CMPT353 D100

Final Project

How Travel Impacts NHL Game Results

Derek Huang, 301441129

Professor – Greg Baker

TAs – Shahrzad Mirzaei, Fatemeh Movafagh, Wanying Tian

August 13, 2025

Executive Summary

This project examines whether travel burden is associated with team performance in the NHL. Using schedule data and arena coordinates to compute leg-by-leg travel distance, and extracting features such as rest days, consecutive away games, and time zone jumps, I explore schedule equity and if the distance traveled correlates with win percentage. From my exploration of data between 2022-2025 (3 NHL seasons), there is no suggested correlation between win percentage and travel distance. However, a key finding is that home teams win more on average than away teams for all 3 seasons analyzed. While distance traveled doesn't have a strong correlation, away teams are still more likely to lose. Travel distance could still be an attributing factor to that statistic.

Motivation

For this project, I wanted to work with the NHL API data, but it seemed like many common projects already focused on player and team skill metrics (e.g., predictions, player stats). Travel is an ever-present constraint but rarely reviewed in sports analytics. While anecdotally, long flights and hard travel conditions may influence performance, there are few analyses that quantify its effects. This project aims to quantify travel burdens and test whether travel distance and rest patterns measurably change win probability.

Problem Statement

The core question that I explored was **does travel-related fatigue affect game outcomes in the NHL?** Therefore, the H0 is: Travel distance has no effect on win probability. And the Ha is: Greater travel distance for away teams is associated with lower win probability. During the process of cleaning the data and exploring the transformed data, I also stumbled across some interesting secondary questions including:

- Is travel equity uneven across teams?
- Do rest days mitigate any distance related effect?
- Do long road trips have a stronger effect on game performance?

Data and Sources

Sources:

1. NHL API (Zmalski / NHL API Reference): https://github.com/Zmalski/NHL-API-Reference?tab=readme-ov-file
2. Kaggle Dataset for stadium data: https://www.kaggle.com/datasets/logandonaldson/sports-stadium-locations
3. Team-abbreviation mapping: A compact, semi-manual CSV to normalize team names to abbreviations (32 rows for 32 teams): *team_abbrevs.csv*

Since the NHL API does not directly expose arena coordinates, the kaggle dataset is used to look up coordinates for each arena. However, this mapping wasn't perfect as the Kaggle dataset only provided columns for team name and arena coordinates. This meant that I used the team name as the merge key, which caused problems with special venues (e.g., Stadium Series, NHL Winter Classic) which will be discussed later on in the limitations section.

The data processing step involved 3 key steps including extraction, cleaning, and feature extraction. Each step was split into separate files for modularity and scalability. Before starting off with the data extraction, I explored the various endpoints and the vast data options through reading the documentation and a

python notebook. After retrieving data from the endpoint, I used json_normalize function to explore the data further.

Extraction

The extraction files are named as *get_nhl_schedule.py, get_nhl_standings_data.py,* and *get_nhl_stadium_data.py*. The first file takes an input of a season string in the form of 'yyyyyyyy' (e.g., 20242025), and outputs the schedule data for the specific season.

Combine/Cleaning/Data Validity

This step was done implicitly in the *combine_data.py* file. This step normalized team identifiers in the stadium dataset using *team_abbrevs.csv*. Additionally, to capture neutral/alternate sites, I manually inserted the rows for stadiums that did not exist in the kaggle dataset. This overrode the team name join which would have assigned incorrect coordinates to special venues. Additionally, I wrote a script *validate_data.py,* which makes sure every column and row is accounted for after a merge. The output of this step is *merged_schedule.csv,* and *nhl_standings_merged.csv*.

Feature Extraction

After the first two steps, I use *generate_features.py* to perform the feature extraction step. This step involves creating new columns that will be used in the analysis section. The features that I produced are:

1. Travel distance (km): Haversine distance between consecutive venues for the same team. For this calculation, I needed to duplicate the original dataset (originally 1 row for 1 game) so that each team owned a unique row of a game.
2. Consecutive away games: Running count for away-game streaks
3. Rest days: This was calculated with the formula of game_date1 - game_date2 - 1, since the game_date shouldn't be considered as a rest day (e.g., rest days between 01-01-2025 and 01-03-2025 is 1).
4. Goal difference: To measure the severity of a game result. For example, an 8-1 loss is worse than a 1-0 loss.
5. Is_away flag: To be used in the analysis to quickly identify rows of traveling teams.

Analysis Techniques

The clean dataset enabled descriptive and statistical analysis. Visualizations and analysis was performed through some trial and error in a Jupyter notebook called *analysis_notebook.ipynb*. I explored win percentage in two ways. The first was overall win% by travel distance, which pooled all teams together. The second was a steam-specific travelling win%, which is a more granular view of who performs better or worse while travelling. Travel was measured both quantitatively through continuous distance between venues and qualitatively through the is_away flag. To contextualize travel effects further, I incorporated consecutive away games and days of rest prior to each game.

Through trial and error, I found that using bins was the most effective way to analyze distance. At first, when grouping by exact distance, it resulted in many noisy visuals with the 0/1 win rates for many exact distances. Binning distances yielded clearer, more stable estimates of win% and goal differential. Ultimately, the binning approach led to more readable figures which were better suited for comparing patterns across various conditions.

Most of the analysis was done using group by to compress the main feature dataframe which was extracted from *team_games_features.csv*. Along with that, the win/loss flag in the database made it easy to simply calculate the win% for each distance bin.

Findings and results

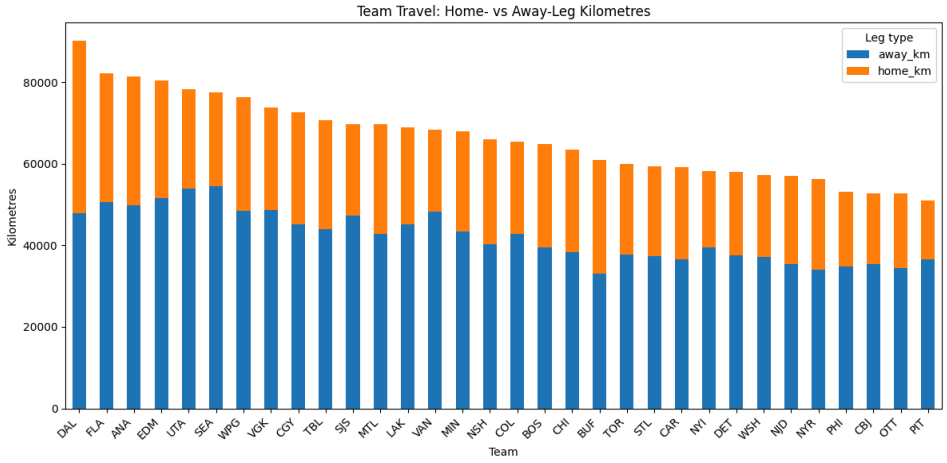

Fig 1: Stacked Bar Plot for Total Distance Traveled per Team (2024-2025 Season)

This figure show us the total kilometres traveled by each team throughout the 2024-2025 season, split into home and away travel distances. The chart highlights the disparity in the schedule making where the difference between Dallas' distance traveled and Pittsburgh's distance traveled throughout the season is close to 40,000 kilometres or a 43.4% difference in total travel distance. Similarly, the travel disparity for away-leg travel is a 39.3% difference. While this confirms travel inequity across the league, further analysis is required to determine whether these disparities create a competitive advantage.

season	is_away	win_rate
20222023	False	0.5236
20222023	True	0.4764
20232024	False	0.5412
20232024	True	0.4588
20242025	False	0.5625
20242025	True	0.4375

To explore whether travel disadvantage appears at a broad level, the table above compares win rates for home and away teams across the 3 seasons analyzed. In every season, the away teams won less often than the home team, with the home win rate advantage ranging from ~4.7 to 12.5 percentage points. This alsgns with the well established home ice advantage in hockey and suggests traveling teams are at a disadvantage. But it does not tell us whether the distance traveled is a key factor in this effect.

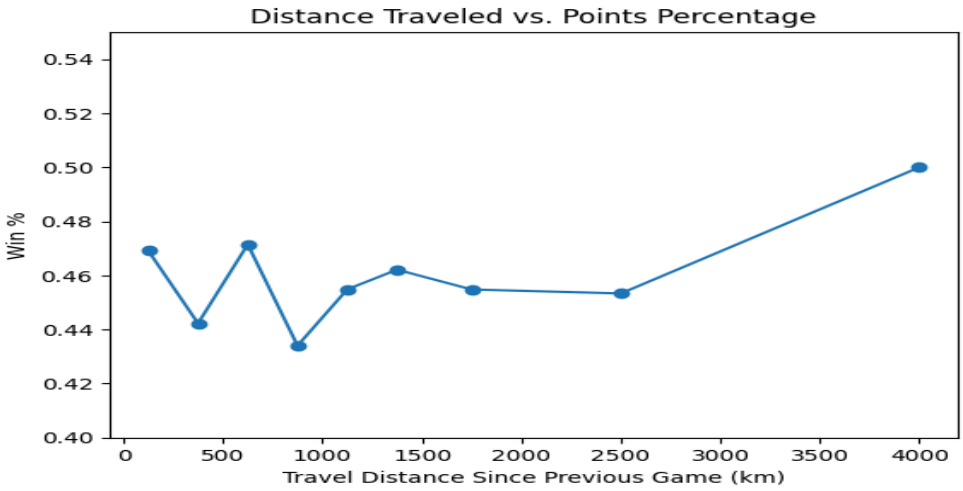

Fig 2: Travel Distance Bins vs. Overall Win % (2022-2025)

This figure shows the aggregated win percentages grouped by travel distance categories. No clear trend emerges, suggesting distance alone does not predict outcomes. I did a similar visual and analysis for goal difference for each travel distance bin which produced nearly identical line shape, reinforcing that the absence of a relationship is consistent across both win rate and scoring margin.

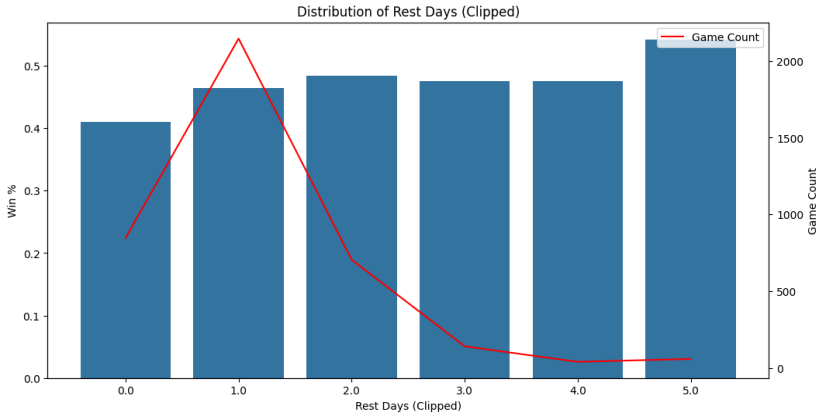

Fig 3: Travel Rest Day Impact vs. Overall Win % (2022-2025)

This figure demonstrates that rest is a more important factor than raw distance. Teams with more rest while traveling tend to win more often. While the NHL schedule generally provides at least one day of rest between games, back-to-back games are still common. The data suggests that when teams travel with more rest, their probability of winning increases, suggesting that recovery time may mitigate some of the fatigue associated with travel.

Figure 4 below examines how consecutive away games interact with travel distance to affect win percentage. Games were grouped both by travel distance bins and by the number of consecutive away games in the current road trip. The results suggest that win rates tend to decrease as the number of consecutive away games increase, particularly for games involving medium travel distances. However, the trend is not perfectly consistent across all distance classes, indicating that while road trip length may amplify travel fatigue, other contextual factors such as opponent quality, rest days, and schedule timing also play important roles.

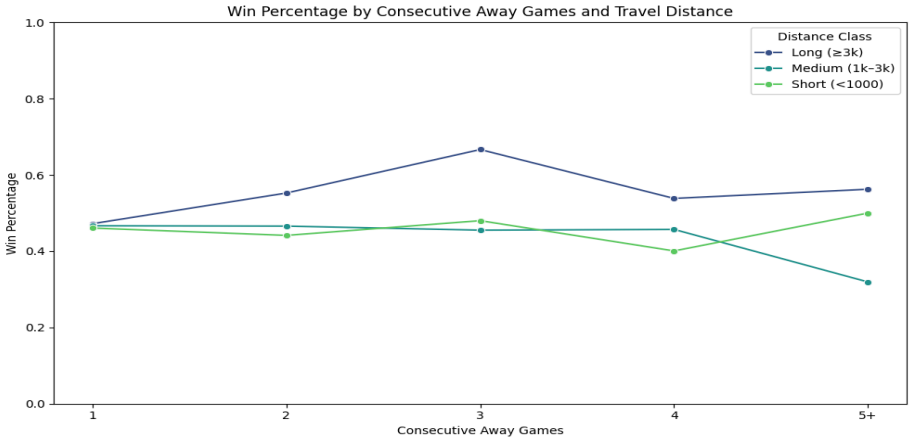

Fig 4: Consecutive Away Games and Grouped Travel Distance vs. Overall Win % (2022-2025)

Lastly, to formally assess whether distance traveled influences game outcomes, I performed a point-biserial correlation test between travel distance and game results (win=1, loss=0) for away games. The correlation coefficient was 0.012 and the p-value was 0.454 indicating no statistically significant relationship. This supports the visual findings from the figures that travel distance, on its own, does not meaningfully affect win probability.

Conclusion

Ultimately, this analysis finds no statistically significant relationship between travel distance and game results in the NHL. However, away teams consistently underperform compared to home teams, confirming the presence of home ice advantage. While the league schedule produces substantial inequities in distance traveled between teams, these disparities do not translate to consistent competitive advantages solely based on distance. Instead, the findings suggest that recovery time and structure of road trips (e.g., back-to-backs, long road trips) have a stronger influence on performance.

Limitations

The analysis faced several limitations that constrain the interpretation of results. Firstly, the NHL game results are not independent, with factors like opponent quality and team strength creating correlation within the data that violate the independence assumption of many parametric tests. Second, the analysis did not incorporate the other contextual factors that were related to game results, which would have allowed for a more nuanced analysis of how travel interacts with competitive balance. Similarly, time zone changes were not included even though they were in the dataset, despite personal experience suggesting they might impact a person's physiology. If I had more time, I would prioritize time zone changes in future analysis. Third, since the stadium dataset was joined by team abbreviation, special cases like Stadium Series or temporary arena relocations (e.g., Arizona Coyotes playing in a college arena) were not accounted for. While manual corrections needed to be made in the data cleaning step, a fully comprehensive dataset would reduce the need for manual intervention. Lastly, the study was limited to only three NHL seasons. Expanding the dataset to include additional seasons would increase statistical power and reveal stronger patterns that persist over time. However, doing so would require significant data cleaning and preprocessing because of the stadium dataset limitation.

Accomplishment Statement

- Built a modular Python data pipeline to process over 10,000 rows of NHL schedule and geospatial data from three seasons to engineer travel-related features.
- Applied statistical testing and data visualization (Matplotlib, Seaborn) to evaluate performance trends, uncovering insights that confirmed home-ice advantage and quantified the impact of rest on win probability.