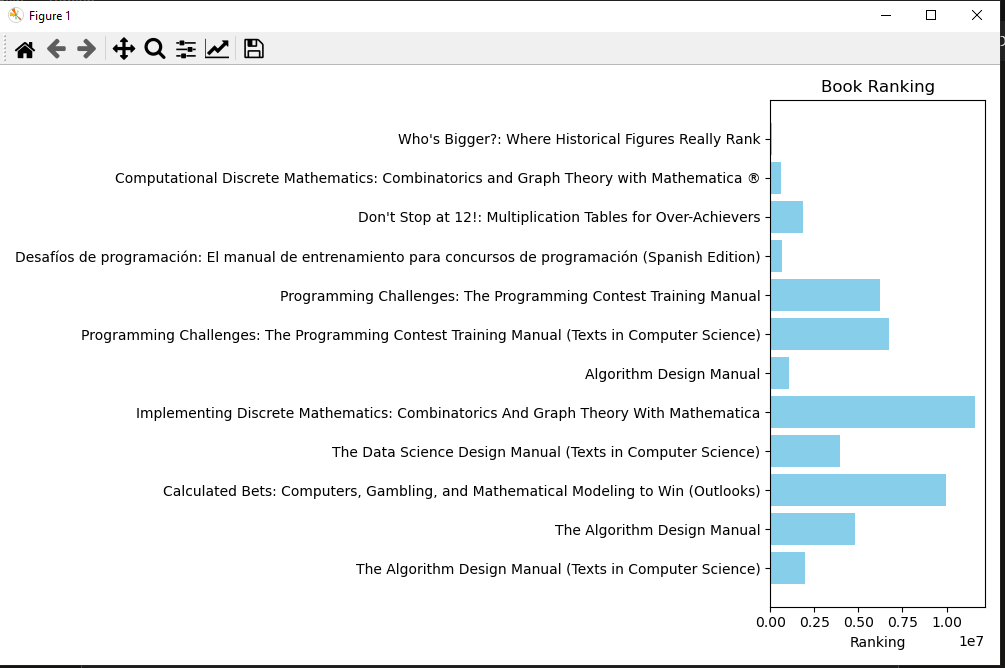
1. **Data Analysis and Visualization**:
   * Plot the rank of all of Skiena's books over time. Use appropriate visualization tools like line graphs or scatter plots.
   * Analyze the trend and discuss any patterns you observe.

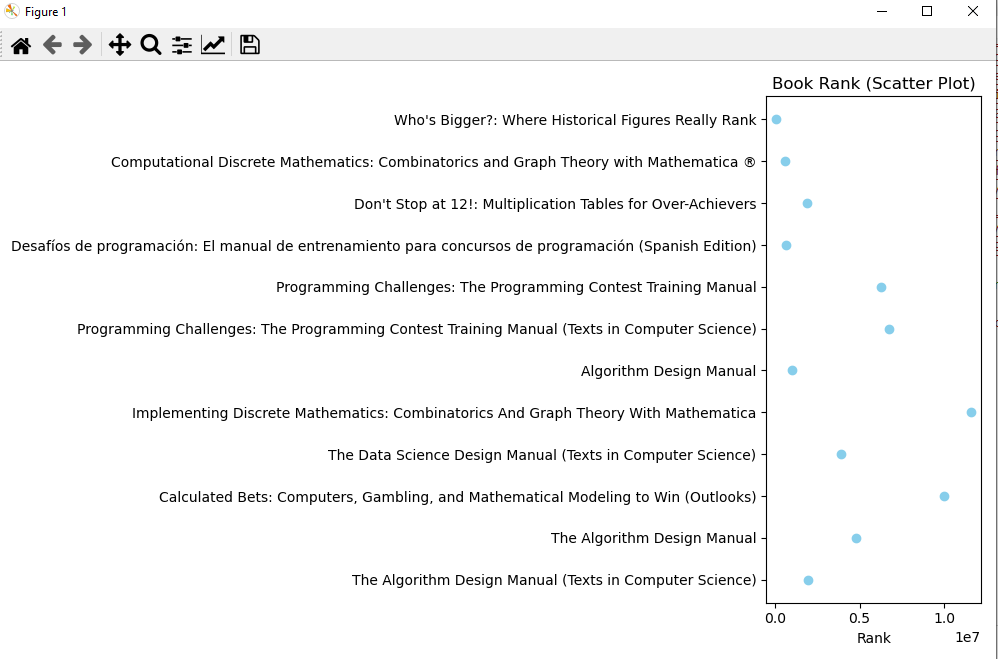
The next piece of the code takes the information of all the current books and utilizes a function visualize to plot the data graphically. It is important to plot the inverse of the seller ranking as if you just use the basic seller ranking it would show the worst selling as if it has the highest sales. So the lowest selling rank should be the most books sold. Below is the code that constructs the visualize function.

Below is the code snippet that defines the "visualize" function, offering a clear and concise representation of how the graphical visualization is constructed.



After the code is run it results in two graphs, one bar graph and one scatterplot graph. It is worth noting that the original assignment expressed overtime as a requirement however at the request of the instructor this is instead a one-time instance of the rankings at the current moment. As such each of the graphs express the ranking in real time as the code is run and does not track movement overtime (as this would require a longer time-frame of analysis)





As can be seen in the above figures we can see the difference in ranking between the best sellers and the worst sellers. This information will be further analyzed in part 3.

1. **Recommendation**:
   * Based on the rank trends, recommend which one of Skiena's books should be the next one to purchase.
   * Justify your recommendation.
   * Discuss if you believe any of these books would be an appropriate gift for a friend.

Analyzing the trends in book popularity reveals valuable insights into which books are gaining more attention and which ones are less popular. While I personally wouldn't recommend any of these books as a gift to a friend, if compelled to make a suggestion, the book of primary interest would likely be "Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica." Given its popularity, it stands out as potentially the best choice, presumably due to its overarching popularity when compared with the other books.

However, it's important to note that these books, being largely academic and centered around data science and mathematics, may not be universally suitable as gifts for friends. The content's technical nature might not align with the general preferences or interests of a friend. Therefore, when considering these books as gifts they may not cater to a broader audience's taste and a gift should probably not include homework problems.

In summary, while the most popular book may pique interest due to its high ranking, it's essential to assess the recipient's preferences before considering any of these academic and technical titles as potential gifts.

**Question 2: Player Ranking in Sports**

**Objective**: Devise a ranking system to determine the best players in a chosen sport based on historical statistical records.

*Tasks*:

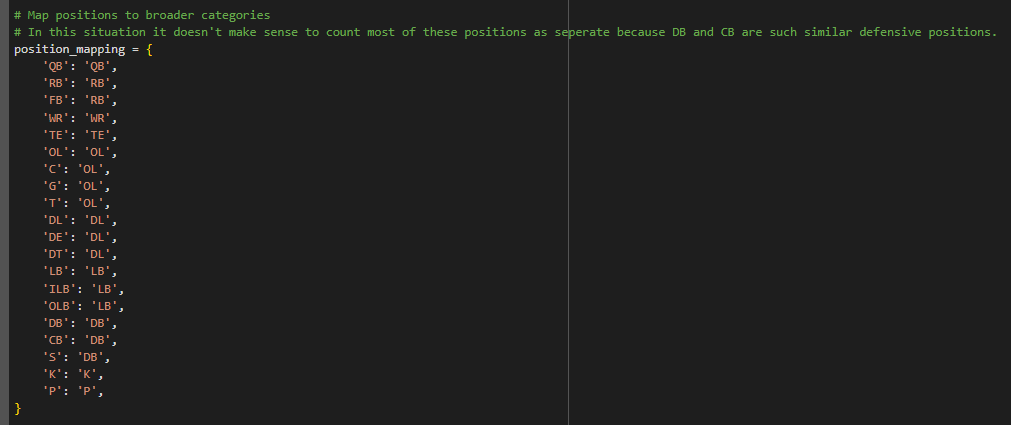
1. **Data Collection**:
   * Choose a sport: baseball, football, basketball, cricket, or soccer.
   * Identify a reliable data set or source with historical statistical records for all major participants in your chosen sport.
   * Tools recommended: Python, pandas, web scraping libraries if necessary.

The chosen dataset focuses on points scored for each football position, offering a comprehensive perspective spanning back to 1922. This historical coverage enables the ability to have an all-time ranking, without the limitations of a narrower Super Bowl-era only analysis. I will be leveraging the power of Python and Pandas, I aim to delve into the dataset, conducting thorough analyses and operations to unveil valuable insights. By opting for a dataset with detailed scoring information dating back several decades, we anticipate uncovering trends, variations, and patterns that span the entire history of football and not just more recent information on the way the game is played today. Seeing the entire history will allow for more interesting analysis as it will show trends over time showing how the game as changed and evolved with rule changes and to suit the taste of the public watching the game.

1. **Ranking System Development**:
   * Develop a system or algorithm to rank players at each position.
     + Define the criteria or metrics you'll use to rank the players (e.g., batting average for baseball, goals scored for soccer).
     + Discuss any weighting factors you'll apply to different metrics.

Players will be organized by their position groups, such that players on defense are organized together e.g. CB, DB, SS, S are all organized together and labled just as DB. Because each of these positions is very similar in terms of point scored and many players change between these positions in their careers. Whereas things like Linebacker are labeled together for example inside and outside linemen and tackles are all labeled similarly. This grouping can be seen below.

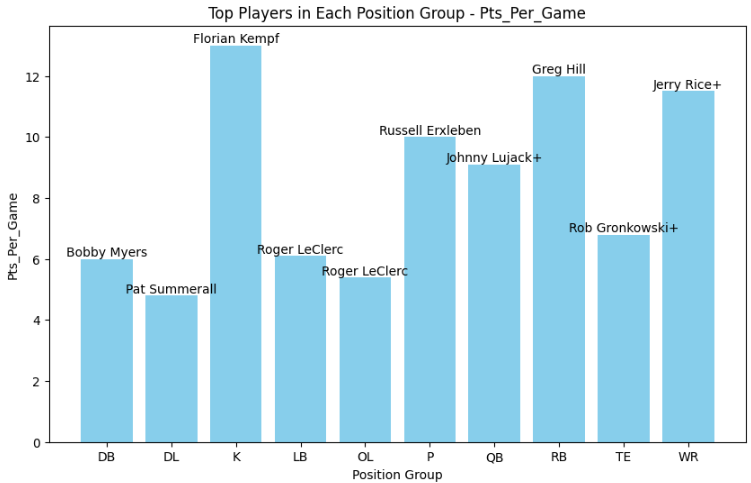
To condense the data and reduce the amount and difficulty of analysis and capture the inherent similarities between certain positions, players will be organized into position groups. For instance, positions on the defensive side, such as CB, DB, SS, and S, will be collectively labeled as DB, recognizing the analogous nature of their scoring patterns. This grouping approach accommodates the fluidity in player roles, acknowledging that many athletes transition between these positions over the course of their careers. Such as a player like Von Miller who has played in various roles on the defense. Similarly, positions like Linebacker will share a common label, encompassing inside and outside linemen, as well as tackles, simplifying the comparative analysis. This reduction can be seen in the image below, so readers are able to understand what category a specific player may fall into.



The ranking system adopts a points-per-game approach, providing a standardized measure for comparing players within each position group. This approach ensures fairness in evaluations, especially when considering the dataset's limitations. By focusing on points scored per game within specific position groups, the system facilitates meaningful comparisons and allows for a comprehensive exploration of performance dynamics across different positions. It is worth noting comparing players between positions would not make sense because of the way the game is played it only makes sense to compare players within a position group.

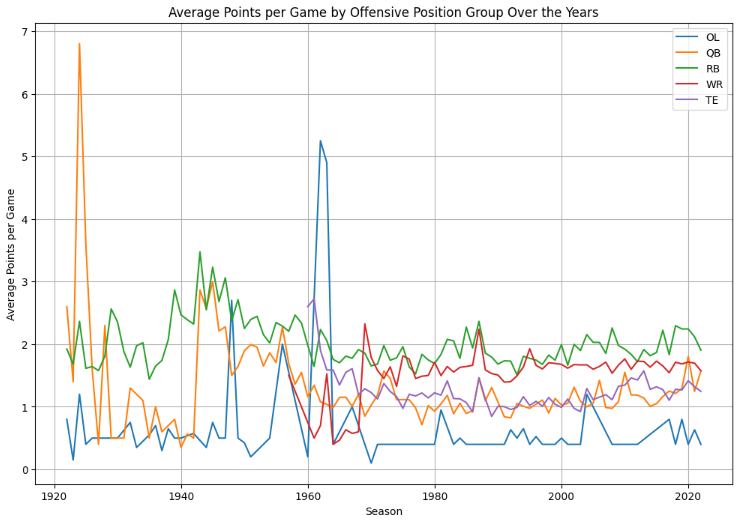
1. **Data Analysis and Visualization**:
   * Apply your ranking system to the data set.
   * Visualize the top players at each position using appropriate charts or graphs.
   * Discuss any interesting patterns or anomalies you observe in the rankings.

The data then undergoes a series of modifications and alterations using Pandas. The objective is to identify the top player at each position, leveraging the power of data manipulation and analysis facilitated by Pandas. This allows us to discover the top player at each position which can be seen in the graphic below.

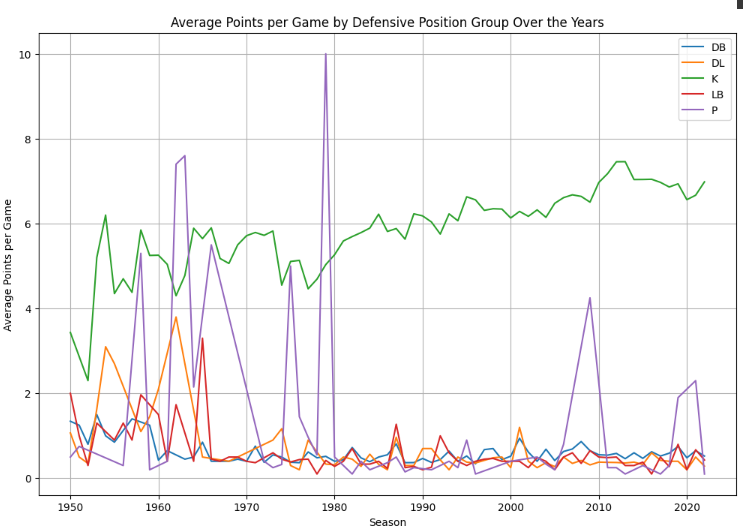


We can see each position and how many points they scored and who scored them at each position.

Next we compare information of scoring over the years by the best scorer at each position.



Offense



Defense

These last two are probably the most interesting of the two as they show how positions score over the years and how the game changes. At the beginning we can see a lot of chaos and a run heavy game with QB and RB scoring more than usual. Then we can see later the kickers begin to score more as the game changes more towards promoting field goals and extra points. We can also see the years the goalposts were moved back to decrease field goals and the adaptation thereof. We can also see points scored by defense and special teams players are generally low.

1. **Report & Discussion**:
   * Prepare a report summarizing your methodology, analysis, and conclusions.
   * Discuss the limitations of your ranking system and potential improvements.

Limitation wise this dataset has a lot of them. First, it really promotes scoring and does not count other traits. This can be an issue for players who have great pass block stats or something but don’t get pick-sixes. It’s also worth noting players play multiple positions, such as kick returner and DB, which gives them increased opportunities to score. Generally though we can see the big names in football here and make it clear that some of these players are truly great for their time. It would be nice to have a more complete metric possibly like fantasy points scored, because this dataset does cause issues. For example a running back may have a great game but if another back is used for goal-line drives the former running back may not score nearly as much as the latter.

**Submission Guidelines**:

1. **Code**: Include well-commented code for both questions. Ensure it's readable and modular.
2. **Report**: A comprehensive report discussing your methods, findings, and conclusions for both questions.
3. **Visualization**: Include relevant plots, charts, or graphs in your report.

**Assessment Criteria**:

1. **Code Quality** (20%): Readability, modularity, and efficiency.
2. **Analysis Depth** (40%): Depth and clarity of analysis, including the justification of choices made.
3. **Visualization Quality** (20%): Relevance and clarity of visualizations.
4. **Report Quality** (20%): Overall presentation, structure, and coherence of the report.