Data Paper

Alec Millman

LIS 4220- Data Curation

Winter Quarter 2022

3.10.22

Abstract

This paper concerns a dataset for a graduate school project spanning two sub-disciplines; the subject matter of the data was National Hockey League (NHL) team records by season, some high-level team statistics beyond Win-Loss record, and head coach(es). The dataset was amalgamated from publicly available, structured data on the website Hockey-Reference into a single table stored in both csv and open Excel (xlsx) formats. The provenance of the data was spurred by the academic project, played to the creator's interests, but also may prove useful for actual real-world insight into success at the NHL level, and perhaps tangential insight into business management strategy. The dataset is by no means comprehensive for either avenue of analytical focus; it is intended to begin a cursory exploratory analysis to investigate whether there is signal in the noise.

Details	
Subject Area	Business Strategy
Specific Subject Area	Sports Management
Type of Data	Table (xlsx); csv
Acquisition of Data	Copy & Paste csv format of publicly available structured data, uploaded to OpenRefine, cleaned and transformed to set up for open-ended analysis
Data Format	Structured
Source Data Challenges	Data had extraneous characters in one column; columns with string values that would make usability for analysis too cumbersome
Source Data Solutions	Transformed source data by: removing extraneous characters in team name fields; split string-value columns into workable number of individual columns
Source Data Location	NHL Teams & Other Hockey Teams. (n.d.). Hockey-Reference.Com. Retrieved February 6, 2022, from https://www.hockey-reference.com/teams/index.html
Dataset Accessibility	Public repository, Google Drive: [*totallyrealURL*].dotcom

Overview

Context

This dataset was compiled as an LIS graduate school project spanning 2 sub-disciplines , Data Curation and Scripting for Large Databases. The main purpose of the dataset, and its creation, was to practice building a skillset in a graduate program, using an interest of the creator to make the subject matter relevant and more accessible. The initial question that spurred the dataset was to determine if there is signal in the data regarding coaching tenure relative to team success in the NHL. The dataset and subsequent analysis is by no means meant to be comprehensive; it is high-level, cursory data that is intended for initial, exploratory analysis to determine if any signal is there and warrants further analysis. If so, the dataset will likely need to have further versions with additional data points (columns) and likely reconciliation with other linked data.

Any external users of this dataset, or subsequent versions, should have some familiarity with the sport of hockey as well as the NHL and its history, though it is not required and a legend for the dataset is provided here (see Legend section below). It is useful, though not imperative, to know league history facts not involved in the data set, i.e. that a team has moved locations, changed mascot names, or both. This could be remedied by reconciling with external linked data, but this step was intentionally left out (among others; see below Potential & Limitations section).

Reuse Potential

Potential Audience

Then intended audience for this is simply the creator and its purpose is purely for academic purposes. However, the conceptual possibility that some insight is gleaned makes prudent a data storage plan and (this) description of the working dataset.

Future Use

Figure1

Reuse Conditions

Reuse conditions fall under the Creative Commons guidelines (“About CC Licenses,” n.d.); the intent is for it to be open for additional analysis, critique, and to be expanded upon so long as the original work is credited and subsequent work is used for non-commercial purposes and is shared under the same conditions.

Versioning

External contributors may also store work in the folder. Uploaded files from said contributors should be labeled with their name in the file's title.

Versioning shall follow standard semantic versioning (MAJOR-MINOR-PATCH) format.

E.g.

- [*file_name, contributor_name*] v.1
- [*file_name, contributor_name*] v.1.1
- [*file_name, contributor_name*] v.1.1.1
- [*file_name, contributor_name*] v.1.1.2
- [*file_name, contributor_name*] v.2
- [*file_name, contributor_name*] v.2.1
- [*file_name, contributor_name*] v.2.1.1

Etc

Data Overview

Repository Location

[*totallyrealURL*]dotcom

Description of Dataset File

Original Dataset:
File Name: NHL_season_by_season_coaches_v.1
Kind: Microsoft Excel Workbook (.xlsx)
Size: 121,908 bytes (123 KB on disk)
Created: 2.26.22
Compiled by: Alec Millman

Working (Current) Dataset:
File Name: NHL_season_by_season_coaches_v.1
Kind: Microsoft Excel Workbook (.xlsx)
Size: 121,908 bytes (123 KB on disk)
Created: 2.26.22
Compiled by: Alec Millman

Data Legend

The original dataset consists of 21 columns (A-U in Excel), representing mostly typical NHL team season statistics with some transformations made by splitting values into multiple columns (see below section Methodology -> Procedure). The data does NOT include playoff records or other statistics; the only column pertaining to the postseason is the "Playoff Result" column (string value of text as opposed to the numerical values in other columns). All data types are numerical except where noted.

A) Season [string of text]

The NHL season (spanning 2 calendar years) to which the subsequent values pertain, dating back to 1967-68

Note: The entire 2004-05 season was canceled due to disputes between labor and ownership. The league returned in the following season with significant rule changes (applicable here in the advent of the shootout, see below)

Additionally, some seasons have fewer than the current number of regulation season games (82) due to a variety of factors like the season being shorter in the past, global events, etc.

B) Team [string of text]

NHL franchise to which the subsequent values belong and the preceding season column pertain

Note: Not each team has a record corresponding to each season back to 1967-68; some franchises came into existence more recently than that year (the majority)

C) GP

The number of regular season (not playoffs) games played by the team in the given season

D) W

Regular season Wins for the specific team in the given season

E) L

Regular Season Losses for the specific team in the given season

Note: Denotes losses in regulation beginning in the 2005-2006 (see T column below)

F) T

Regular Season ties

Note: ties were eliminated beginning with the 2005-06 season with the advent of the shootout deciding regular season games tied after regulation time and a five-minute overtime. A shootout win registered in the Win column; a shootout (or overtime) loss registered in a newly created "OTL" column (shown here as OL, see below)

G) OL

Overtime and shootout losses

Note: only contains values after 2005-2006

H) PTS

Regular season points accumulated in the given season

Note:

1967 to 2003-04: 2pts for a win, 1pt for a tie, 0pts for a loss in either regulation or overtime

2005-present: 2pts for a win of any kind (including overtime and shootout), 1 pt for reaching overtime (an overtime or shootout Loss), 0pts for a regulation Loss

I) PTS%

The team's points accumulated during the given season out of the total points available (determined by number of games in a season)

J) SRS

Simple Rating System, a statistic calculated by the website hosting the source data

Note: calculated by using a team's goal differential (goals for vs. goals against; not in dataset) and Strength of Schedule. 0 denotes league average for the given season

K) SOS

Strength of Schedule, based on quality of opponents (i.e. their success and supporting statistics)

L) Finish [string of text]

Regular season result within their respective Division (see below)

M. Playoff Result [string of text]

Result of team's postseason for seasons where team qualified based on regular season performance

N) Coach_1 [string of text]

Name of the team's head coach for the given season, only coach if not removed from position mid-season

O) Coach_1_Rcrd

The Win-Loss-Tie (or overtime/shootout losses after '05-06) of the team's first (or only) head coach of the given season

P) Coach_2 [string of text]

Name of a team's 2nd head coach if a coaching change was made midseason (many rows left blank)

Q) Coach_2_Rcrd

The Win-Loss-Tie (or overtime/shootout losses after '05-06) of the team's 2nd head coach of the given season (many rows blank)

R) Coach_3 [string of text]

Name of a team's 3rd head coach of a given season (many rows blank)

S) Coach_3_Rcrd

The Win-Loss-Tie (or overtime/shootout losses after '05-06) of the team's 3rd head coach of the given season (many rows blank)

T) Division [string of text]

The division in which the team played in the given season, subsets of Conferences (see below)

Note: Divisional alignment, and the names of the Divisons, has changed many times over the NHL's history

U) Conference [string of text]

The Conference in which the team played in the given season

Note: though the names of the Conferences has changed, teams have switched conferences very rarely

Methodology

Acquisition of data

Data was aggregated by drilling down into each team's main franchise history page on the website hockey-reference.com. Each franchise's team history table on the site page was converted to csv format, copied and then pasted into OpenRefine to be compiled in a single structured table and then cleaned.

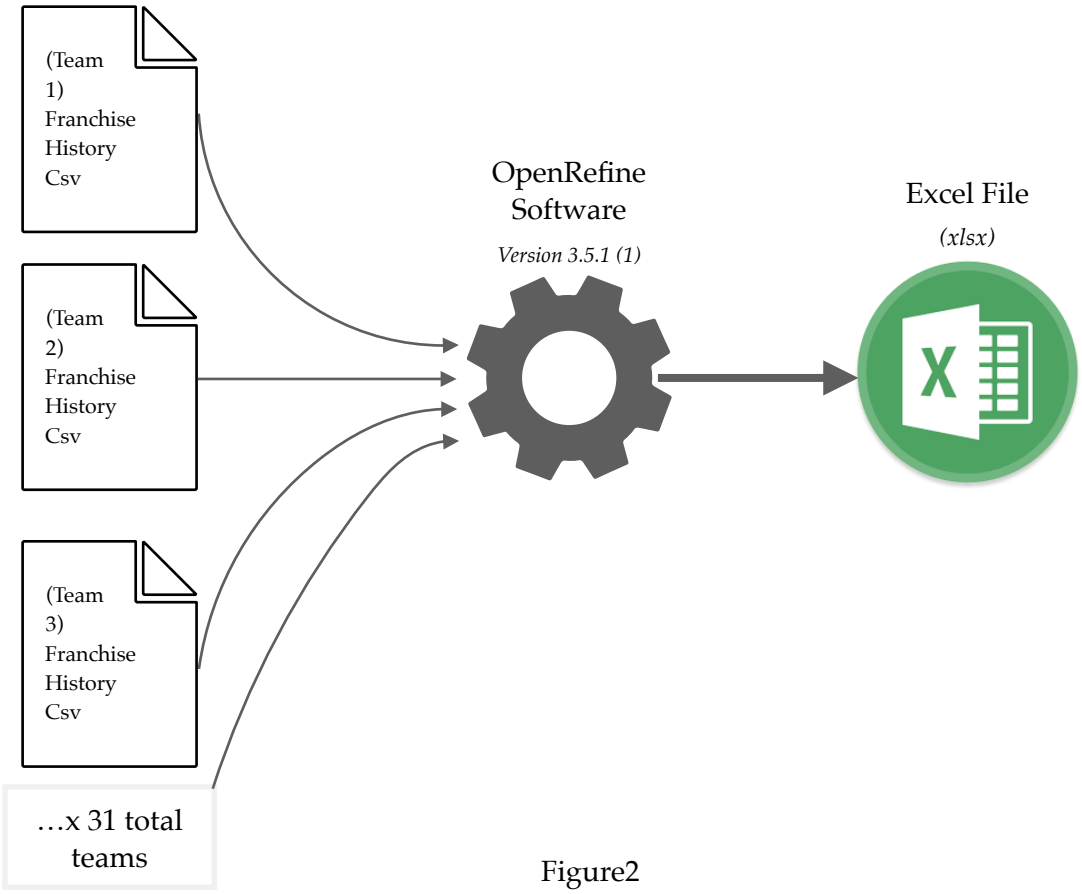

Figure2

Procedure

Data Cleaning Steps- (OpenRefine Software)

1) "Team" column was transformed, removing an extraneous character from some cells-

> An asterisk in the column denoted a team making the playoffs in a given season. There is an entire column dedicated to this (Playoff Result) and thus the denotation is superfluous. Additionally, some teams have different spelling of their team nickname over their history, so they were normalized into a standardized spelling.

2-5) When an instance of multiple coaches for a single team in a single season occurs, they were grouped together in a text string in a single column.

> Step 2 was to separate the coaches and their respective records from each other using ")". Subsequent steps were to separate the coaches respective records from accompanying their names in a column to a column that contained only the record (1 instance of this resulted in 3 new columns given the most head coaches for a single team in a season is 3).

6-20) Self-evident in Figure2

> steps 11 & 12 were to remove two completely blank columns that seem to have been added in the C&P process

0. Create project
1. Mass edit 1248 cells in column Team
2. Split 1251 cell(s) in column Coaches into several columns by separator
3. Split 1251 cell(s) in column Coaches 1 into several columns by separator
4. Split 235 cell(s) in column Coaches 2 into several columns by separator
5. Split 28 cell(s) in column Coaches 3 into several columns by separator
6. Text transform on 1251 cells in column Coaches 1 2: grel:("("+value)
7. Text transform on 1251 cells in column Coaches 1 2: value.replace("(","")
8. Text transform on 1016 cells in column Coaches 1 2: value.replace(")","")
9. Text transform on 207 cells in column Coaches 2 2: value.replace(")","")
10. Text transform on 28 cells in column Coaches 3 2: value.replace(")","")
11. Remove column Column 18
12. Remove column Column 19
13. Remove column Lg
14. Rename column Coaches 1 1 to Coach_1
15. Rename column Coaches 1 2 to Coach_1_Rcrd
16. Rename column Coaches 2 1 to Coach_2
17. Rename column Coaches 2 2 to Coach_2_Rcrd
18. Rename column Coaches 3 1 to Coach_3_Rcrd
19. Rename column Coach_3_Rcrd to Coach_3
20. Rename column Coaches 3 2 to Coach_3_Rcrd

Figure3

Potential & Limitations in Use of Dataset

Limitations

The dataset is not meant to be a comprehensive. Indeed, there are many more transformations that could have been done in OpenRefine, or even basic Excel calculations; calculations like creating specific columns of winning % for each individual coach in a given season (when a team has multiple head coaches). There is also potential for additional statistics to be quite useful to this dataset, either through reconciling existing fields in the table with linked data or making transformations in Excel before finalizing the first working dataset. However, a deliberate choice was made to make the dataset as granular-yet-basic as possible. As it was originally compiled to be a starting point for an exploratory analysis, further detail was not yet needed; taking a additional steps like clustering a team's various permutations (different locales, nicknames, etc. over time) or further transforming columns might end up restricting

what analysis might be done by a future user by not allowing them to slice the data as they see fit.

Potential Use

Though it was meant purely as an academic exercise for an audience of one, there is value in this dataset in that, to the creator's knowledge, there is no previously existing, publicly available single list of all of this subject matter. Initial analysis of the data has not been performed as of the publishing of this paper but should some signal be detected, those actions like adding calculated columns or reconciling with linked data would likely be useful.

Future analysts may use the dataset as a starting point for their own analysis on the efficacy of head coaches and their contribution to NHL franchise success.

Sources

About CC Licenses. (n.d.). *Creative Commons*. Retrieved February 27, 2022, from https://creativecommons.org/about/cclicenses/

NHL Teams & Other Hockey Teams. (n.d.). Hockey-Reference.Com. Retrieved February 6, 2022, from https://www.hockey-reference.com/teams/index.html

OpenRefine Version 3.5.1 (1)