Project 1: Final Report

I. Introduction.

1. Data understanding.

Our dataset originates from multiple reputable sources, including Statistics Canada, the NHL team list endpoint, and the NHL API. Our investigation delves into the intriguing realm of professional ice hockey demographics. Our dataset serves a critical role in assessing the enduring relevance of Gladwell's observations within the contemporary landscape of the National Hockey League (NHL). By analyzing player birthdate data extracted from the NHL API, we aim to elucidate whether similar patterns persist among Canadian-born NHL players. Additionally, our analysis seeks to uncover broader demographic trends and potential biases within the NHL player population.

2. Key Questions:

Within the interest in NHL league, we confront two pivotal questions that bear significant implications for team dynamics and strategic decision-making:

- The first inquiry delves into the validity of Malcolm Gladwell's assertion regarding the dominance of players born in the first three months of the year. Gladwell's hypothesis posits that individuals born in these early months hold a statistical advantage in sports, a concept often referred to as the "relative age effect." Through rigorous analysis of player birthdates in the National Hockey League (NHL), we endeavor to ascertain whether Gladwell's claims hold true within the context of professional ice hockey.

- The second question at the forefront of our investigation pertains to the utilization of local talent by NHL teams. As sports franchises aim to cultivate a sense of community and identity among their fan base, the recruitment and retention of local players can serve as a strategic asset. By scrutinizing team rosters and player demographics, we seek to elucidate whether NHL teams exhibit a propensity towards selecting players from their immediate geographic regions or whether talent acquisition transcends local boundaries.

II. Question 1: Is there any bias regarding the birth city?

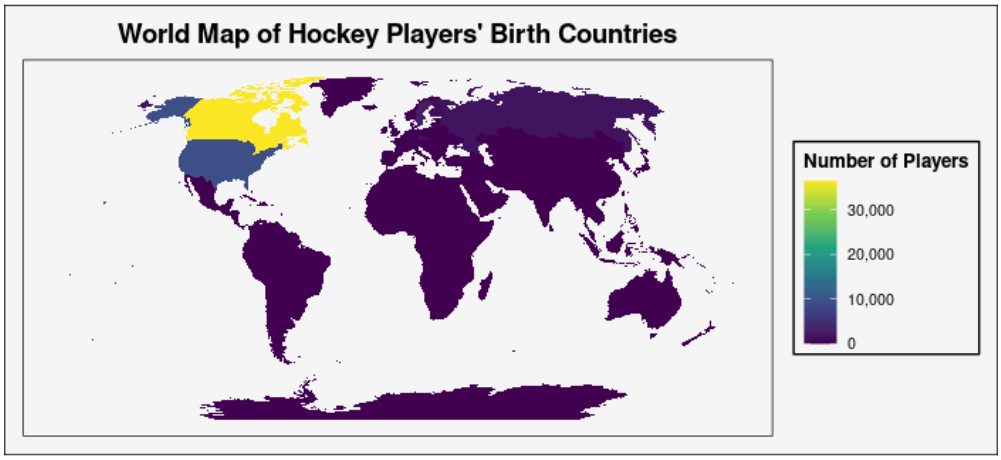
1. Introduction:

To examine potential biases in the recruitment processes of teams in the National Hockey League (NHL), a comprehensive analysis must be undertaken, focusing on multiple dimensions. This should include an evaluation of demographic data, scouting reports, and hiring trends over a significant period. Key demographic factors such as race, nationality, and age could be investigated to discern any patterns that may suggest biases. Furthermore, the evaluation of scouting reports could reveal if certain demographics are consistently described in biased terms, which might affect their chances of being recruited. Historical data on team compositions and recruitment policies can also provide insight into whether there has been a progression toward inclusivity or if certain biases have persisted. Additionally, interviewing former and current scouts and players could give more context to the quantitative data, helping to identify any unconscious biases in player evaluation and team selection. This multifaceted approach will enable a more accurate assessment of whether recruitment practices in the NHL are biased and, if so, in what ways.

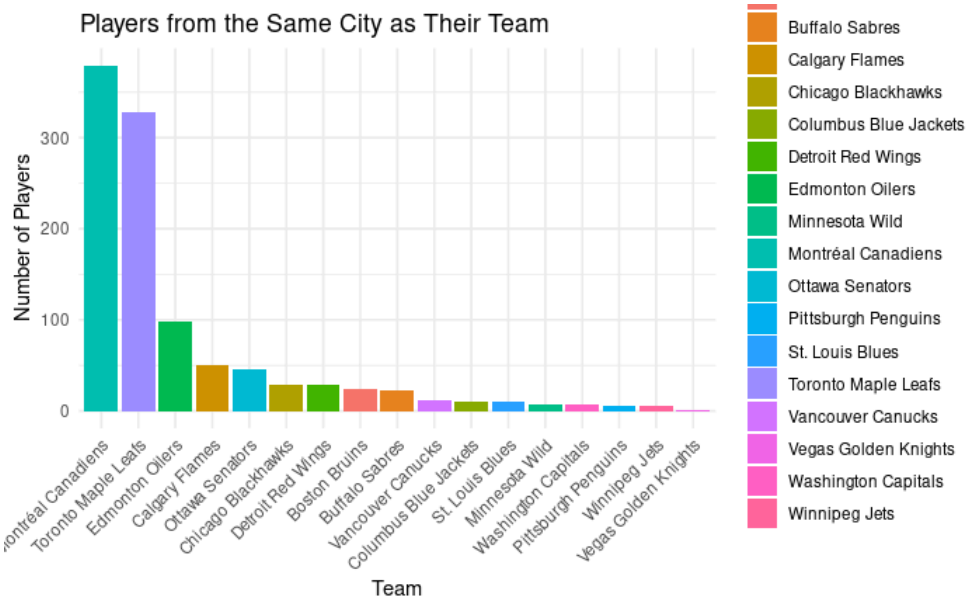
2. Approach

To rigorously investigate potential biases in the NHL's recruitment process, our analysis will be conducted in two main phases. Firstly, we will assess whether there is a preference for players from certain countries across the league. This will involve examining the nationalities of players who have been drafted and those who make it to professional levels, comparing these figures against the global pool of available talent. We will also look into the presence and performance of international scouting networks to see if there are disparities in how players from different regions are scouted and valued. Secondly, we will explore whether individual teams show a bias towards local players, specifically those born in the same city or region as the team. This part of the analysis will include reviewing historical recruitment data, player hometowns, and the geographic distribution of player origins within each team's roster. By examining these aspects, we can better understand how local biases might influence team compositions and if such preferences affect the overall diversity and talent distribution within the league. This thorough, two-pronged approach will help us identify and contextualize any biases in recruitment practices, providing a clearer picture of inclusivity and fairness in the NHL.

3. Analysis



We used the ggplot2 and sf libraries to visualize the distribution of NHL players' birth countries on a world map, also employing the dplyr package for data manipulation. We started by loading a medium-scale map of world countries as a simple features object. We assumed the existence of an NHL rosters dataset, which we grouped by birth country to count the number of players from each. We then merged this player data with the world map data, explicitly setting zero counts to NA to highlight them visually. We chose ggplot2 for plotting, applying the Viridis color scale for aesthetic clarity and using comma-separated labels. We opted for a Robinson projection to enhance the map's visual appeal. We also made various aesthetic adjustments to the plot to enhance its readability and overall presentation.



We start by loading team data from a CSV file named "NHL.csv" into a data frame team_data. Next, we merge this with another data frame nhl_rosters that presumably contains player rosters. The merging is based on matching 'team_code' in the rosters with 'abbreviation' in the team data. After merging, we view the first few rows of the combined data using the head function to check the initial rows for correctness. We then add a new column same_city to the

merged data to indicate whether a player's birth city matches the team's city. This is done using the mutate function, where we compare the venue/city with birth_city and assign 1 if they match, otherwise 0. Subsequently, we summarize the data to count how many players per team are from the same city as their team. This involves grouping by the team name and summing up the same_city flags. Teams with no players from their city are then filtered out. Finally, we create a bar plot to visualize the number of players who are from the same city as their team. This plot is crafted using ggplot2, arranging teams by the number of such players, and applying minimal theme styling for clarity. Labels for the axes and the plot title are set appropriately, and the x-axis labels are rotated for better visibility.

4. Discussion

The first figure, a world map adorned in a spectrum of colors, serves as a vivid indicator of the birth countries of hockey players around the globe, with the intensity of color corresponding to the density of players from each region. A striking dominance of darker hues blankets the northern hemisphere, particularly over North America and Northern Europe. This conspicuous pattern underscores a potential geographical bias, hinting at a concentration of talent emerging from these latitudes. Such a trend may stem from the deep-seated cultural roots of hockey in these regions, the availability of ice and facilities, or the structured development and training programs that have historically cultivated the skills of young players in these colder climates.

Conversely, the second figure, a bar chart, provides a granular look at the local recruitment biases within individual NHL teams. The chart's bars leap upward for certain teams, graphically representing a significant number of players whose birth cities coincide with their team's location. This phenomenon could suggest a preferential bias toward homegrown talent, a facet of recruitment that may reflect the teams' marketing strategies aimed at strengthening community support by featuring local stars. Moreover, it could also be reflective of the teams' investment in local sports programs that generate a pipeline of talent directly accessible to them.

Together, these two figures weave a narrative of local and global influences that shape the NHL's team compositions. They point towards a multi-layered fabric of recruitment strategies, one that intertwines the threads of regional hockey culture, the nurturing of local talent, and international scouting mechanisms. This visualization provokes a comprehensive analysis of the recruitment process, raising questions about the fairness and scope of talent searches, the potential for overlooked talent in less-represented regions, and the impact of local allegiances on the international stage of professional hockey.

III. Question 2: Whether Malcolm Gladwell's assertion is still valid?

1. **Introduction:**

Our interest in exploring the impact of birth month on professional ice hockey stems from its potential to challenge existing perceptions and practices within the sport. As we initially delved into the dataset, we encountered discussions surrounding Malcolm Gladwell's hypothesis, which suggests that individuals born in the first three months of the year may possess a statistical advantage in sports due to developmental factors. This notion intrigued us and sparked a curiosity to delve deeper into its validity within the context of the NHL.

Understanding whether birth month significantly influences players' success could revolutionize how talent is identified, nurtured, and supported within the ice hockey community. If Gladwell's assertion holds true, it could prompt a reevaluation of scouting methods, training programs, and team composition strategies. Conversely, if our analysis reveals otherwise, it could challenge long-held beliefs and pave the way for more nuanced discussions about factors contributing to athletic achievement.

2. **Approach:**

In our approach, we intend to utilize two distinct types of plots to address our research question effectively. Firstly, we will employ a line chart to illustrate the overall trend of birthdate distributions among NHL players over the years. This type of plot is ideal for visualizing temporal trends as it allows us to observe how birthdate patterns have evolved annually. By plotting the percentage of first-quarter-born players we can identify any consistent patterns or deviations over time, providing valuable insights into the dynamics of player recruitment and selection.

Secondly, to assess the validity of Malcolm Gladwell's assertion and its impact on individual NHL teams, we will utilize a bar chart. This chart will display the proportion of how many seasons did they have dominant number of first-quarter-born players. By aggregating the data at the team level and visually comparing the proportions across teams, we can discern any disparities or trends that may indicate the influence of Gladwell's assertion. The bar chart's clear and concise presentation of categorical data makes it an excellent choice for highlighting variations among different groups, allowing us to evaluate the assertion's impact on a team-by-team basis. Additionally, color mapping or face can be incorporated into this chart to further enhance its visual clarity and interpretability.

3. Analysis:

```
## Merging data so it will include only useful columns.
```{r}
  nhl_rosters <- read.csv("nhl_rosters.csv")

  unique_team_codes <- unique(nhl_rosters$team_code)
  nhl_rosters <- read.csv("nhl_rosters.csv")
  nhl_data <- read.csv("nhl_player_births.csv")

  # Merge the datasets on player_id
  merged_data <- merge(nhl_rosters, nhl_data, by = "player_id")

  # Select and rename columns from the merged dataset
  middle_data <- subset(merged_data, select = c("player_id", "team_code", "season", "birth_month"))
  colnames(middle_data) <- c("player_id", "team_code", "season", "birth_month")

  # Print the resulting middle data
  print(middle_data)

```
```

Merging data and removing distracting columns

```
```{r}
library(dplyr)
result <- middle_data %>%
  group_by(season) %>%
  summarise(total_players = n(),  # Count total players in the season
            first_3_months = sum(birth_month %in% 1:3)) %>%  # Count players born in the first 3 months
  mutate(percent_first_3_months = (first_3_months / total_players) * 100)  # Calculate percentage

# Print the result
print(result)

```
```

New data frame calculating percentage of first-quarter-born players.

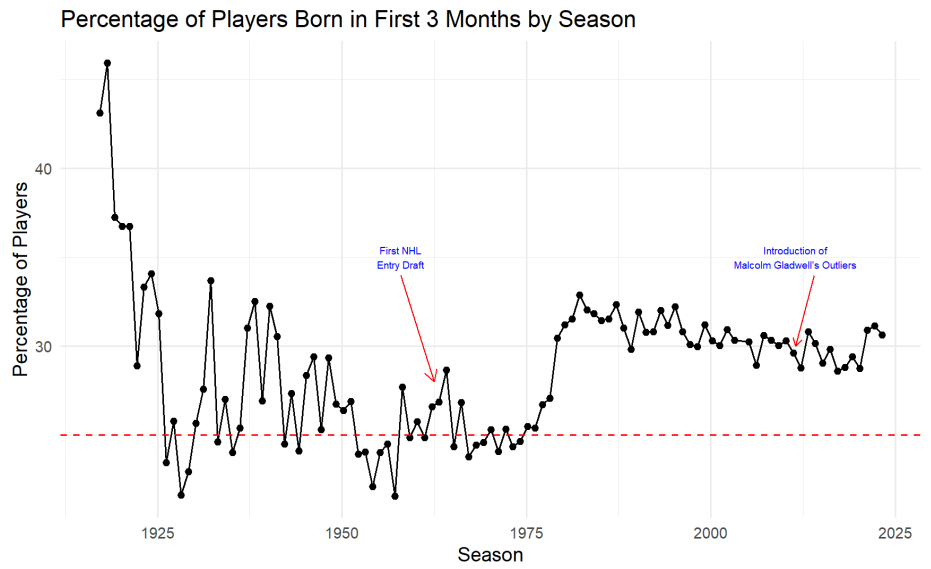
To begin our analysis, we will first merge the two data files using a unique identifier such as the player_ID. This will allow us to combine relevant information from both datasets into a single, comprehensive dataset. Additionally, we will remove any extraneous columns that are not pertinent to our analysis, streamlining the dataset for easier handling and interpretation. Once we have merged and cleaned the dataset, we will create a new dataframe that calculates the percentage of players born in the first quarter of the year (January to March) compared to the total number of players. Now, the data is ready to be visualized:

```r
## How dominant the first-quarter-born through the history?
```{r}
# Load the required library
library(ggplot2)


# Plot the line chart
ggplot(result, aes(x = season / 10000, y = percent_first_3_months)) +
  geom_line() +
  geom_point() +
  labs(title = "Percentage of Players Born in First 3 Months by Season",
       x = "Season",
       y = "Percentage of Players") +
  geom_hline(yintercept = 25, color = "red", linetype = "dashed") +
  annotate("text", x = 2011.5, y = 35, label = "Introduction of\nMalcolm Gladwell's Outliers", color = "blue", size = 2) +
  geom_segment(aes(x = 2014, y = 34, xend = 2011.5, yend = 30), color = "red", arrow = arrow(length = unit(0.07, "inches")), size = 0.3) +
  annotate("text", x = 1958, y = 35, label = "First NHL\nEntry Draft", color = "blue", size = 2) +
  geom_segment(aes(x = 1958, y = 34, xend = 1962.5, yend = 28), color = "red", arrow = arrow(length = unit(0.1, "inches")), size = 0.3) +
  theme_minimal()

```
```

Code block for the line chart



Plot showing the trend of first-quarter-born player

```r
```{r}
library(dplyr)
library(ggplot2)

library(ggplot2)

# Count the number of seasons for each team
team_season_counts <- result %>%
  group_by(team_code) %>%
  summarize(total_seasons = n_distinct(season))

# Count the number of seasons where first_3_months_players > 25 for each team
team_exceed_25_counts <- result %>%
  filter(percent_first_3_months > 25) %>%
  group_by(team_code) %>%
  summarize(exceed_25_seasons = n_distinct(season))

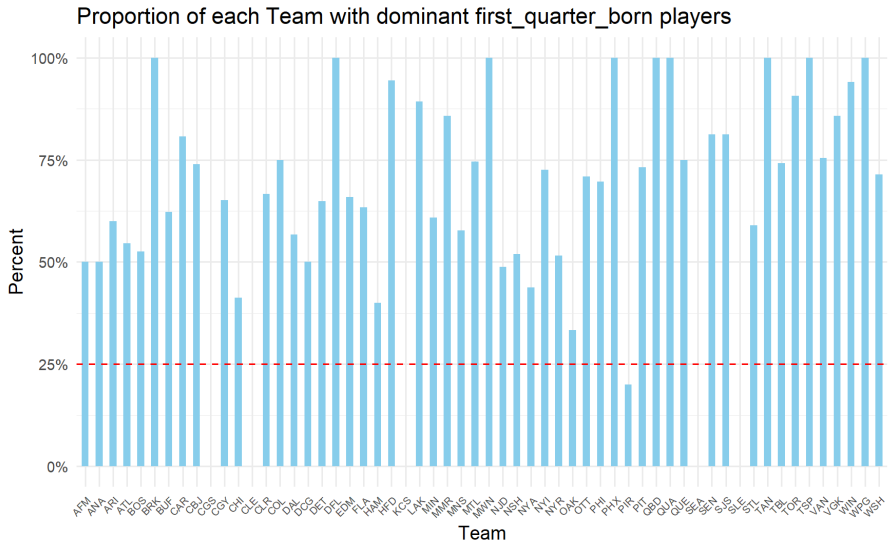
# Merge the two datasets
team_counts <- merge(team_season_counts, team_exceed_25_counts, by = "team_code", all.x = TRUE)

# Fill NA values with 0 (for teams where no seasons have first_3_months_players > 25)
team_counts[is.na(team_counts)] <- 0

# Print the result
print(team_counts)


```
```

New data counting the number of seasons the first-quarter-born numbers dominant



Resulting chart

4. Discussion:

In the first graph, we address the primary concern of whether Malcolm's assertion holds true throughout history. We observe a consistent pattern where the number of players born in the first quarter of the year consistently comprises more than a quarter of the team. However, after analyzing the graph and doing some research, we figure out the reason why the graph comes from being very fluctuating to stabilizing at around 30% by the end, which is attributed to the NHL Draft event. The establishment of the NHL Draft in 1964 marked a significant turning point in player recruitment practices, bringing stability to team compositions. Prior to this event, the absence of a formal draft process led to erratic player acquisition methods, resulting in pronounced fluctuations in the proportion of first-quarter-born players. This observation underscores the profound impact of the NHL Draft on shaping the demographics of NHL teams over time.

In the second graph, we aim to investigate whether Malcolm's assertion holds true uniformly across all teams or if it is applicable to only a subset of teams. Upon examination, we find that most teams exhibit a prevalence of first-quarter-born players, with more than a quarter of their roster falling into this category. Additionally, we observe intriguing outliers where certain teams boast a 100% representation of first-quarter-born players, underscoring the dominance of individuals born within the first three months of the year. This observation suggests a widespread trend across teams, further

supporting the validity of Malcolm's assertion regarding the significance of birth month in player recruitment and team composition.