Fair Play: Examining Pay Equity and Rise of International MLB Players

Analyzing Trends in Player Performance and Compensation

Will Crouch

2026-01-08

1. Introduction

Over the past several decades, Major League Baseball (MLB) has witnessed a dramatic increase in the number of international players. As these athletes contribute significantly to the league's success, questions arise about whether they are compensated equitably compared to their domestic counterparts.

This project explores the intersection of international representation and pay equity in MLB from 1985 to 2016. Using the `Lahman` database and data from Baseball Reference, we can analyze trends in player demographics, salary distributions, and performance metrics. In addition, there will be exploration of the full database, with regards to overall performance metrics and award history to compare the proportion of top international-born players.

This project investigates whether international players are paid fairly relative to their contributions, whether country of origin correlates with top annual performers, and how historical and economic factors shape these outcomes over time.

From a personal perspective, I have been surrounded by baseball from an early age, whether it be a player or fan. My first baseball game was at the old Busch Stadium in St. Louis, in which the St. Louis Cardinals played the Atlanta Braves. I was 2 years old and my dad stuffed my ears with cotton to handle the crowd noise. From then, I was hooked. Always mimicking popular players' batting stances, keeping journals of baseball statistics, and learning how to score baseball games properly. Once I was able to play in our public leagues, I strived to be the best second baseman and relief pitcher on my team - I'd like to believe I managed this goal, but perhaps a picture might show it best:

Growing up in Georgia, I've always been a fan of the Braves, witnessing countless records and Hall of Fame players during the peak of the late 90s/early 2000's Braves dynasty. My dad, brother, and I would often attend multiple games a year, mostly at the old Turner Field in Atlanta.

Playing baseball as a kid was central to my identity and I've remained a baseball fan since, always catching games at PNC Park when the Braves visit (and usually win against) the Pirates.

The findings aim to provide insights into the systemic disparities that may exist in player compensation and highlight opportunities for policy recommendations that foster fairness for international-born players. By examining trends in pay and performance across countries of origin, this analysis contributes to a broader understanding of equity in professional sports and its implications for future policy.

2. Outline of Project

Below is the outline of the project, along with relevant research questions that I hope to answer.

Introduction

- **Focus**: Investigate trends in MLB player demographics, performance, and pay equity.
- **Objectives**:
  - Investigate the rise of international players in MLB.
  - Compare pay equity between international and domestic players.
  - Explore the impact of international status on performance metrics.

Methods

- **Dataset Loading and Cleaning**:
  - Loaded `Lahman` database, merging WAR data from Baseball Reference.
  - Corrected country names and consolidated them into parent nations.
  - Trimmed down dataset to remove entries with missing values that would be necessary for my analyses

Analysis 1: Demographics and Debut Trends

- **Research Questions**:

  - How has the proportion of international MLB players changed over time?
  - Do international players debut at younger ages than domestic players?
  - What do their performance metrics look like compared to domestic players?

Analysis 2: Pay Equity

- **Research Questions**:
  - Is there a significant difference in salary based on WAA for international vs. domestic players?
  - Have these differences changed over time?

Conclusion

- **Summary**:
  - Highlight findings on demographics, performance metrics, and pay equity.
  - Discuss implications for MLB policies.
- **Call to Action**: Advocate for equitable pay structures and highlight the importance of further research into these trends.

3. Lahman Database and Baseball Reference (BR)

This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2023. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875.

This database was created by Sean Lahman, who pioneered the effort to make baseball statistics freely available to the general public. What started as a one man effort in 1994 has grown tremendously, and now a team of researchers have collected their efforts to make this the largest and most accurate source for baseball statistics available anywhere. A preview of the `People` dataset from `Lahman` is provided as an example. Other datasets from `Lahman` that will be used include `Teams`, `AwardsPlayers`, and `Salaries`.

```
# loading People dataset as an object for ease of manipulation
players_lahman_only <- People

# preview of the People dataset
head(players_lahman_only)
```

Lahman

```
##    playerID birthYear birthMonth birthDay  birthCity birthCountry birthState
## 1 aardsda01      1981         12       27     Denver          USA         CO
## 2 aaronha01      1934          2        5     Mobile          USA         AL
## 3 aaronto01      1939          8        5     Mobile          USA         AL
## 4  aasedo01      1954          9        8     Orange          USA         CA
## 5  abadan01      1972          8       25 Palm Beach          USA         FL
## 6  abadfe01      1985         12       17  La Romana         D.R.  La Romana
##   deathYear deathMonth deathDay deathCountry deathState deathCity nameFirst
```

```
## 1        NA         NA       NA         <NA>       <NA>      <NA>     David
## 2      2021          1       22          USA         GA   Atlanta      Hank
## 3      1984          8       16          USA         GA   Atlanta    Tommie
## 4        NA         NA       NA         <NA>       <NA>      <NA>       Don
## 5        NA         NA       NA         <NA>       <NA>      <NA>      Andy
## 6        NA         NA       NA         <NA>       <NA>      <NA>  Fernando
##   nameLast        nameGiven weight height bats throws      debut   bbrefID
## 1  Aardsma      David Allan    215     75    R      R 2004-04-06 aardsda01
## 2    Aaron      Henry Louis    180     72    R      R 1954-04-13 aaronha01
## 3    Aaron       Tommie Lee    190     75    R      R 1962-04-10 aaronto01
## 4     Aase   Donald William    190     75    R      R 1977-07-26  aasedo01
## 5     Abad    Fausto Andres    184     73    L      L 2001-09-10  abadan01
## 6     Abad Fernando Antonio    235     74    L      L 2010-07-28  abadfe01
##    finalGame  retroID  deathDate  birthDate
## 1 2015-08-23 aardd001       <NA> 1981-12-27
## 2 1976-10-03 aaroh101 2021-01-22 1934-02-05
## 3 1971-09-26 aarot101 1984-08-16 1939-08-05
## 4 1990-10-03 aased001       <NA> 1954-09-08
## 5 2006-04-13 abada001       <NA> 1972-08-25
## 6 2021-10-01 abadf001       <NA> 1985-12-17
```

Seems we have data that is either incomplete or inaccurate, so we must clean the `birthCountry` data to reflect the proper designations. For the purposes of consistency, territories and constituent countries are merged into their "parent" country, while errors in spelling and/or acronyms are fixed for ease of reading.

There is an argument that cultural identity of many Latin American and Caribbean territories/constituent countries does not match that of the parent nation. However, by combining these values, we are able to look at countries of origin more broadly and accounting for players whose smaller sample size could otherwise suffer from low power. Nevertheless, it is important to be mindful of cultural identity when discussing individual players and the importance of one's heritage at the MLB level, for which there are many events that highlight these differences.

Special Case: Because American Samoa, Puerto Rico, U.S. Virgin Islands, and Guam are territories of the United States, they could be excluded from the international-born player analyses, as they would be merged into the USA category.

I want to do justice to prolific players who come from Puerto Rico and not dilute their contributions under the USA umbrella. Puerto Rico yields a strongly represented proportion of MLB players; therefore, in order to increase power while maintaining cultural sensitivity. Players from American Samoa, Guam and the US Virgin Islands are a much smaller proportion, constituting only 17 players (1 from American Samoa, 14 from V.I. and 2 from Guam), but will still fall under this case for consistency.

```
players_lahman_only <- players_lahman_only %>%
  mutate(birthCountry = case_when(
    birthCountry == "D.R." ~ "Dominican Republic",
    birthCountry =="Ukriane" ~ "Ukraine",
    birthCountry == "CAN" ~ "Canada",
    birthCountry == "P.R." ~ "Puerto Rico",
    birthCountry== "Viet Nam" ~ "Vietnam",
    birthCountry == "Bohemia" ~ "Czech Republic",
    birthCountry == "England" ~ "United Kingdom",
    birthCountry =="Curacao" ~ "Netherlands",
    birthCountry == "Aruba" ~ "Netherlands",
    birthCountry == "Hong Kong" ~ "China",
    birthCountry == "Taiwan" ~ "China",
```

```
    birthCountry == "V.I." ~ "U.S. Virgin Islands",
    TRUE ~ birthCountry))

# determine if there are missing values for birth country
colSums(is.na(players_lahman_only))
```

```
##     playerID    birthYear   birthMonth     birthDay    birthCity birthCountry
##            0           94          261          404          141           48
##   birthState    deathYear   deathMonth     deathDay deathCountry   deathState
##          404        10799        10799        10799        10805        10844
##    deathCity    nameFirst     nameLast    nameGiven       weight       height
##        10806           31            0            0          843          762
##         bats       throws        debut      bbrefID    finalGame      retroID
##         1215         1010          274           56         1671           58
##    deathDate    birthDate
##        10800          404
```

Because there is missing data in the `birthCountry` column, it complicates the analysis somewhat. I would have two options in this case:

1) Individually fix each data point: By referencing other sources, such as Baseball Reference, I could manually fix each data point; however, with 55 players missing data about their birth country, this would take a long time and runs the risk of data entry error.

2) Exclude the data points with missing data: By removing these values, it provides an easy method to move forward; however, when removing data, it is important to determine how many data points would be lost **and** if the removal compromises the analyses. If we make the assumption that all NA values in `birthCountry` are, in fact, international, we can then determine how much data we would be losing.

```
# create list of players with missing birthCountry data
missing_country <- players_lahman_only %>%
  filter(is.na(birthCountry))

# determine if there are duplicate entries in this list
missing_players <- unique(missing_country$playerID)

# determine if there are playerID's that actually have correct values in other years, but had an NA due
missing_players_dupes <- players_lahman_only %>%
  filter(playerID %in% missing_players)
```

Given the missing data only constitutes 55 unique players, assuming that all NA values **could** be international, I felt that a potential loss of 0.262% was reasonably within bounds to exclude from the full dataset.

Baseball Reference Additionally, the `Lahman` database includes an ID number to be linked to publicly available sabermetrics data (conducted by the Society for American Baseball Research, but popularized by the movie "Moneyball") on the popular baseball database, Baseball Reference (BR). Sabermetrics look beyond classic statistics, to determine player efficiency and performance, using composite values from complex (yet published) equation modeling. As such, I have pulled the Wins Above Average values for all players in the `Lahman` database and merged the corresponding data in my analyses.

Wins Above Average, according to BR, is "a statistical measure that defines a player's worth in terms of his contribution as compared to the average major league player. WAA is strongly correlated to team performance, that is the sum of WAA by all of a team's players will almost always represent its final record."

While `Lahman` is native to R, BR's database is available as .csv files, which will be merged with `Lahman` to conduct my analyses, re: player performance and relative compenstation. This column will be listed as `WAA` in the merged dataset.

```
# import batting data from BR
batting_br <- read.csv("batting.csv")

# previews of the batting data from BR
head(batting_br)
```

```
##     name_common age mlb_ID player_ID year_ID team_ID stint_ID lg_ID PA  G  Inn
## 1 David Aardsma  22 430911 aardsda01    2004     SFG        1    NL  0 11 10.7
## 2 David Aardsma  24 430911 aardsda01    2006     CHC        1    NL  3 43 53.0
## 3 David Aardsma  25 430911 aardsda01    2007     CHW        1    AL  0  2 32.3
## 4 David Aardsma  26 430911 aardsda01    2008     BOS        1    AL  1  5 48.7
## 5 David Aardsma  27 430911 aardsda01    2009     SEA        1    AL  0  3 71.3
## 6 David Aardsma  28 430911 aardsda01    2010     SEA        1    AL  0  4 49.7
##   runs_bat runs_br runs_dp runs_field runs_infield runs_outfield runs_catcher
## 1     0.00       0       0          0         0.00          0.00         0.00
## 2    -0.90       0       0          0         0.00          0.00         0.00
## 3     0.00       0       0          0         0.00          0.00         0.00
## 4    -0.29       0       0          0         0.00          0.00         0.00
## 5     0.00       0       0          0         0.00          0.00         0.00
## 6     0.00       0       0          0         0.00          0.00         0.00
##   runs_good_plays runs_defense runs_position runs_position_p runs_replacement
## 1            0.00            0          0.00            0.00               0
## 2            0.00            0          0.01            0.46               0
## 3            0.00            0          0.00            0.00               0
## 4            0.00            0          0.00            0.14               0
## 5            0.00            0          0.00            0.00               0
## 6            0.00            0          0.00            0.00               0
##   runs_above_rep runs_above_avg runs_above_avg_off runs_above_avg_def   WAA
## 1            0.0            0.0               0.0                  0  0.00
## 2           -0.4           -0.4              -0.4                  0 -0.04
## 3            0.0            0.0               0.0                  0  0.00
## 4           -0.2           -0.2              -0.2                  0 -0.02
## 5            0.0            0.0               0.0                  0  0.00
## 6            0.0            0.0               0.0                  0  0.00
##   WAA_off WAA_def   WAR WAR_def WAR_off WAR_rep  salary pitcher teamRpG  oppRpG
## 1    0.00   -0.01  0.00   -0.01    0.00    0.00  300000       Y 4.67092 4.67092
## 2   -0.04   -0.01 -0.04   -0.01   -0.04    0.00    NULL       Y 4.85675 4.86675
## 3    0.00    0.00  0.00    0.00    0.00    0.00  387500       Y 4.85895 4.85895
## 4   -0.02    0.00 -0.02    0.00   -0.02    0.00  403250       Y 4.67400 4.70400
## 5    0.00    0.00  0.00    0.00    0.00    0.00  419000       Y 4.79788 4.79788
## 6    0.00    0.00  0.00    0.00    0.00    0.00 2750000       Y 4.44684 4.44684
##   oppRpPA_rep oppRpG_rep pyth_exponent pyth_exponent_rep waa_win_perc
## 1     0.08651    4.67092         1.890             1.890       0.5000
## 2     0.09085    4.86457         1.912             1.913       0.4990
## 3     0.08422    4.85895         1.912             1.912       0.5000
## 4     0.08092    4.69650         1.893             1.894       0.4970
## 5     0.08302    4.79788         1.905             1.905       0.5000
## 6     0.07567    4.44684         1.864             1.864       0.5000
##   waa_win_perc_off waa_win_perc_def waa_win_perc_rep        OPS_plus TOB_lg
## 1           0.5000           0.5000           0.5000            NULL  0.000
```

```
## 2           0.4990           0.5000           0.4998 -100.0000000000  0.694
## 3           0.5000           0.5000           0.5000             NULL  0.000
## 4           0.4970           0.5000           0.4992 -100.0000000000  0.345
## 5           0.5000           0.5000           0.5000             NULL  0.000
## 6           0.5000           0.5000           0.5000             NULL  0.000
##   TB_lg
## 1 0.000
## 2 0.896
## 3 0.000
## 4 0.434
## 5 0.000
## 6 0.000
```

Because the playerIDs on BR differ from those of `Lahman`, we have to link the BR data with a corresponding value. Fortunately, `Lahman` includes a reference ID in the `People` dataset, allowing for easy merging.

Special Case: However, with the introduction of the Designated Hitter rule to the American League in 1973, pitchers in the American League do not have Batting WAA beyond 1973; whereas, pitchers batted in the National League until 2020. Therefore, we will only analyze offensive players because there would be separate values for pitchers, while batters would only have one (barring position players who came in to pitch during a blowout and/or a lack of available pitchers due to injury).

You might ask why the values cannot be combined to capture overall contributions to the team; however, because pitchers are notoriously weak batters and batters are notoriously poor pitchers, we do not want an extremely negative value on either end to skew a composite value to reduce their overall contributions (in their main role) to the team.

```
# renaming ID column in BR to match the merge value in Lahman
names(batting_br)[names(batting_br) == "player_ID"] <- "bbrefID"
names(batting_br)[names(batting_br) == "WAA"] <- "WAA_batting"


# merging datasets together with respect to batting
final_df <- left_join(players_lahman_only, batting_br, by = "bbrefID", "year_id") %>%
  filter(pitcher != "Y")
```

```
## Warning in left_join(players_lahman_only, batting_br, by = "bbrefID", "year_id"): Detected an unexpe
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 57634 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

4. Analysis 1: Demographics and Trends in Performance of International Players

In this section, we will explore demographics of the MLB, along with trends related to debuts in the MLB and performance metrics. We will also compare these data with domestic players to determine diferences (if any) between the two groups.

```
# filtering for list of international-born players
intl_players <- final_df %>%
```

```
  filter(birthCountry != "USA") %>%
  select(playerID, nameFirst, nameLast, birthCountry)

summary(unique(intl_players$playerID))
```

Demographics

```
##    Length     Class      Mode
##      1560 character character
```

After filtering the number of international players from the larger `People` dataset, I determined that there are 2857 unique players listed in `Lahman` who meet criteria for "internationally-born" MLB players, keeping in line with the Special Case, as mentioned above.

```
# list of countries of origin
countries_of_origin <- intl_players %>%
  distinct(playerID, birthCountry) %>%
  group_by(playerID)

# count the number of countries represented
dist_countries <- table(countries_of_origin$birthCountry)
dist_countries <- as.data.frame(dist_countries)
names(dist_countries) <- c("Country", "Count")
unique(dist_countries$Country)
```

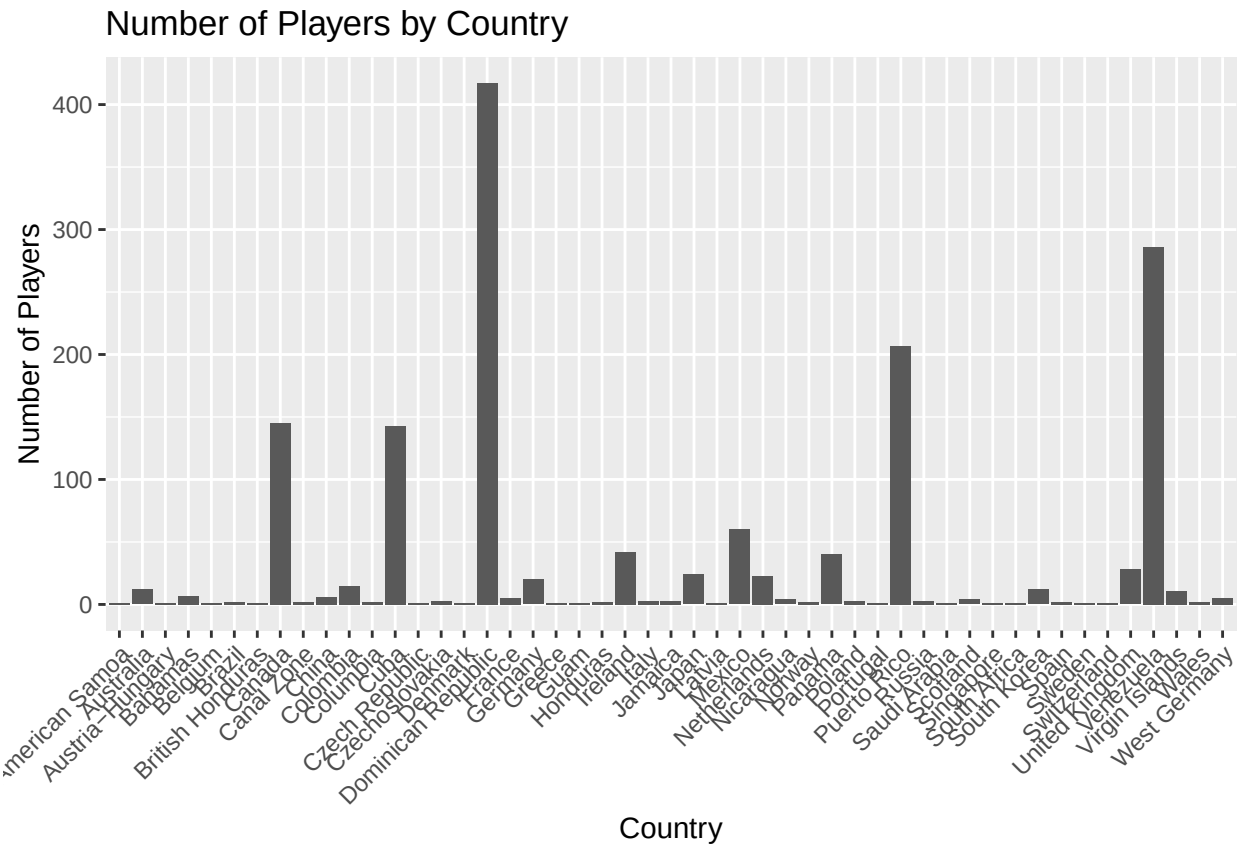
```
##  [1] American Samoa     Australia          Austria-Hungary    Bahamas
##  [5] Belgium            Brazil             British Honduras   Canada
##  [9] Canal Zone         China              Colombia           Columbia
## [13] Cuba               Czech Republic     Czechoslovakia     Denmark
## [17] Dominican Republic France             Germany            Greece
## [21] Guam               Honduras           Ireland            Italy
## [25] Jamaica            Japan              Latvia             Mexico
## [29] Netherlands        Nicaragua          Norway             Panama
## [33] Poland             Portugal           Puerto Rico        Russia
## [37] Saudi Arabia       Scotland           Singapore          South Africa
## [41] South Korea        Spain              Sweden             Switzerland
## [45] United Kingdom     Venezuela          Virgin Islands     Wales
## [49] West Germany
## 49 Levels: American Samoa Australia Austria-Hungary Bahamas Belgium ... West Germany
```

There are 53 "countries" represented by international-born players. Including the USA for our other analyses later on, there are 54 total countries represented in the `Lahman` database.

Strangely enough, there is an entry in `Country` as "Atlantic Ocean". Initially, I assumed this was a typo, in lieu of an island in the Atlantic Ocean. However, upon searching the playerID (porraed01), it was determined that Ed Porray was, in fact, born at sea.

```
# plot the results of the count
ggplot(dist_countries,
       aes(x = Country,
           y = Count)) +
  geom_bar(stat = "identity") +
  labs(
```

```
    title = "Number of Players by Country",
    x = "Country",
    y = "Number of Players") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Unfortunately, this is not a very useful graph, largely because there is such a large range in values on the y-axis. We can instead display this information as a table for easier viewing.

```
dist_countries <- dist_countries %>%
  arrange(desc(Count))

kable(dist_countries,
        caption = "Distribution of Countries Represented by MLB Players")
```

Table 1: Distribution of Countries Represented by MLB Players

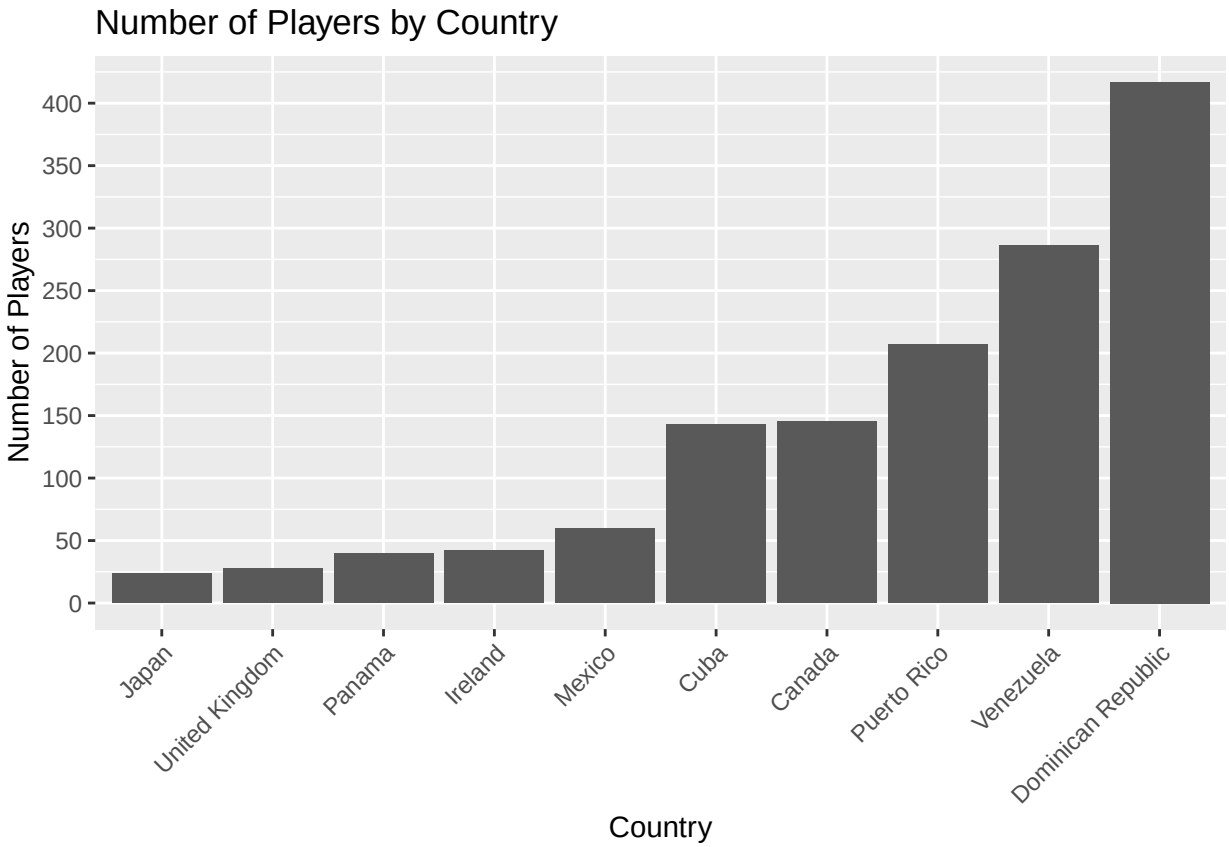
Country	Count
Dominican Republic	417
Venezuela	286
Puerto Rico	207
Canada	145
Cuba	143
Mexico	60
Ireland	42
Panama	40

Country	Count
United Kingdom	28
Japan	24
Netherlands	23
Germany	20
Colombia	15
Australia	12
South Korea	12
Virgin Islands	11
Bahamas	7
China	6
France	5
West Germany	5
Nicaragua	4
Scotland	4
Czechoslovakia	3
Italy	3
Jamaica	3
Poland	3
Russia	3
Brazil	2
Canal Zone	2
Columbia	2
Honduras	2
Norway	2
Spain	2
Wales	2
American Samoa	1
Austria-Hungary	1
Belgium	1
British Honduras	1
Czech Republic	1
Denmark	1
Greece	1
Guam	1
Latvia	1
Portugal	1
Saudi Arabia	1
Singapore	1
South Africa	1
Sweden	1
Switzerland	1

If we only focus on the top 10 results, we get a plot with better resolution, as seen below.

```
dist_countries_top10 <- dist_countries %>%
  arrange(desc(Count)) %>%
  slice_head(n = 10)

ggplot(dist_countries_top10,
       aes(x = reorder(Country, Count),
           y = Count)) +
  geom_bar(stat = "identity") +
```

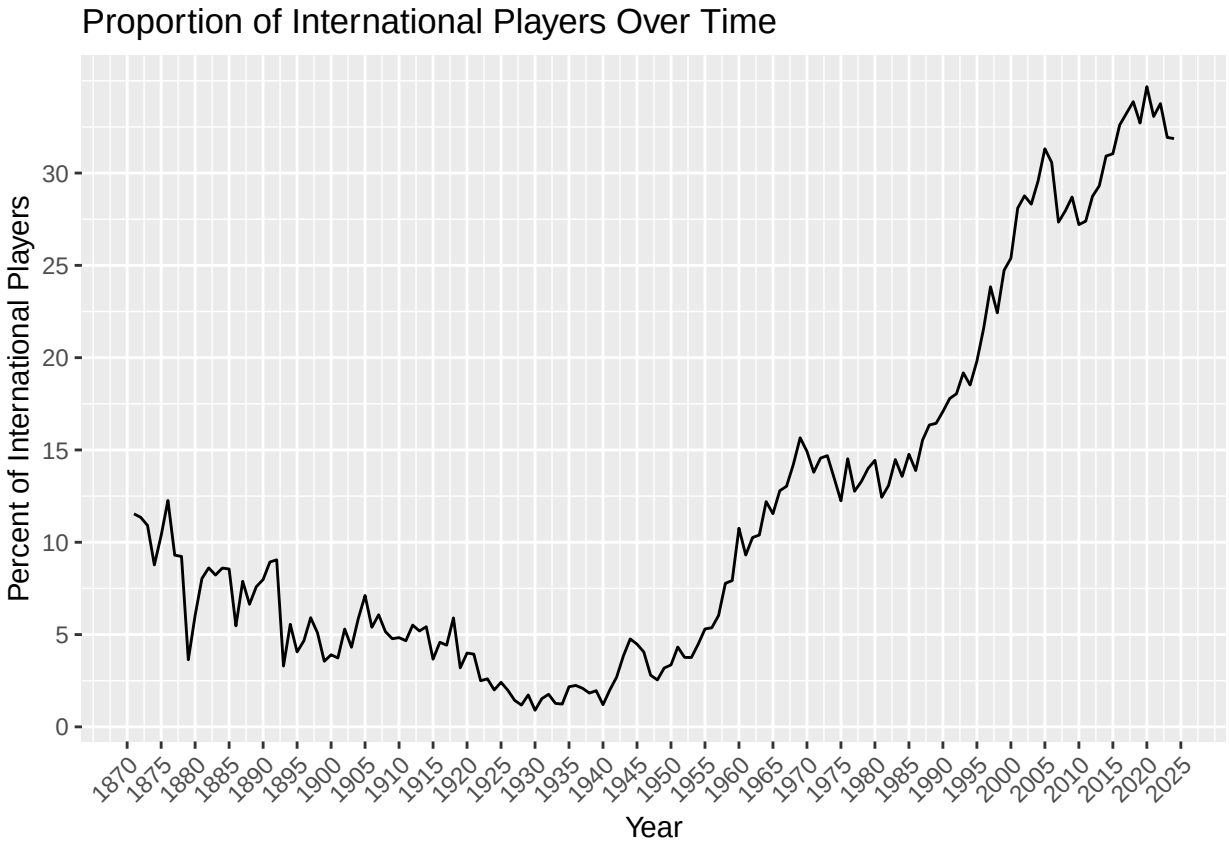
```
  labs(
    title = "Number of Players by Country",
    x = "Country",
    y = "Number of Players") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(
    breaks = seq(0, max(dist_countries_top10$Count), by = 50))
```



We can also view the proportion of international players in the MLB through a historical lens. Many factors have contributed to this growth, such as:

- Expansion of scouting efforts to international markets, particularly in Latin America in the 1980's, with the most notable being the Dominican Republic's baseball academies.
- Globalization of baseball to countries around the world. As baseball becomes more popular, careers in the MLB are more accessible. Additionally, baseball careers represent a path to economic prosperity, particularly from low-income countries such as Venezuela, Cuba, and the Dominican Republic.
- Changes in MLB international policy, such that the international draft age floor was lowered to 16 and expansion of talent recruitment from baseball leagues in overseas markets, such as those in Japan. These changes increased in the mid-1980's and led to a large increase in international players throughout the 90's and 00's.
- Finally, US domestic policies have strong influence on access for internationally-born players. Immigration policies, particularly targeted towards Latin American and Carribean nations, led to a decrease in numbers in the late 50's/60's. As diplomatic relations with countries flucuate, such as with Cuba and Venezuela in the latter half of the 20th century, we also see more barriers to recruitment and participation in American sports.

```r
remove_na_year_birth <- final_df[final_df$birthCountry != "USA" & !is.na(final_df$birthCountry) & !is.na

intl_players_per_year <- aggregate(x = remove_na_year_birth$playerID,
                                   by = list(Year = remove_na_year_birth$year_ID),
                                   FUN = length)
colnames(intl_players_per_year)[2] <- "Count"

total_players_per_year <- aggregate(x = final_df$playerID,
                                    by = list(Year = final_df$year_ID),
                                    FUN = length)
colnames(total_players_per_year)[2] <- "Count"

proportion_intl_per_year <- intl_players_per_year %>%
  left_join(total_players_per_year, by = "Year") %>%
  mutate(proportion_intl = (intl_players_per_year$Count / total_players_per_year$Count)*100) %>%
  select(Year, proportion_intl)

colnames(proportion_intl_per_year)[2] <- "Percent"

ggplot(proportion_intl_per_year, aes(x = Year, y = Percent)) +
  geom_line() +
  labs(title = "Proportion of International Players Over Time",
       x = "Year",
       y = "Percent of International Players") +
  scale_x_continuous(breaks = seq(1870, max(proportion_intl_per_year$Year)+2, by = 5)) +
  scale_y_continuous(breaks = seq(0, max(proportion_intl_per_year$Percent), by = 5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Debut Ages One of the contributing factors for international participation was the reduction in age restrictions for the international draft. By lowering the age to 16, younger players are able to be recruited earlier and often debut earlier than their domestic counterparts.

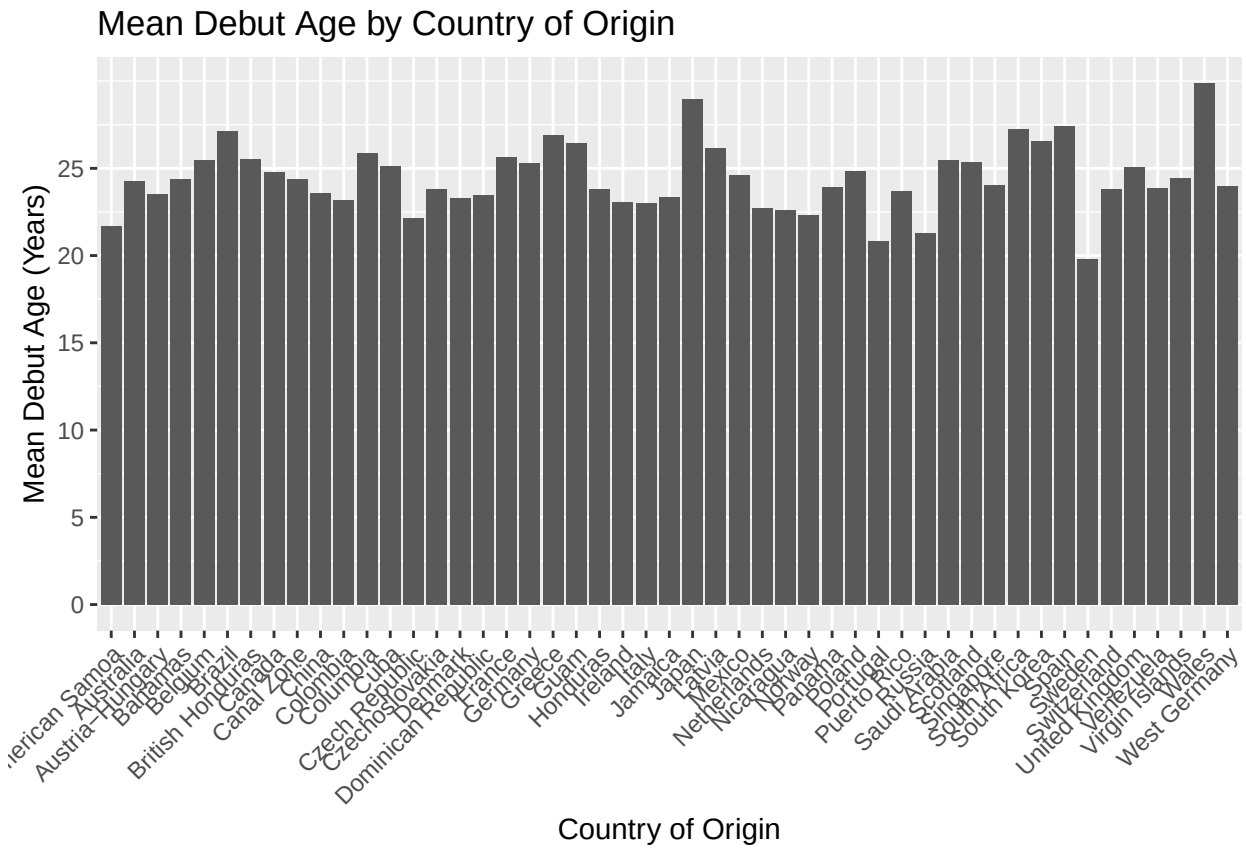
On the domestic side, with the improvement in development teams in college baseball, younger players who are scouted during high school are opting to spend 3-4 years in college to tap into otherwise-inaccessible development for those immediately declaring for the MLB draft. By delaying their recruitment, domestic players are able to develop their skills, increase their market value and draft stock, and gain educational funding.

Unfortunately, for international players, these opportunities are limited and are not common pathways. Therefore, market expansion, international recruitment, and the international draft age have allowed for earlier access to MLB resources and subsequently earlier debut dates.

However, how much do these timelines actually differ beyond anecdotal evidence by MLB pundits?

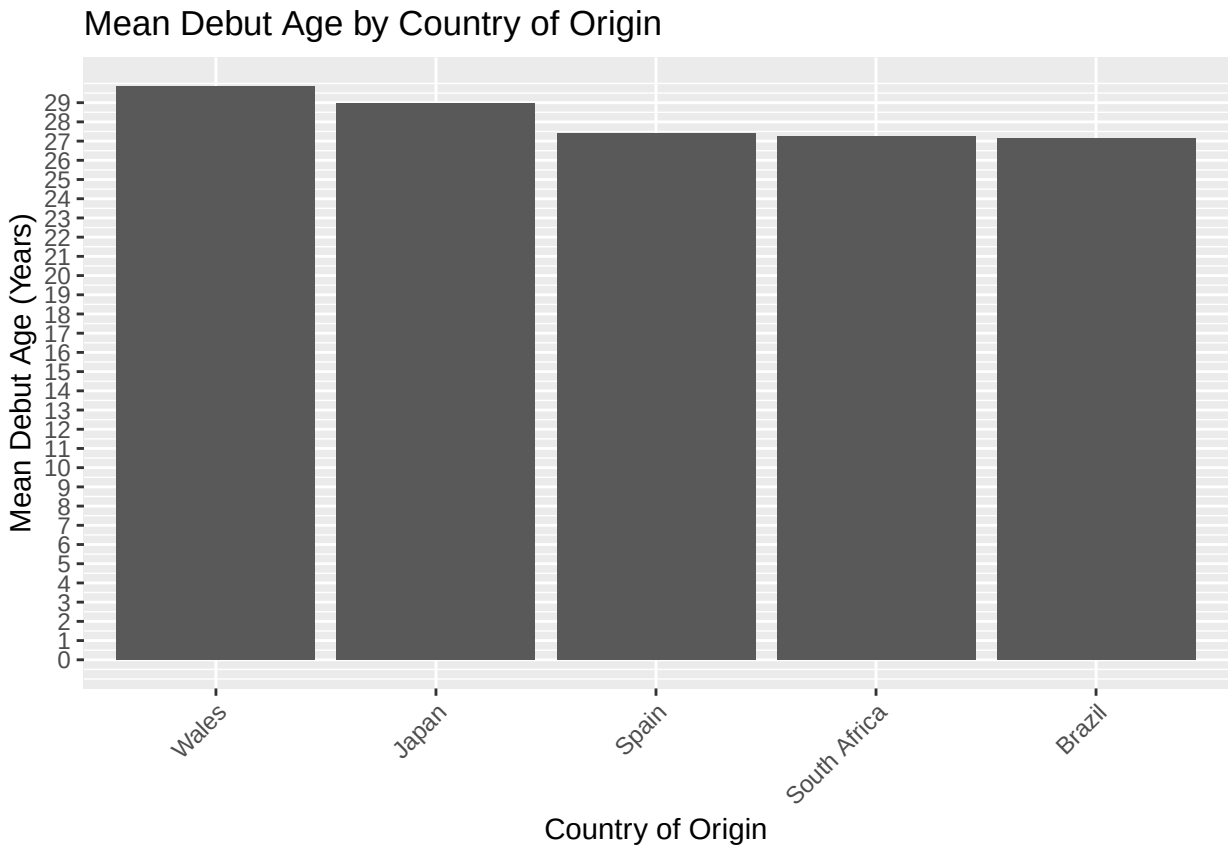
```
# pulling debut data and player birth date
intl_players_debut <- final_df %>%
  filter(birthCountry != "USA") %>%
  select(playerID, birthCountry, birthDate, debut)

# converting string data into date format and calculate days into years
debut_age_summary <- intl_players_debut %>%
  mutate(debut = as.Date(debut, format = "%Y-%m-%d"),
         birthDate = as.Date(birthDate, format = "%Y-%m-%d")) %>%
  filter(!is.na(birthCountry), !is.na(birthDate), !is.na(debut)) %>%
  mutate(debut_age = as.numeric(difftime(debut, birthDate, units = "days")) / 365.25)
```

```r
# assign debut age to each unique player ID
debut_age_summary <- aggregate(x = debut_age_summary$debut_age,
                               by = list(playerID = debut_age_summary$playerID),
                               FUN = mean)
# rename default "x" to debutAge
colnames(debut_age_summary)[2] <- "debutAge"

# combine birthCountry data
debut_age_summary <- left_join(intl_players_debut, debut_age_summary, by = "playerID")

# select distinct player ID's with age at debut
debut_age_summary_comb <- debut_age_summary %>%
  select(playerID, birthCountry, debutAge) %>%
  distinct() %>%
  group_by(birthCountry)

# create mean age of debut by country
mean_debut_age_by_country <- aggregate(
  debutAge ~ birthCountry,
  data = debut_age_summary_comb,
  FUN = mean,
  na.rm = TRUE
)

# rename default "data" into avg_debutAge
colnames(mean_debut_age_by_country)[2] <- "avg_debutAge"

# plotting mean debut age by country of origin
ggplot(mean_debut_age_by_country, aes(x = birthCountry, y = avg_debutAge)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Mean Debut Age by Country of Origin",
    x = "Country of Origin",
    y = "Mean Debut Age (Years)"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(breaks = seq(0, max(mean_debut_age_by_country$avg_debutAge), by = 5))
```



Again, this plot isn't very useful because it does not illustrate the granularity of the differences between each country. This data could be better displayed as a subset of data and table once again. Additionally, displaying the number of players per country could help contextualize the plot.

```
# select top 5 countries by average age of debut
top5_mean_debut_age_by_country <- mean_debut_age_by_country %>%
  arrange(desc(avg_debutAge)) %>%
  slice_head(n = 5)

# plot top 5 countries by average age of debut
ggplot(top5_mean_debut_age_by_country, aes(x = reorder(birthCountry, -avg_debutAge), y = avg_debutAge))
  geom_bar(stat = "identity") +
  labs(
    title = "Mean Debut Age by Country of Origin",
    x = "Country of Origin",
    y = "Mean Debut Age (Years)"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(breaks = seq(0, max(mean_debut_age_by_country$avg_debutAge), by = 1))
```




```
# sorting from greatest age at debut to lowest
mean_debut_age_by_country <- mean_debut_age_by_country %>%
  arrange(desc(avg_debutAge))

# rename columns for better clarity
colnames(mean_debut_age_by_country)[2] <- "avg_debutAge"
colnames(mean_debut_age_by_country)[1] <- "Country"

# Average of all international-born players
round(mean(mean_debut_age_by_country$avg_debutAge),2)
```

```
## [1] 24.45
```

```
# table representation
kable(mean_debut_age_by_country,
        caption = "Average age of international MLB players at debut")
```

Table 2: Average age of international MLB players at debut

Country	avg_debutAge
Wales	29.86311
Japan	28.94536
Spain	27.41410
South Africa	27.26899

Country	avg_debutAge
Brazil	27.13210
Greece	26.88843
South Korea	26.58362
Guam	26.47502
Latvia	26.17933
Columbia	25.89049
France	25.62355
British Honduras	25.54141
Belgium	25.47570
Saudi Arabia	25.46475
Scotland	25.33881
Germany	25.29803
Cuba	25.14483
United Kingdom	25.05133
Poland	24.83687
Canada	24.77687
Mexico	24.59462
Virgin Islands	24.45150
Canal Zone	24.36140
Bahamas	24.35866
Australia	24.24367
Singapore	24.01917
West Germany	23.96988
Panama	23.95407
Venezuela	23.89071
Honduras	23.83710
Czechoslovakia	23.82478
Switzerland	23.81109
Puerto Rico	23.67476
China	23.57883
Austria-Hungary	23.53457
Dominican Republic	23.47295
Jamaica	23.35387
Denmark	23.28816
Colombia	23.19087
Ireland	23.06818
Italy	23.01346
Netherlands	22.74011
Nicaragua	22.60027
Norway	22.33676
Czech Republic	22.15743
American Samoa	21.67009
Russia	21.29409
Portugal	20.82957
Sweden	19.82204

Now that we have the average age of debut for each country and the overall average for international players, we can compare with domestic player debuts.

```
# creating dataset of all USA-born players
us_players_debut <- final_df %>%
  filter(birthCountry == "USA") %>%
```

```r
  select(playerID, birthCountry, birthDate, debut)

# cleaning the subset of the data to ensure proper calculation of age of debut
us_players_summary <- us_players_debut %>%
  mutate(debut = as.Date(debut, format = "%Y-%m-%d"),
         birthDate = as.Date(birthDate, format = "%Y-%m-%d")) %>%
  filter(!is.na(birthCountry), !is.na(birthDate), !is.na(debut)) %>%
  mutate(debut_age = as.numeric(difftime(debut, birthDate, units = "days")) / 365.25)

us_debut_age_summary <- aggregate(x = us_players_summary$debut_age,
                                  by = list(playerID = us_players_summary$playerID),
                                  FUN = mean)
# rename "x" to more representative column name
colnames(us_debut_age_summary)[2] <- "debutAge"

# calculation of USA debut ages
round(mean(us_debut_age_summary$debutAge),2)
```

```
## [1] 24.32
```

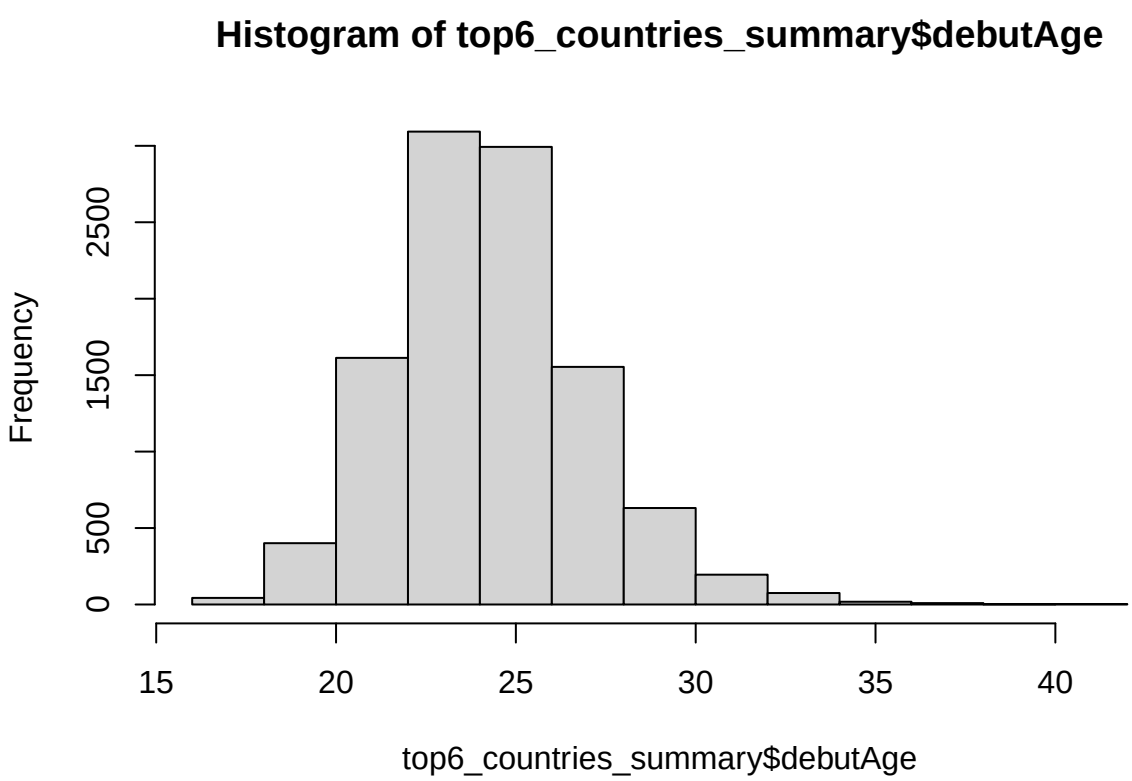
Average age of international-born players at debut: 24.45

Average age of US-born players at debut: 24.32

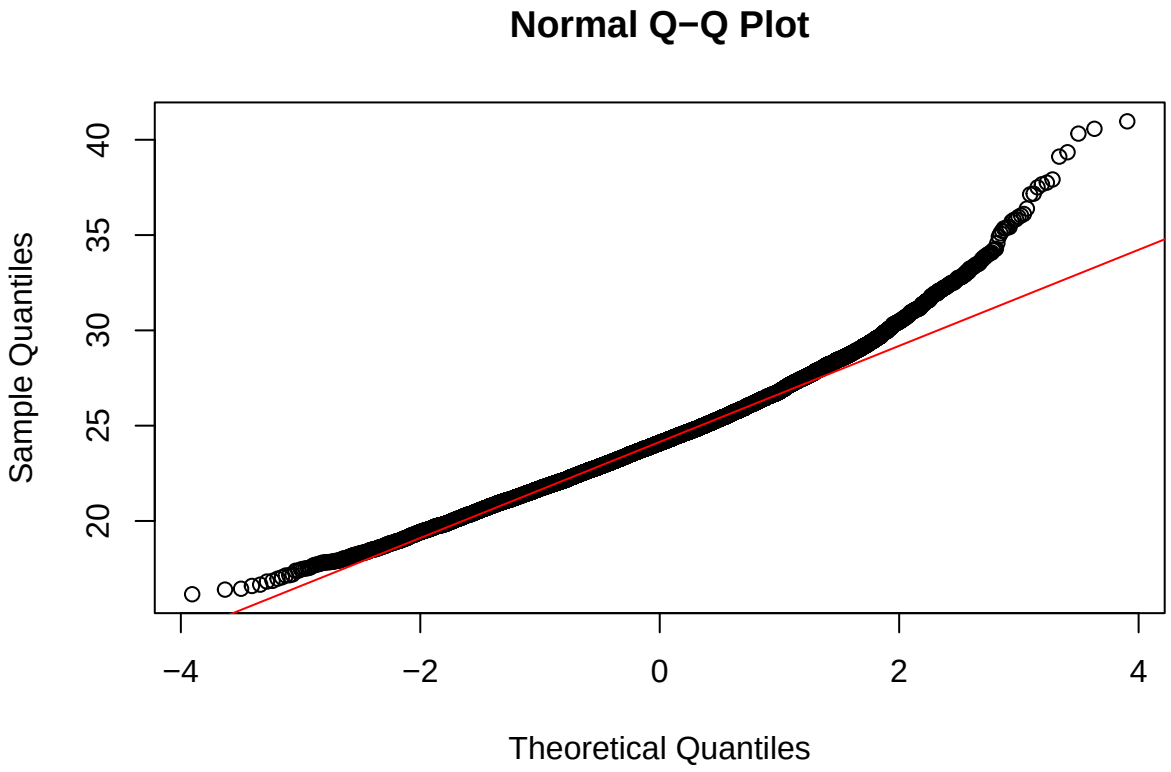
The average age of US-born players' debuts turns out to be lower than international-born players; however, the overall average of international players' debuts could be strongly skewed by countries with smaller sample sizes. Therefore, we can find more meaningful and statistically sound results by selecting a subset of countries with higher sample sizes (e.g., $n < 50$). By identifying the top 5 countries, plus the US, we can look at analyses in other sections easily.

```r
top6_countries <- dist_countries %>%
  arrange(desc(Count)) %>%
  slice_head(n =5)

us_players <- final_df[final_df$birthCountry == "USA", ]
unique_us_players <- us_players[!duplicated(us_players$playerID), ]
total_us_players <- nrow(unique_us_players)
us_row <- data.frame(Country = "USA",
                     Count = total_us_players)

top6_countries_raw <- rbind(top6_countries, us_row) %>%
  arrange(desc(Count)) %>%
  pull(Country)

top6_countries_data <- final_df %>%
  filter(birthCountry %in% c(top6_countries_raw))

top6_countries_debut_age <- top6_countries_data %>%
  mutate(debut = as.Date(debut, format = "%Y-%m-%d"),
         birthDate = as.Date(birthDate, format = "%Y-%m-%d")) %>%
  filter(!is.na(birthCountry), !is.na(birthDate), !is.na(debut)) %>%
  mutate(debut_age = as.numeric(difftime(debut, birthDate, units = "days")) / 365.25)

top6_countries_summary <- aggregate(x = top6_countries_debut_age$debut_age,
```

```
                                  by = list(playerID = top6_countries_debut_age$playerID),
                                  FUN = mean)

colnames(top6_countries_summary)[2] <- "debutAge"

top6_countries_summary <- top6_countries_summary %>%
  left_join(select(top6_countries_debut_age, playerID, birthCountry), by = "playerID") %>%
  distinct()

# approximate normality based on shape of curve
hist(top6_countries_summary$debutAge)
```




```
qqnorm(top6_countries_summary$debutAge)
qqline(top6_countries_summary$debutAge, col = "red")
```




```
# normality, skew and kurtosis tests for a large number of entries that exceed limits of Shapiro-Wilk t

ad.test(top6_countries_summary$debutAge)
```

```
##
##  Anderson-Darling normality test
##
## data:  top6_countries_summary$debutAge
## A = 32.349, p-value < 0.00000000000000022
```

```
skewness(top6_countries_summary$debutAge)
```

```
## [1] 0.6161591
```

```
kurtosis(top6_countries_summary$debutAge)
```

```
## [1] 1.337877
```

```
# Because there was evidence of skewed data, I log transformed and re-ran the normality/skew/kurtosis t
trans_top6_countries_summary <- log(top6_countries_summary$debutAge)

ad.test(trans_top6_countries_summary)
```

```
##
##  Anderson-Darling normality test
##
## data:  trans_top6_countries_summary
## A = 6.1537, p-value = 0.000000000000004068
```

```r
skewness(trans_top6_countries_summary)
```

```
## [1] 0.1801251
```

```r
kurtosis(trans_top6_countries_summary)
```

```
## [1] 0.5281931
```

```r
# ANOVA of debut ages of 6 countries with most players
anova_debut <- aov(trans_top6_countries_summary ~ birthCountry, data = top6_countries_summary)
summary(anova_debut)
```

```
##                 Df Sum Sq Mean Sq F value           Pr(>F)
## birthCountry     5   0.82 0.16336   13.52 0.00000000000036 ***
## Residuals    10624 128.41 0.01209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Tukey to assess which pairs differ significantly
TukeyHSD(anova_debut)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = trans_top6_countries_summary ~ birthCountry, data = top6_countries_summary)
##
## $birthCountry
##                                       diff          lwr          upr     p adj
## Cuba-Canada                      0.012367916 -0.025443770  0.050179601 0.9382504
## Dominican Republic-Canada       -0.051675147 -0.082879643 -0.020470651 0.0000352
## Puerto Rico-Canada              -0.044486787 -0.079309143 -0.009664432 0.0037061
## USA-Canada                      -0.018906137 -0.046267661  0.008455388 0.3600733
## Venezuela-Canada                -0.034880701 -0.067767974 -0.001993429 0.0301712
## Dominican Republic-Cuba         -0.064043062 -0.094488582 -0.033597543 0.0000000
## Puerto Rico-Cuba                -0.056854703 -0.090998597 -0.022710809 0.0000310
## USA-Cuba                        -0.031274052 -0.057766732 -0.004781372 0.0100004
## Venezuela-Cuba                  -0.047248617 -0.079416640 -0.015080593 0.0004100
## Puerto Rico-Dominican Republic   0.007188360 -0.019453760  0.033830479 0.9726955
## USA-Dominican Republic           0.032769010  0.017089114  0.048448906 0.0000000
## Venezuela-Dominican Republic     0.016794446 -0.007263376  0.040852267 0.3482059
## USA-Puerto Rico                  0.025580651  0.003563953  0.047597348 0.0119768
## Venezuela-Puerto Rico            0.009606086 -0.018988575  0.038200747 0.9311666
## Venezuela-USA                   -0.015974565 -0.034781777  0.002832647 0.1490038
```

```r
anova_means_debut <- aggregate(debutAge ~ birthCountry, data = top6_countries_summary, FUN = function(x)
  arrange(desc(debutAge))

colnames(anova_means_debut)[2] <- "Age"
colnames(anova_means_debut)[1] <- "Country"

kable(anova_means_debut,
      caption = "Average ages of international players at debut, with USA")
```

Table 3: Average ages of international players at debut, with USA

Country	Age
Cuba	25.14483
Canada	24.77687
USA	24.32181
Venezuela	23.89071
Puerto Rico	23.67476
Dominican Republic	23.47295

There appear to be group differences! For the sake of our analyses, let us focus on the pairs that contain the United States and assess the p-values. Based on the Tukey results, we are able to determine that the debut ages of players from the Dominican Republic (23.46) and the US (24.31), Cuba (25.14) and the US, as well as the US (24.31) and Puerto Rico (23.67) are significantly different. Any reationships in which the p-value is less than 0.05 represent significant differences between the means.

Therefore, we can conclude that international players **in some countries** debut in the MLB at significantly different ages (alpha = 0.05), such that players from the Dominican Republic and Puerto Rico debut earlier than the US who debut earlier than players from Cuba.

Perfomance Metrics Because we built a dataset that already aggregated the groups of countries we are interested in because they have sufficient N's to power our analyses, we can now build a model to learn about differences in Wins Above Average between international-born players in the 5 most represented countries and compare with WAA of players in the United States.

```
top6_countries_data$WAA_batting <- as.numeric(top6_countries_data$WAA_batting)
```

```
## Warning: NAs introduced by coercion
```

```
intl_waa_batting <- top6_countries_data %>%
  filter(birthCountry != "USA") %>%
  select(playerID, WAA_batting, year_ID, birthCountry) %>%
  mutate(WAA_batting = case_when(
    WAA_batting == "NULL" ~ NA,
    TRUE ~ WAA_batting))

us_waa_batting <- top6_countries_data %>%
  filter(birthCountry == "USA") %>%
  select(playerID, WAA_batting, year_ID, birthCountry) %>%
  mutate(WAA_batting = case_when(
    WAA_batting == "NULL" ~ NA,
    TRUE ~ WAA_batting))

mean(intl_waa_batting$WAA_batting, na.rm= TRUE)
```

```
## [1] -0.03185703
```

```
mean(us_waa_batting$WAA_batting, na.rm = TRUE)
```

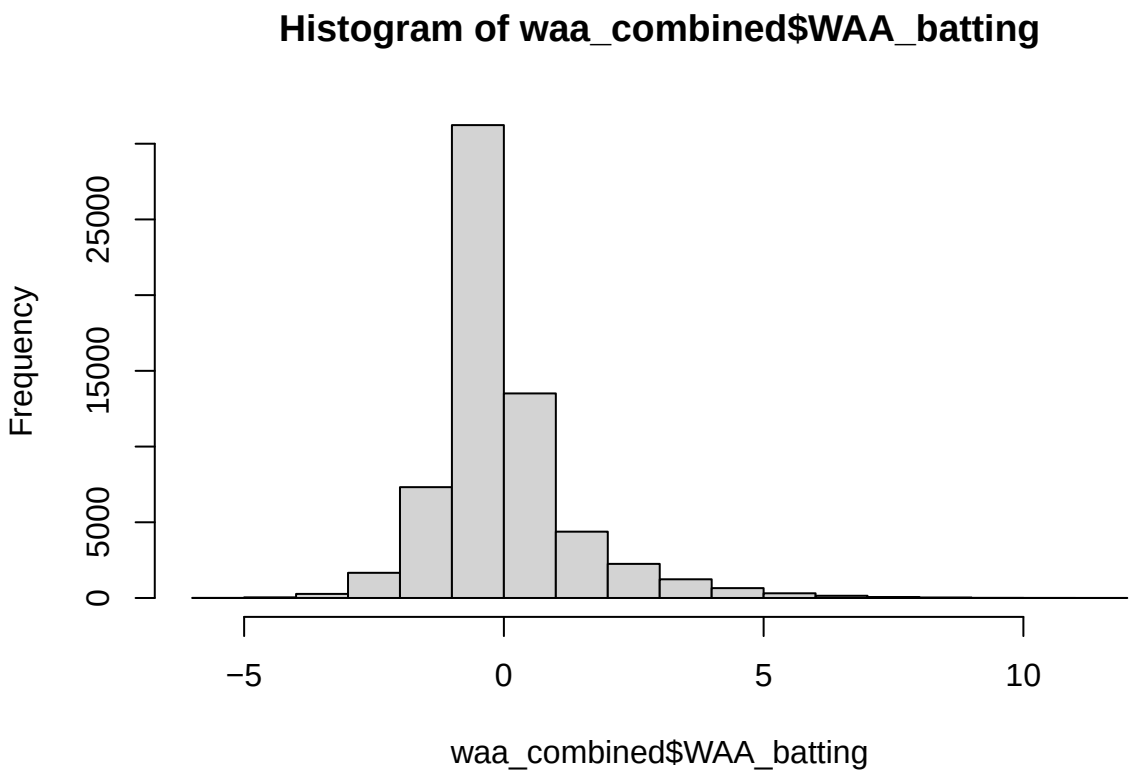
```
## [1] 0.01568563
```

Average WAA of international-born players: -0.03

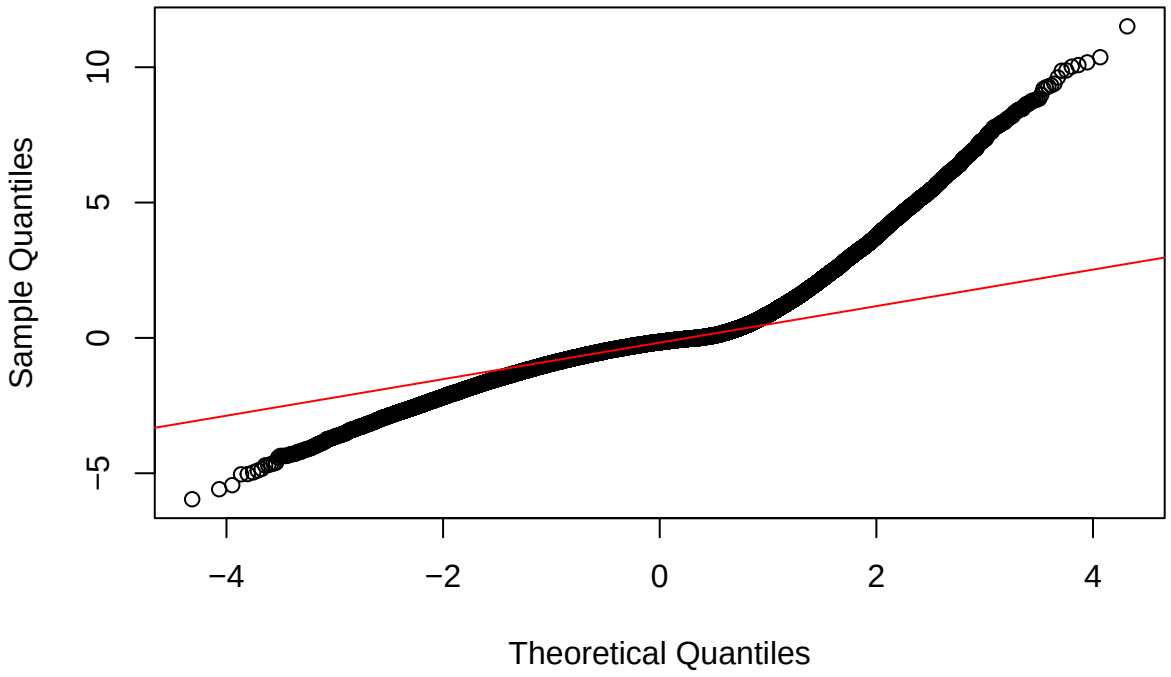
Average WAA of US-born players: 0.02

The average WAA of US-born players' turns out to be more positive than international-born players. However, in order to determine exact group differences, we need to conduct ANOVA.

```
waa_combined <- bind_rows(intl_waa_batting, us_waa_batting)
waa_combined <- na.omit(waa_combined)

hist(waa_combined$WAA_batting)
```




```
qqnorm(waa_combined$WAA_batting)
qqline(waa_combined$WAA_batting, col = "red")
```




```
# normality, skew and kurtosis tests for a large number of entries that exceed limits of Shapiro-Wilk t

ad.test(waa_combined$WAA_batting)
```

```
##
##  Anderson-Darling normality test
##
## data:  waa_combined$WAA_batting
## A = 2544.1, p-value < 0.00000000000000022
```

```
skewness(waa_combined$WAA_batting)
```

```
## [1] 1.664772
```

```
kurtosis(waa_combined$WAA_batting)
```

```
## [1] 5.885549
```

```
# cube root transformation
waa_combined$WAA_batting_cube <- sign(waa_combined$WAA_batting) * abs(waa_combined$WAA_batting)^(1/3)

ad.test(waa_combined$WAA_batting_cube)
```

```
##
##  Anderson-Darling normality test
##
## data:  waa_combined$WAA_batting_cube
## A = 2826.2, p-value < 0.00000000000000022
```

```r
skewness(waa_combined$WAA_batting_cube)
```

```
## [1] 0.6004131
```

```r
kurtosis(waa_combined$WAA_batting_cube)
```

```
## [1] -1.021476
```

```r
# ANOVA of debut ages of 6 countries with most players
anova_waa <- aov(WAA_batting_cube ~ birthCountry, data = waa_combined)
summary(anova_waa)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## birthCountry     5     10  2.0922   2.746 0.0174 *
## Residuals    63115  48082  0.7618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Tukey to assess which pairs differ significantly
TukeyHSD(anova_waa)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = WAA_batting_cube ~ birthCountry, data = waa_combined)
##
## $birthCountry
##                                       diff         lwr        upr     p adj
## Cuba-Canada                     -0.035491112 -0.16360252 0.09262030 0.9693990
## Dominican Republic-Canada       -0.087470540 -0.19636765 0.02142657 0.1984172
## Puerto Rico-Canada              -0.085181649 -0.20091432 0.03055102 0.2884811
## USA-Canada                      -0.052769657 -0.15127558 0.04573627 0.6470244
## Venezuela-Canada                -0.098465256 -0.21193487 0.01500435 0.1321436
## Dominican Republic-Cuba         -0.051979428 -0.14730995 0.04335110 0.6291561
## Puerto Rico-Cuba                -0.049690537 -0.15276031 0.05337924 0.7428273
## USA-Cuba                        -0.017278545 -0.10054165 0.06598456 0.9916416
## Venezuela-Cuba                  -0.062974143 -0.16349618 0.03754789 0.4752949
## Puerto Rico-Dominican Republic   0.002288891 -0.07561764 0.08019542 0.9999994
## USA-Dominican Republic           0.034700883 -0.01407199 0.08347375 0.3265251
## Venezuela-Dominican Republic    -0.010994716 -0.08549792 0.06350849 0.9983295
## USA-Puerto Rico                  0.032411992 -0.03015208 0.09497606 0.6795081
## Venezuela-Puerto Rico           -0.013283606 -0.09746316 0.07089595 0.9976970
## Venezuela-USA                   -0.045695599 -0.10396707 0.01257587 0.2217660
```

```r
anova_means_waa <- aggregate(WAA_batting ~ birthCountry, data = waa_combined, FUN = function(x) mean(x,
  arrange(desc(WAA_batting))

colnames(anova_means_waa)[2] <- "WAA"
colnames(anova_means_waa)[1] <- "Country"

kable(anova_means_waa,
      caption = "Average WAA of international players at debut, with USA")
```

Table 4: Average WAA of international players at debut, with USA

Country	WAA
Canada	0.1322016
USA	0.0156856
Cuba	-0.0273319
Dominican Republic	-0.0344375
Puerto Rico	-0.0374493
Venezuela	-0.0816348

There appear to be no group differences, such that the WAA values amongst all countries selected are not significantly different. This makes sense as it is a metric assessing the "average" player; therefore, we should expect to see values that are closer together, indicating that no one country produces players that are significantly better than another.

Countries with strong baseball infrastructures (like the Dominican Republic, Venezuela, or Cuba) may produce similarly high-quality players compared to domestic U.S. players, reducing the variance in WAA scores across countries. Additionally, it is reassuring that the recruitment model appears to select for talent, rather than specific nationalities, which could indicate bias.

Alternate concerns could be sample size or contextual factors that impact WAA. To address the former, we selected the countries with the most representation in the MLB, to improve power of analyses. As for the latter, it is impossible to predict all positive and negative factors that impact WAA over a given season, particularly with season-ending injuries, cultural differences, and sociopolitical events.

Finally, because we only analyzed offensive players, we are missing data about pitchers, some of whom have excellent WAA over the course of their long careers. Further analyses regarding pitching WAA will be completed at a future time.

5. Analysis 2: Pay Equity

Because we built a dataset that already aggregated the groups of countries we are interested in because they have sufficient N's to power our analyses, we can now build a model to learn about differences in Wins Above Average between international-born players in the 5 most represented countries and compare with WAA of players in the United States.

```
all_salary <- Salaries

colnames(all_salary)[1] <- "yearID"
colnames(intl_waa_batting)[3] <- "yearID"

intl_salary <- intl_waa_batting %>%
  left_join(all_salary, by = c("playerID", "yearID")) %>%
  filter(!is.na(salary))
```

```
## Warning in left_join(., all_salary, by = c("playerID", "yearID")): Detected an unexpected many-to-ma
## i Row 1116 of `x` matches multiple rows in `y`.
## i Row 66 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```r
nrow(intl_salary)
```

```
## [1] 3106
```

```r
length(unique(intl_salary$playerID))
```

```
## [1] 500
```

According to the read.me published by Sean Lahman in his most recent package, states that Salary data has not been updated since 2016. Additionally, data from seasons in which a player is injured, remains a free agent, or did not complete the entire season in the MLB (assigned to the minor leagues).

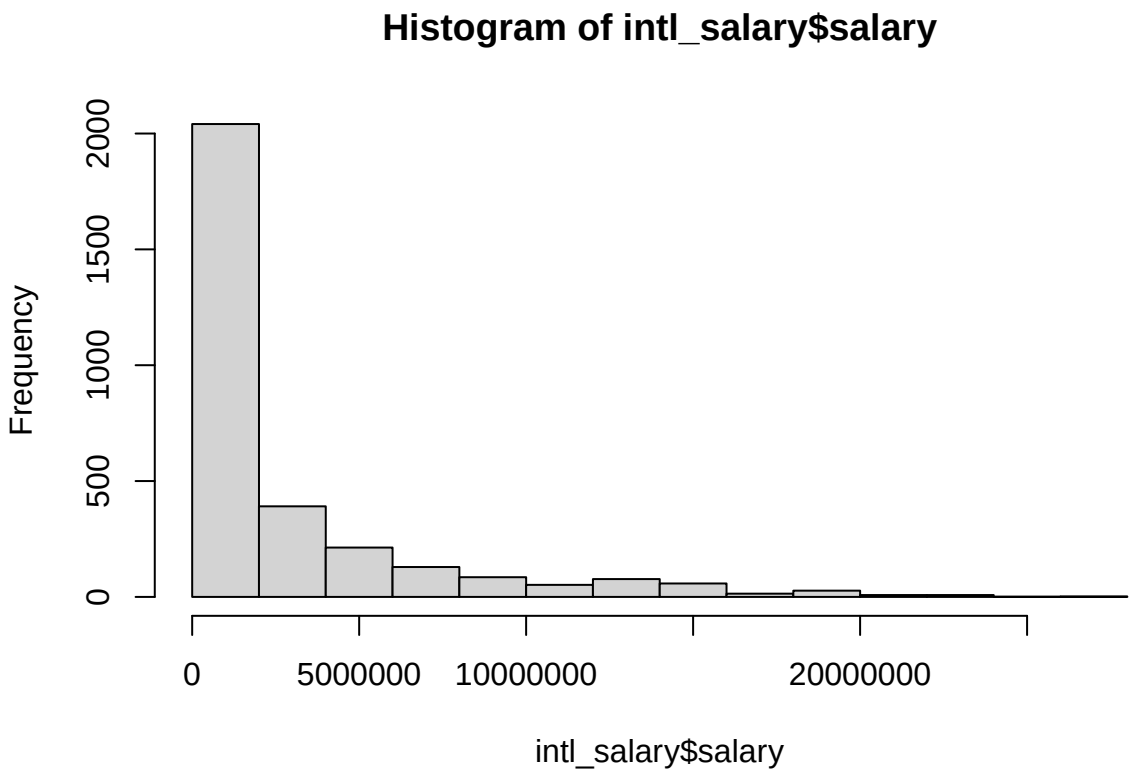
Therefore, the data that will be analyzed for the remainder of this project will only include players with complete salary data between the years of 1985 and 2016. Players who do not meet this criteria will be excluded from the analyses, bringing our N to 3106.

International Player Salary EDA Important to note that 3106 represents the sum of all seasons in which international-born players completed a full season between 1985 and 2016 - NOT the number of unique playerIDs, of which there are 500.

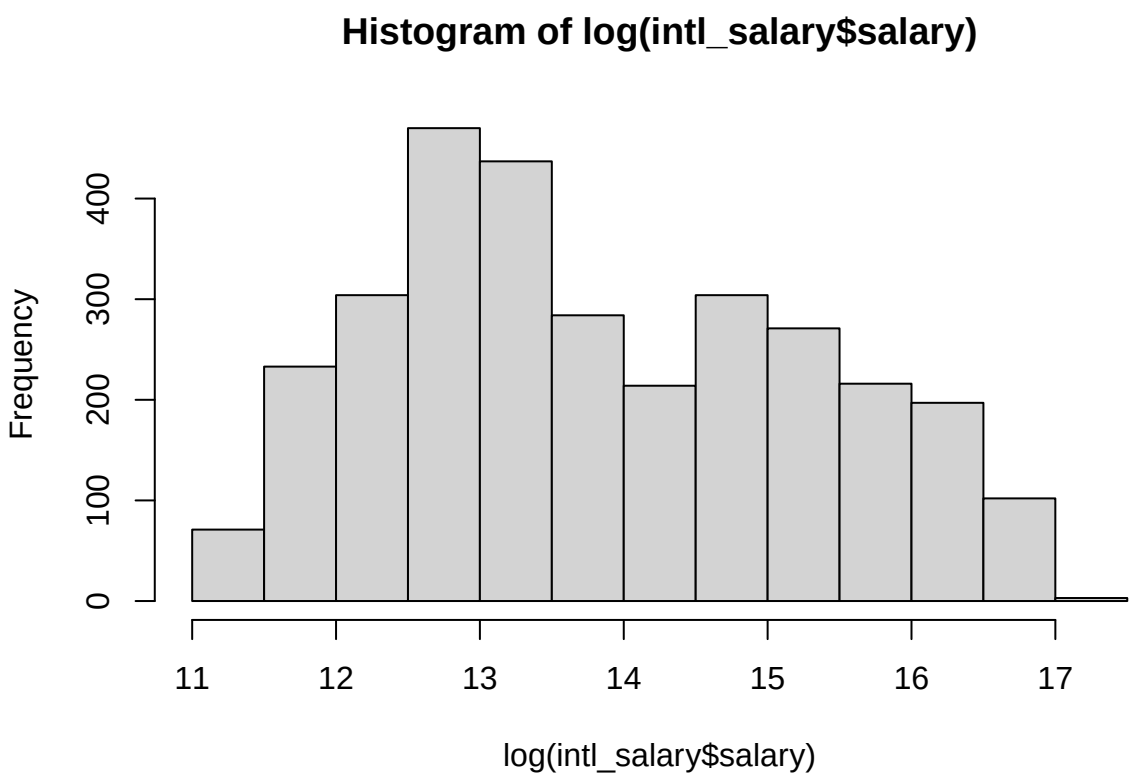
```r
summary(intl_salary$salary)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    60000   330500   775000  2809113  3333333 28000000
```

```r
# compare salary distribution for international players
hist(intl_salary$salary)
```



```r
# log transformation
hist(log(intl_salary$salary))
```



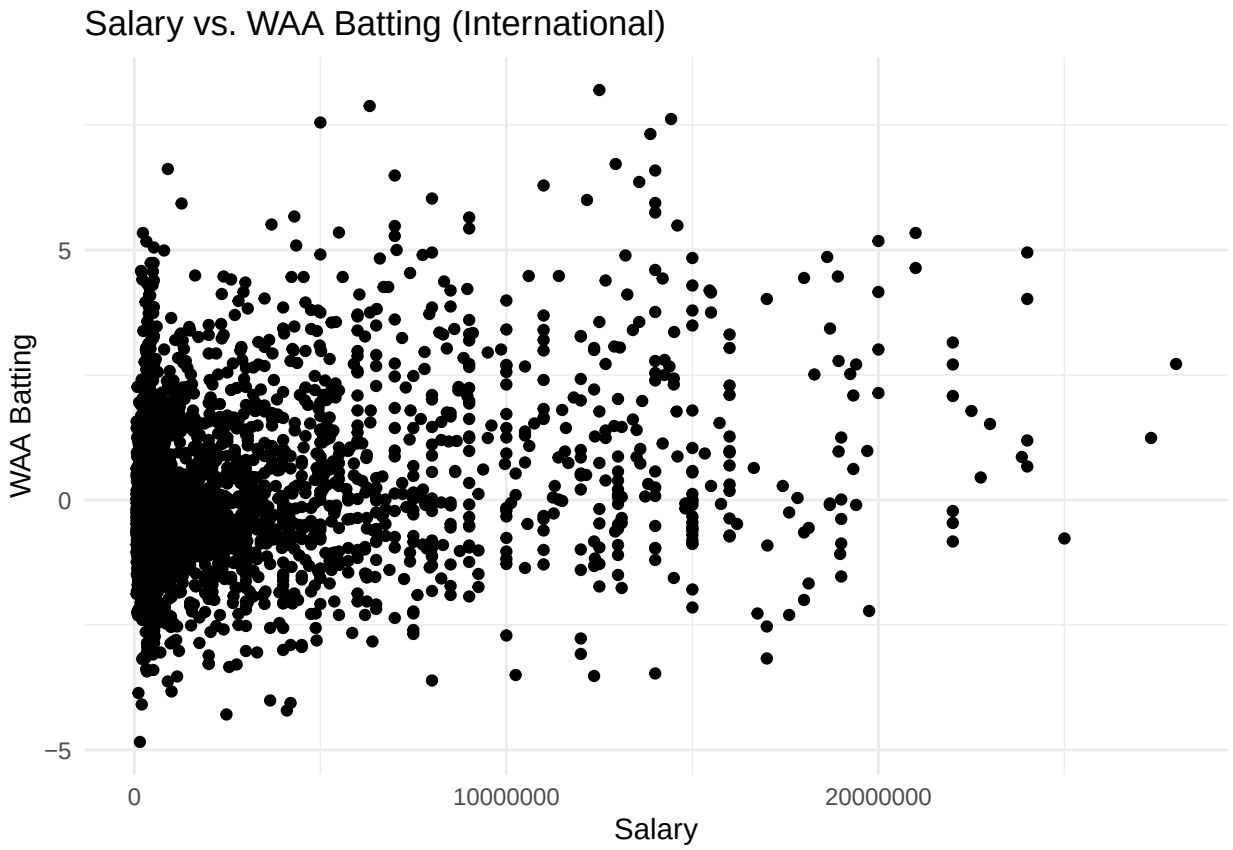
Clearly there is quite a range of salaries, to the point that most distributions of salary are heavily skewed to the right, which makes sense, given that the most outstanding players are few and far between. Players on the lower end might constitute a lack of playing time, short careers, and/or poor market value.

When log transformed, we get something closer to a bimodal distribution, as other methods of transformations were unsuccessful.

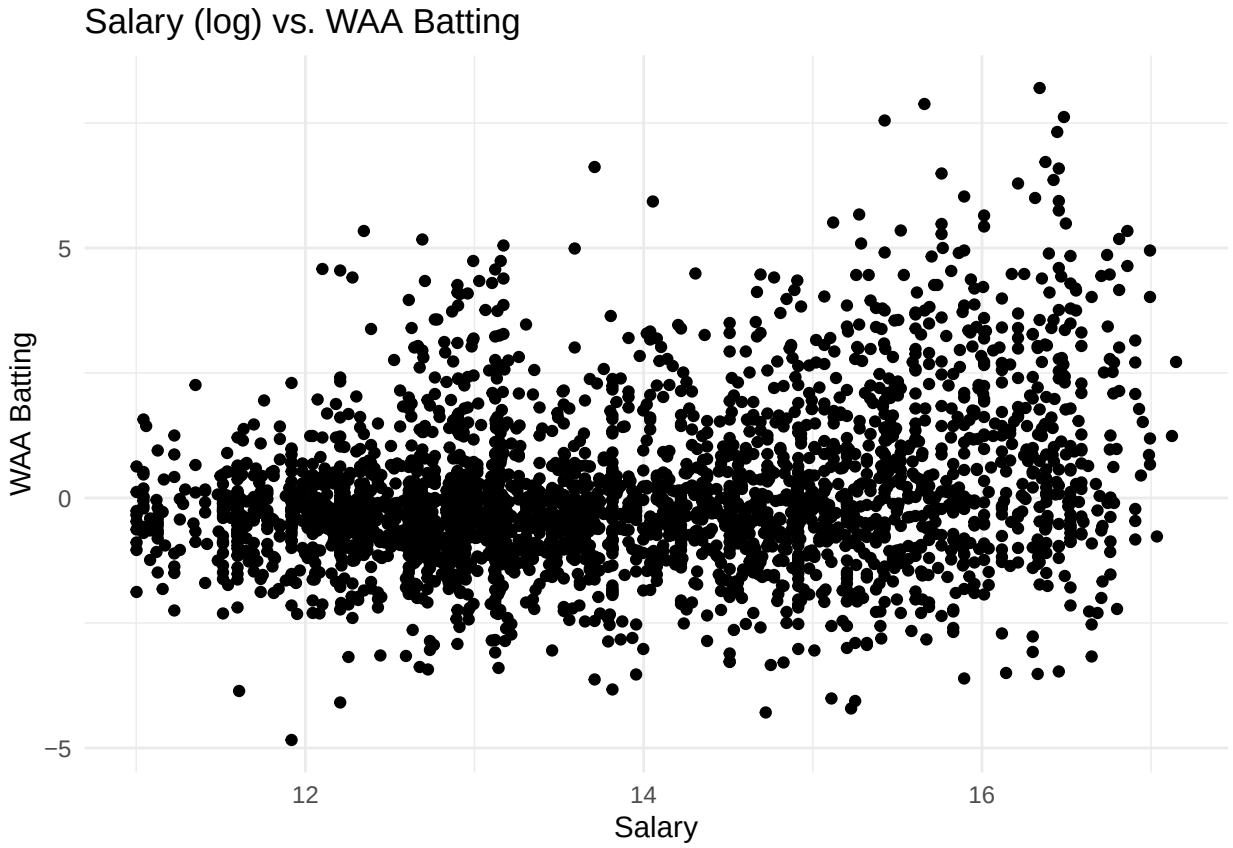
```r
cor_waa_salary <- cor(log(intl_salary$salary), intl_salary$WAA_batting, use = "complete.obs", method =
print(cor_waa_salary)
```

```
## [1] 0.2624735
```

```r
ggplot(intl_salary, aes(x = salary, y = WAA_batting)) +
  geom_point() +
  labs(title = "Salary vs. WAA Batting (International)",
       x = "Salary",
       y = "WAA Batting") +
  theme_minimal()
```




```
ggplot(intl_salary, aes(x = log(salary), y = WAA_batting)) +
  geom_point() +
  labs(title = "Salary (log) vs. WAA Batting",
       x = "Salary",
       y = "WAA Batting") +
  theme_minimal()
```



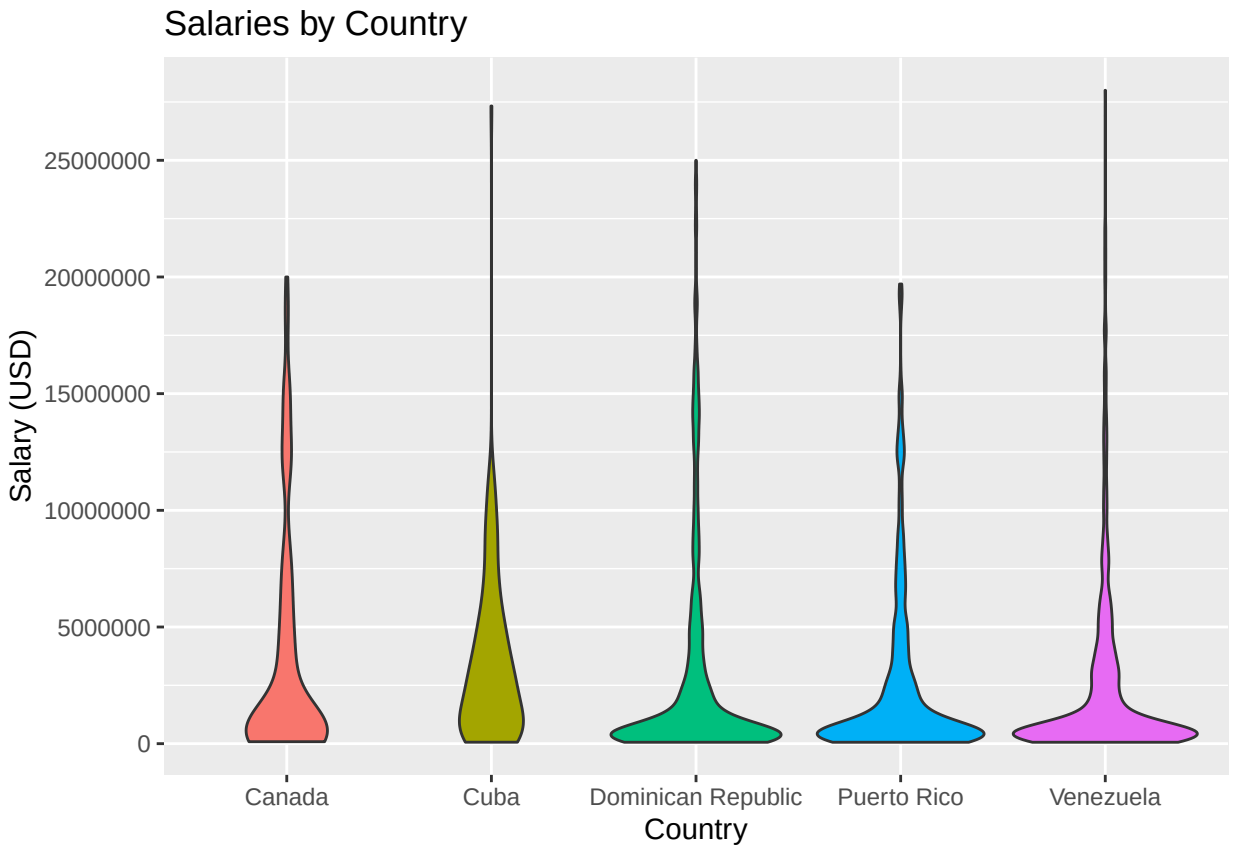

```
lm_waa_salary <- lm(salary ~ WAA_batting,
                    data = intl_salary)
summary(lm_waa_salary)
```

```
##
## Call:
## lm(formula = salary ~ WAA_batting, data = intl_salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7036209 -2275362 -1494455   642180 23599853
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  2758316      73567   37.49 <0.0000000000000002 ***
## WAA_batting   782159      47305   16.53 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4096000 on 3104 degrees of freedom
## Multiple R-squared:  0.08094,    Adjusted R-squared:  0.08065
## F-statistic: 273.4 on 1 and 3104 DF,  p-value: < 0.00000000000000022
```

When looking at international players as a monolith, we can see that there is a very weak positive correlation between WAA and salary. Even after a log transformation and assessment with a linear model, there does not appear to be a strong (if any) relationship between WAA and salary.

```
qplot(x = birthCountry, y = salary,
      data= intl_salary,
      geom = "violin", fill = birthCountry) +
  labs(title = "Salaries by Country",
       x = "Country", y = "Salary (USD)") +
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0, max(intl_salary$salary), by = 5000000))
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once per session.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Again, visualization by country also shows that there are very few superstars and the distribution of salaries reflects this trend when international players are combined together. In particular, it is clear that Cuba appears to have more heterogeneity in their pay; whereas, the latter three on the plot show that there is some evidence for superstars skewing the salary distribution heavily.

With that said, let's take a quick look at the US players and then compare with international-born players.

```
all_salary <- Salaries

colnames(all_salary)[1] <- "yearID"
```

```r
colnames(us_waa_batting)[3] <- "yearID"

us_salary <- us_waa_batting %>%
  left_join(all_salary, by = c("playerID", "yearID")) %>%
  filter(!is.na(salary))
```

Domestic Player Salary EDA

```
## Warning in left_join(., all_salary, by = c("playerID", "yearID")): Detected an unexpected many-to-ma
## i Row 3330 of `x` matches multiple rows in `y`.
## i Row 36 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```r
nrow(us_salary)
```

```
## [1] 11465
```

```r
length(unique(us_salary$playerID))
```

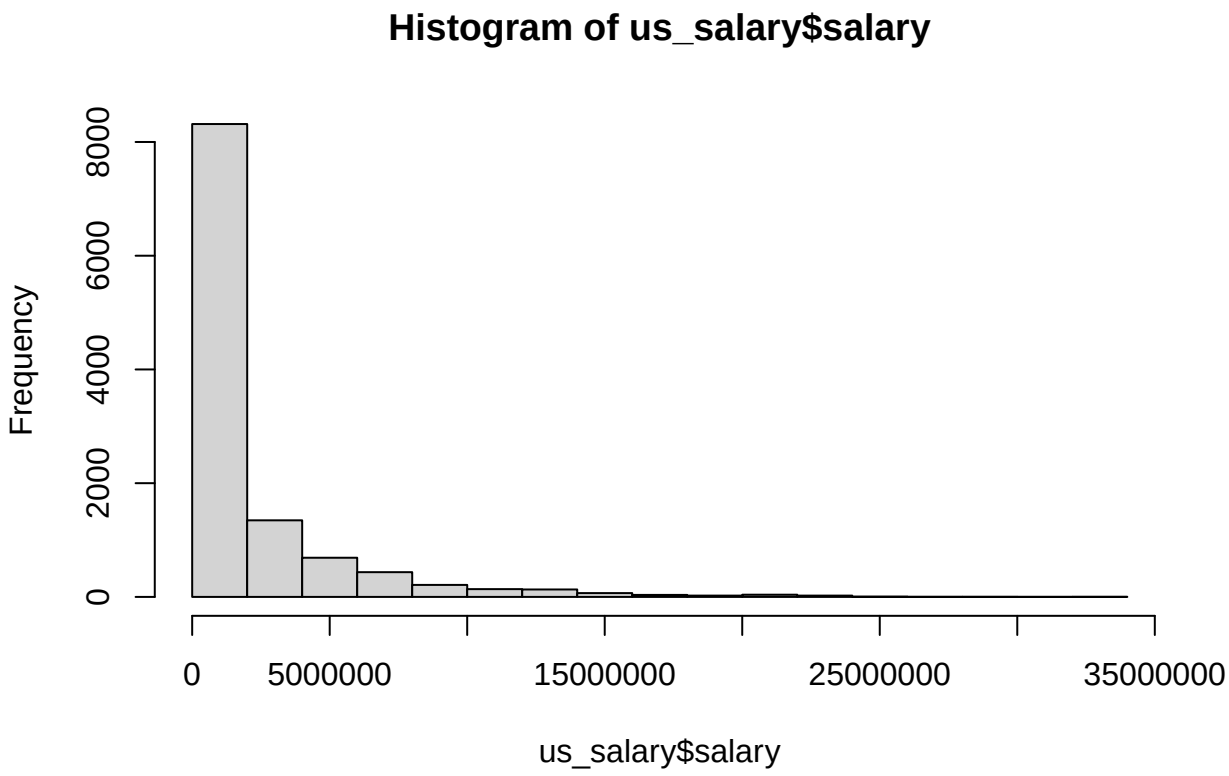
```
## [1] 1847
```

Important to note that 11465 represents the sum of all seasons in which international-born players completed a full season between 1985 and 2016 - NOT the number of unique playerIDs, of which there are 1847.

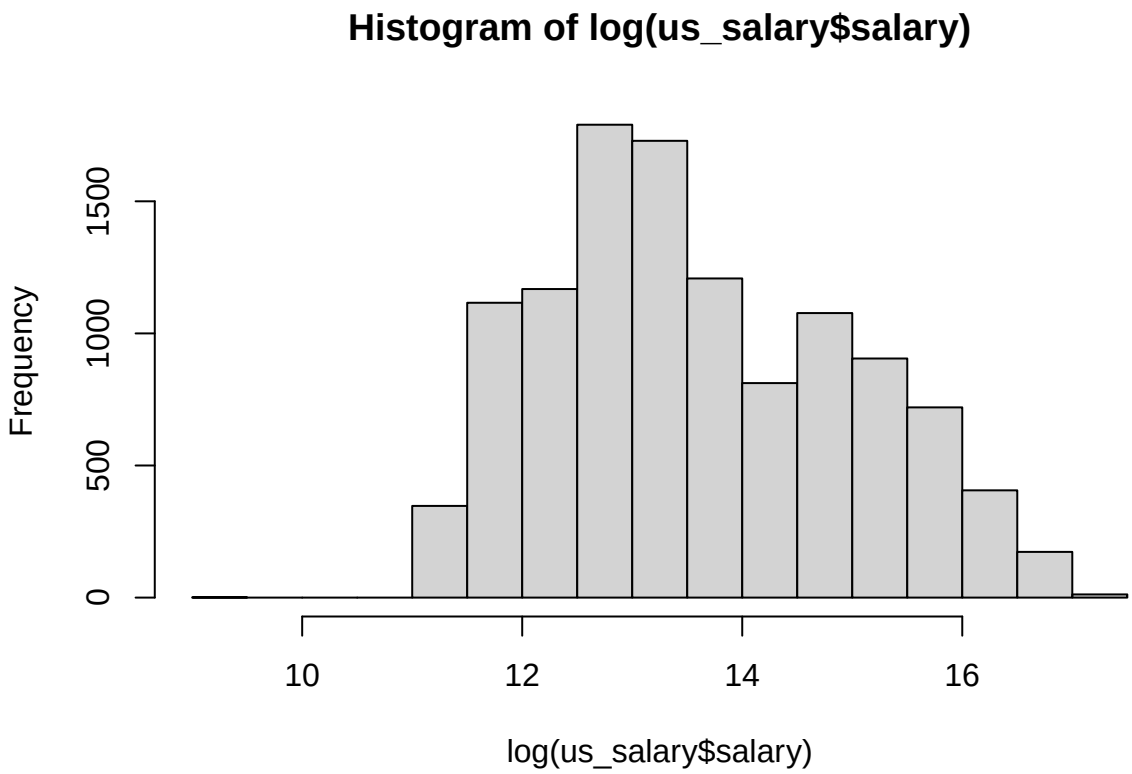
```r
# explore the range of values
summary(us_salary$salary)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0   300000   600000  2096058  2373439 33000000
```

```r
# compare salary distribution for international players
hist(us_salary$salary)
```




```r
# log transformation
hist(log(us_salary$salary))
```



Clearly there is quite a range of salaries, to the point that most distributions of salary are heavily skewed to

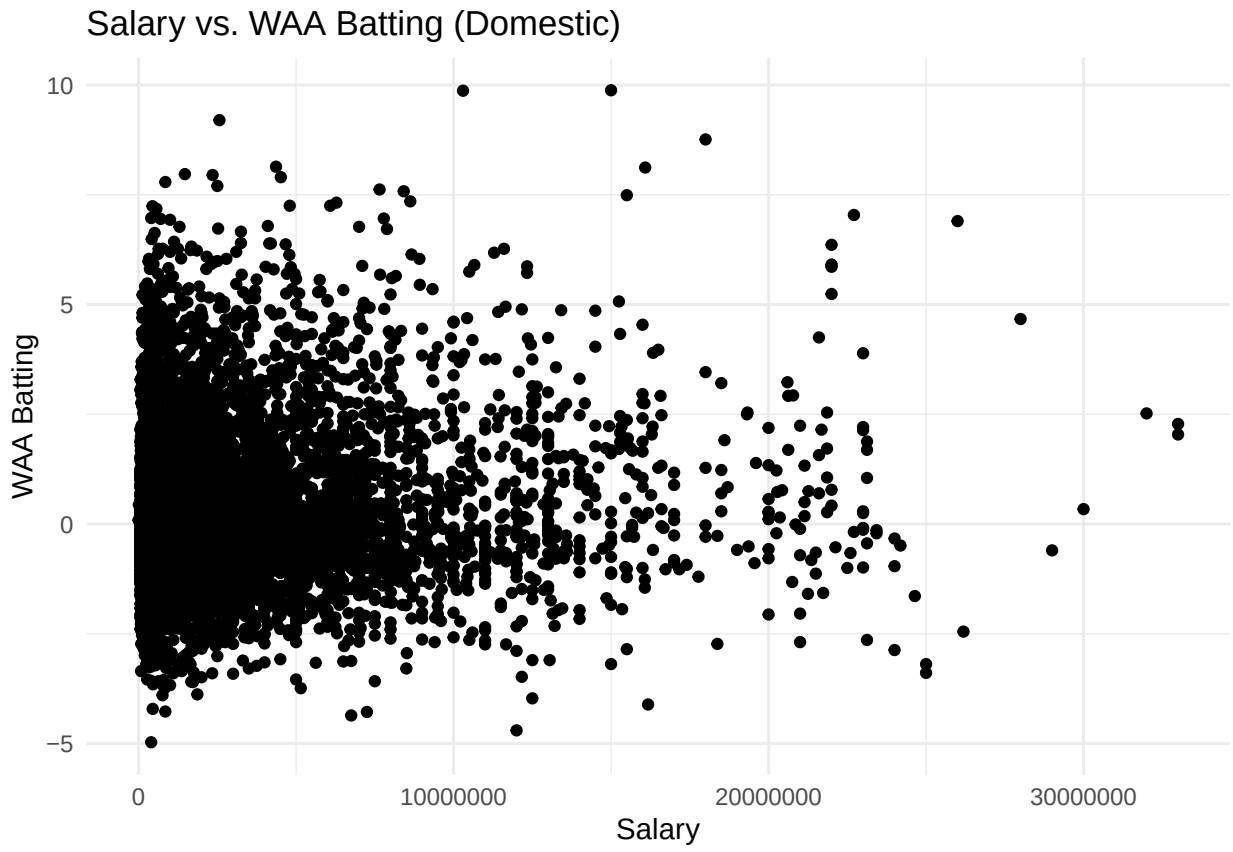
the right, which makes sense, given that the most outstanding players are few and far between. Players on the lower end might constitute a lack of playing time, short careers, and/or poor market value.

When log transformed, we get something closer to a bimodal distribution, as other methods of transformations were unsuccessful.

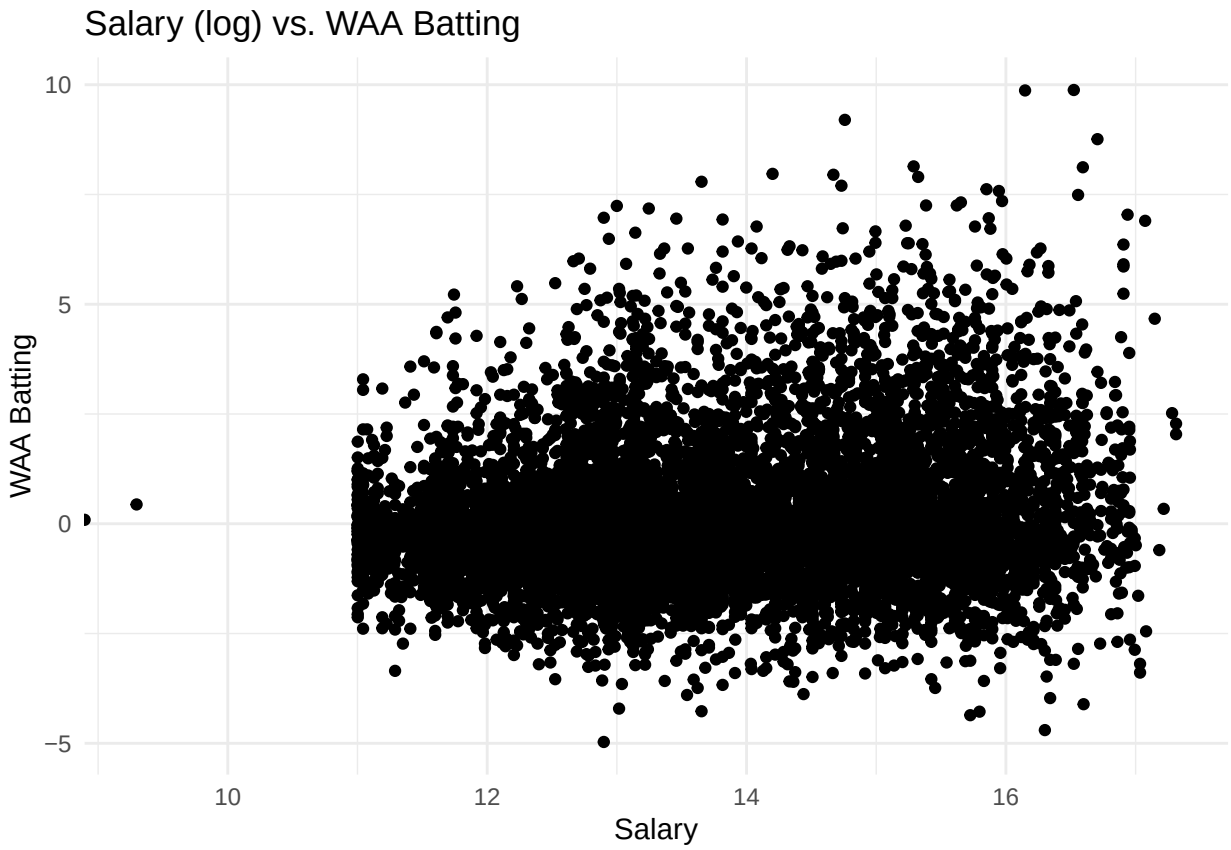
```
cor_waa_salary_dom <- cor(us_salary$salary, us_salary$WAA_batting, use = "complete.obs", method = "pears
print(cor_waa_salary_dom)
```

```
## [1] 0.1627957
```

```
ggplot(us_salary, aes(x = salary, y = WAA_batting)) +
  geom_point() +
  labs(title = "Salary vs. WAA Batting (Domestic)",
       x = "Salary",
       y = "WAA Batting") +
  theme_minimal()
```




```
ggplot(us_salary, aes(x = log(salary), y = WAA_batting)) +
  geom_point() +
  labs(title = "Salary (log) vs. WAA Batting",
       x = "Salary",
       y = "WAA Batting") +
  theme_minimal()
```

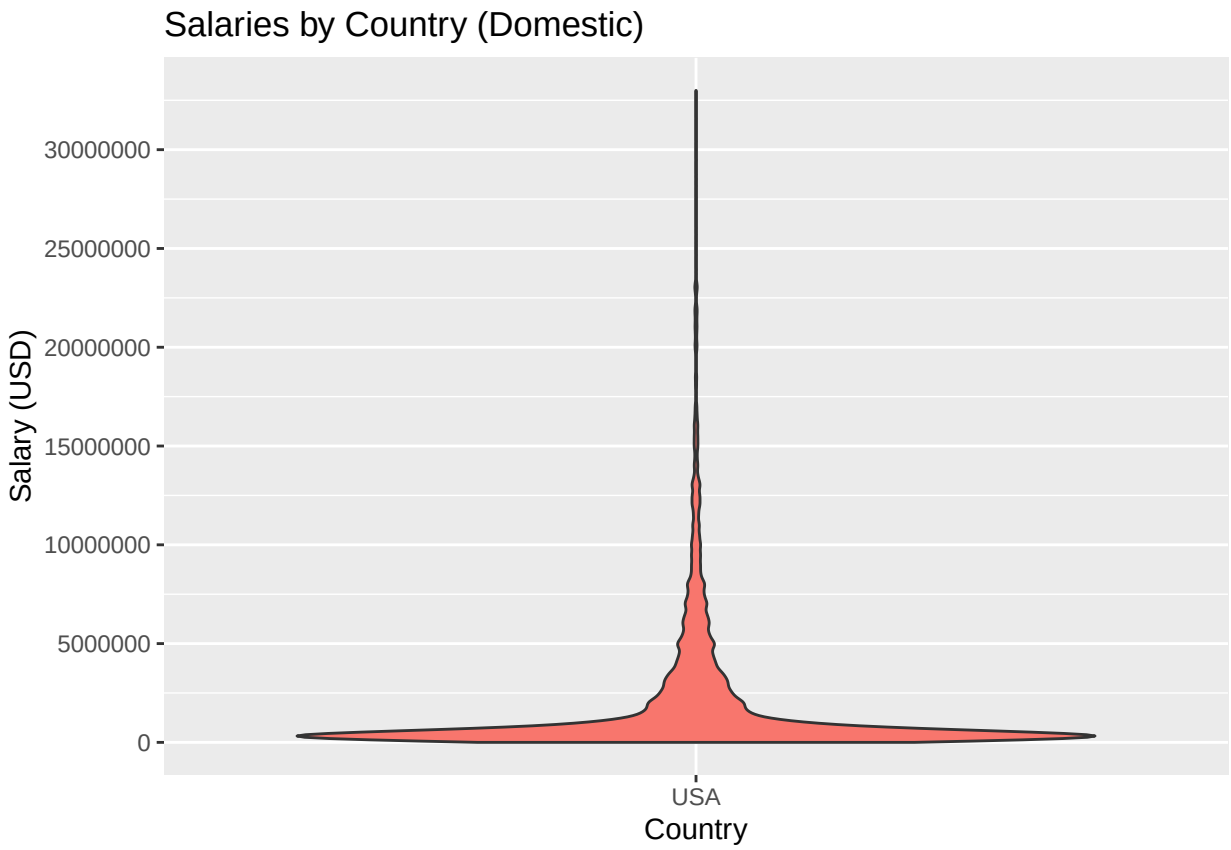



```
lm_waa_salary_dom <- lm(salary ~ WAA_batting,
                        data = us_salary)
summary(lm_waa_salary_dom)
```

```
##
## Call:
## lm(formula = salary ~ WAA_batting, data = us_salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4354137 -1730167 -1278426   267261 30162030
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  2069636      31687   65.31 <0.0000000000000002 ***
## WAA_batting   376635      21320   17.66 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3389000 on 11463 degrees of freedom
## Multiple R-squared:  0.0265, Adjusted R-squared:  0.02642
## F-statistic: 312.1 on 1 and 11463 DF,  p-value: < 0.00000000000000022
```

```
qplot(x = birthCountry, y = salary,
      data= us_salary,
      geom = "violin", fill = birthCountry) +
```

```
  labs(title = "Salaries by Country (Domestic)",
       x = "Country", y = "Salary (USD)") +
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0, max(us_salary$salary), by = 5000000))
```



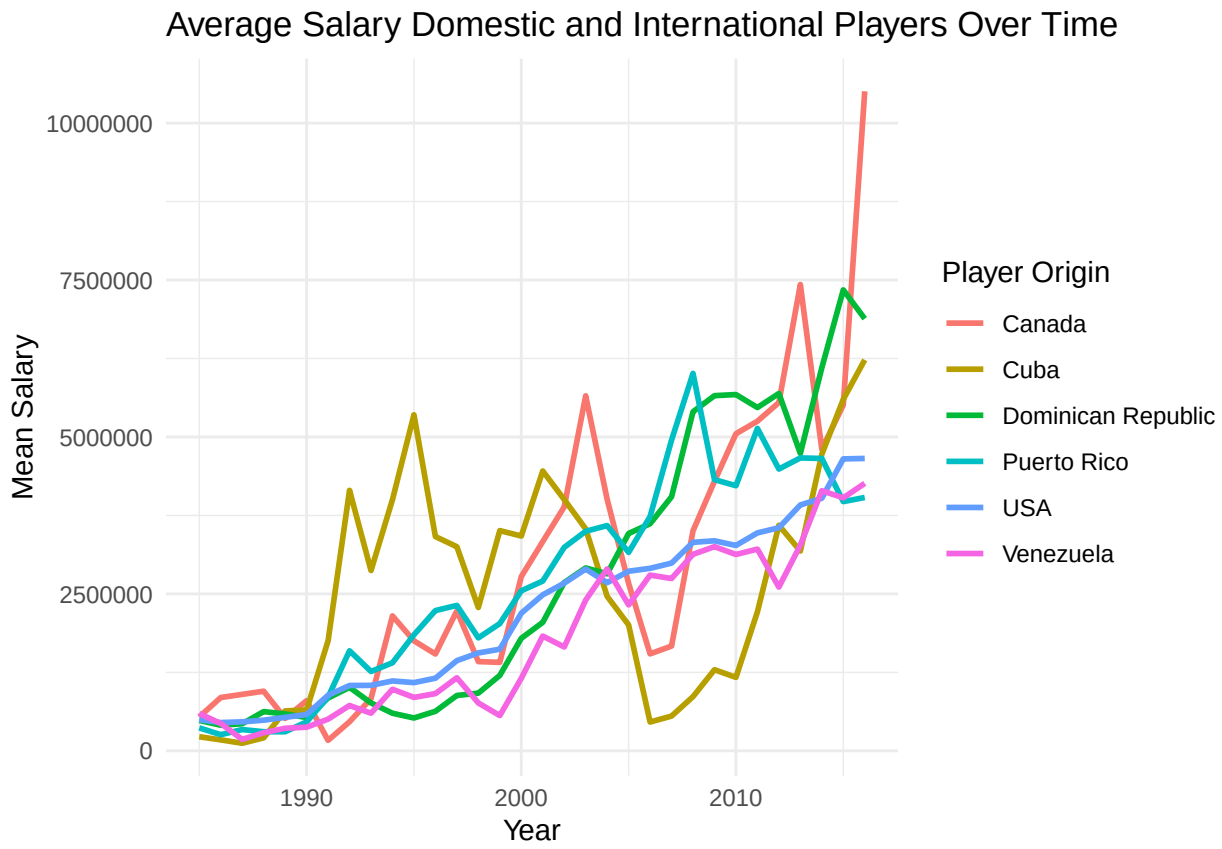
When looking at domestic players as a monolith, we can see that there is a very weak positive correlation between WAA and salary. Even after a log transformation and assessment with a linear model, there does not appear to be a strong (if any) relationship between WAA and salary among domestic players, so we see a similar trend to international players, despite the larger sample size.

When shown as a violin plot, we see a very large range amongst players, but it is important to note that there are nearly 4x the number of US players, compared to international born players, creating more opportunities in pay disparity.

Performance Metrics and Salary over Time Lastly, I want to assess whether there are significant differences in pay with regards to WAA, controlling for time spent in the league, between international and domestic players. We will use ANCOVA to create a model to assess the relationship between these factors and whether there is indeed a pay discrepancy between domestic and international players

```
# combine intl and us salaries
salary_comb <- bind_rows(intl_salary, us_salary)

# average salary by country
mean_salary_by_origin <- aggregate(
  salary ~ yearID + birthCountry,
  data = salary_comb[!is.na(all_salary$salary), ],
```

```r
  FUN = mean
)

# trends over the years
salary_trends <- aggregate(
  salary ~ yearID + birthCountry,
  data = mean_salary_by_origin,
  FUN = function(x) mean(x, na.rm = TRUE)
)

# plotting changing trends in pay
ggplot(salary_trends, aes(x = yearID, y = salary, color = birthCountry)) +
  geom_line(linewidth = 1) +
  labs(
    title = "Average Salary Domestic and International Players Over Time",
    x = "Year",
    y = "Mean Salary",
    color = "Player Origin"
  ) +
  theme_minimal()
```



Using this line graph, we can assess the change in average pay per player from each country of interest over a period of time. Something interesting to note is the large spike in the Canadian data, which corresponds to significant contract signings, such as Freddie Freeman, Justin Morneau and Joey Votto, that propelled the average salary for Canadian-born players extremely high, combined with a shrinking number of Canadian players.

```r
# prepare dataset for ancova analysis
ancova_data <- salary_comb %>%
  filter(!is.na(salary), !is.na(birthCountry), !is.na(yearID), !is.na(WAA_batting))

ancova_data$birthCountry <- as.factor(ancova_data$birthCountry)

ancova_data <- ancova_data %>%
  filter(birthCountry %in% c("USA", "Dominican Republic", "Venezuela", "Japan", "Cuba", "Canada"))

# run ancova and return results
ancova_model <- aov(salary ~ birthCountry + yearID + WAA_batting + birthCountry:yearID, data = ancova_da
summary(ancova_model)
```

```
##                         Df             Sum Sq           Mean Sq F value
## birthCountry             4   1489737350671074   372434337667768   34.52
## yearID                   1  24449616722622284 24449616722622284 2266.46
## WAA_batting              1   6487281242068914  6487281242068914  601.37
## birthCountry:yearID      4    785803297791827   196450824447957   18.21
## Residuals            13824 149127473670361536    10787577667127
##                                    Pr(>F)
## birthCountry        < 0.0000000000000002 ***
## yearID              < 0.0000000000000002 ***
## WAA_batting         < 0.0000000000000002 ***
## birthCountry:yearID  0.00000000000000623 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# post-hoc analyses to look at pairwise relationships
pairwise <- emmeans(ancova_model, pairwise ~ birthCountry)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

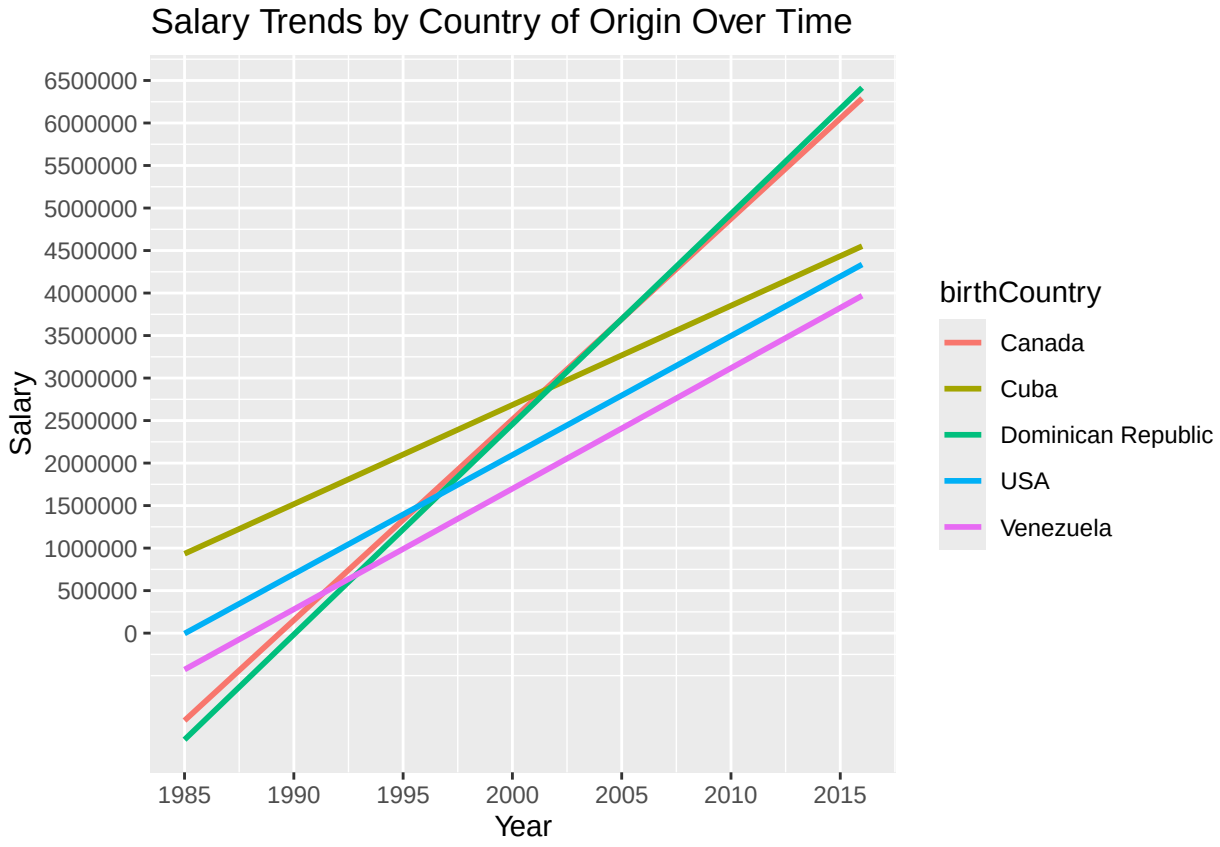
```r
summary(pairwise)
```

```
## $emmeans
##  birthCountry        emmean     SE    df lower.CL upper.CL
##  Canada             2383759 307000 13824  1781600  2985918
##  Cuba               2651008 297000 13824  2068548  3233468
##  Dominican Republic 2643197  94200 13824  2458556  2827838
##  USA                2184128  30700 13824  2123857  2244400
##  Venezuela          1857129 134000 13824  1595437  2118822
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast                      estimate     SE    df t.ratio p.value
##  Canada - Cuba                 -267249 427000 13824  -0.625  0.9710
##  Canada - Dominican Republic   -259438 321000 13824  -0.807  0.9285
##  Canada - USA                   199631 309000 13824   0.647  0.9673
##  Canada - Venezuela             526630 335000 13824   1.572  0.5156
##  Cuba - Dominican Republic        7811 312000 13824   0.025  1.0000
```

```
##  Cuba - USA                        466880 299000 13824   1.563  0.5214
##  Cuba - Venezuela                  793879 326000 13824   2.437  0.1057
##  Dominican Republic - USA          459069  99100 13824   4.633 <0.0001
##  Dominican Republic - Venezuela    786068 163000 13824   4.812 <0.0001
##  USA - Venezuela                   326999 137000 13824   2.387  0.1189
##
## P value adjustment: tukey method for comparing a family of 5 estimates
```

```r
# plot interaction effects
ggplot(ancova_data, aes(x = yearID, y = salary, color = birthCountry)) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(title = "Salary Trends by Country of Origin Over Time",
         x = "Year",
         y = "Salary") +
  scale_x_continuous(breaks = seq(min(ancova_data$yearID), max(ancova_data$yearID)+4, by = 5)) +
  scale_y_continuous(breaks = seq(min(ancova_data$salary), max(ancova_data$salary), by = 500000))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Main Effects:

The p-value for birthCountry is extremely small, indicating a statistically significant difference in mean salaries across countries of origin after controlling for other variables. The F-statistic (34.52) suggests a strong effect of country of origin on salary.

yearID also shows a highly significant effect, meaning that salaries have significantly changed over the years. The very high F-statistic (2266.46) highlights the impact of time on salary trends.

WAA_batting is significant with an F-statistic of 601.37, showing that better performance correlates with higher salaries.

Interaction Effects:

The interaction term birthCountry:yearID is significant, meaning the relationship between salary and year differs across countries. This implies that salary growth trends over time are not uniform across player origins.

Insights from Pairwise Comparisons:

The following pairs were considered to be significant: Dominican Republic-USA and Dominican Republic-Venezuela, such that (at some point) the discrepancies in pay between these pairs became significantly different. We can see this difference illustrated in the interaction plot, such that the difference in average pay could be considered statistically significant. For pair 1, we can conclude that players from the Dominican Republic have achieved a much higher rate of pay compared to USA players. For pair 2, we can conclude that players from the Dominican Republic have achieved a much higher rate of pay compared to Venezuelan players. Both conclusions make sense as they have the greatest distance between the slopes of the lines as time advances.

6. Conclusions and Future Directions

Based on the results summarized through this project, there appear to be differences between domestic and international players, perhaps not in the way originally imagined. Anecdotal evidence and logical assumptions make it easy to overlook contextual factors that drive significant effects once you burrow into the numbers.

We've witnessed how the proportion of international players has expanded in the MLB, we've looked at performance metrics amongst different groups, as well as how those performance metrics predict the rate of compensation and whether country of origin significantly impacts this relationship.

There is still work to be done to improve the quality of life for internationally-born players, such as greater access to social support, early childhood education, financing for housing/transportation/family planning - just to name a few. Furthermore, access to U.S. college education provides an important pipeline for personal and professional development; increasing those opportunities for international players could lead to even greater market value, while also gaining valuable years of academic and athletic resources.

7. References

Baseball Reference. (2024). Data. Retrieved December 7, 2024, from https://www.baseball-reference.com/data/

Baseball Reference. (2024). Wins Above Average (WAA) – Batting statistics. Retrieved December 7, 2024, from https://www.baseball-reference.com/about/waa_batting.shtml

Chavez, L. (2018). The influence of international talent on MLB salaries. Baseball Research Journal, 45(2), 12–28. https://doi.org/10.1177/0034355218764759

Hernandez, B. (2020). Understanding the global expansion of Major League Baseball. Journal of Sports Economics, 21(4), 340–355. https://doi.org/10.1177/1527002520903835

Lahman, L. (2024). Lahman Baseball Database. Retrieved December 7, 2024, from https://cran.r-project.org/web/packages/Lahman/Lahman.pdf

Major League Baseball. (2024). Major League Baseball statistics. Retrieved from https://www.mlb.com/stats

Tingley, D. (2019). R for data science (2nd ed.). O'Reilly Media. https://r4ds.had.co.nz

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer. https://ggplot2.tidyverse.org