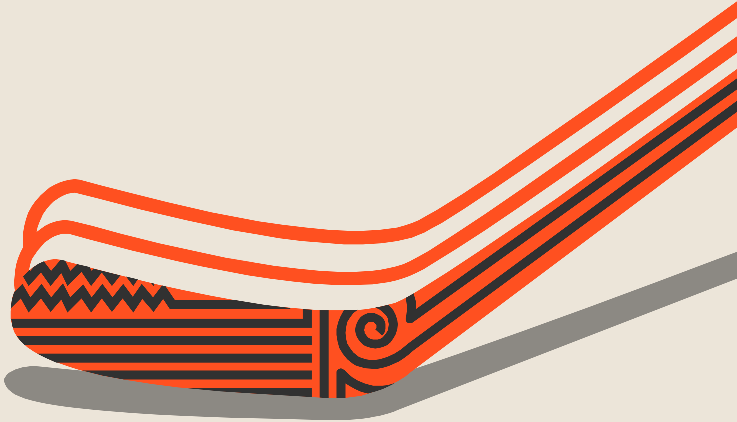
BTMA 431 Final
NHL Statistics

Armaan Atwal
Vikram Brar
Rupin Sehgal
William Spencer

TABLE OF CONTENTS



01

Introduction

Introduction

- The motivation behind our analysis on NHL statistics came from the desire to measure performance in the NHL using player and team data
- Our goal was to find insights within the scraped data that could help develop ideas that can help increase the performance of both players and teams
- Some examples of analysis conducted include models that identify significant key metrics and the relationship they have with other metrics.
- The insights developed from the completed analysis were then transformed in to visual representations to simply the analysis and deliver it in an appealing form
- Recommendations and conclusions are drawn from the analysis and subsequent visuals

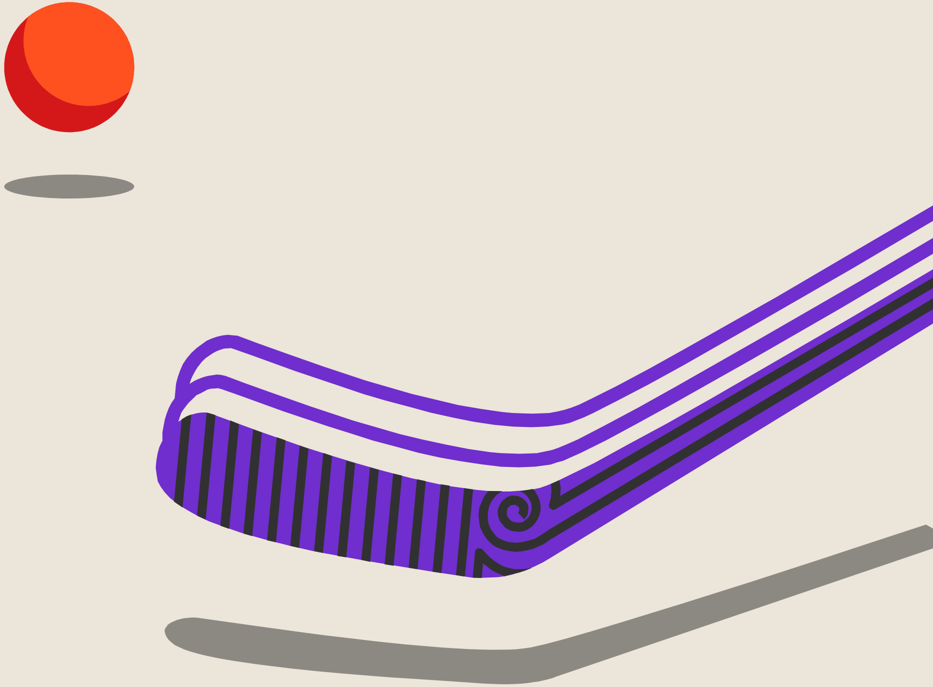
Beneficiaries from our study?

Hockey enthusiasts

NHL Team Management

Hockey Bettors and Agencies

02
Data Scraping & Cleaning



DATA COLLECTION & SCRAPING

01
GOAL
Collect player and team statistics

02
HTML
Downloaded the webpages' HTML into the Global Env.

03
CSS
Used the CSS selector to find our tables
'#Stats = Player Data'
'#active_franchises'

04
TABLES
Used dplyr, rvest, and tidyverse for implementations and initial table setup

Data Cleaning - The big issues

Missing Data

Some entries were missing values

NA	80
94	105
172	NA

Duplicate Data

Both datasets had duplicate data

Anaheim Ducks	NHL
Anaheim Ducks	NHL
Mighty Ducks of Anaheim	NHL
Arizona Coyotes	NHL
Arizona Coyotes	NHL
Phoenix Coyotes	NHL

Wrong Data Type

After scraping player Data all columns were character type when some should have been numeric

```
     Player         Age
"character" "character"
```

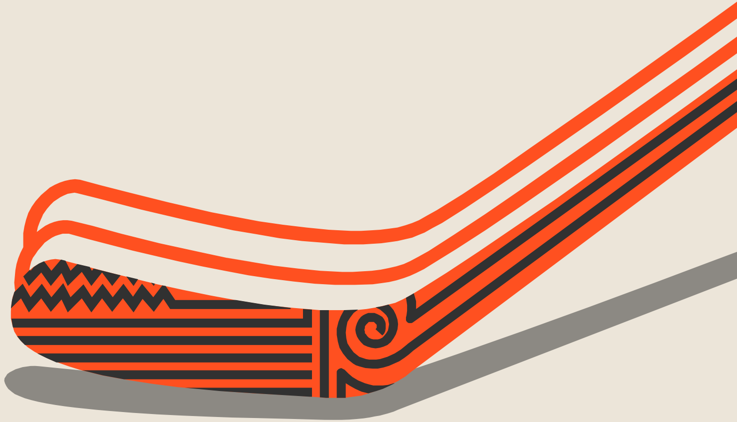
Teams Data Cleaning

01 Replaced all NA cells from the data and with 0

02 Find and all teams that stopped playing before 2023 and replace data with NA

03 Find all duplicate teams and replace the older team with NA

04 Remove all NA rows (teams that no longer exist, and duplicate teams)



Player Data Cleaning

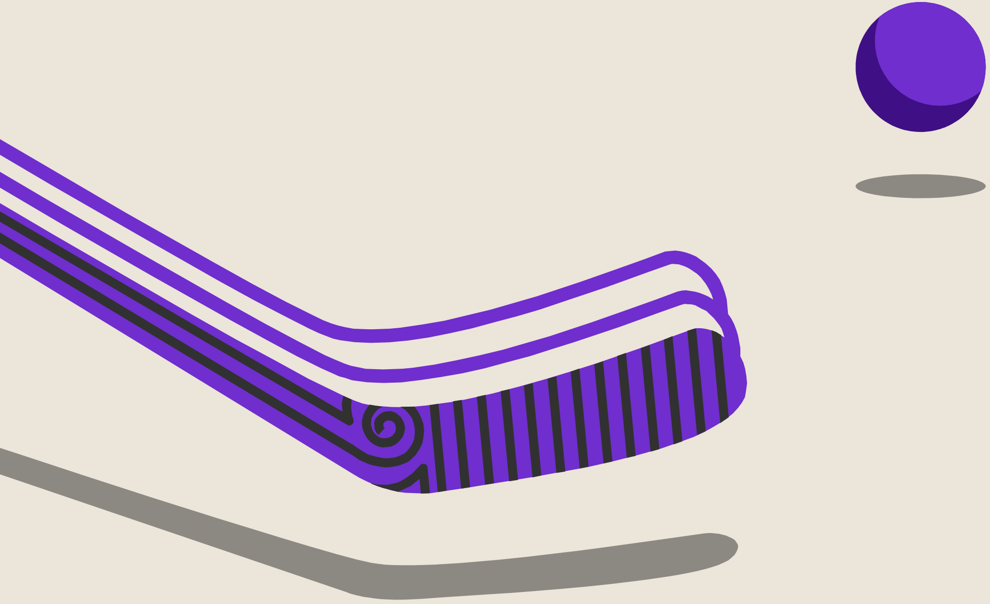
- Step 1
  - Column names
- Step 2
  - Remove duplicated players
- Step 3
  - Remove duplicate headers
- Step 4
  - Find all empty cells and fill with 0
- Step 5
  - Check data, and correct data types for all numeric columns

							Scoring
1	Rk	Player	Age	Tm	Pos	GP	G
2	1	Nicholas Abruzzese	23	TOR	C	2	0
3	2	Noel Acciari	31	TOT	C	75	13
4	2	Noel Acciari	31	STL	C	54	10

1	Rk	Player	Age	Tm	Pos	GP
159	Rk	Player	Age	Tm	Pos	GP

491	51.2
0	
97	35.3
0	

```
     Player         Age
"character" "character"
```



03

Analysis

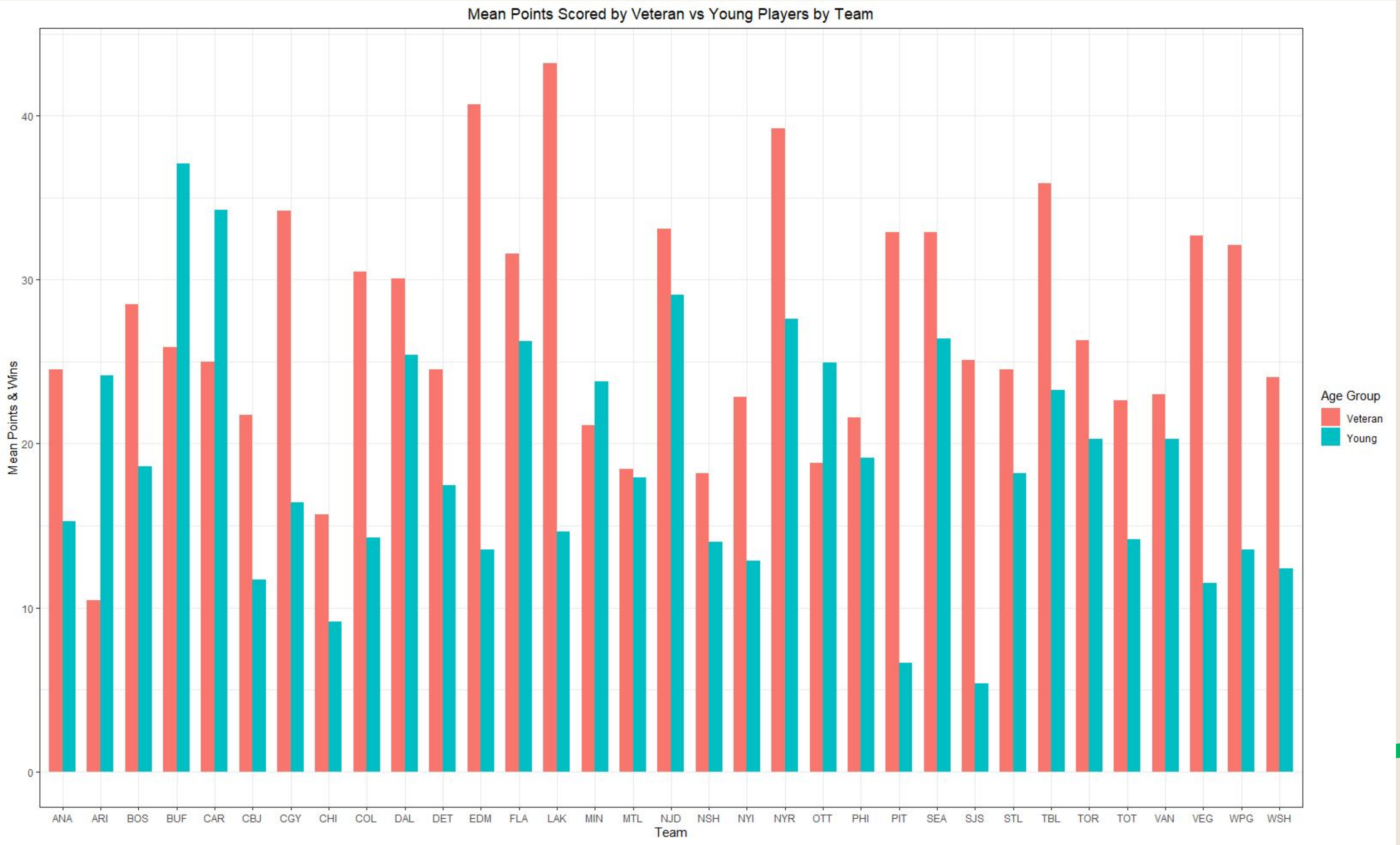
Q1

VETERAN VS YOUNG PLAYERS IMPACT ON A TEAM'S PERFORMANCE

Q1: 01 Data Setup

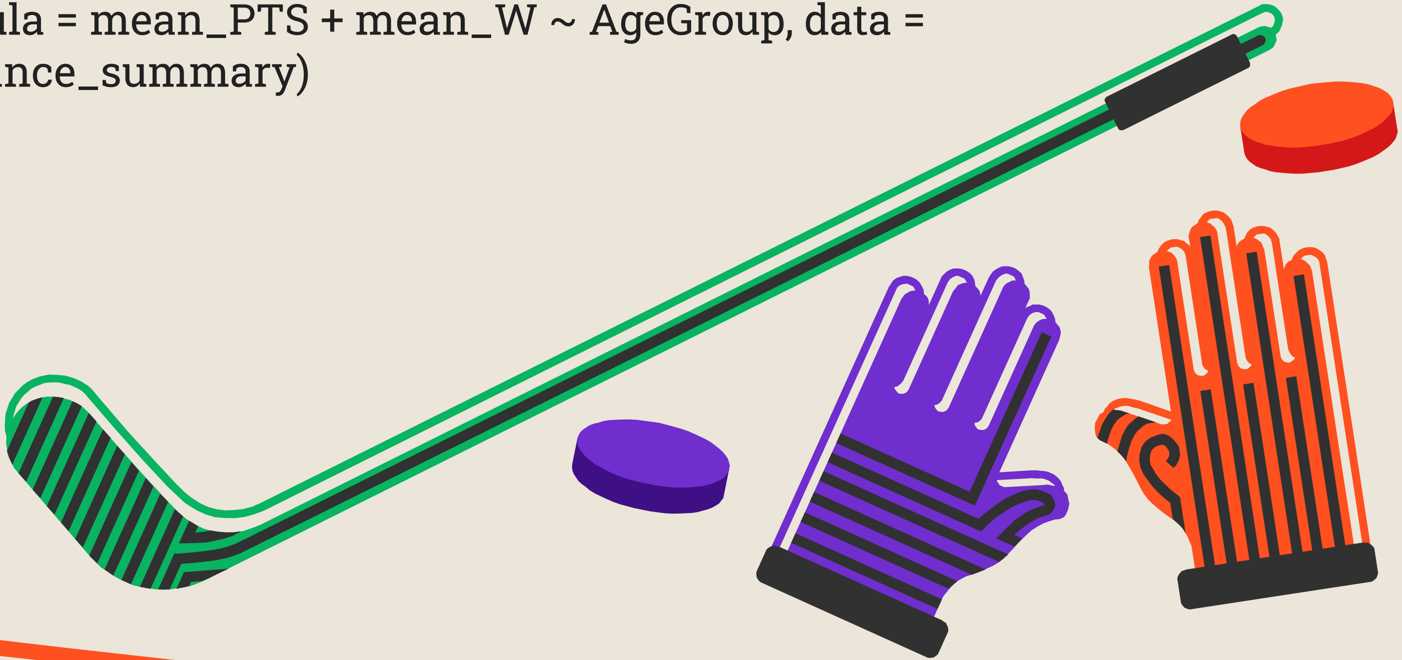
- **Step 1**
  - Mutated a new column that classifies Age Groups >25 as 'Veteran' and <25 as 'Young'
- **Step 2**
  - Grouped age classified players with their respective team, average points, and wins

Q1: 02 Data Visualization



Q1: 03 MODEL SETUP

- **LM MODEL USED**
  - Predicting mean wins and Points using Age Groups
  - lm(formula = mean_PTS + mean_W ~ AgeGroup, data = performance_summary)



Q1: 04 MODEL INSIGHTS

- **LM MODEL SUMMARY**
  - Both Age groups are significant in predicting the output (P values <= 0.05)
    - Coefficient of Veteran Players: 42.03
    - Coefficient of Young Players: -13.11
  - R squared: 0.29
    - 29% of the variation in mean wins and points can be explained by age
    - Indication of factors other than age group impacting performance

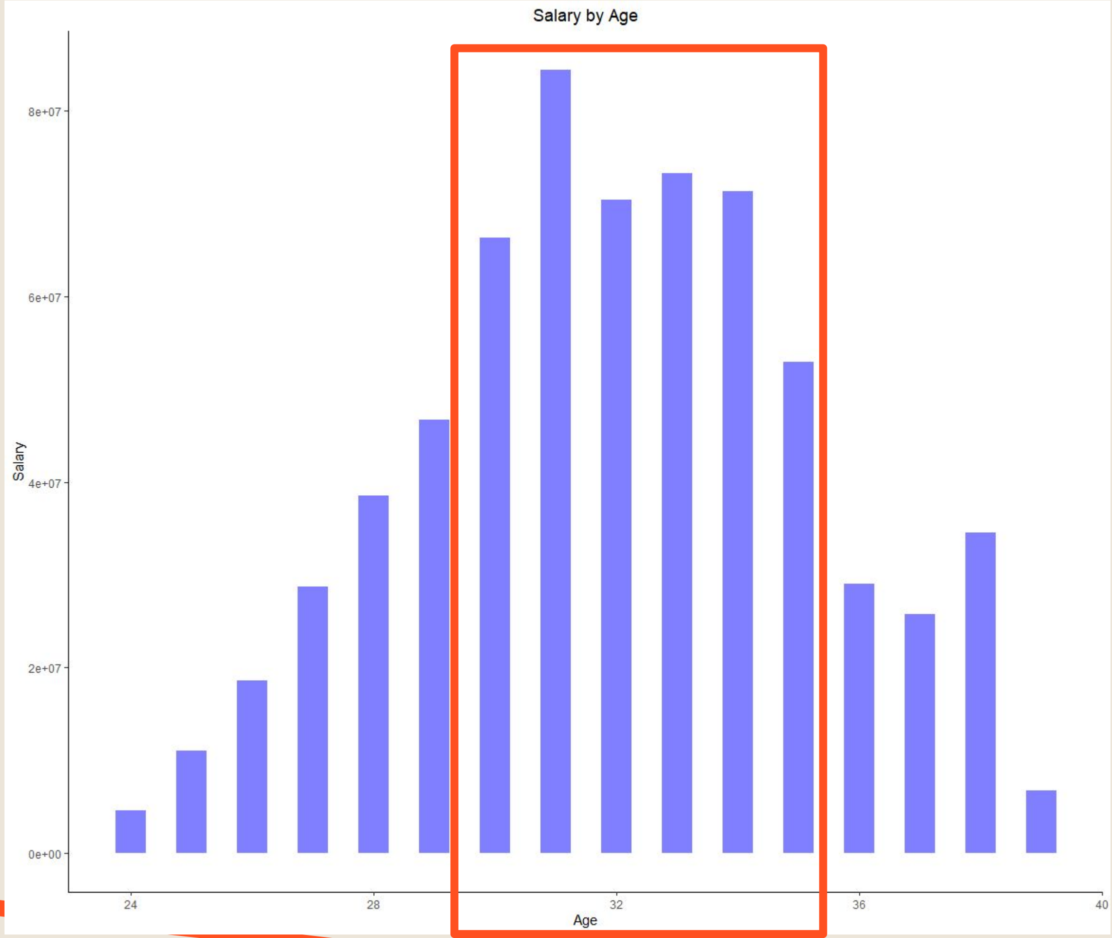
Q1 PART 2

What are the key determinants of a player's potential earnings?

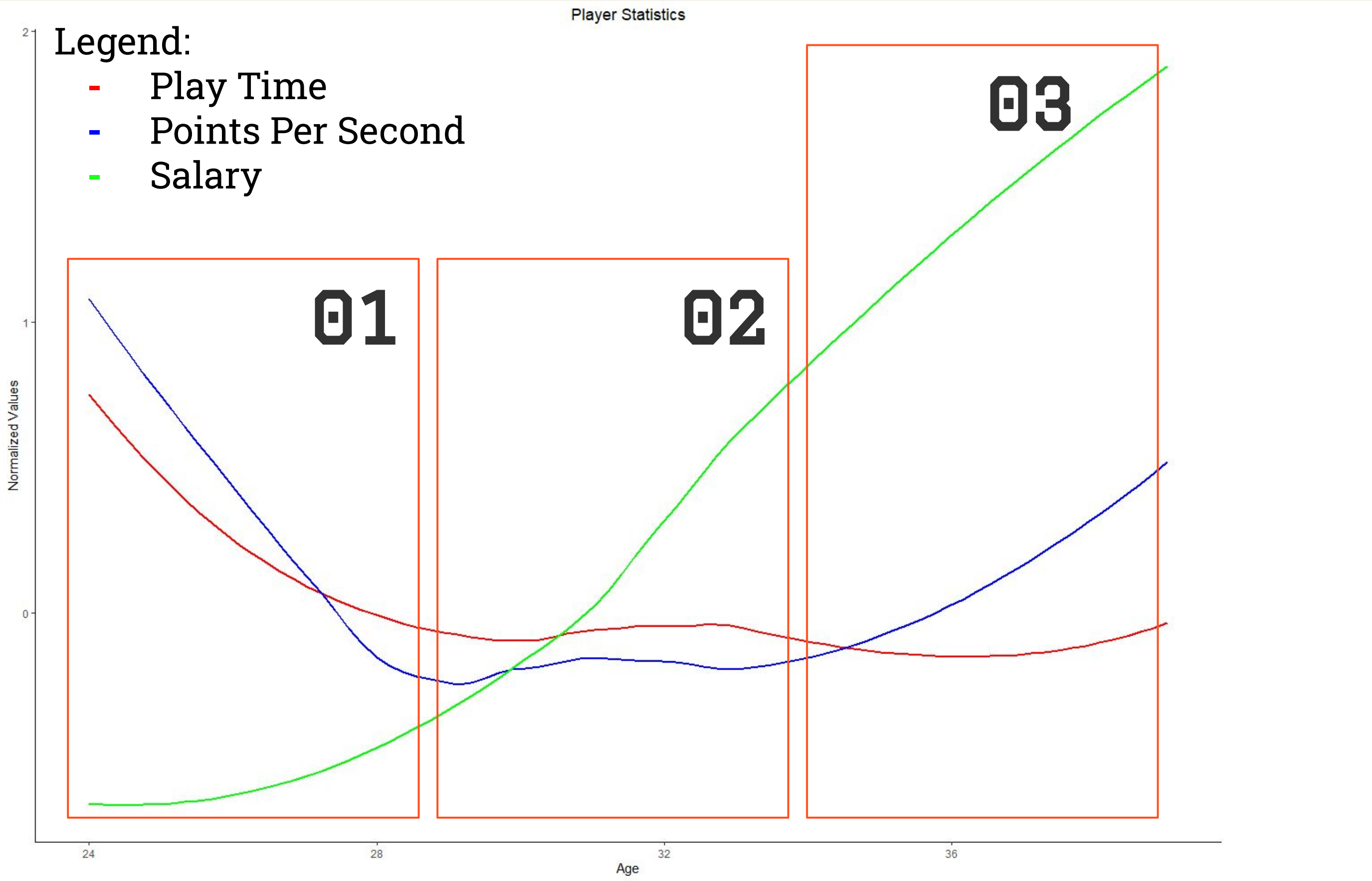
Q1.2: 01 Data Setup

- **Step 1**
  - Formatting columns needed correctly
    - E.g. transforming ATOI from mins and secs (%M%S) to only secs (%S)
- **Step 2**
  - Mutating a new column to represent points scored per second
    - To help analyze a players efficiency on ice
  - Supplement scraped data with Player Salaries
    - Used Merge to combine data, using Player Name as the reference id

Q1.2: 02 Data Visualization



Q1: 02 Data Visualization



Q1.2: 03 MODEL SETUP

- **LM MODEL USED**
  - Predicting salaries using all available predictors
  - lm(formula = playerCleaned.df$Salary ~ ., data = playerCleaned.df[,c(4, 6:17, 19:26)])

Q1: 04 MODEL INSIGHTS

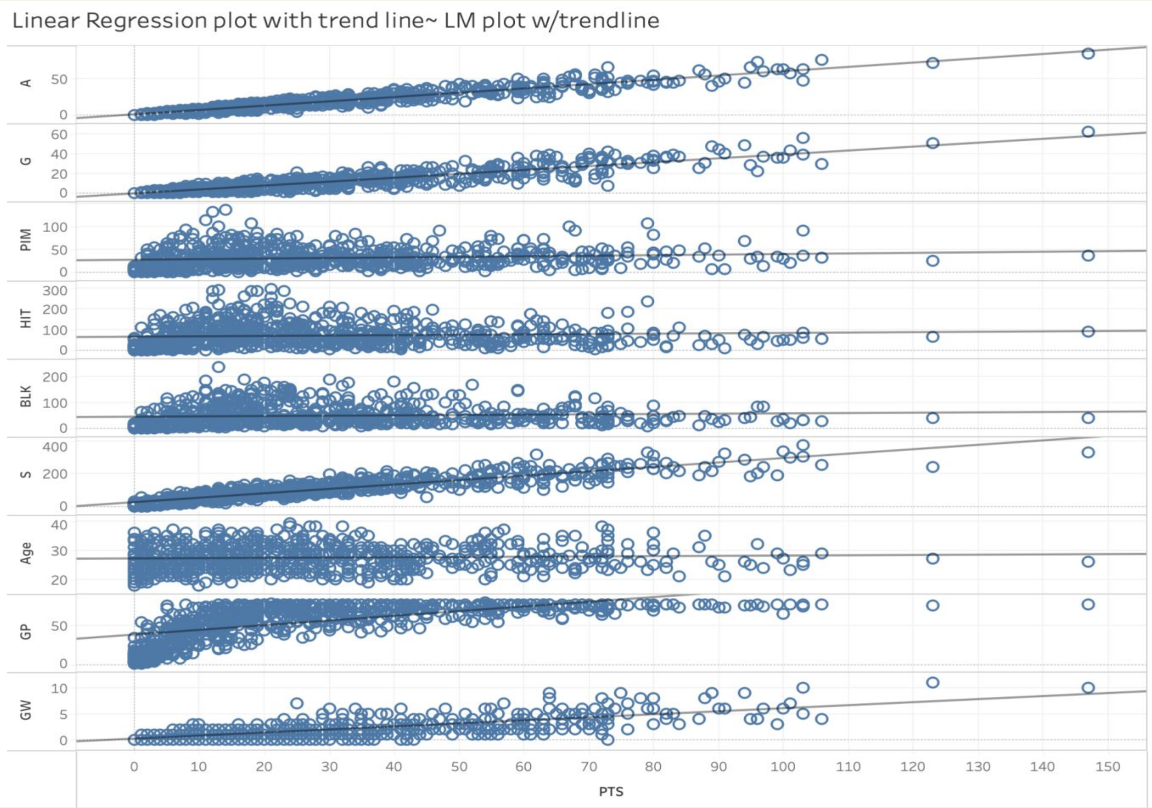
- **LM MODEL SUMMARY**
  - Positional variables do not have an impact on Salary (P values >= 0.05)
  - Age, Goals, +/-, PS, FOW, FOL, secs played all have a significant impact on a player’s salary (P <= 0.05)
    - E.g. The model indicates that with every year a player’s salary increases by $557,500
  - R squared: 0.56
    - 56% of the variation in salary can be explained by our predictors

Q2

What game statistics have a significant impact on a players points ?

LM MODEL USED

- Created a regression model to identify what statistics were reliable predictors of points scored
- To predict Points scored by player we used predictors such as Age, Goals, Etc...
- lm(formula = PTS ~ G + A + GP + PIM + GW + S + BLK + HIT + Age, data = playerCleaned.df)

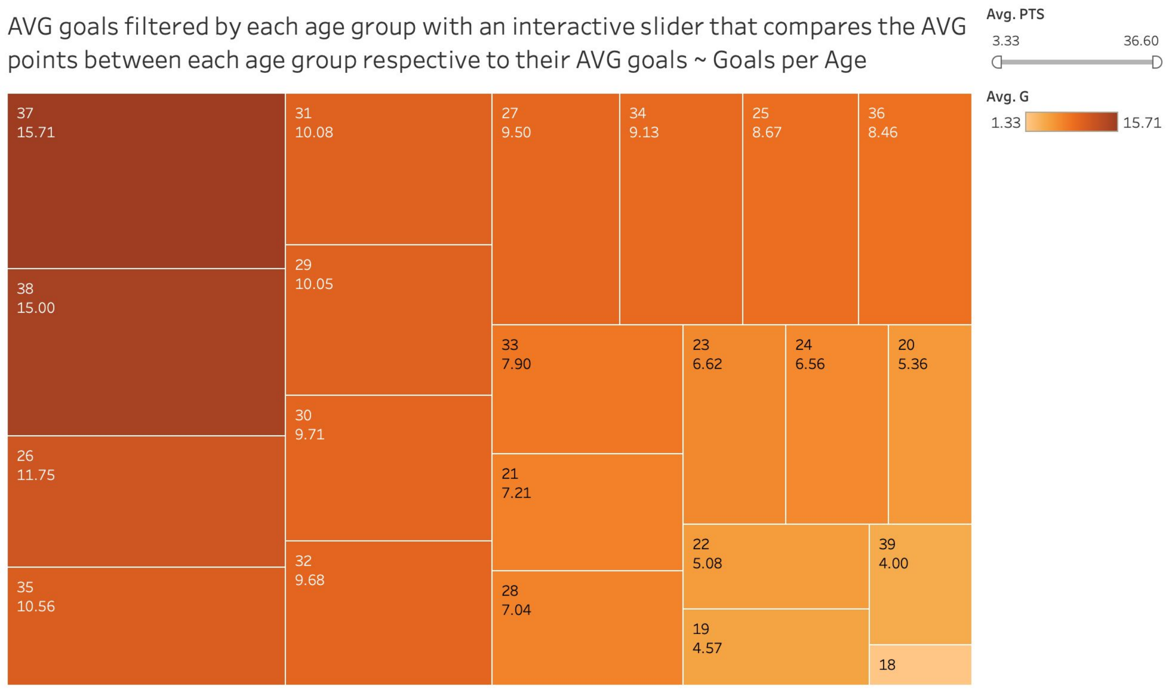


- **LM MODEL SUMMARY**
  - Goals, Assists, Hits, and penalty minutes all were all significant predictors with P<= 0.05.
  - Residual STE: 1.126e^-13
    - Such a small residual error suggests, on average, our models predicted values are very close to the actual values.

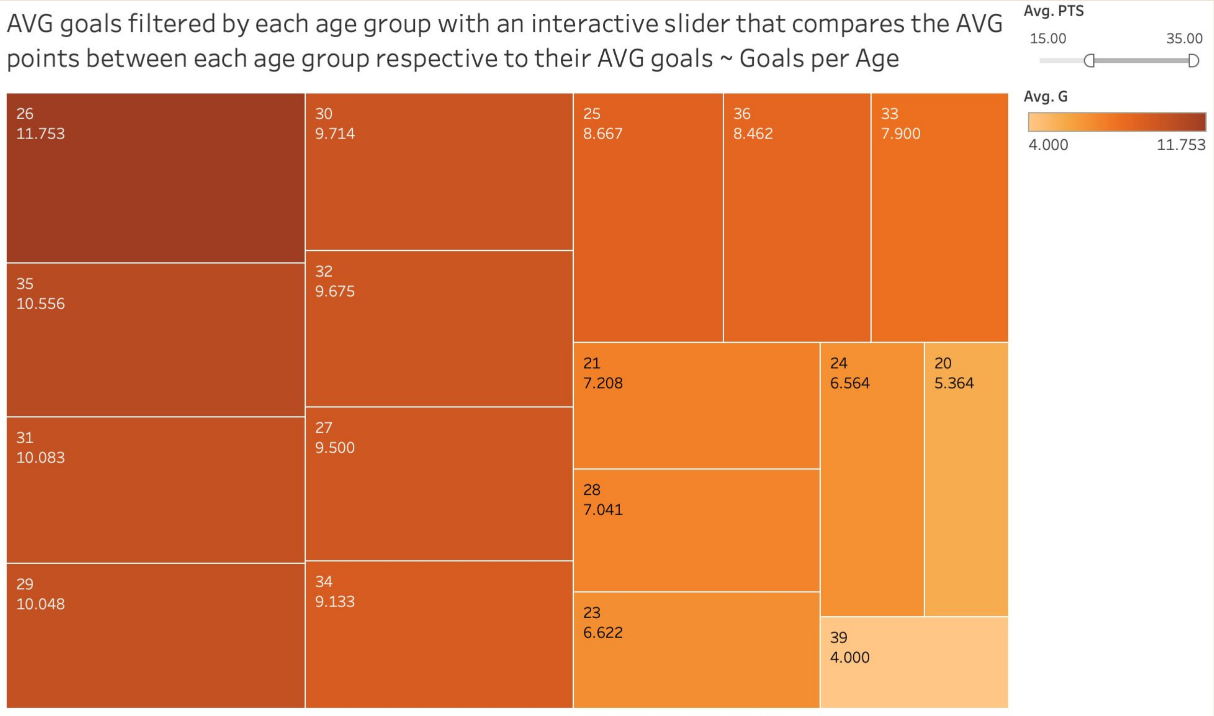
Q2 Part 2

How does the average goals scored per player change across all age groups?

- Heat map showing AVG goals per age
- Slider function in regards to AVG points



- Slider function applied between 15 and 35 points
- Most outliers are now ignored



Q3

Predicting Stanley Cup wins

Predictors Not Used

Franchise

The Team's franchise was unique to each team

League

The only Teams that play for the Stanley cup are in the NHL

Team Inception

This data is best captured by the years played

Final year of Play

This would be 2023 from all Teams

League Champion Wins

This correlates nearly 1:1 to cup wins

Steps

Check predictor Correlation

Strong Correlation

We settled on 0.7 as the break point for strong or weak correlation

Assess Weakly Correlated linearity

Linearity

If predictors had a correlation of less than 0.7 they would be charted to check for non-linear relations

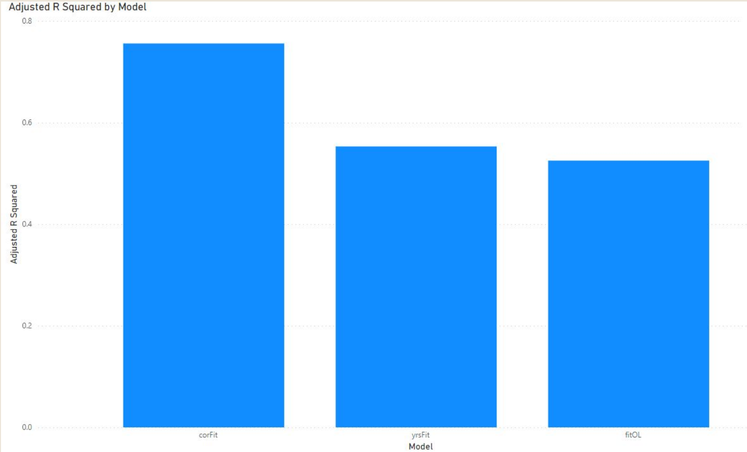
Regression

Model

I would run multiple regressions based on Correlation, and to try and find non-linear relationships

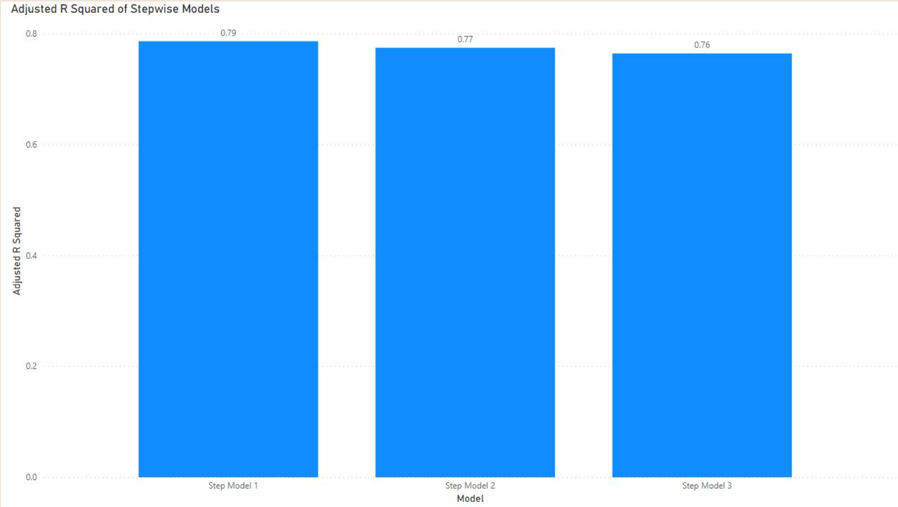
Models

- corFit (St Cup ~ Yrs + GP + W + T + Yrs Plyf + Div)
    - Created with only strongly correlated predictors
- yrsFit (St Cup ~ Yrs)
    - Created with only the statistically significant predictors from corFit
- fitOL (St Cup ~ Yrs + OL + OL^2 + OL^3 + OL^4)
    - Created using yrsFit plus overtime assessed at multiple powers



Stepwise Models

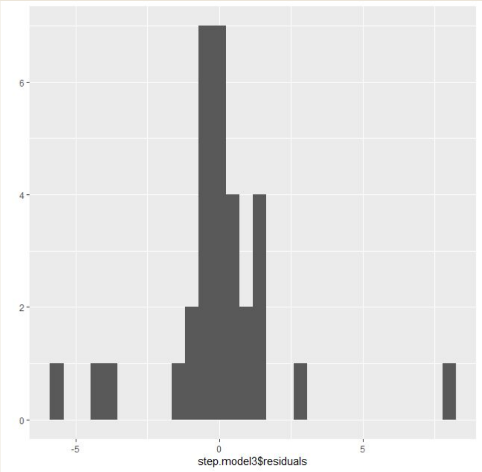
- Step1 (St Cup ~ Div + T + Yrs + Gp + Conf)
    - I used a stepwise function prioritizing AIC, with a forwards and backwards search.
- Step2 (St Cup ~ Div + Yrs + Gp + Conf)
    - I removed the predictor with the largest p-value (Ties)
- Step3 (St Cup ~ Yrs + Gp + Conf)
    - I removed the predictor with the largest p-value (Years finished at top of division). All remaining predictors had a p-value < 5%



Best Model

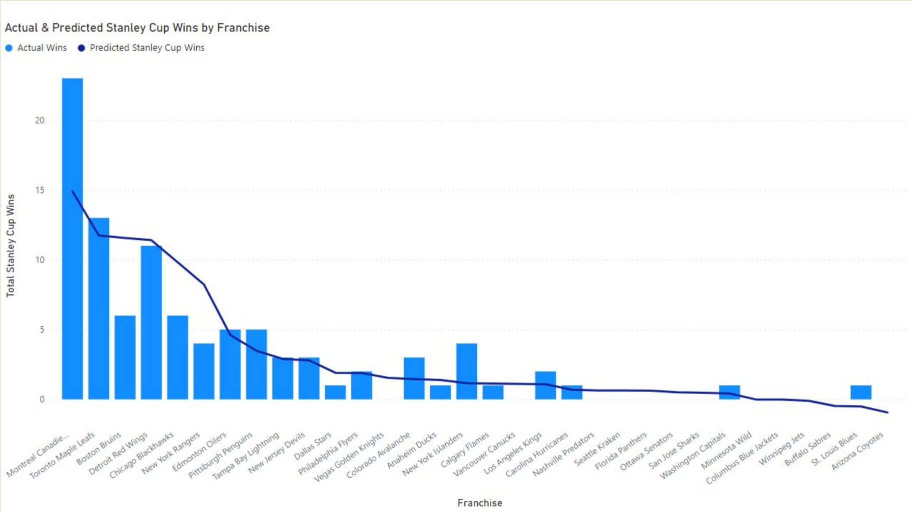
step3

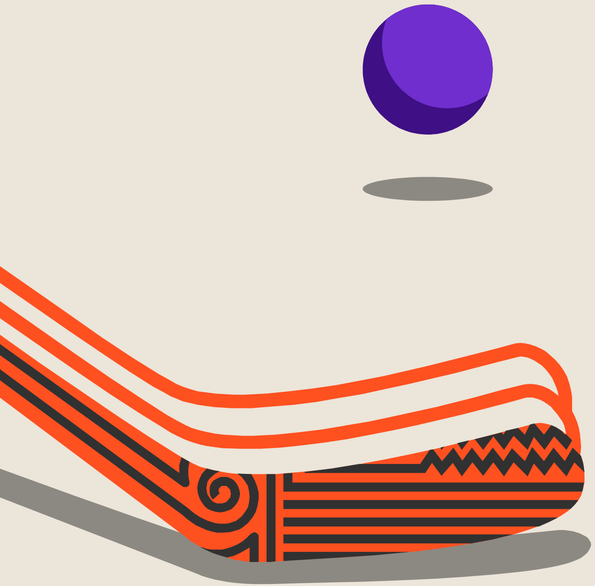
St Cup ~ Yrs*0.85 + GP*-0.01 + Conf*0.79



We decided that step1 was the best model generated because:

- Adjusted R Squared (0.7639)
  - It had the highest adjusted R squared values out of all other statistically significant models
- Residuals
  - The residual error follows a normal distribution





04 Recommendations

Conclusion

Data

More player data from numerous seasons would result in more accurate results

Scraping

Availability of scrapable websites

Model diversity

More models would allow for better comparisons and higher quality analysis

THANKS FOR LISTENING!