Report & Writing Guide

SECTION 1.1 — Research Goal & Workflow Overview

For this project, our group chose something that people argue about all the time in football — whether playing at home actually gives Premier League teams an advantage. We limited the scope to the seasons from 2020 to 2024 so we had a recent and manageable dataset to work with. Because this class focuses on Semantic Web ideas, we tried to rely on Linked Data first before using any spreadsheets or CSVs. That meant collecting club information from DBpedia and pulling the match details from Wikidata to see what was already available online. Along the way we realized that some seasons didn't have every match entry on SPARQL, so we added a backup plan using football-data.co.uk whenever things were missing. After we finished gathering and cleaning everything up, we passed the combined dataset to the teammates who were working on analysis and visualizations. The whole process felt pretty straightforward once we figured out the gaps in the data and how to work around them.

SECTION 1.2 — SPARQL Endpoint Exploration & Query Design

Before writing any actual SPARQL queries, we spent some time getting used to how DBpedia and Wikidata organize their information. On the DBpedia side, we eventually found that the dbo:league property linked to dbr:Premier_League was the simplest way to pull the clubs we needed. While checking a handful of club pages, we noticed different useful fields, like stadium names under dbo:ground and founding dates under dbo:foundingDate. DBpedia sometimes mixes in labels from different languages, so we had to filter for English labels to keep everything consistent.

Wikidata took a bit more effort because matches are modeled differently. We looked at a few example match items and slowly pieced together which properties described which parts. That led us to wdt:P6112 for the home team, wdt:P6113 for the away team, wdt:P585 for the match date, and the goal properties (wdt:P1350, wdt:P1351). To avoid pulling results from other leagues, we used a VALUES block to list only the Premier League seasons we cared about. During testing, we noticed that Wikidata was missing a noticeable number of matches for some years, so our script had to fall back to the CSV source way more often than we expected. We also wrapped optional properties in OPTIONAL to prevent errors when certain details weren't included. After trying a few variations of the queries and fixing some strange edge cases, we settled on three main SPARQL files: one for team information, one for match lists, and one for simple club statistics. These ended up becoming the basis of the rest of the project.

Data Analysis & Hypothesis Design (Section 2.1)

Starting with the topic of sports betting. Then it is specific to the Premier League football. In our project, the goal is to have the main idea of home advantage in the Premier League. We have developed two specific research questions that our collected dataset could address empirically.

Question 1: Are the teams in the Premier League creating a statistically important home advantage throughout the seasons of 2020-2024? Among them, which teams win more, whether home teams or away teams? This research question helps us examine whether the overall pattern of match outcomes is different. It also helps us to know what would be expected by random chance or equal performance. **Question 2**: To what extent does home advantage vary significantly across teams in the Premier League? For which clubs is the home advantage especially high? This question addresses variability at the team level by identifying clubs with unusually strong or weak home advantage effects. It uncovers whether home advantage is an undifferentiated phenomenon or depends on the characteristics of the teams.

In our analysis design, we employ a two-statistical approach to answer these questions completely.

Question 1: First, we start by calculating the overall home win percentage. Then moving to the away win percentage and the last draw percentage across all 1,520 matches in the dataset. We apply a chi-square test for independence to determine whether these proportions differ significantly from a null hypothesis of equally distributed outcomes (33.33% each if no advantage exists). **Question 2:** We compute individual home advantage scores for each team. It helps us to define the difference between home win percentage and away win percentage. We use a paired sample t-test to assess whether teams systematically perform better at home than away across the league. We filter teams to include only those with at least 10 home matches and 10 away matches to ensure sufficient sample sizes for reliable percentage calculations. All statistical tests use an alpha level of 0.05. Finally, we report both test statistics and p-values to assess the strength of evidence against null hypotheses.

This dual approach aggregates league-level testing combined with team-specific analysis, allowing us to establish both the existence and the magnitude of home advantage effects.

Data Extraction, Cleaning, and Validation (Section 2.2)

For the analysis phase of our project, my main responsibility was to prepare the cleaned datasets so they could be used reliably for statistical evaluation of home and away performance in Premier League matches. Although the data was collected earlier from SPARQL endpoints, the focus of my work in this section was on ensuring that the dataset was in a form suitable for meaningful analysis. I used **Python**, along with libraries such as **pandas**, **NumPy**, and **SciPy**, because these tools provide efficient ways to organize the data, compute summary statistics, and validate analytical assumptions.

Before any analysis could be performed, I verified that all numerical fields-such as goals scored, goals conceded, and win/loss counts-were correctly converted into numeric types. I also checked for missing or inconsistent values and applied filtering steps to remove rows with invalid dates or incomplete team information. These steps were important because statistical calculations, such as computing win rates or comparing average goals, require clean and consistent inputs. To keep the analysis reproducible, all preprocessing steps were done programmatically in Python so they can be repeated at any time using the same code.

Validation was an important part of this workflow. I made sure that basic analytical assumptions were satisfied-for example, verifying that home and away team identifiers matched across datasets, and ensuring that each match had valid goal values before being included in comparisons. I also checked that grouping operations (such as grouping by home vs. away results) produced expected counts, which helped confirm that the dataset matched the structure we intended to analyze.

Mark – Visualization & Interpretation (Section 3)

To represent the results of our investigation, I utilized Python's **Matplotlib** and **Seaborn** libraries to generate PNG images. The data was first aggregated into dictionary structures within our make_charts.py script, which then produced two distinct figures via the functions **plot_team_variance** and **plot_overall_stats**. These PNG images show a visual representation of the Premier League teams' home advantage, and a ranking system of Premier League teams with the highest home advantage to the lowest. I specifically applied a green-to-red color palette to clearly distinguish between positive home advantages and negative performance outliers. These static formats were chosen for their compatibility with our GitHub repository and presentation slides.

The function, **plot_overall_stats**, produced Figure 1, which reveals a 10% home advantage across the league. Although this may not seem significant, visualizing it provides a tangible edge, turning Sports Betting from pure gambling into educated

decision-making. Conversely, the function, **plot_team_variance (**Figure 2) highlights that this advantage is not universal; it exposes that Watford performed worse at home. I suspect this anomaly stems from factors like empty stadiums or general squad underperformance. In conclusion, visualization is a powerful tool for Sports Bettors, as it highlights specific anomalies that raw data might hide.

FIGURE 1

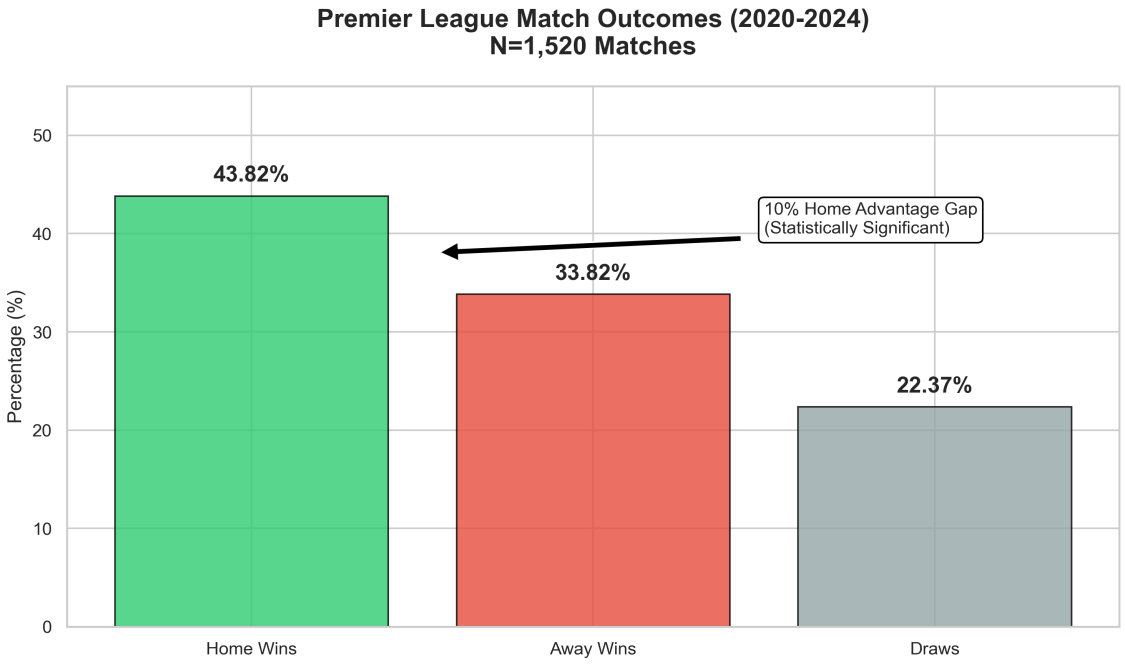

FIGURE 2