Boston Bruins Data Engineer Applicant Coding Challenge: Connor Young

Code: Python with Jupyter Notebooks, Pandas, urllib, matplotlib, numpy, no additional data sources

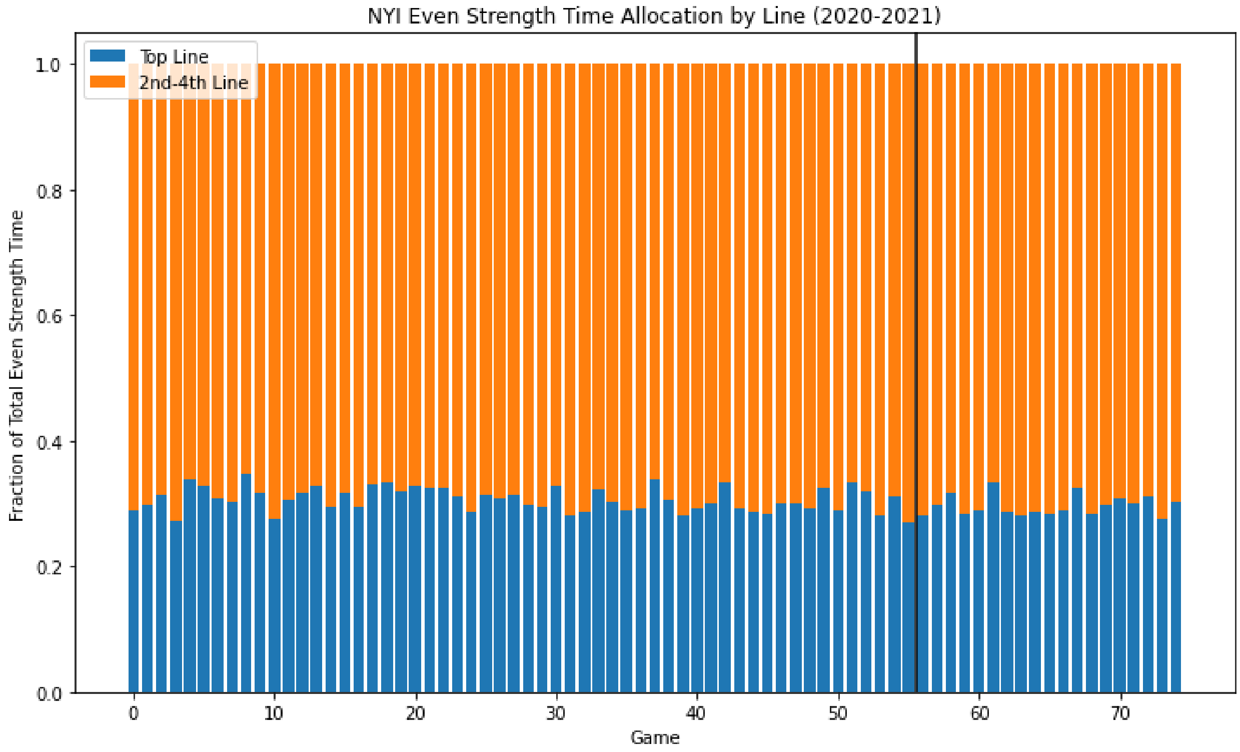
Task 1

I approached this task by initially exploring the shift charts and game-level data from the NHL API to see what I had available to work with. Although the shift charts provide more granular data on individual shift start and end times, I did not need this level of granularity to answer the question of which teams give the most ice time to their top line. The game-level data, under the 'boxscore' section, provides the total time on ice (TOI) for each player in the game, as well as their TOI broken down into even strength, power play, and shorthanded situations. Therefore, I was able to pull the total even strength TOI for each player in each game directly from this source to answer the given question.

One important thing to note is that the prompt specified that we only care about 5 on 5 data, and even strength could include 4 on 4 or 3 on 3 situations. If I were to try to account for these, I would need to use penalty and goal events from the play-by-play data to track the current strength state of the game at all points in time, and then use the granular shift-level data alongside that to determine the amount of time that each player spent on the ice at these different points. The relative lack of commonality of these strength states made me feel okay with assuming that their impact would be negligible on the results, especially given the time constraints for this project and the challenge of this alternate approach.

Using my even strength TOI approach, the next choice I had to make was how I would classify a team's top line. Obviously, and according to the prompt, the top line would consist of only forwards, so I removed defensemen and goalies from the data. With only forwards remaining, the top line could be classified as the three players that received the most even strength TOI, which is what I chose to do. This approach is not completely foolproof, as it could be affected by individual players that get stuck on the ice without their normal line in situations like icing or long shifts in the defensive zone, or just players that tend to take longer or shorter shifts in general. Again, however, it seemed the most reasonable approach to me given the time constraints of the project. In addition, a team may shuffle lines throughout a game, which makes it even more difficult to determine what the top line actually is. One other key choice that I made was to ignore preseason games since these very rarely represent the actual line-use strategy a team will pursue when they care about winning and that many top players will have reduced or no ice time in them.

I calculated the fraction of total even strength time that each forward was on the ice for and took the top three as my top line. Then, I summed the fractions of all three, which represented the fraction of total even strength time that the top line was responsible for. Once I had this working for both teams in a single game, I automated my script to perform these same calculations for all regular season and (for 2020-2021) playoff games. I then aggregated the results by team and calculated the average top line usage fraction for each team over their entire season, separating regular season and playoffs. I also visualized the usage throughout the course of the season for each team. An example is shown below for a team that made the playoffs, where the black line indicates the end of the regular season.



I then put the season-long usage average for each team into a table for each season and sorted them (by regular season usage) to determine which teams relied most heavily on their top line. I also combined the two seasons into one overall table and calculated the change between seasons.

I found that (unsurprisingly), Edmonton relied most heavily on their top line in both the 2020-2021 and 2021-2022 seasons, with it taking up 34.255% and 35.085% of their even strength ice time respectively. Overall, in 2020-2021 the five most top-heavy teams were Edmonton, Toronto, Pittsburgh, Chicago, and Winnipeg. In 2021-2022 the five most top-heavy teams were Edmonton, Winnipeg, Colorado, Chicago, and Vancouver. Across both seasons combined the five most top-heavy teams were Edmonton, Winnipeg, Colorado, Toronto, and Chicago. In 2020-2021 the total range of regular-season-long average top line usage was from 30.025% to 34.255% and in 2021-2022 was from 29.836% to 35.085%. I was a bit surprised at how small these ranges were and expected more variation from team to team, but that speaks to the importance of forward depth given the physical constraints of how much any player can play. Looking at teams that made the playoffs in 2020-2021, there was no clear trend for how their top line usage changed vs. the regular season, with exactly half of the teams increasing usage and half decreasing. I also looked at the average usage during the regular season of teams that made the playoffs vs. teams that didn’t to see if there were any trends there. These values were 31.749% and 31.317% respectively. I was slightly surprised by this as one might expect playoff teams to utilize their top line less, indicating more depth, but I could also see how higher usage indicates better top-end talent which also is impactful in team success. I think the key takeaway in terms of this is that most successful teams will not rely too heavily or too lightly on their top line, indicating good top-end talent but also good depth. There are obviously exceptions, but this is supported by teams like Toronto, Edmonton, and Colorado that rely very heavily on these top lines struggling in the playoffs if those lines can’t produce as well as teams like Ottawa, Philadelphia, and the Islanders that are on the other end of the spectrum having difficulty scoring in or even making the playoffs without dominant top-end talent. Tampa Bay and Vegas were two very good teams that are around the middle of the pack in terms of top line usage.

I would like to explore the relationship between top line reliance and finishing position of teams more if I had more time, as well as some other areas. One of these other topics would be the alignment between the percentage of even strength TOI of a top line with the percentage of their point contribution and the percentage of salary cap relative to the rest of the forward group. This would be a way to gauge who is getting the best return from their top line in terms of ice time or money invested in them. I would also like to look at the breakdown of even strength TOI of the 2nd-4th lines and whether there is a large drop between any lines or if they are all closer together. In this task we took top-heavy to mean only top line usage, but a team could also be top-heavy by relying on their top two lines much more than their bottom two and looking at the total TOI breakdown by line might reveal slightly different results than just top vs. other three. It would also be interesting to look at the consistency of top line players from team to team and which teams tend to shuffle their personnel more or less and how that relates to their average top line TOI percentage. I also looked at the change in top line usage percentage per team between the two seasons and would like to explore this further and how it relates to roster moves or improvement in standings. Finally, I think it would be interesting to perform this same exercise with defensive pairings and see how the breakdown of TOI percentage for forward lines compares to that of D-pairs or if there is a wider variation from team to team.

To run my code, you'll need to run 2020-2021 scrapers then graph the data before the 2021-2022 scrapers due to reused variable names:

- Run cells 3.1 and 3.2 (~3.5 mins and ~20 seconds respectively) and then everything under 4.1 to get 2020-2021 data
- Run cell 3.3 (~90 seconds) then everything under 4.2 to get 2021-2022 data
- Run everything under 4.3 to compare across seasons

I didn't run into any significant issues while doing this task, the only thing was an HTTP 403 Forbidden error while trying to scrape the shift-level data which was easily resolved by setting a known browser user agent in the request. I had my season-long average calculation for the 2021-2022 season affected by canceled games still counting as games when dividing by number of games and the top line having 0% of ice time, but I caught this and corrected it easily as well.

Task 2

I would write a Python script that retrieves the necessary data from the NHL API (similarly to Task 1 but more extensively) and other sources to fill all the fields in my database. I would use an AWS Lambda function to trigger this script to run at a specified time and write the results to an Amazon S3 bucket. Then, I would use Lambda again to trigger an AWS Glue job that would extract the data from the bucket, process it, and load it into a data storage solution like Amazon Redshift where it can be queried as desired.

For our use-case, I think it makes the most sense to have the Python script trigger at the completion of all games from the previous day and then have the Glue job trigger once all data from the day is loaded into the S3 bucket. We could schedule this to happen at something like 3am Eastern each night when all games will be finished (barring extremely long playoff games). Alternatively, we could have the script trigger upon completion of any single game and optionally have the Glue job do so as well, which would allow us quicker access to data rather than having to wait for all games to finish. However, I would assume that it is generally not necessary to have the data for analysis that quickly, especially given the fact that Lambda charges based on number of requests as well as runtime, so this option would likely be more expensive.

My pipeline will extract from the publicly accessible NHL API (specifically play-by-play) and the NHL shift charts at minimum. It would also ideally be able to have access to non-publicly accessible data sources to track events that the NHL API currently does not contain. In terms of which events will be tracked, from the public sources I will track game start and end time, period start and end time, shift start and end time for each player, faceoffs, shots on net, missed shots, blocked shots, takeaways, giveaways, hits, stoppages, goals, and penalties. The key events that the NHL API does not contain that I would like to track as well and I know some entities chart by hand are passes, zone exits, zone entries. In addition to just tracking event info itself, I would also want each event row to contain information about the previous event, namely what it was and the time since it occurred, which would be valuable in detecting things like rebound shots and (combined with passes) one-timers that obviously impact the quality of a scoring chance. Also related to scoring chance quality would be shot velocity and location on net, which is not publicly available information.

The schema of my database is shown graphically in the tables further below, I figured this to be an easier way to present and understand it than describing it fully in words. There are related tables to the nhl_events table, namely the game_info and on_ice tables.

The game_info table is related to nhl_events through the game_id column and contains general information about the game that events are from, such as the teams, location, final score, team stats, and more.

With the on_ice table, which is related to nhl_events through the time_id column, we could include information about who is on the ice for each event by utilizing the shift charts from the NHL API. Since each event is associated with the time into a period that it occurs and the shift charts tell us the start and end time of each shift, we can determine which players were on the ice at the time that the event occurred.

nhl_events		
Column Name	**Type**	**Example/Notes**
game_id	Integer	2020020001 (Foreign Key to game_info table)
event_idx	Integer	1
event_id	Integer	20200200011 (Primary Key)
event_type	Varchar	Goal
secondary_type	Varchar	Snap Shot
event_description	Varchar	Sidney Crosby (1) Snap Shot, assists: none
player1_id	Integer	8471675
player1_name	Varchar	Sidney Crosby
player1_team	Char	PIT
player1_type	Varchar	Scorer
player2_id	Integer	8479394
player2_name	Varchar	Carter Hart

player2_team	Char	PHI
player2_type	Varchar	Goalie
x_coordinate	Float	-79.0
y_coordinate	Float	4.0
period	Integer	2
period_type	Varchar	Regular
period_time	Time	03:39
period_time_remaining	Time	16:21
time_id	Integer	202002000110339 (Foreign key to on_ice table)
event_datetime	Datetime	2021-01-13T23:34:41Z
strength	Varchar	Power Play
home_goals	Integer	2
away_goals	Integer	2
game_winner	Boolean	false
empty_net	Boolean	false
primary_assister_id	Integer	8479122
secondary_assister_id	Integer	8477365
penalty_severity	Varchar	Minor
penalty_minutes	Integer	2
time_since_last_event	Time	00:17
last_event	Varchar	Shot

These examples would obviously not all coincide simultaneously, for example if it were a ‘Goal’ event_type we would have null values for penalty_severity and penalty_minutes, all fields are just filled for demonstration. event_id is a unique identifier that combines the game_id and event_idx into one integer to be used as a primary key.

on_ice		
Column Name	**Type**	**Example/Notes**
time_id	Integer	202002000110339 (Primary Key, Foreign Key to nhl_events table)
away_player1_id	Integer	8477365
…	…	…
away_player6_id	Integer	(In case of empty net or delayed penalty)
away_goalie_id	Integer	8479394
home_player1_id	Integer	8477365
…	…	…
home_player6_id	Integer	(In case of empty net or delayed penalty)
home_goalie_id	Integer	8476268

time_id is a unique identifier that combines the game_id, period, and period_time into one integer to be used as a primary key so that we can track the players on the ice at any given time in any given game to determine who is on the ice when an event occurs. There will be a row in this table for each second of every game that is built from the shift chart API.

game_info		
Column Name	**Type**	**Example/Notes**
game_id	Integer	2020020001 (Primary Key, Foreign Key to nhl_events table)

season_id	Integer	20202021
game_type	Varchar	Regular
start_time	Datetime	2021-01-16T00:00:00Z
end_time	Datetime	2021-01-16T02:33:06Z
away_team_id	Integer	5
away_team_name	Varchar	Pittsburgh Penguins
away_team_abbrev	Char	PIT
away_team_division_id	Integer	25
away_team_division_name	Varchar	MassMutual East
away_team_conference_id	Integer	6
away_team_conference_name	Varchar	Eastern
Same for home team as above	Same as above	-
game_location	Varchar	Philadelphia
game_venue	Varchar	Wells Fargo Center
time_zone	Char	EST
away_active1_id	Integer	8470642
…	…	…
away_active19_id	Integer	8479376
away_goalie_id	Integer	8471679
Same for home team as above	Same as above	-
game_length	Varchar	Regulation
away_goals	Integer	3
away_pim	Integer	14
etc.	etc.	More away team game stats
home_goals	Integer	2
home_pim	Integer	18
etc.	etc.	More home team game stats

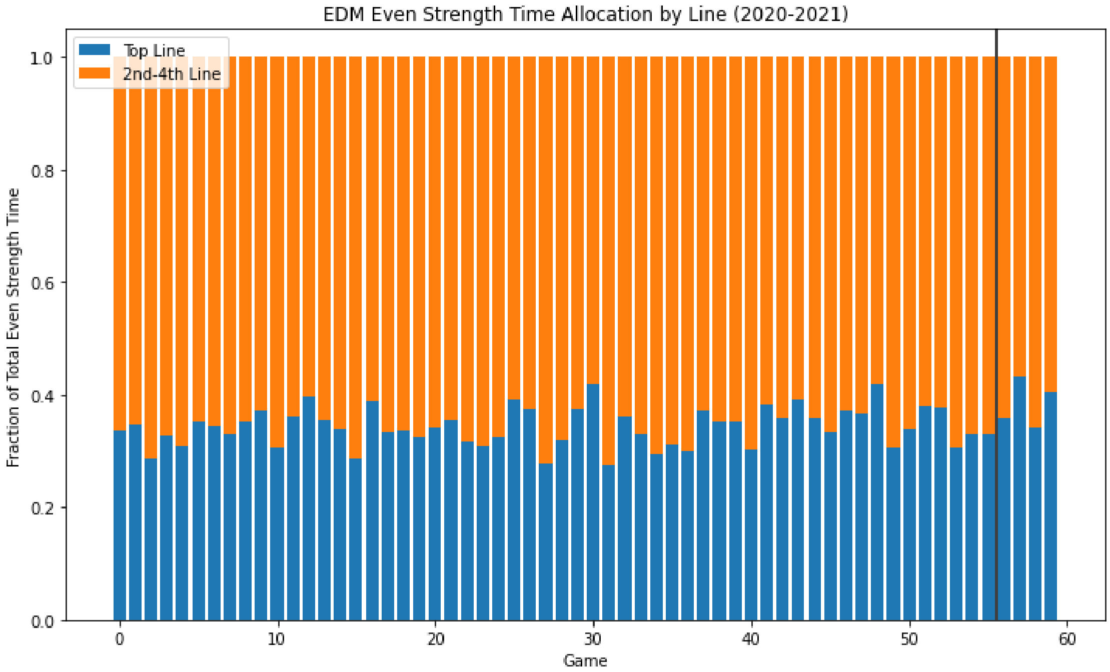
This table could obviously include far more team stats from the game as indicated by the etc. rows but these are omitted for sake of time and space. I also omitted obvious columns such as home team duplicates of away team information for the same reasons.

We could also include a couple more related tables. One could be a team_info table that uses a team_id column as a primary key, relates to away_team_id and home_team_id in game_info, and stores basic information about a team such as its roster, team stats, historical/biographical info, and more. Another could be a player_info table that uses a player_id column as a primary key, relates to any of the player ID fields in the nhl_events and game_info tables, and stores basic information about a player such as their physical attributes, stats, assigned ratings/evaluations, and more. However, both of these tables are not as directly critical to the nhl_events table and as such are omitted for time and space.

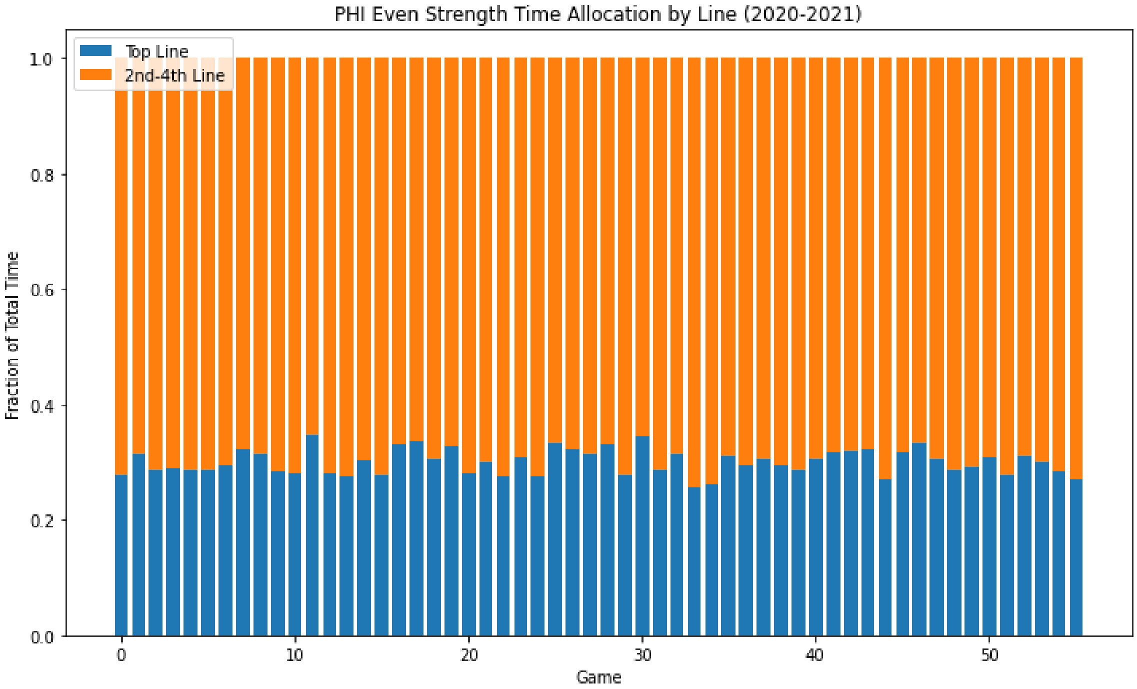
Appendix

Some more examples of top line usage visualization for various teams (all teams shown in code):

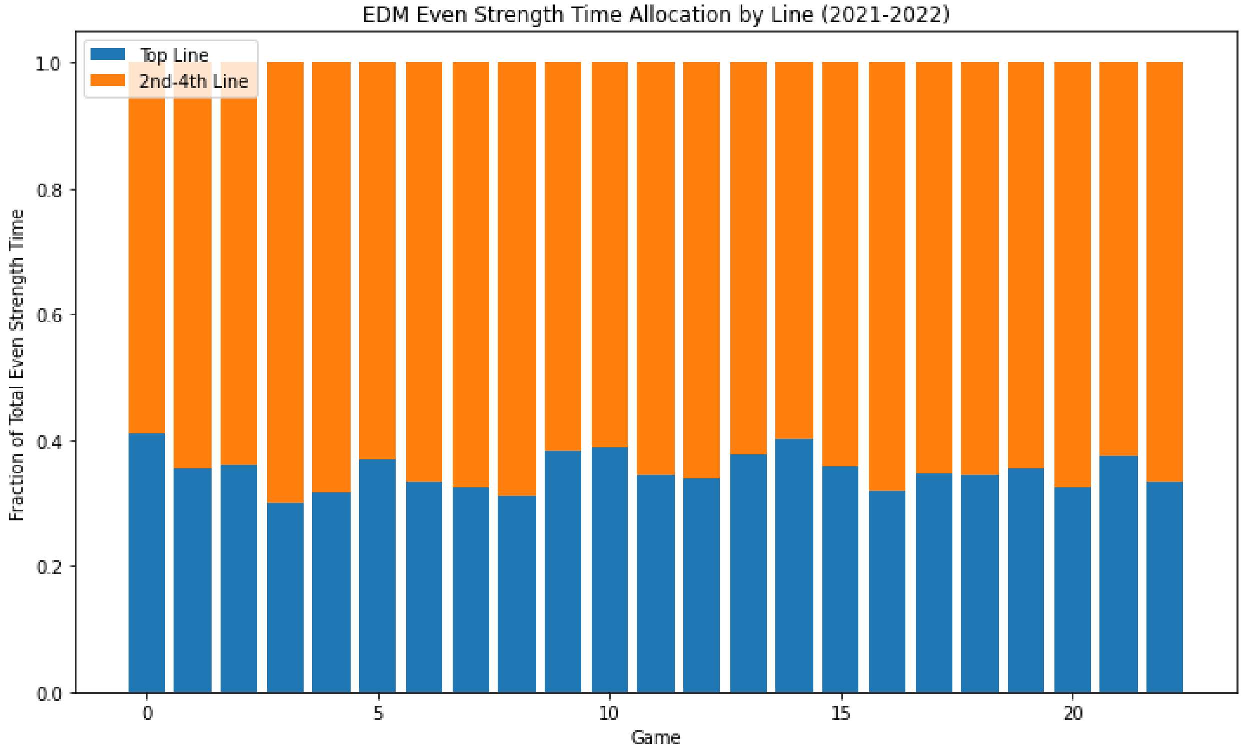
Edmonton – highest average top line usage in 2020-2021 (34.255%)



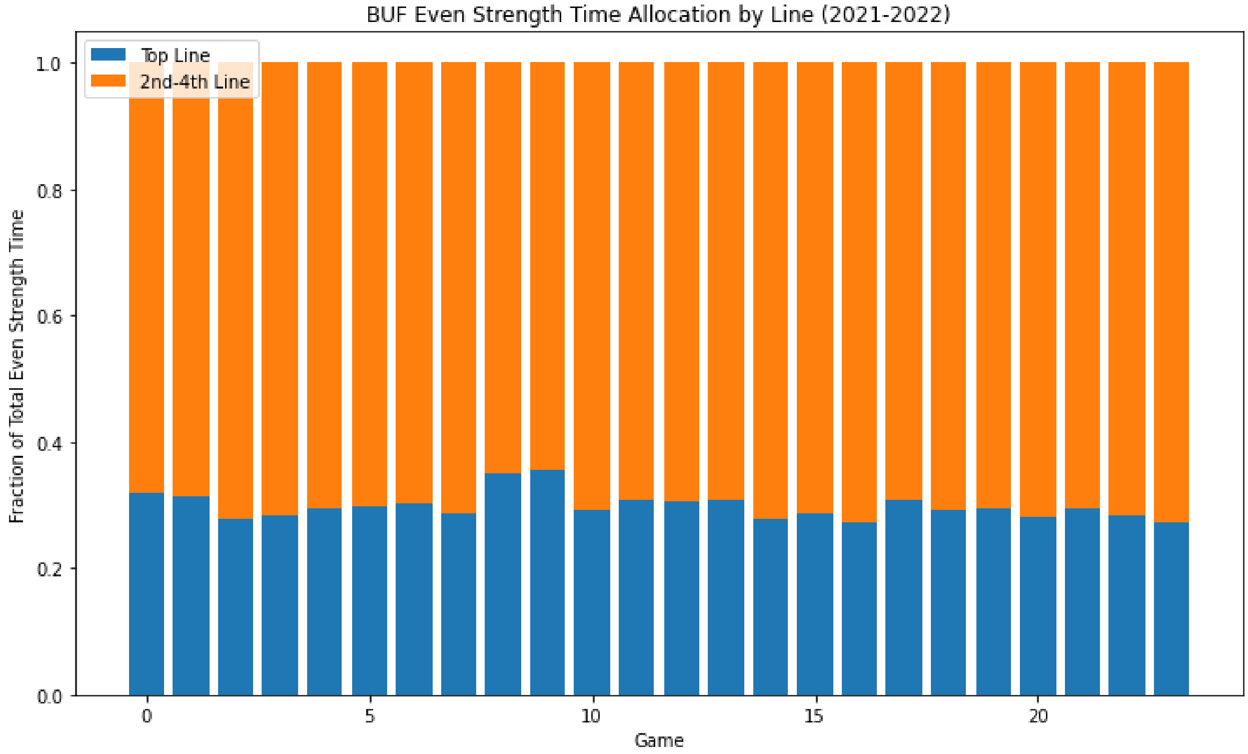
Philadelphia – lowest average top line usage in 2020-2021 (30.025%)



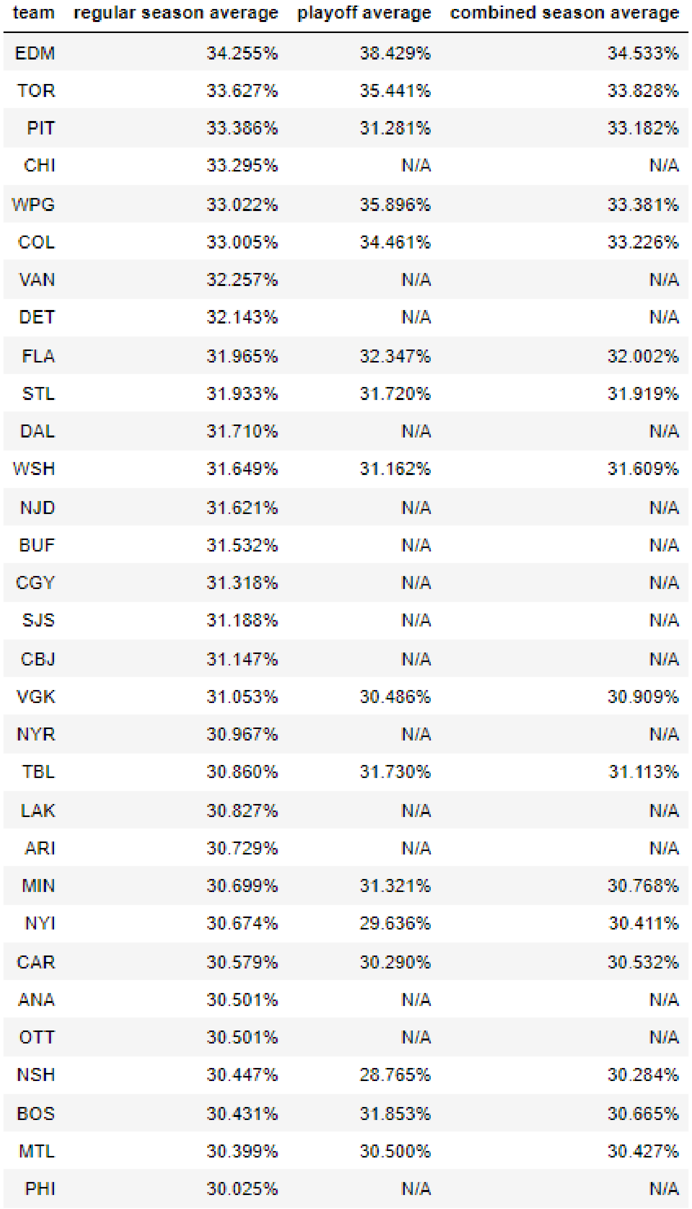
Edmonton – highest average top line usage thus far in 2021-2022 (35.085%)



Buffalo – lowest average top line usage thus far in 2021-2022 (29.836%)

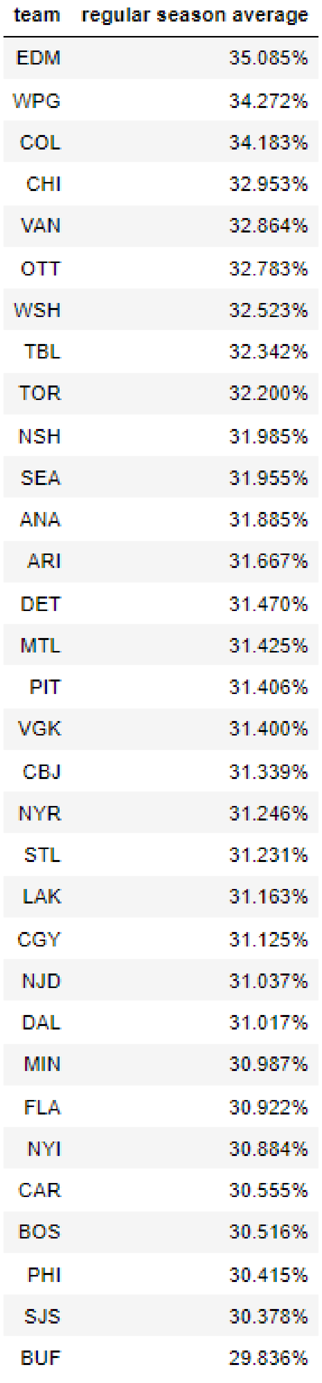


Average top line usage by team for 2020-2021 season, ordered descending by regular season average:



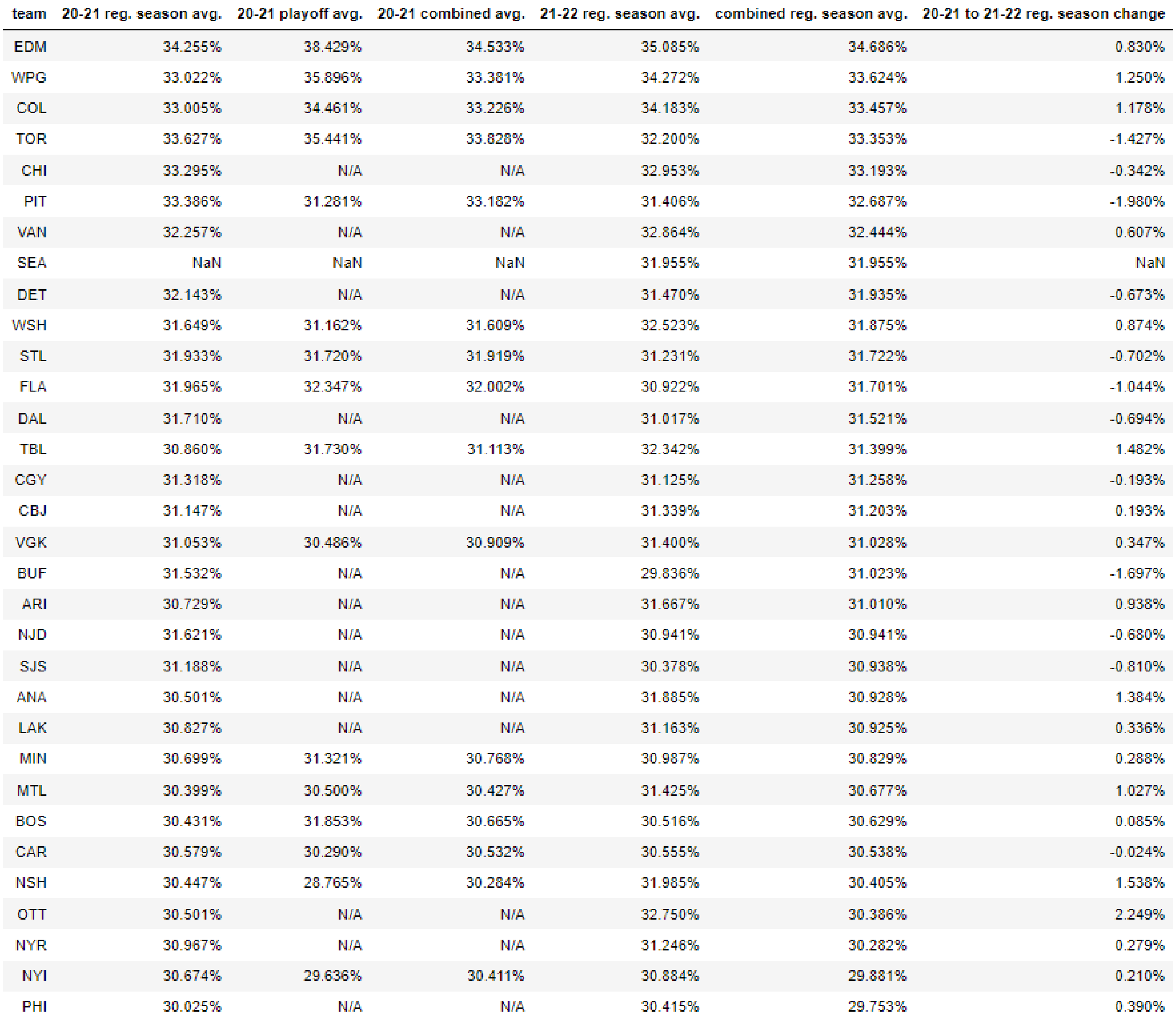
team	regular season average	playoff average	combined season average
EDM	34.255%	38.429%	34.533%
TOR	33.627%	35.441%	33.828%
PIT	33.386%	31.281%	33.182%
CHI	33.295%	N/A	N/A
WPG	33.022%	35.896%	33.381%
COL	33.005%	34.461%	33.226%
VAN	32.257%	N/A	N/A
DET	32.143%	N/A	N/A
FLA	31.965%	32.347%	32.002%
STL	31.933%	31.720%	31.919%
DAL	31.710%	N/A	N/A
WSH	31.649%	31.162%	31.609%
NJD	31.621%	N/A	N/A
BUF	31.532%	N/A	N/A
CGY	31.318%	N/A	N/A
SJS	31.188%	N/A	N/A
CBJ	31.147%	N/A	N/A
VGK	31.053%	30.486%	30.909%
NYR	30.967%	N/A	N/A
TBL	30.860%	31.730%	31.113%
LAK	30.827%	N/A	N/A
ARI	30.729%	N/A	N/A
MIN	30.699%	31.321%	30.768%
NYI	30.674%	29.636%	30.411%
CAR	30.579%	30.290%	30.532%
ANA	30.501%	N/A	N/A
OTT	30.501%	N/A	N/A
NSH	30.447%	28.765%	30.284%
BOS	30.431%	31.853%	30.665%
MTL	30.399%	30.500%	30.427%
PHI	30.025%	N/A	N/A

Average top line usage by team for 2021-2022 season, ordered descending by regular season average:



team	regular season average
EDM	35.085%
WPG	34.272%
COL	34.183%
CHI	32.953%
VAN	32.864%
OTT	32.783%
WSH	32.523%
TBL	32.342%
TOR	32.200%
NSH	31.985%
SEA	31.955%
ANA	31.885%
ARI	31.667%
DET	31.470%
MTL	31.425%
PIT	31.406%
VGK	31.400%
CBJ	31.339%
NYR	31.246%
STL	31.231%
LAK	31.163%
CGY	31.125%
NJD	31.037%
DAL	31.017%
MIN	30.987%
FLA	30.922%
NYI	30.884%
CAR	30.555%
BOS	30.516%
PHI	30.415%
SJS	30.378%
BUF	29.836%

Average top line usage by team for both seasons, ordered descending by combined regular season average:



team	20-21 reg. season avg.	20-21 playoff avg.	20-21 combined avg.	21-22 reg. season avg.	combined reg. season avg.	20-21 to 21-22 reg. season change
EDM	34.255%	38.429%	34.533%	35.085%	34.686%	0.830%
WPG	33.022%	35.896%	33.381%	34.272%	33.624%	1.250%
COL	33.005%	34.461%	33.226%	34.183%	33.457%	1.178%
TOR	33.627%	35.441%	33.828%	32.200%	33.353%	-1.427%
CHI	33.295%	N/A	N/A	32.953%	33.193%	-0.342%
PIT	33.386%	31.281%	33.182%	31.406%	32.687%	-1.980%
VAN	32.257%	N/A	N/A	32.864%	32.444%	0.607%
SEA	NaN	NaN	NaN	31.955%	31.955%	NaN
DET	32.143%	N/A	N/A	31.470%	31.935%	-0.673%
WSH	31.649%	31.162%	31.609%	32.523%	31.875%	0.874%
STL	31.933%	31.720%	31.919%	31.231%	31.722%	-0.702%
FLA	31.965%	32.347%	32.002%	30.922%	31.701%	-1.044%
DAL	31.710%	N/A	N/A	31.017%	31.521%	-0.694%
TBL	30.860%	31.730%	31.113%	32.342%	31.399%	1.482%
CGY	31.318%	N/A	N/A	31.125%	31.258%	-0.193%
CBJ	31.147%	N/A	N/A	31.339%	31.203%	0.193%
VGK	31.053%	30.486%	30.909%	31.400%	31.028%	0.347%
BUF	31.532%	N/A	N/A	29.836%	31.023%	-1.697%
ARI	30.729%	N/A	N/A	31.667%	31.010%	0.938%
NJD	31.621%	N/A	N/A	30.941%	30.941%	-0.680%
SJS	31.188%	N/A	N/A	30.378%	30.938%	-0.810%
ANA	30.501%	N/A	N/A	31.885%	30.928%	1.384%
LAK	30.827%	N/A	N/A	31.163%	30.925%	0.336%
MIN	30.699%	31.321%	30.768%	30.987%	30.829%	0.288%
MTL	30.399%	30.500%	30.427%	31.425%	30.677%	1.027%
BOS	30.431%	31.853%	30.665%	30.516%	30.629%	0.085%
CAR	30.579%	30.290%	30.532%	30.555%	30.538%	-0.024%
NSH	30.447%	28.765%	30.284%	31.985%	30.405%	1.538%
OTT	30.501%	N/A	N/A	32.750%	30.386%	2.249%
NYR	30.967%	N/A	N/A	31.246%	30.282%	0.279%
NYI	30.674%	29.636%	30.411%	30.884%	29.881%	0.210%
PHI	30.025%	N/A	N/A	30.415%	29.753%	0.390%