Wharton UNIVERSITY of PENNSYLVANIA Sports Analytics and Business Initiative

Wharton High School Data Science Competition

ICE HOCKEY PERFORMANCE PREDICTIONS

Presented by Google Gemini

WORLD HOCKEY LEAGUE

2026 Workbook

The purpose of this workbook is to allow you to develop your approach and work through Phase 1 as a team. However, the Student Team Leader is responsible for submitting your team's Phase 1 answers in SurveyMonkey Apply.

Competition Prompt

You’re the analytics staff for the World Hockey League. Your group is preparing for the World Hockey League (WHL) Tournament, and the WHL commissioner has a set of tasks for you. You will explore the numbers from the current WHL season to make predictions of what teams will likely succeed in the tournament. In this competition, your group will step into the role of sports data scientists to predict tournament matchups and dive deeper to uncover features underlying team performance.

Your challenge is to formulate your responses based only on the provided, fictional World Hockey League data set. While these data are meant to reflect real-world ice hockey, they do *not* represent any actual teams or leagues.

**Note that all data is simulated and does not represent real-world ice hockey teams or outcomes.*

How to Succeed

Success in this competition is about showcasing your skills in three key areas:

1. **Accuracy:** How effective are your matchup predictions and rankings of team depth? How clear was your visualization of the impacts of team depth on game outcomes?
2. **Methodology:** How well-thought are your quantitative approaches to creating matchup predictions and offensive depth rankings? How clear and accurate is your visualization? Can you explain your approach clearly and justify your choices with sound reasoning?
3. **Communication:** Do your findings tell a compelling story, both visually and in your presentation?

This is more than a numbers game—it’s a chance to turn data into strategy, strategy into predictions, and predictions into high-level sports analytics grounded in real-world ice hockey dynamics.

Main Competition Tasks

The competition will take part in 3 phases:

Phase 1: Main Competition

Analyze stats from games across a season of the World Hockey League to rank teams and predict which teams would win in hypothetical matchups, submitted via online platform SurveyMonkey Apply.

Phase 2: Semifinals (top Student Teams invited to participate)

Creation of a short slide deck that describes and visualizes the team’s methods and findings from Phase 1. Additionally, Student Teams will explore effects and impacts of line strength disparity. Slide deck will be submitted via SurveyMonkey Apply.

Phase 3: Finals (top 5 Student Teams invited to participate)

Presentation of the Phase 2 submission material to a panel of judges during a virtual meeting.

Competition Structure

About the World Hockey League

Your group will be crunching the numbers from the most recent season of the World Hockey League. You will receive one season of data, including 32 teams and 82 games per team, yielding 1,312 total games. For each game, you will see multiple rows depicting game stats broken down by matchups of home and away teams' offensive lines (first and second) and defensive pairings (first and second). Note that there are no changes in team or line quality over the course of the season, and the regular season is representative of the playoffs.

See the educational modules for more information.

Phase 1: Main Competition

The competition heats up as you prepare to dive into the action of a WHL tournament! Your mission: as an analytics group, crunch the numbers to provide an outlook for the WHL tournament, using the provided in-season data.

Your strategy and predictions will be submitted through the online platform SurveyMonkey Apply. This is where the groundwork is laid, and the sharpest solutions will advance to the next stage.

Phase 1a: Team Performance Analysis

Before diving into Phase 1a, first create a league table outlining teams' overall standings for the season following the online educational module using your method of choice (R, Python, Google Sheets, GenAI/LLM). The league table will not be submitted on its own, but it is a necessary first step to creating team power rankings and predicting game outcomes for your Phase 1a submission.

Create team power rankings

Based on season data, rank the 32 teams using the underlying team performance. But this isn't just about win-loss records. Your rankings should reflect the *overall strength and quality* of the teams.

Submission: Power ranking of all 32 teams submitted by Student Team Leader via SurveyMonkey Apply

Predict game outcomes and win probabilities
For the first round of the World Hockey League tournament, predict the win probability for the home team for 16 matchups.

Submission: Winning probabilities of home teams in 16 matchups submitted by Student Team Leader via SurveyMonkey Apply

Phase 1b: Line Performance Analysis

Quantify team offensive line quality disparity
Hockey teams use multiple offensive lines and defensive pairings. The first offensive line is often more productive than the secondary lines — meaning they tend to generate more scoring opportunities. Your task is to quantify how large that disparity is for each team.

First, for each offensive line on each team, form an *offensive performance measure* based on expected goals (xG). Consider accounting for differences in time on ice so teams or lines that play more don't automatically appear better, and defensive matchups since tougher opponents can affect performance.

Second, for each team, compare performance of the first line to the secondary line by calculating a ratio of the first line's offensive performance measure to that of the secondary line. This ratio represents the team's *offensive line quality disparity*. Which teams have the largest offensive line quality disparity? Rank teams from largest offensive line quality disparity to smallest.

Submission: A list of the Top 10 teams with the largest offensive line quality disparity ratios (ranked 1 through 10) submitted by Student Team Leader via SurveyMonkey Apply

Phase 1c: Data Visualization

Communicate the impact of offensive team line quality disparity
The WHL commissioner wants to know if teams with more evenly-matched offensive lines are more likely to succeed. Your task is to provide insight into how offensive line quality disparity (calculated in 1b) may relate to team strength (calculated in 1a) in a data visualization.

Your visualization should:

- Clearly communicate your main message or finding.
- Use labeled axes, titles/subtitles, captions, and legends to make your graphic understandable without additional explanation.
- Reflect sound data practices as described in the educational modules.

Submission: single PNG file upload (max file size 5MB) titled with your team name without spaces (e.g., DataScienceTeam.png) submitted by Student Team Leader via SurveyMonkey Apply

Phase 1d: Tell Your Story - Summarize Your Methodology

Approach this as if your group is presenting its findings to the WHL commissioner, breaking down how your insights can shape strategy and drive success while telling a clear story. Your task is to **craft a concise but comprehensive explanation of your approach**—how you analyzed the raw data, identified key drivers of WHL outcomes, and predicted matchups.

Ensure your explanation is detailed enough that another analytics group could replicate your methodology, showcasing the rigor and transparency of your work.

Your summary should cover these key points:

1. Process:
   - How did you clean or transform the raw data before analysis? Please describe it in around 50 words.
   - Did you create any additional variables? Please describe in around 25 words.
2. Tools and Techniques:
   - What software tools did you use? [Select all that apply]
   - How did you use the selected software, and for what purposes? Please describe in around 50 words.
   - What statistical methods did you employ? Please describe in around 100 words.
3. Your Predictions:
   - 1a: How did you determine power rankings and matchup win probabilities? Please describe in around 50 words.
   - 1b: How did you approach summarizing offensive line quality disparity? Please describe in around 50 words.
   - 1c: What data visualization choices did you make to clarify the numbers and emphasize takeaways? Please describe in around 50 words.
4. Your Insights:
   - How did you assess your model performance? Please describe in around 50 words.
   - Did you use Generative AI tools (ChatGPT, Gemini, etc.)? If so, explain how– describe each stage where AI tools came into the process and how your group approached using them. Please describe in around 50 words.

Submission: Form question text responses submitted by Student Team Leader via SurveyMonkey Apply

Data Roadmap

Competition Data and Resources provided to participants:

To develop accurate rankings and predictions, your team will work with multiple data sets containing game statistics, team performance metrics, and matchup details. By analyzing these data sets, your team can identify trends, evaluate team strength, and explore ice hockey performance. Individual files are linked below:

Primary data set, one season of games:
Each row is the game's summary for a team - line and line pairings with descriptors of the game outcome. Each game has approximately 18 rows for the different pairings. Full season n=1,312 games.

- Game data set (whl_2025): Primary data set, with game level line summary data
- WHSDSC_2026_DataDictionary: Data dictionary for fields in the game box scores file
- WHSDSC 2026 Glossary: Glossary of terms

Tournament teams to predict:

- WHSDSC_Rnd1_matchups: Round 1 game matchups to predict in Phase 1

We encourage you to download and save the provided data sets. However, this PDF is provided to ensure access to the entire Phase 1 competition prompt, which contains important competition details, including data set descriptions, links, and key guidelines.

Additionally, all competition materials, including the **data files and project workbook**, will remain accessible throughout the competition for reference.

Use the links below to download all of the files mentioned above. **Files are provided in two ways**:

- **Box WHSDSC 2026 Hockey** folder (full link URL: https://upenn.box.com/v/wsabi-hsdsc-2026)

- **Google Drive WHSDSC_2026_CompetitionPackage** folder (full link URL: https://drive.google.com/drive/folders/1WMQFKMW7yEixEfZCtTMTtS2K80IyGnRM?usp=sharing)

Educational Modules

We are happy to provide you with some resources to get started.

This competition isn't about how much you know about ice hockey or stats about your favorite team. However, we've created this video library so you can learn more about what terms mean and how they may relate to the competition. Our video library is a tool for you to reference throughout the competition.

Click the Video Library to access a playlist of all videos.

Educational Modules
Educational video from Google Gemini
Overview of WHL & game data
Team organization
Hockey stats: shots, xG & goals
League Table: Overview
League Table: Google Sheets (w/ quick dataviz)
League Table: R (w/ quick dataviz)
League Table: Python (w/ quick dataviz)
League Table: GenAI/LLM (w/ quick dataviz)
Probability: logistic & ELO
Confounding
Data Visualization

These materials were developed for educational purposes associated with the Wharton High School Data Science Competition only. Please do not reuse, reproduce, or distribute outside of your school.