Players in the NFL, NBA, MLB, & NHL

By Sandra Froonjian, Solito Reyes, and Matt Sadowski

Essential Questions:

1. Is player height correlated to ability in the NBA? Is so, what is the optimal height for NBA players?

2. Is player age correlated to ability in the NFL, NBA, MLB, and NHL? If so, what is the optimal age for each sport?

3. Are there specific towns/areas that players in the NFL, NBA, MLB, and NHL are more likely to come from? If so, where?

4. Are there specific colleges that players in the NFL, NBA, MLB, and NHL are more likely to attend? Is so, which colleges?

Motivation to Answer These Questions:

1. Matt and Solito are big sports fans.
2. It is a multi-billion dollar industry full of statistics
3. It is an extremely competitive industry, most teams now have dedicated analytics departments dedicated to finding any trends that may provide an edge over the competition.
4. The questions we are are looking to answer could help decide contracts and know where to look to recruit players.

Brief Overview of Findings:

- Able to find hot spots of where players in each sport were born in and attended college
- Not able to find strong correlations between age/height vs. ability.

Data Exploration and Cleanup Process

See Jupyter Notebooks

1. Import data from API- www.mysportsfeeds.com
2. Remove players whose ages was 1-2 years old (incorrect data)
3. Convert object columns into floats
4. Run GeoCode to find coordinates
5. Delete rows with NaN values for college, birth city, birth state, latitude, or longitude

Cleanup & Data Exploration Process

```
msf = MySportsFeeds(version="2.0")
msf.authenticate(sportskey, "MYSPORTSFEEDS")
output = msf.msf_get_data(league='nba',season='2018-2019-regular',feed='seasonal_player_stats',
num_records = len(output["playerStatsTotals"])
for x in range(num_records):
    stats_df.loc[x, 'Name'] = output["playerStatsTotals"][x]["player"]["firstName"] + " " + out
    stats_df.loc[x, 'Age'] = output["playerStatsTotals"][x]["player"]["age"]
    stats_df.loc[x, 'Birth City'] = output["playerStatsTotals"][x]["player"]["birthCity"]
    stats_df.loc[x, 'Birth Country'] = output["playerStatsTotals"][x]["player"]["birthCountry"]
    stats_df.loc[x, 'College'] = output["playerStatsTotals"][x]["player"]["college"]
    stats_df.loc[x, 'Height (in)'] = output["playerStatsTotals"][x]["player"]["height"]
    stats_df.loc[x, 'Ability Score'] = output["playerStatsTotals"][x]["stats"]["rebounds"]["reb
```

	Age	Number of Players of that Age
15	2	17
17	19	9
12	20	26
8	21	42
5	22	68

```
# removes rows where age was 2
age_count = age_count.loc[age_count["Age"] > 2]
age_count


# converts rows to numbers instead of objects
height_df['Height (in)'] = pd.to_numeric(height_df['Height (in)'])
height_df['Ability Score'] = pd.to_numeric(height_df['Ability Score'])
```

```
# loops through every row and changes the height from ft'in" to just inches
for index, row in height_df.iterrows():
    H_feet = height_df.loc[index, 'Height (in)'].split("'")[0]
    H_inch = height_df.loc[index, 'Height (in)'].split("'")[1].split("\"")[0]
    H_inches = int(H_feet) * 12 + int(H_inch)
    height_df.loc[index, 'Height (in)'] = H_inches

height_df
```

Using GeoCode to Find Coordinates of Hometowns and Colleges

```
# itterate through every index and row of df
for index, row in birth_df.iterrows():
# sets target place to the name of each player's birth location
    target_place = "{0},{1}".format(row['Birth City'], row['Birth Country'])
# sets parameters for url
    params = {
        "address": target_place,
        "key": gkey
    }
# if geocode can find the coordinates for the college, proceed, if not, output a print statemen
    try:
        base_url = "https://maps.googleapis.com/maps/api/geocode/json"
# extracts contents of API
        response = requests.get(base_url, params = params).json()
# sets variable to the path to take within the API dictionary
        results = response['results'][0]["geometry"]["location"]
# adds the lat and lng of each location to the df
        birth_df.loc[index, 'Birth Lat'] = results["lat"]
        birth_df.loc[index, 'Birth Lng'] = results["lng"]
    except IndexError:
        print("Can't find coordinates of town... skipping")
```

	Unnamed: 0	Name	Birth City	Birth Country	Birth Lat	Birth Lng	College	College Lat	College Lng
0	0	Justin Abdelkader	Muskegon, MI	USA	43.234181	-86.248392	NaN	NaN	NaN
1	1	Pontus Aberg	Stockholm	Sweden	59.329323	18.068581	NaN	NaN	NaN
2	2	Pontus Aberg	Stockholm	Sweden	59.329323	18.068581	NaN	NaN	NaN
3	3	Vitaly Abramov	Chelyabinsk	Russia	55.164442	61.436843	NaN	NaN	NaN
4	4	Noel Acciari	Johnston, RI	USA	41.820520	-71.512617	NaN	NaN	NaN

```
# removes any rows where birth cities/countries lat/lng were NaN
birth_df = birth_df[birth_df['Birth Lat'].notna()]
birth_df = birth_df[birth_df['Birth Lng'].notna()]
birth_df
```

Analysis Process

NFL - created a production metric based on total yards per player for each position plus total TD*70

MLB - cross-reference player data with WAR information from baseball-reference.com

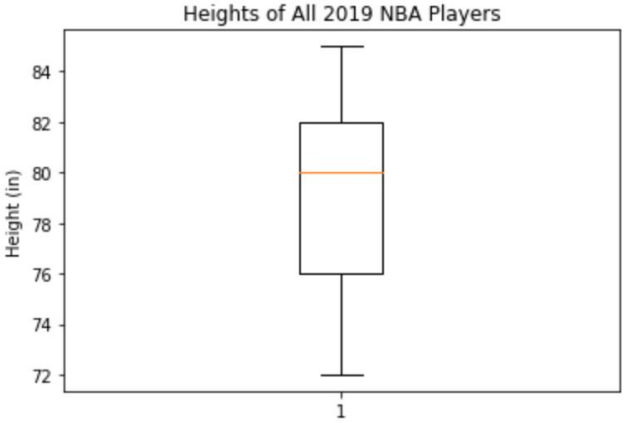
NBA- combined points, assists, rebounds, blocks, and steals for each players

NHL-

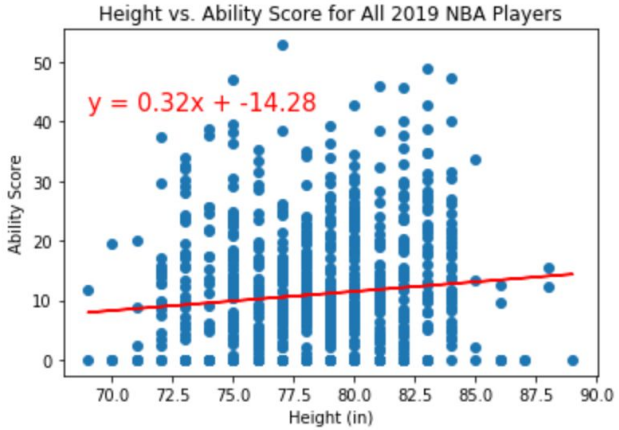
- Non-goalies: used points metrics for players (goals + assists)
- Goalies: used goals against average + (save percentage * 10)

Conclusions

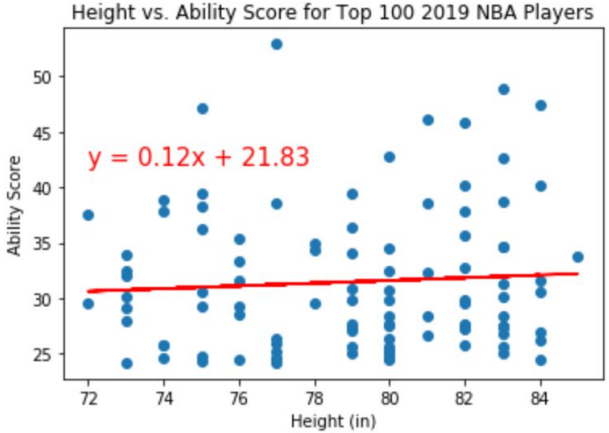
Is player height correlated to ability in the NBA? If so, what is the optimal height?



The r-squared value is: 0.010513050309248833



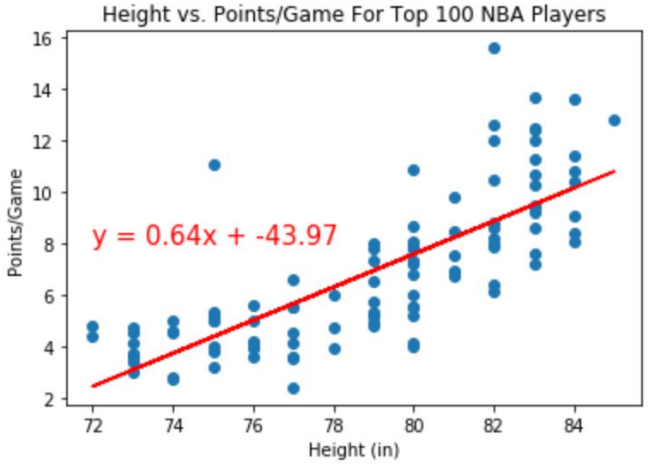
The r-squared value is: 0.004507402780407751



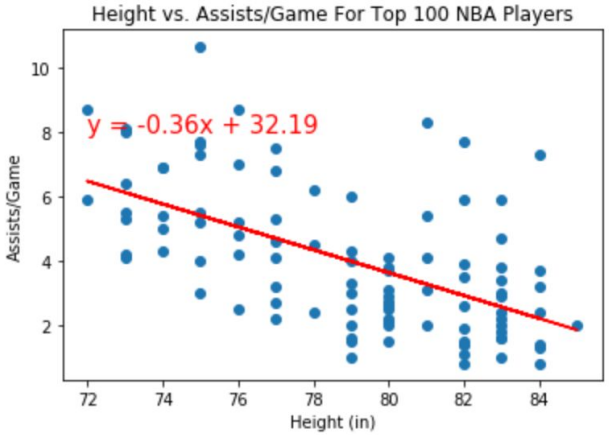
The median age of 2019 NBA players is about 80 in. (6’8”). There is no definitive correlation between player height and ability* for both all players and top 100 players, as the r-squared values of both plots are very close to 0. Therefore, an optimal height cannot be determined. However, no players with a height greater than 85 in. have high ability scores (above about 15).

*ability was calculated by summing each player’s points, assists, rebounds, blocks, and steals

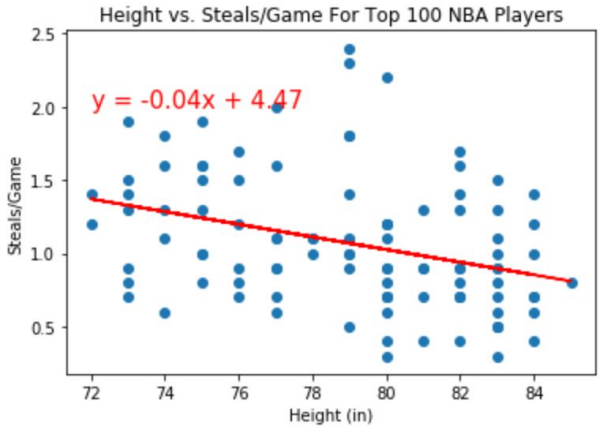
The r-squared value is: 0.5880239251984185



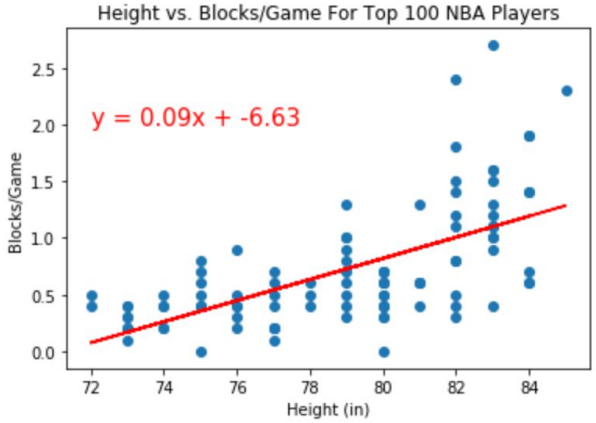
The r-squared value is: 0.3308596237333154



The r-squared value is: 0.11324236051985086



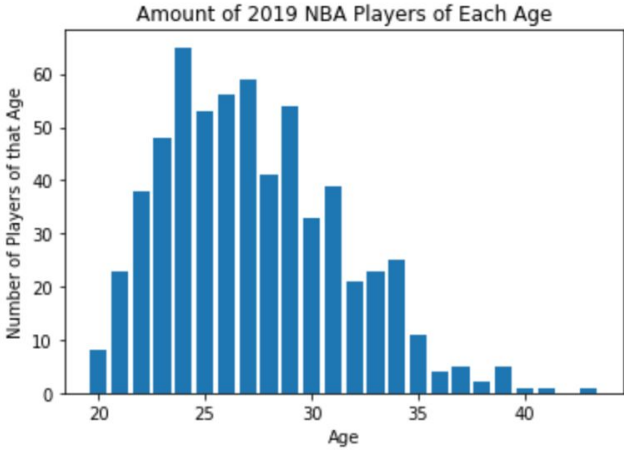
The r-squared value is: 0.39010195915685175



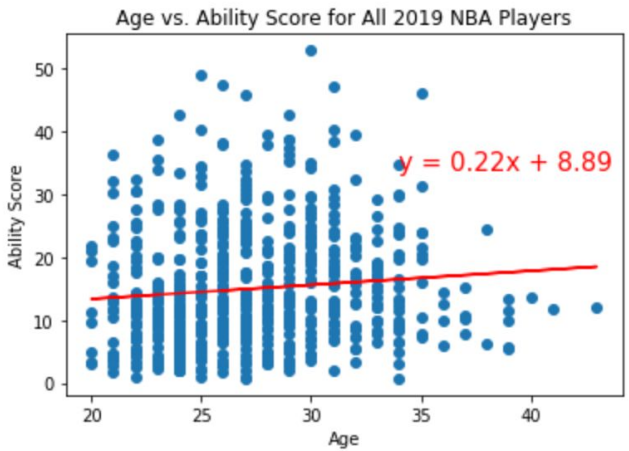
When breaking down ability score, a stronger correlation between height and ability can been see. There is a positive correlation between height vs. points and height vs. block, while there is a negative correlation between height vs. assists and heights vs. steals. This makes sense because if a player is making more points, they're probably going to be making less assists.

Conclusions

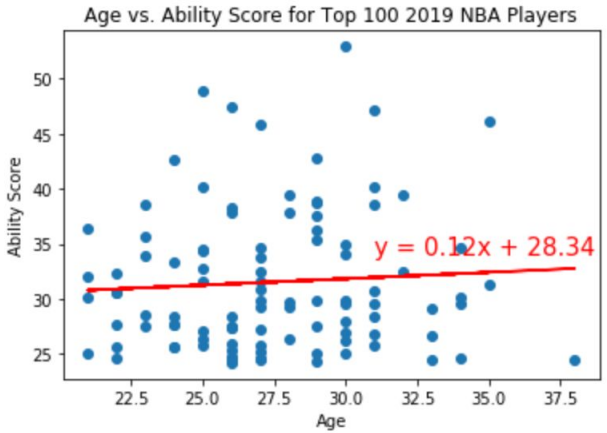
Is player age correlated to ability in the NBA? If so, what is the optimal age?



The r-squared value is: 0.00947471196943705



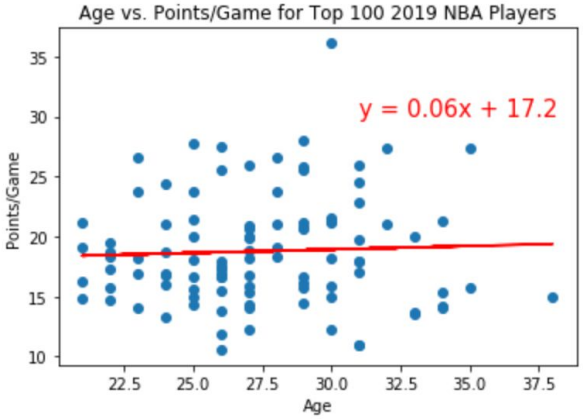
The r-squared value is: 0.0044050091120690785



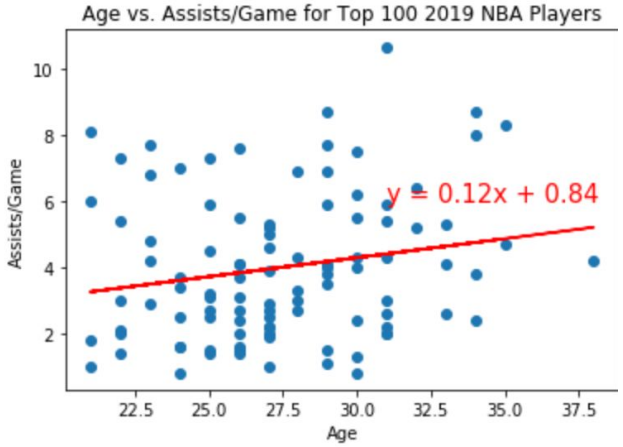
Most NBA players are between 22-31 years old. There is no definitive correlation between player age and ability* for both all players and top 100 players, as the r-squared values of both plots are very close to 0. Therefore, an optimal age cannot be determined. However, after age 38, no players have high ability scores (above 12).

*ability was calculated by summing each player's points, assists, rebounds, blocks, and steals

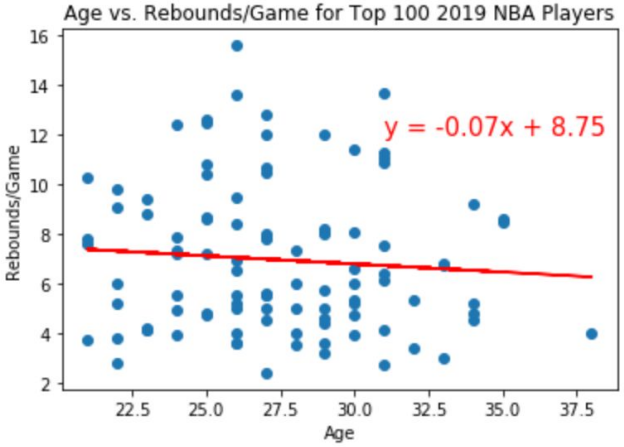
The r-squared value is: 0.002084010247404446



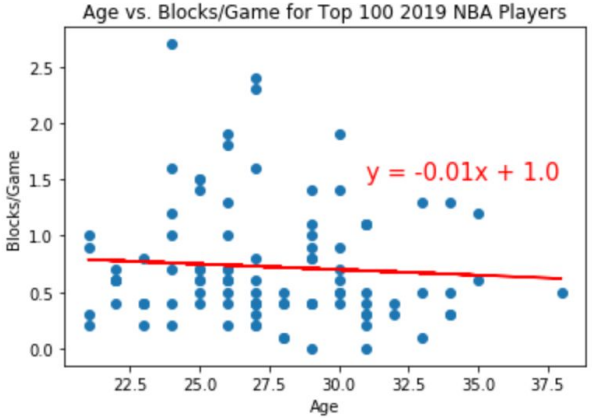
The r-squared value is: 0.0367480676393909



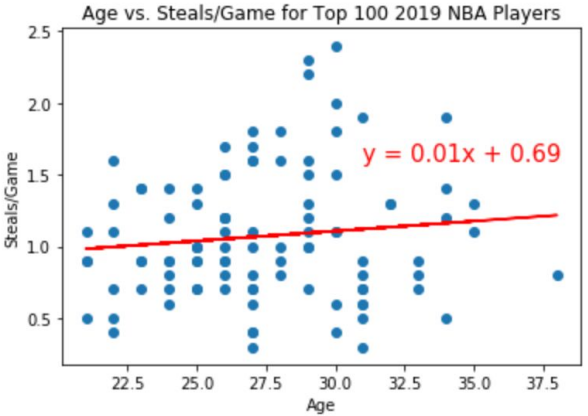
The r-squared value is: 0.006540752135741553



The r-squared value is: 0.004701746876120729



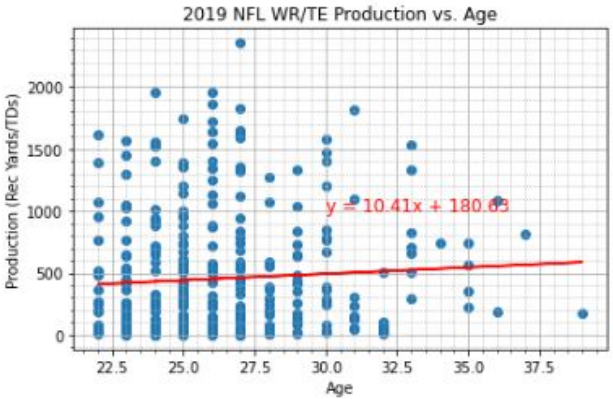
The r-squared value is: 0.012427968214639856



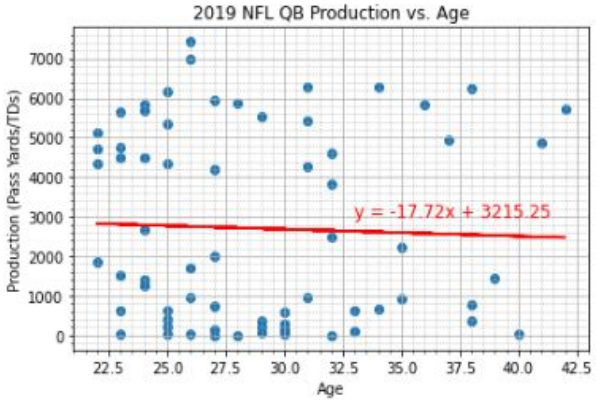
Conclusions

Is player age correlated to ability in the NFL? If so, what is the optimal age?

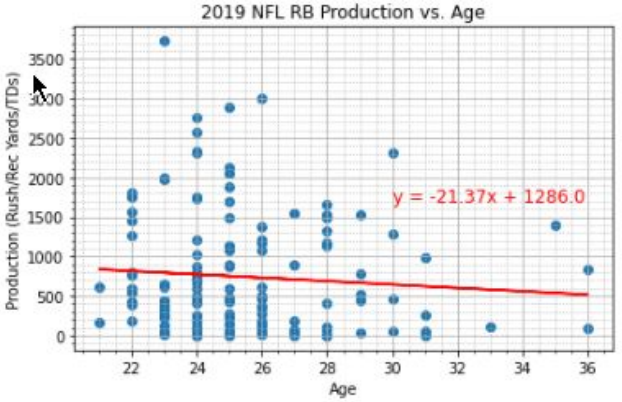
The r-squared is: 0.0040417909231381865
The correlation is: 0.06

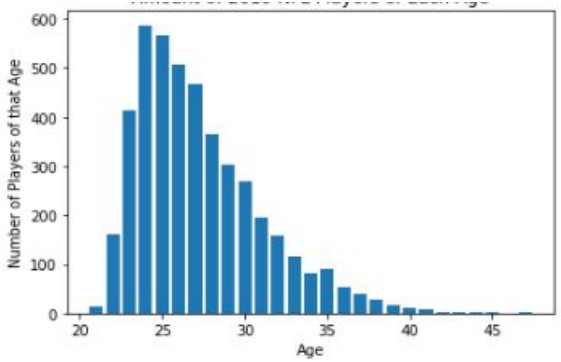


The r-squared is: 0.0014539331339513337
The correlation is: -0.04



The r-squared is: 0.006198669022436674
The correlation is: -0.08

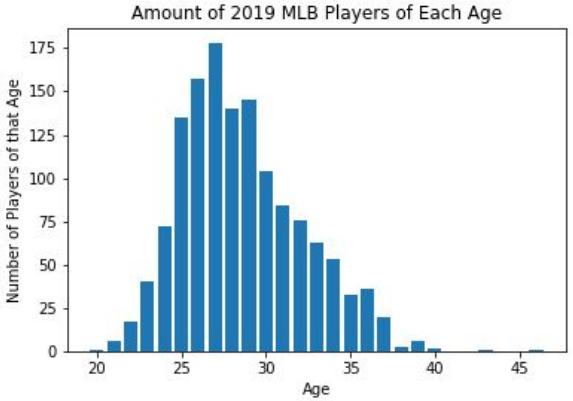




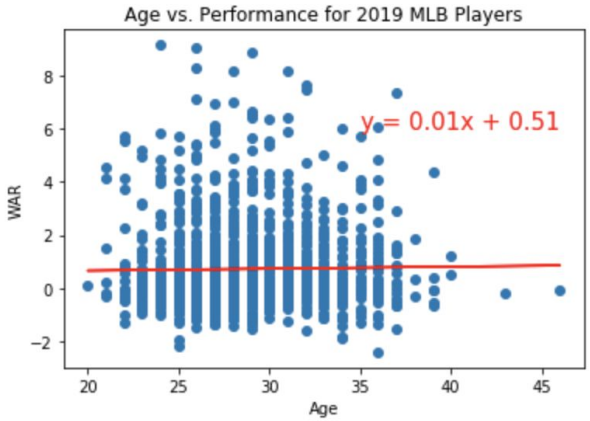
When checking player age across positions in the NFL, there are only slight correlations to production and age. In looking at the data for QBs, the scatter plot shows that there are a number of players who have performed near the top of the league in their late 30's and even into their 40's. This means that QBs can be more productive as they age whereas RB, WR and TE are generally more productive when they are younger. For RBs 23-26 seems to be the most productive age and for WR/TE 24-27.

Conclusions

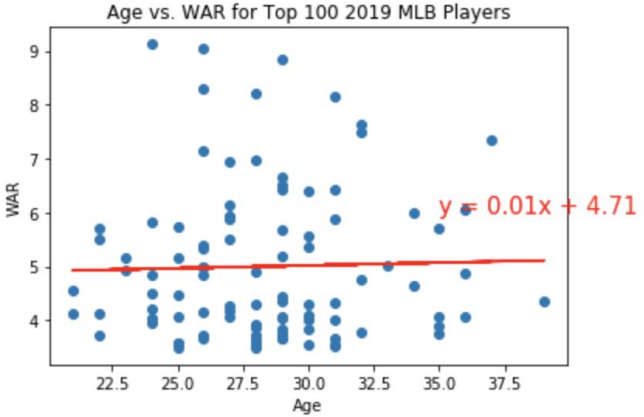
Is player age correlated to ability in MLB? If so, what is the optimal age?



The r-squared value is: 0.0003013279623995339



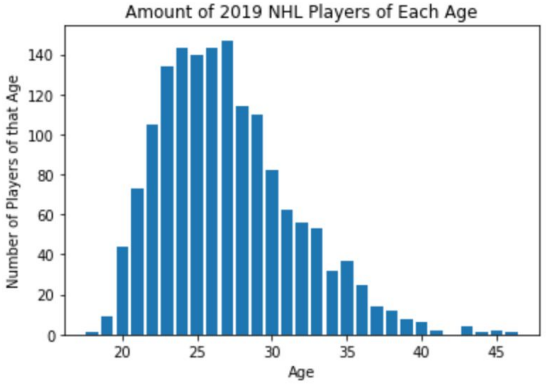
The r-squared value is: 0.0007703148384587155



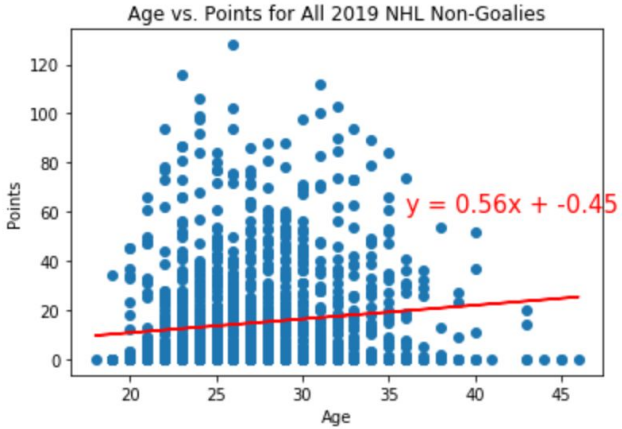
While analyzing a single season (2019) for the MLB, we were able to determine that age had no significant correlation with player performance for a single season. To determine the optimal age for performance in MLB, each player's career would have to be analyzed over time in order to determine the natural age for peak performance in baseball. Based on the median and mode of ages for 2019 MLB players, the common age of MLB players is in the 25-29 age range.

Conclusions

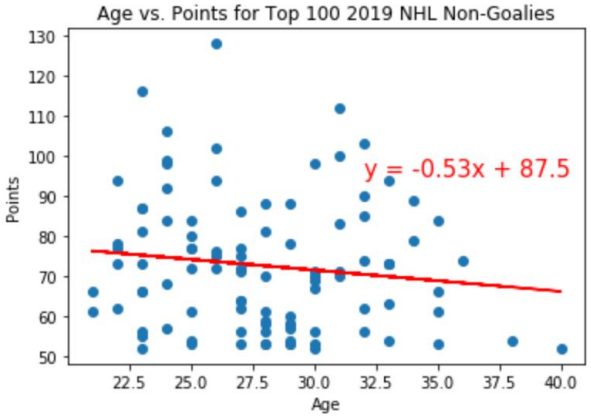
Is player age correlated to ability in the NHL? If so, what is the optimal age?



The r-squared value is: 0.013872639467567937



The r-squared value is: 0.017453778723969086

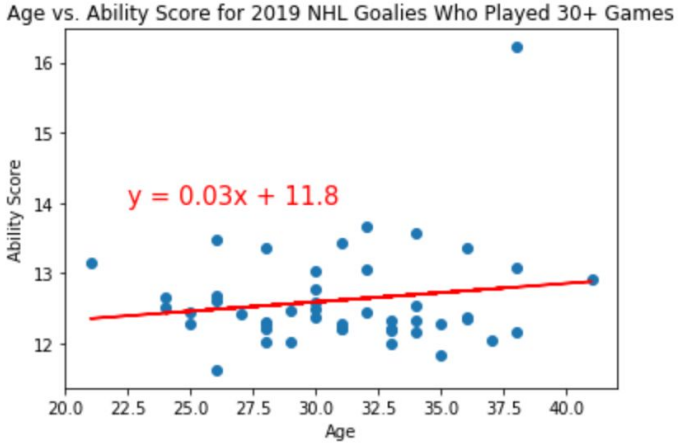


Most NHL players are between 22-29 years old. There is no definitive correlation between player age and points* for both non-goalies and top 100 non-goalies, and there is no correlation between player age and ability** for top goalies, as the r-squared values for all plots are very close to 0.

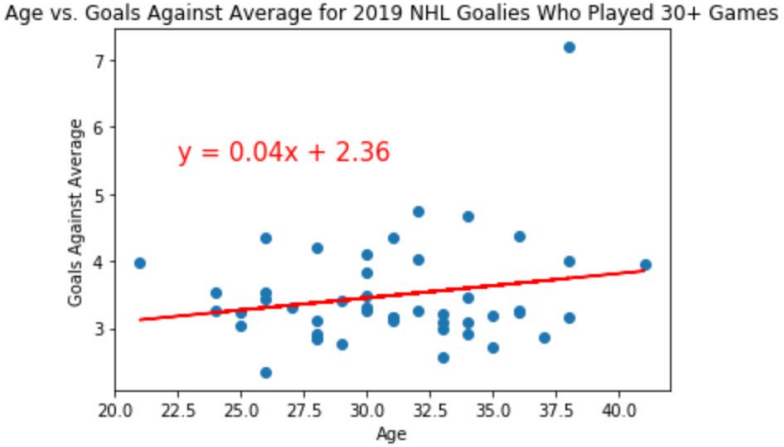
*points = goals + assists

**ability = goals against average + (save percentage * 10)

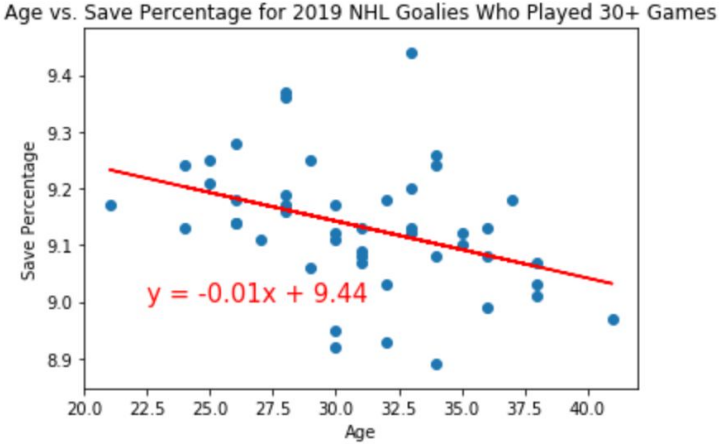
The r-squared value is: 0.025813952987953304



The r-squared value is: 0.041876060908195234

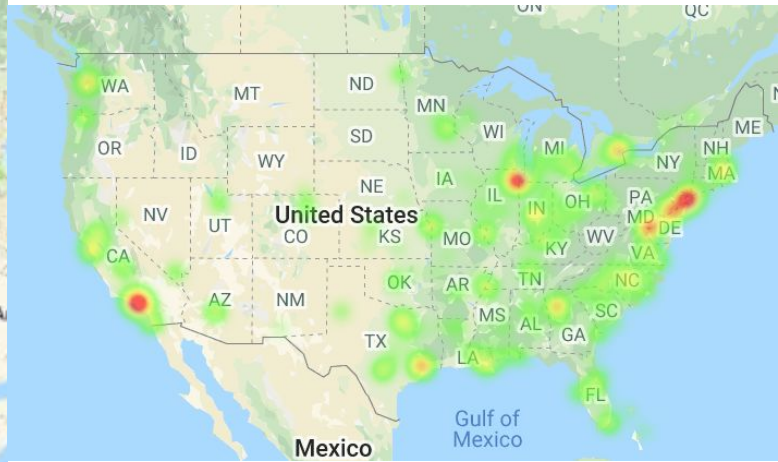


The r-squared value is: 0.14879047634115197



Conclusions

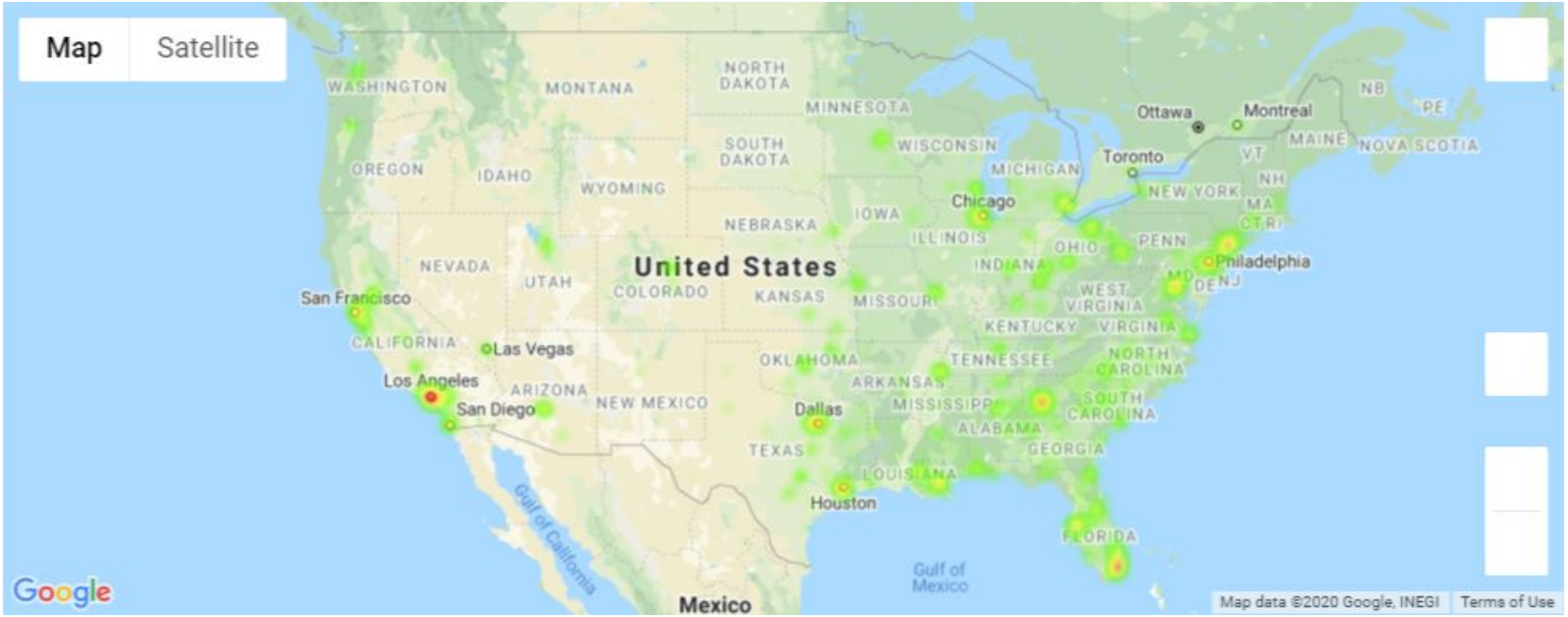
Are there specific towns/areas that players in the NBA are more likely to come from? If so, where?



The areas that NBA players are the most likely to be born in are southern California, NY/NJ, Chicago, and southeastern Texas.

Conclusions

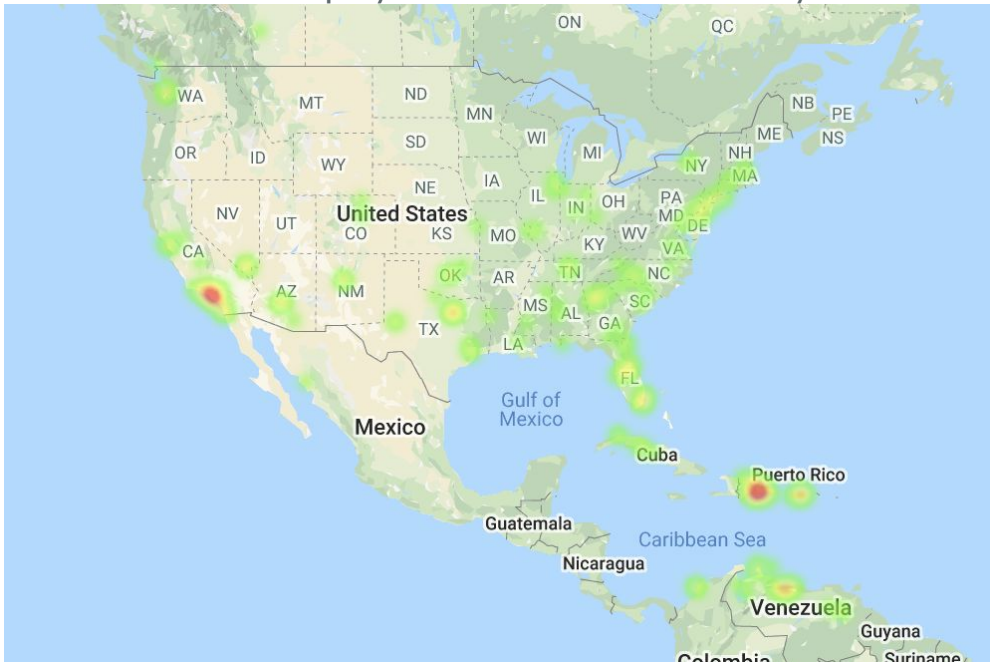
Are there specific towns/areas that players in the NFL are more likely to come from? If so, where?



There are definite hot spots in the Northeast corridor of the US, as well as Southern California, Florida, Georgia and the Eastern parts of Texas.

Conclusions

Are there specific towns/areas that players in MLB are more likely to come from? If so, where?



Most MLB players are from Southern California, the Dominican Republic, Venezuela, and Puerto Rico.

Conclusions

Are there specific towns/areas that players in the NHL are more likely to come from? If so, where?

The areas that NHL players are the most likely to be born in are Toronto, Montreal, New England, Minnesota, British Columbia, Alberta, and Sweden. Mostly though, they are from Toronto.

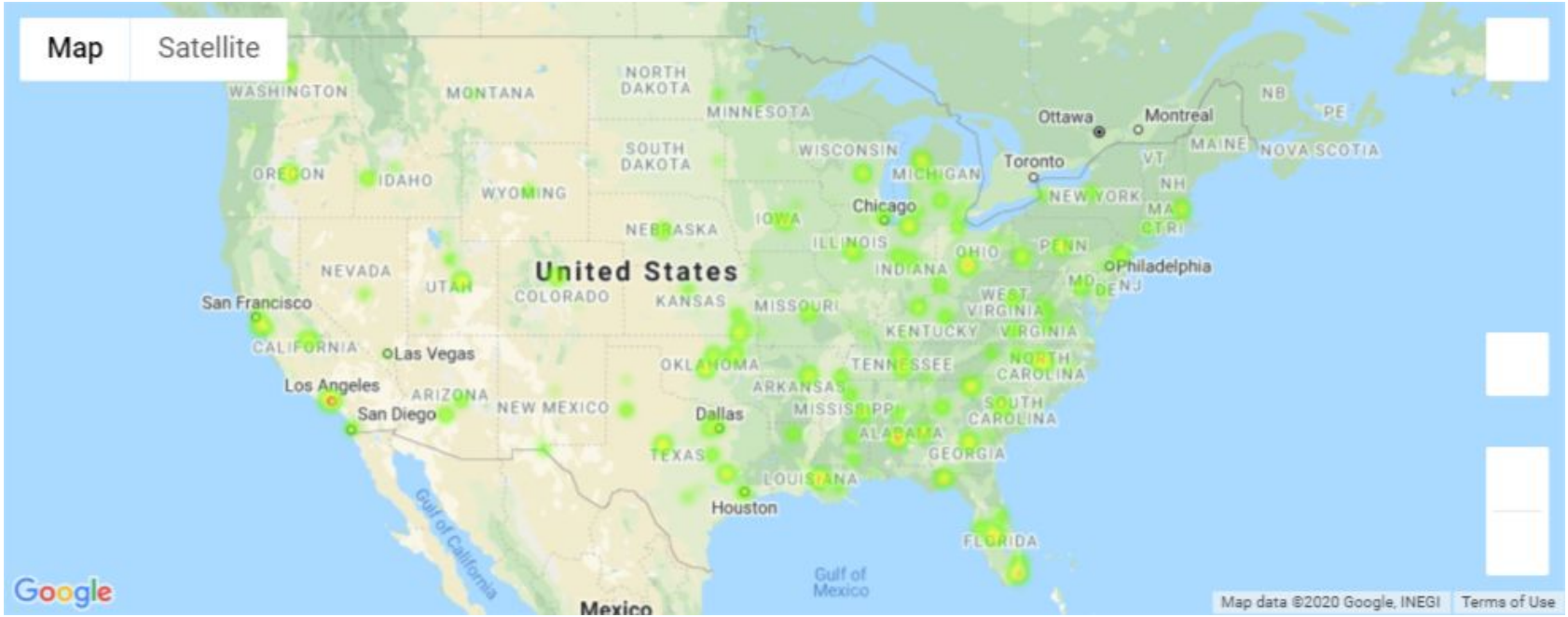
Conclusions

Are there specific colleges that players in the NBA are more likely to attend? Is so, which colleges?

NBA players are most likely to attend college in southern California, Kentucky, and North Carolina.

Conclusions

Are there specific colleges that players in the NFL are more likely to attend? Is so, which colleges?



NFL Players are most likely to attend college in Southern California, Alabama, Georgia, Louisiana, Florida. There are also small pockets in NC, Oregon, PA and Ohio.

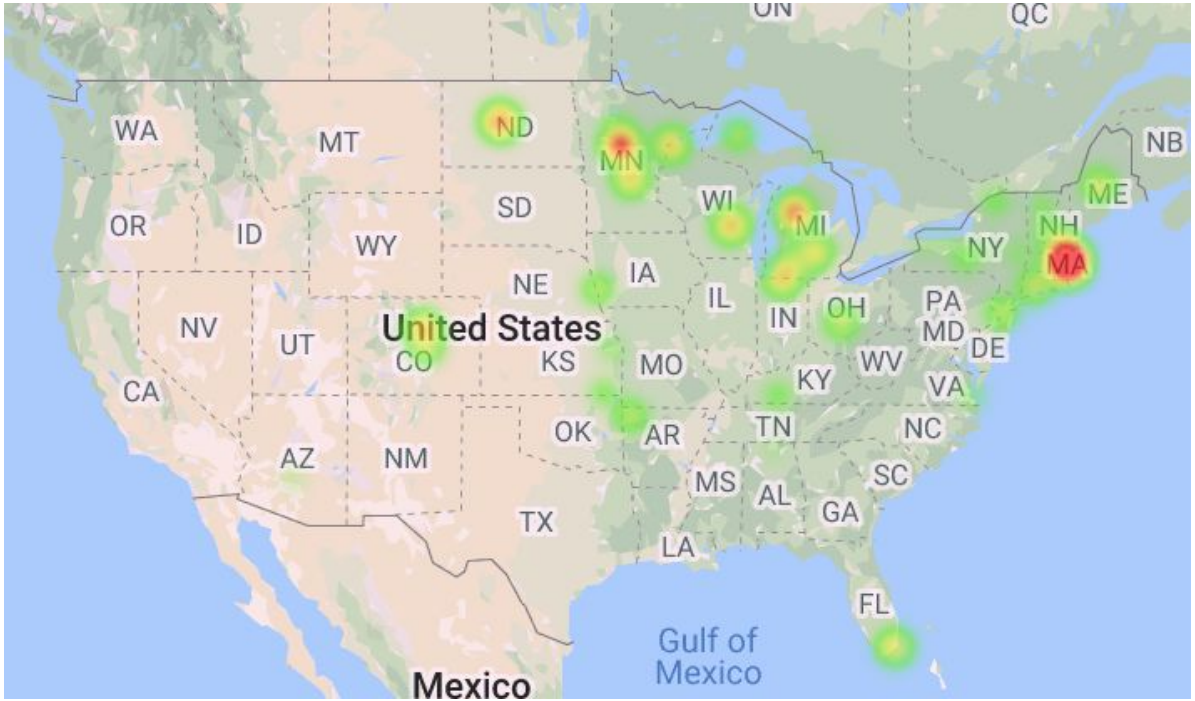
Conclusions

Are there specific colleges that players in MLB are more likely to attend? Is so, which colleges?

Most of the MLB players that attended college went to Cal State Fullerton, Vanderbilt, LSU, and Mississippi State

Conclusions

Are there specific colleges that players in the NHL are more likely to attend? Is so, which colleges?



NHL players are most likely to attend college in southern Massachusetts, Minnesota, North Dakota, and Michigan.

Implications - Age/Height vs. Ability

- No definitive correlations between player age and ability for the NBA, NFL, MLB, or NHL. Age should not be a factor used to predict:
  - how well a player will perform
  - what their contracts should be like
  - if they should be resigned or not.

- Stronger correlations between player height and points/blocks- can be used by coaches to assign positions and roles for players.

Implications - Birth Locations

The areas that most professional athletes of each sport come from tells us that those areas most likely have better programs and organized youth sports for that particular sport. Also, in those hot spots, that particular sport is more part of the culture, so kids are more encouraged, exposed, and/or drawn to the sport.

This information could be helpful for college scouts looking to recruit new players by showing where top players are most likely to be found.

Implications - Colleges

The areas that most professional athletes of each sport attend college tells us that those colleges most likely have a well-developed program for that sport, a history of winning teams, and the top coaches.

This information could be helpful for scouts from the professional leagues looking to recruit new players by showing which colleges top players are most likely to be found.

Difficulties & Additional Questions

- API did not include colleges NHL players attended
  - Used data from https://www.stateofhockey.com/news_article/show/1010827 instead
- In most sports the games have become very specialized so it can be very difficult to compare players across different positions
- If we had more time:
  - Create metrics to better evaluate players across specific positions in each sport.
  - Analyze individual players' stats over the course of their career to see what age they peaked at.

Any Questions

?