Effects of Player Statistics and Team Predictors on NHL Salaries

Matthew McAnear
Department of Statistics, University of Michigan

Anthony Paolillo
Department of Statistics, University of Michigan

Keywords:
NHL, salary, hierarchical modeling

Abstract:
When NHL players and teams are negotiating contracts, player statistics are an obvious bargaining point. Less obvious are the characteristics inherent to individual teams. In this paper, we examine whether certain teams have different average salaries, controlling for individual statistics contributions and statistics. Specifically, we explored if the effective tax rate of a team's players increases or decreases the expected salary of a player, given their statistics. We found that while certain teams have a lower average salary compared to others, effective tax rates are not predictive of this change, and player statistics remain far and away the most important predictor of player salary.

Overview

Every fan wants their favorite team to sign the best players. Many blame the cheapness of their team's owner when they miss out on a star free agent, or when they cannot trade for top talent. But is it that simple? With 32 teams each located in as many cities spread across two countries, certain players may prefer a specific climate, or opt for a state with lower taxes. Some want to set themselves up best for success and playing time, while others just want the best chance to win a Stanley Cup. The goal of our project is to determine if teams have to pay players more based on these factors; given a certain level of hockey performance, can teams expect to sign a player for the same pay?

In order to accomplish our goal, we have split the project into two parts. We first sought out to collect both player statistics and contract details since 2010. Then, we analyzed the data in order to determine which statistics tend to more accurately represent a player's salary based on their position. Once we had a baseline model for determining a player's worth in terms of salary, we then tested this model's coefficients across our desired factors in our hierarchical model. This allowed us to see the different effects that our player's stats had on salary based on a given team, position, and tax rate.

Data

Our data for this project comes from two primary websites that are central repositories of hockey statistics: Moneypuck.com[1] for all player statistics, and Capfriendly.com[2] for all salary data. Data was scraped from these resources using a custom web scraper[3]. Using these sources, we were able to gather the following information for each individual player:

1. **Player-based statistics** - These stats include basic offensive statistics such as goals, points, assists and hits, as well as defensive statistics such as penalty minutes, hits, and shots blocked. We also included expected statistics, such as expected goals, on/off ice expected goals, etc, as many hockey teams are beginning to adopt advanced analytical strategies and weighing projections more than production.
2. **Contract Features** - The length of contract, type of contract, and clauses included (no-move clause, no trade clause, etc).
3. **Team Metadata** - The effective tax rate for the given player's team, state/province of the team, country in which they play, and whether the teams were part of the "Original Six" (the first six teams that formed the NHL). Teams are organized neatly into two conferences of two divisions each. Each division has 8 teams, roughly geographically clustered.

[1] Tanner, "MoneyPuck.Com -Download Datasets."
[2] CapFriendly, "CapFriendly - NHL Salary Caps"; CapFriendly, "Income Tax Calculator."
[3] Richardson, "BeautifulSoup"; Reitz and Contributors, "Requests: HTTP for Humans"; McAnear, "Hockey_salaries."

Between the two datasets, there are roughly ~11,000 player contract and season pairs spanning back from the 2009-2010 season. There are several players who have no salary information, and many players who have salary information, but no recorded statistics with an NHL team across the two datasets. That being said, the vast majority of NHL players are accounted for.

Merging the two datasets requires merging on names, which can be difficult, especially due to Unicode characters used in certain names. For that reason, normal misspellings, accent marks, and translation inconsistencies reduce the total size of the dataset. Despite these issues, of the roughly 12,500 players in our dataset with NHL statistics, we have salary information on around 11,000, for a missing rate of around 8%.

We decided to use data starting in the 2016-2017 NHL season after evaluating our data in order to give us a large sample size that was still reflective of the current day decisions. In addition, we decided to narrow down our analysis to standard one-way contracts only. Being that our goal is to determine which factors most significantly affect a player's free agency decision, we felt that the inclusion of entry-level contracts and two-way contracts would confound our analysis across different parts. For example, all players entering the NHL via the draft are subject to entry-level contracts, and their length depends on their current age. However, these contracts all have a very low maximum salary, meaning that many of the top young players will significantly outperform their salary in comparison to a player on a standard one-way contract. Thus, the omission of these contracts allows us to better predict salary and solely focus on our research question of why players choose to sign with certain teams.

At the conclusion of the cleaning, we were left with 3,744 rows and 50 columns, with each row representing a player's statistics for a specific season. We then manually added two columns, "Log Salary" (the log of a player's salary) and "xGoals_diff" (onIceXGoals% - offIcexGoals%), as we determined in our preliminary analysis that these would be measures we wanted to include. From here, we now had the sufficient data to attack our research question and begin to analyze trends within both a player's statistics and contract details.

EDA

Our project goal revolves around the idea of comparing a player's actual salary to what we believe he should be paid. Therefore, it is essential that we first decipher which player statistics best determine the salary of a player. Additionally, it is important to understand how each player's salary is distributed across team, position, conference, etc. Differences across these would allow us to then use our hierarchical model to deem if they are significant. With these goals in mind, our exploratory research could serve as a perfect basis for our model.

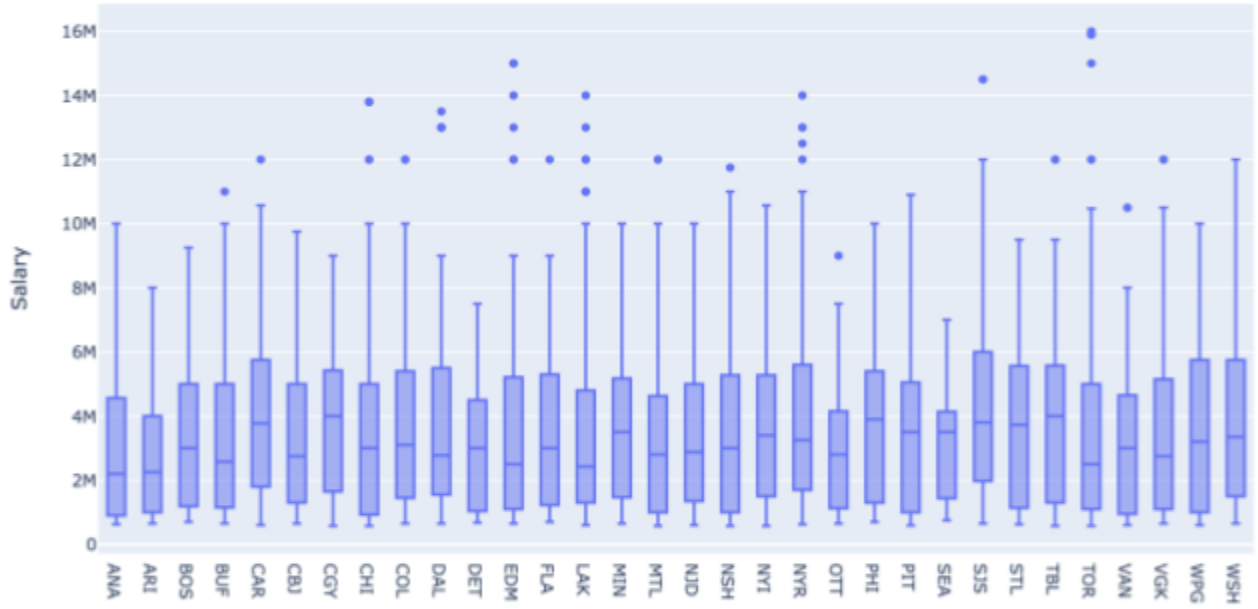

Figure 1

We first wanted to see trends in our salary data based on different categorical variables of a player. This would allow us to get a baseline as to how salary tends to vary between teams, position, production, etc. Figure 1 above shows the distribution of salary across all 32 NHL teams.

The plots, along with an ANOVA test, allow us to verify that the distribution of salary is not consistent across the teams. It also may appear that certain teams (ex. EDM) have many outliers that could be an issue; however, these points reflect the same player over the course of multiple years, so the production associated with the salary amount will be consistent.

With the knowledge now that salary was different amongst teams, we wanted to see how this salary variable was altered. It is important for us to understand which factors best predict a salary, as these stats will be ones that teams know that they can "buy" in free agency. Also, it will give us a method of comparison to compare our estimated salary to what the player actually received, and see which factors ultimately persuaded the player to sign where he did. Figure 2 shows the average value of four statistics (Points, Hits, Expected Goals Difference, and Faceoffs Won) for each of the 32 teams plotted against the average salary per player. This helped us see relationships between our stats and salary. It seems clear that the higher

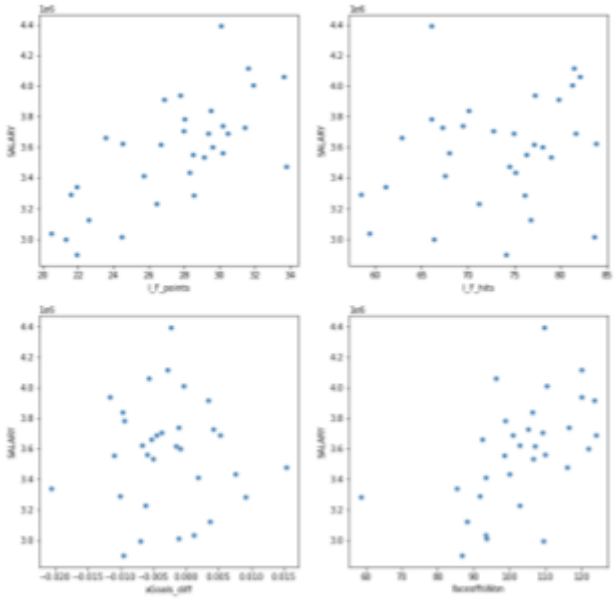

Figure 2

average salary, the more points a player records, but a stat such as hits seems to have no correlation with salary.

As mentioned above, this now gives us a way for teams to see how to best improve their team. It's clear that if a team wants a skater with more points, they should pay more, whereas if they want more hits, paying more might not guarantee more production. Lastly, the plots back up the idea that production and salary are not consistent between the 32 teams, furthermore incentivizing a reason for our analysis.

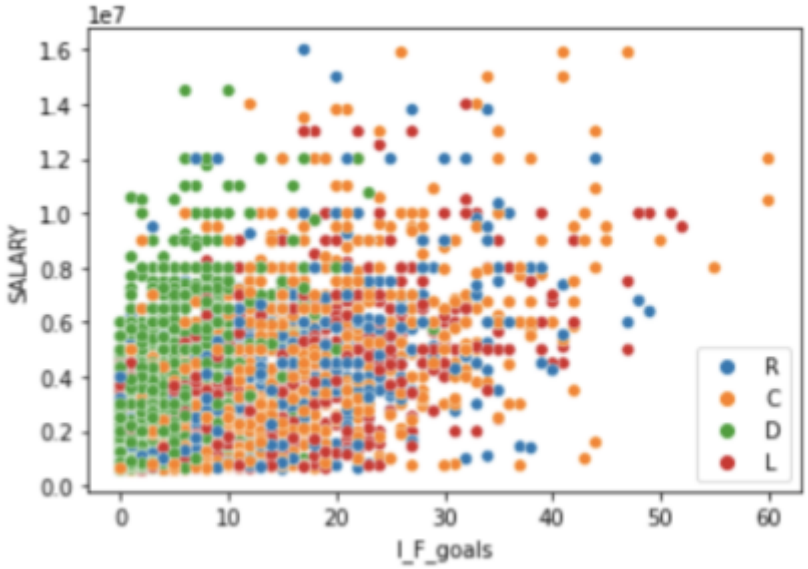

Figure 3

While our first two figures show differences between teams, we now needed to look at other groupings in order to fill out our final model. Figure 3 is a scatter plot of goals vs salary based on the position of a player (Right Wing, Center, Defenseman, Left Wing). There are some key points that are important to draw from the plot. For starters, certain stats (goals in this case) vary based on position, so our model for predicting salary will have to include a positional hierarchy. With this though, we see that when grouping by defensemen and forwards, the highest paid in each group tend to be the most productive in goals.

To more fully understand all of our predictor variables and how they relate to salary, we evaluated their correlation with salary. Figure 4 shows a heatmap between Salary and 17 different variables in our dataset. From here, we were able to further down which statistics were to be used for predicting the salary of a player. It also helped us address collinearity. For example, points are goals+assists, with points being higher correlated with salary, we dropped goals and assists.

So with clear differences between salary across teams and position, as well as understanding of how we can predict the salary of a player, we have a foundation

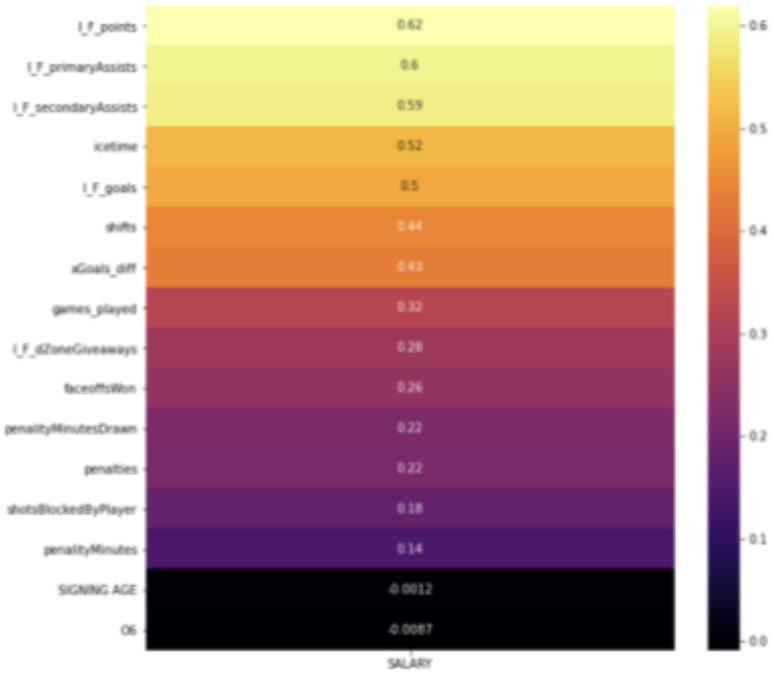

Figure 4

for our final model. However, there were now a couple more things to address. For starters, games played and icetime are highly correlated with all of our statistics; those who play more will generally have more goals, points, etc. In order to address this, we modified our statistics into “per 60 minute” measurements, as a hockey game is a total of 60 minutes in gametime. This allows us to evaluate the efficiency of our players independent of the amount of time they spend on the ice. The only exception to this standardization is the expected goal differential, since this is already a percentage relative to the time on ice.

Our analysis also showed us that the distribution of each statistic differs from year to year; it is possible that 100 points in one season would rank a player 1st in the league, whereas the next year, 100 points is not as impressive. Therefore, in order to control for this year-to-year noise, we standardized each player’s statistic relative to the median measure and IQR of the statistic for the given year. This robust scaling method is more applicable, as our distributions for our statistics are skewed, so using any transformation dependent on the mean would not be as effective. While this may affect the interpretation of our coefficients in our model, we deemed it necessary due to the changing landscape of the league and the fluctuations in our statistics from year to year. This will increase the accuracy of our predictions of salary and the overall performance of the model.

With a foundational understanding of what determines a player’s salary, along with adjusted statistics, we could now move to our model in order to determine which of our original factors have a significant effect on a player’s free agency decision.

Final Model

We tried several iterations of our hierarchical model before settling on the model shown below.

Figure 5

Notably, there are few team-level statistics and no contract features. There were three factors that we were planning to explore, but two had to be dropped due to sampling issues and correlation:

- Original-Six - This variable was of interest, but also ended up being correlated with higher tax rates, and for that reason made sampling more difficult for no real benefit in understanding.
- Country - We had initially added a variable for whether a team was located in Canada or the United States, but this had no impact on salary through EDA and also had almost no impact during modeling. Therefore, we dropped it for simplicity.

This left only the tax rate as a variable of interest at the team level. From there, we decided to allow other systematic differences among the teams to be captured by the intercept term. Contract features were omitted completely to keep the model simple and parsimonious as a first iteration. Though we omitted all contract features, it remains an area for future research and refinement.

At the player level, we set up a separate hierarchy by position to estimate the global impact of points, defensive zone giveaway, expected goals differential, penalty minutes, shots blocked, and hits (each per 60). Finally, this allows us to estimate the hypothetical salary that a player would receive on each NHL team, given their stats. We selected the Laplace prior on each of these statistics, so that position by position, irrelevant statistics can be discarded.

To aid in fitting, we implemented a non-centered parameterization[4]. This, along with some hyperparameter tuning, eased sampling significantly. Though omitted here, model diagnostics agreed with this approach; we had Gelman-Rubin statistics of 1.0 for all parameters of interest, adequate autocorrelation plots and sensible traceplots, with very few divergences (1 in 4000 iterations).

Results

Our relatively simple model does not show any obvious patterns with regards to the "Original Six" or location. There are differences, to be sure, but no team's intercept term (average salary) in our model is far and away less than others. That being said, a conspiratorially-minded hockey fan could look at Toronto's average and remark that it is suspiciously low, indicating that their players tend to be on franchise-friendly deals.

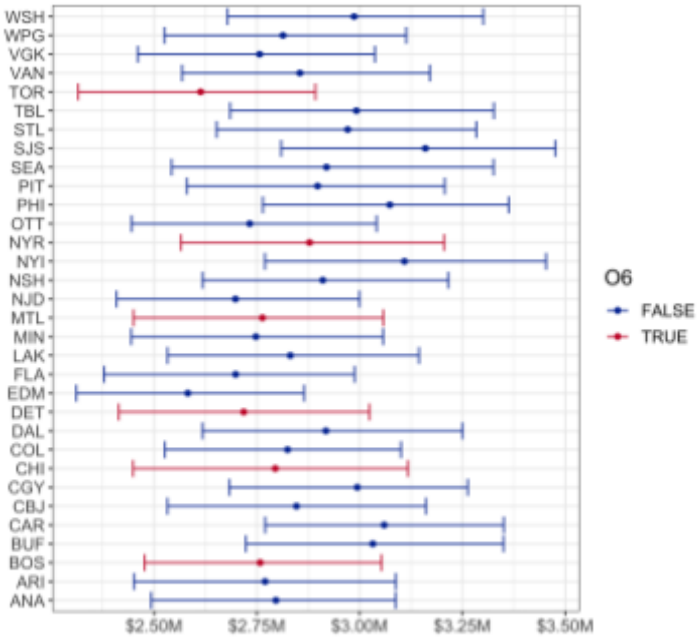

Figure 6
Team intercepts: "Original Six" teams are indicated in red.

[4] "A Primer on Bayesian Methods for Multilevel Modeling — PyMC3 3.11.5 Documentation."

This is not as far-fetched as it sounds. In fact, according to our simulations, the average salary for a standard one-way contract in Toronto is lower than the same player would receive in San Jose, all else equal. And for the Philadelphia Flyers, there is about an 85% chance that they're paying more for an equivalent player than their in-state rival, the Pittsburgh Penguins.

The table below shows important coefficients from our model. The effect of tax rates on player salary is marginal at best. Most likely, it has no practical effect. Surprisingly, what little evidence there is points in the opposite direction that one would expect - there's about a 70% chance that the coefficient for tax rates is less than 0, meaning that as the tax rate a team is subject to increases, the average salary of a player is expected to *decrease*. And low tax teams such as the Tampa Bay Lightning, Florida Panthers, and Dallas Stars more than likely pay more for players than high tax teams such as the Toronto Maple Leafs and the Montreal Canadiens.

	mean	hdi_3%	hdi_97%	ess_bulk	r_hat
tax_beta	-0.262	-1.079	0.594	2454	1.0
I_F_points[R]	0.388	0.321	0.46	4185	1.0
I_F_points[L]	0.318	0.253	0.382	4190	1.0
I_F_points[C]	0.345	0.291	0.401	4075	1.0
I_F_points[D]	0.713	0.613	0.815	4053	1.0
I_F_hits[R]	-0.042	-0.057	-0.025	3350	1.0
I_F_hits[L]	-0.047	-0.059	-0.035	3981	1.0
I_F_hits[C]	-0.065	-0.079	-0.051	3256	1.0
I_F_hits[D]	-0.047	-0.061	-0.031	3429	1.0
shotsBlockedByPlayer[R]	-0.019	-0.073	0.041	3922	1.0
shotsBlockedByPlayer[L]	0.003	-0.048	0.055	4946	1.0
shotsBlockedByPlayer[D]	0.099	0.077	0.121	4132	1.0
shotsBlockedByPlayer[C]	0.006	-0.034	0.043	3910	1.0

That being said, individual statistics play the most important role in determining salary. Looking position by position, individual points (goals + assists) have the largest impact on salary by far, and this effect varies according to position. Forwards have roughly similar coefficients on points per 60, but defensemen are paid much, much more for their offensive ability. Conversely, teams that want their defensemen to contribute significant points will have to pay handsomely for that production.

Next, hits per 60 is significant across all positions, but is actually a negative coefficient. The more physically oriented a player is, the less a team is willing to pay (or the less salary they can command). This is not *necessarily* because it is not a valuable stat. To some degree, hits are a measure of defensive ability. But the reality is that other stats are more valued on defense, as

seen by the positive coefficient on shots blocked (but only for defensemen). For any team looking to expand hitting ability, the trend is to sign cheaper players to fulfill this role, and so the skill commands lower pay as a result.

Conclusion

Our study has two primary findings: first, individual player statistics are the absolute strongest predictor of a player's salary. Not all statistics are created equal, however - points were hugely indicative, while our Laplace priors on other stats effectively set unimportant values to zero. Coefficients on statistics vary greatly across positions.

Second, individual teams have different expected average salaries given a set level of goal production. However, tax rate is not an important factor in predicting whether a particular team can pay less for players. For example, all else being equal, Toronto pays less for high production players than the Dallas Stars, despite Dallas having the lowest effective tax rate in the league. But with the exception of Toronto, other original six teams do not receive an obvious discount from players during salary negotiation.

In the future, we would like to consider more player preferences such as the weather, size of the city, distance from the hometown of a player, and contract specific features (no-move clauses, no-trade clauses, etc). More team-centric factors such as recency of success, strength of the team prior to signing, and expected Stanley Cup winning odds would also be important to consider. The addition of these factors could help us to evaluate typical fan theories in an empirical light for why certain players sign team-friendly deals (if this is in fact common).

The hierarchical model structure gives us a better way to estimate these effects than a typical OLS implementation, notably by having a separate model for team level factors that we estimate at the same time as our player centric model. Next, the flexibility of using robust priors and estimating the effect of player statistics on player salary position-by-position gives us estimates of effects at multiple levels without underestimating the high variance inherent to our problem. This model gives us a flexible way for us to add both team-centric and player-centric features for future investigations.

Finally, and most importantly (as a Flyers fan), I can take solace that my team is so terrible partly because we have to pay more for our goals. Though the randomness of the game and the skill of the players has far more to do with the winning-ness of the teams than their salary negotiation environment, at least it's something.

Contributions

Matthew McAnear contributed most of the code for the initial modeling in PyMC and data gathering. Matt also compiled the results and gathered coefficient estimates from the model. Matt also handled the final model, some of the data gathering, and results section of the paper. Main packages Matt used are included in the paper.

Anthony Paolillo provided valuable domain expertise and helped with designing the model's hierarchical structure, as well as most of the data cleaning and exploratory visualizations to ensure that our data was gathered correctly. Anthony also wrote the intro, conclusion, and most of the data gathering section, as well as the proposal and slides for the presentation.

Bibliography

"A Primer on Bayesian Methods for Multilevel Modeling — PyMC3 3.11.5 Documentation." Accessed April 17, 2023. https://www.pymc.io/projects/docs/en/v3/pymc-examples/examples/case_studies/multilevel_modeling.html.

CapFriendly. "CapFriendly - NHL Salary Caps," April 18, 2023. https://www.capfriendly.com/.

CapFriendly.. "Income Tax Calculator," 2023. https://www.capfriendly.com/income-tax-calculator/brayden-point.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. "Array Programming with NumPy." *Nature* 585, no. 7825 (September 2020): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Kumar, Ravin, Colin Carroll, Ari Hartikainen, and Martin Osvaldo. "ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in Python." *Journal of Open Source Software* 5, no. 55 (December 2020): 1143. https://doi.org/10.21105/joss.01143.

McAnear, Matt. "Hockey Salaries," 2023. https://github.com/mcanearm/hockey_salaries.

Reitz, Kenneth and Contributors. "Requests: HTTP for Humans," 2021. https://docs.python-requests.org/en/master/.

Richardson, Leonard. "BeautifulSoup." *April*, 2007.

Salvatier, John, Thomas V. Wiecki, and Christopher Fonnesbeck. "Probabilistic Programming in Python Using PyMC3." *PeerJ Computer Science* 2 (April 2016): e55. https://doi.org/10.7717/peerj-cs.55.

Tanner, Peter. "MoneyPuck.Com -Download Datasets." Accessed March 25, 2023. https://moneypuck.com/data.htm.

The pandas development team. "Pandas-Dev/Pandas: Pandas." Zenodo, April 2023. https://doi.org/10.5281/zenodo.7794821.