[IS477-FA25-Course-Project-Overview - Google Docs]

IS 477 Project Plan - Siddhu / Colin

Overview:

Our project aims to analyze NFL team salaries and how much each team allocates to each position group (quarterback, wide receivers, offensive line, defensive line, etc.) in relation to team performance in the regular season and playoffs. Our core objective is to see how efficiently teams are allocating their financial investments so that it best translates to on field success.

Because the NFL team uses a rigid salary cap that every team must abide by and cannot exceed, each team has the same financial resources. This makes the NFL the perfect example to study resource allocation and efficiency. Your success is not determined by how much you spend, but by how well you spend it. We want to focus on the patterns and trends that reveal what position groups offer the highest return on investment, or the greatest influence on team success per dollar spent. We also want to see if there are certain spending strategies that consistently correlate to success, such as quarterbacks on rookie contracts.

We plan to use multiple datasets to achieve this. Our first dataset will be a dataset that included every NFL team's spending information on a year by year basis over the last 15 years. It will have how much cap space was used on each player and position. The second dataset will include each team's performance over that same period. It will include regular season wins, playoff performance, and performance metrics that tell us how the offense or defense performed over the course of the season. By integrating these two datasets, we can explore spending efficiency over multiple seasons and see which positional investments lead to the most success both in the regular season and in the playoffs.

Our ultimate goal is to create an ideal allocation model that will demonstrate what percentage of the salary cap each position should take up in order to maximize performance and wins.

Research Question(s):

- How does total team payroll correlate to team performance statistics including wins, point differential, and offensive/defensive efficiency?
- Which teams achieve the highest performance per dollar spent, achieving the best salary efficiency?
- Are there patterns in salary allocation by position or role that correlate with higher team efficiency?
- How has team level salary efficiency changed over the past 10 years?

Team:

To stay aligned on the progress of our project and analysis of salaries correlating to team success, we think it is best to take everything task by task and work together. For certain portions, we will split the work as follows:

Siddhu:

- Data acquisition and web scraping, including python requests, BeautifulSoup, Pandas and API work, along with data integrity checks
- Data integration and storage, including data emerging and joining, schema design and organization, and version control in GitHub
- Visualization and communication, including Matplotlib, seaborn, Plotly, markdown and documentation and narrative storytelling

Colin:

- Data cleaning and preparation, including handling missing values in pandas, string manipulation, understanding data formats and data quality assessment
- Data analysis and statistical reasoning, including exploratory data analysis, efficiency metric creation, and basic statistical reasoning
- Workflow automation and reproducibility, including python scripting, reproducibility practices and metadata provenance

Datasets:

- https://www.spotrac.com/nfl/dallas-cowboys/overview/_/year/2025: Spotrac is an online sports news website and database that contains team payrolls, salary cap breakdowns and player contract info for all major U.S. sports leagues, and is a rich source of salary info for NFL players
- 2023 NFL Standings & Team Stats | Pro-Football-Reference.com: Pro Football Reference is a popular source of in depth data for the NFL, and it contains data for the standings at the regular season going back to when the NFL was formed, and includes the wins and losses for each team, playoff position (divisional, wild card, miss), offensive and defensive ratings and point differentials

Timeline:

Week:

1. Siddhu - We will start with finalizing our project goals and research questions. I will establish a folder structure so that we can effectively organize all of our material.
2. Siddhu- The next step will be data acquisition where I will collect salary cap data from Spotrac using web scraping and collect team performance data from Pro Football Reference.
3. Colin - I will assess and document the data quality. I will then clean and standardize the data sets and handle any missing or inconsistent data.
4. Siddhu - I will integrate the two datasets of different formats using pandas based on an integration schema that defines how the data from both sources align.
5. Colin - I will conduct exploratory data analysis by charting spending patterns per position group. Siddhu- I will help with visualizing different relationships between cap spending and team performance.
6. Colin - I will begin the efficiency analysis by developing efficiency metrics and identifying what spending patterns correlate with the best performance and team success. Siddhu - I will create visualizations to highlight key insights and showcase our conclusions.

7. Colin - I will automate the data pipeline using python scripts. I will also ensure that the entire project is reproducible by documenting our process. Siddhu - I will make sure our project includes proper metadata and data documentation.
8. Siddhu and Colin - We will both work on the final project report and make sure our project sufficiently meets all of the requirements.

Constraints:

- Salaries in the NFL can get complex, and our source (Spotrac) may have limited accuracy of salary data for bonuses, contract restructures or mid season signings, however should still be accurate when looking at a team level
- Teams may have different roster structures, leading to an apples to oranges comparison of pay mix by position group
- Team success metrics are influenced by many variables outside of salary composition, such as injuries, coaching quality and schedule strength
- Politics can play a significant role, where players with a strong following and recognition from the team's fans could be paid disproportionately due to non-performance factors
- Correlation vs. causation: This project only analyzes statistical relationships, it is difficult to prove that one salary mix directly causes team success although statistical relationships are very important
- Due to the amount of teams in the NFL, our sample size is inherently limited to 32, although taking data across several seasons may mitigate this issue

Gaps:

- Web-scraping: the data acquisition process for this project involves scraping data from 32 teams across several seasons, necessitating a repeatable data scraping process that is accurate and not overly burdensome. We will likely need assistance in coming up with the best way to accomplish this.
- Data integration: integrating data from two separate sources may require additional cleaning or transformation beyond our expertise
- Feature engineering: we still need to determine the best way to group positions and normalize salary data relative to team salary cap / total spending
- We considered going down to the player level, and incorporating individual performance but that would require much more complex data acquisition, integration and modeling