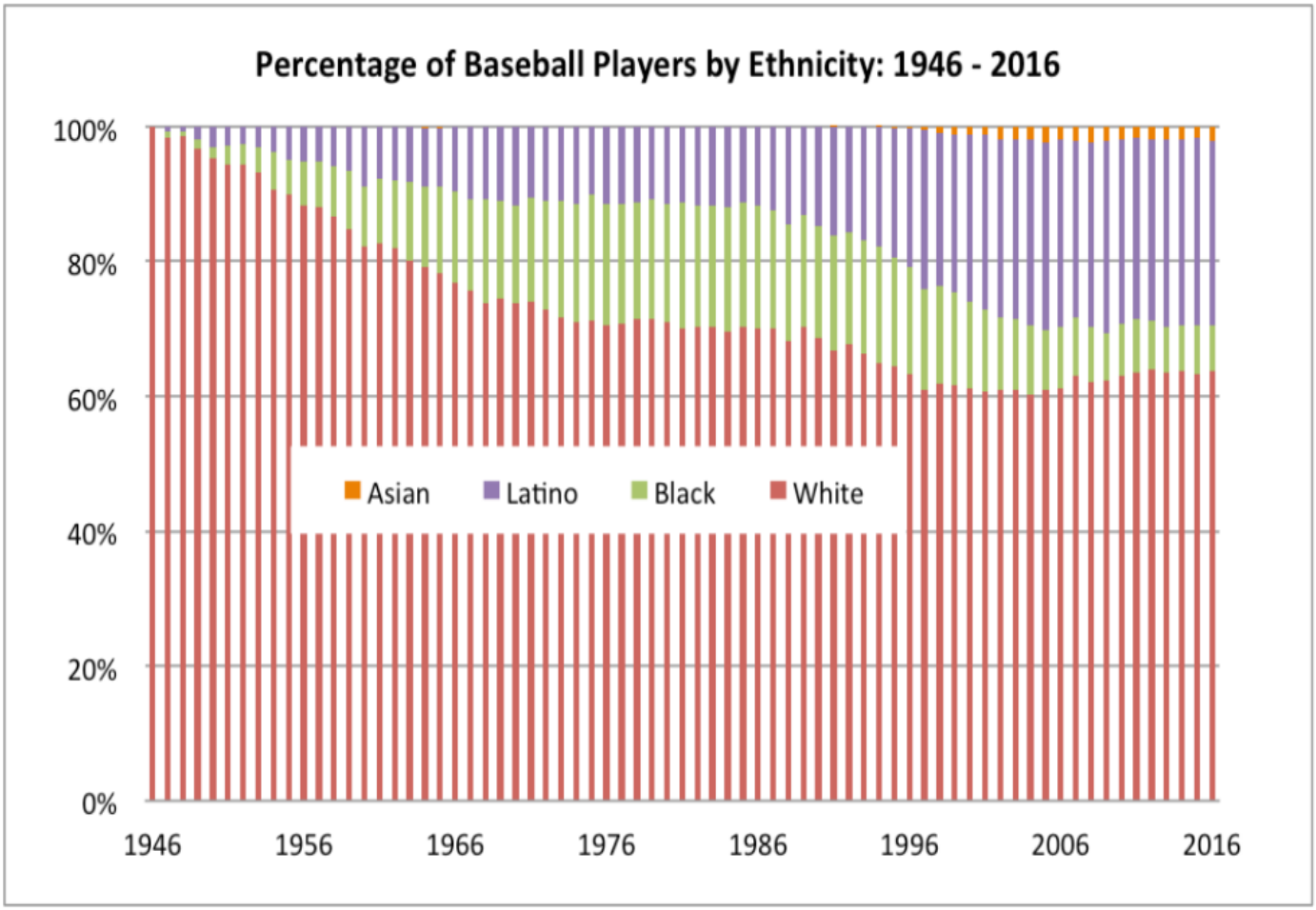
MIDTERM REDESIGN PROJECT

Group 8: Sanika Suhas Dalvi, Snehita Moturu, Thanmayee Akkineni

Dataset: A study to measure the variation in player participation in the years following Jackie Robinson's debut.

We used the dataset published by Mark Armour and Daniel R.Levitt for redesigning this project. There are two datasets and 6 graphs available. However, we decided to go with the first dataset and its subsequent graph for redesigning. This dataset provides us information about the rise and decline of four ethnicity groups, namely Asian, Latino, African American and white, over the years from 1947-2016. However, Mark's study is more inclined toward analyzing the trend of African American ethnicity groups. He wanted to answer the question of how differently other ethnic groups were influenced after the debut of Jackie Robinson.

Initial visualization:



Above is the visualization of the graph that we want to redesign. As we can see, the graph is not very expressive in terms of percentages. The x-axis representing years is clear. However, for y-axis that represents percentages is jumbled up and it is very difficult to predict at the first look of it about which group represents how much percentage. Since most of the occupancy is made by white ethnicity group, the other ethnicity count is not clear to read. Thus, this makes it difficult to analyze and therefore make any future predictions. Also, the visualization is not interactive or comprehendible. The white lines in between of the bars in the graph cause confusion to the eyes. Due to all above reasons, we consider it as a bad graph and want to redesign it to make it more clean, comprehendible and appealing.

Mistakes in the visualization

As we can see from the visualization, the x-axis starts from the year 1946. However, the dataset that was considered by the author is from 1947. In the visualization plotting has been done for 1946 as well. This causes confusion and maybe it has been plotted incorrectly. We have overcome this in our redesign of the above visualization.

Explaining Data:

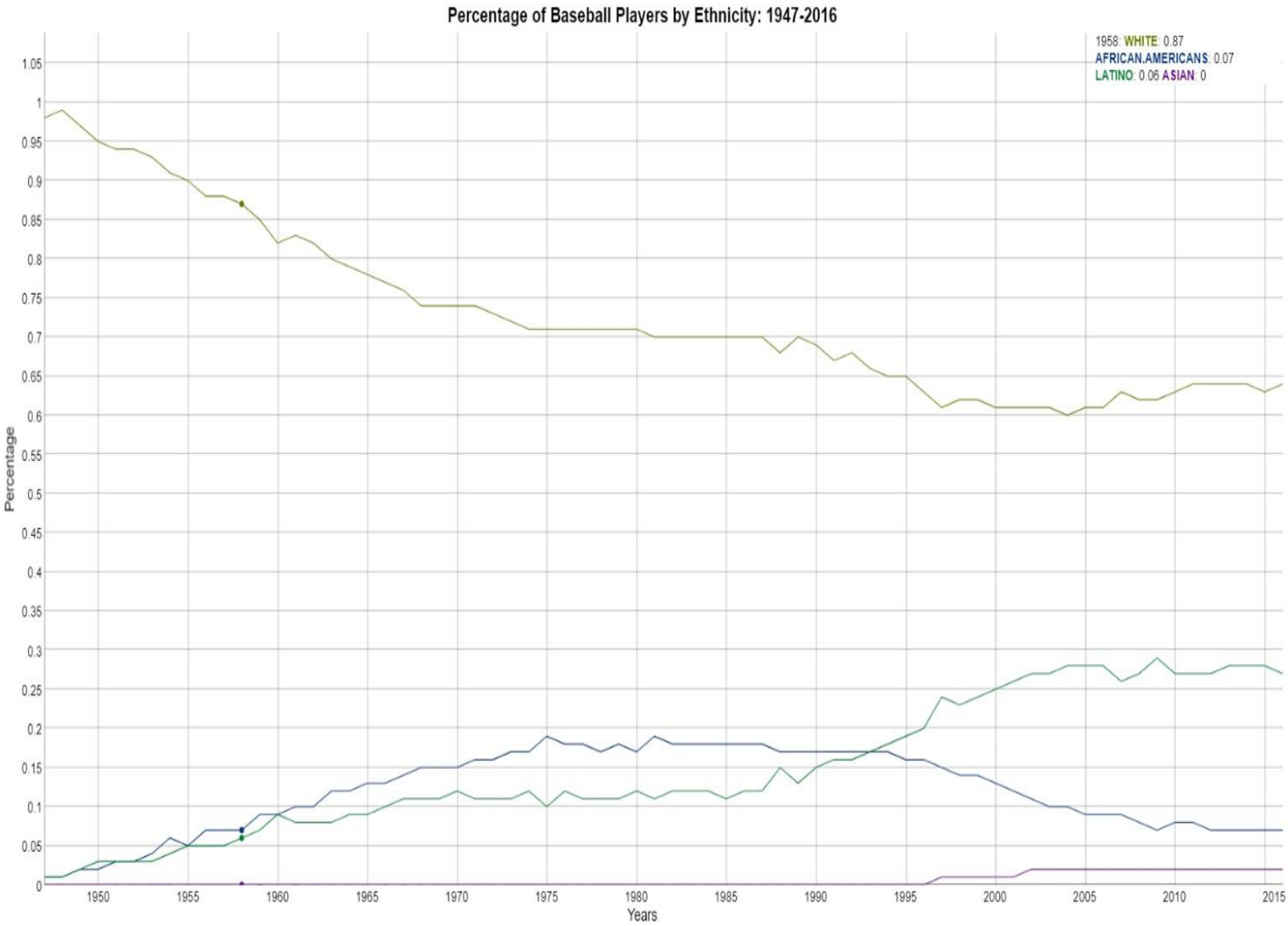
Table 1 – All Players, 1947-2016

YEAR	WHITE	AFRICAN AMERICANS	LATINO	ASIAN
1947	98.3%	0.9%	0.7%	0.0%
1948	98.5%	0.7%	0.7%	0.0%
1949	96.6%	1.5%	1.9%	0.0%
1950	95.3%	1.7%	3.0%	0.0%
1951	94.3%	2.9%	2.8%	0.0%
1952	94.4%	2.9%	2.7%	0.0%
1953	93.3%	3.7%	3.0%	0.0%
1954	90.7%	5.6%	3.7%	0.0%
1955	89.8%	5.2%	5.0%	0.0%
1956	88.2%	6.7%	5.1%	0.0%
1957	88.1%	6.7%	5.2%	0.0%
1958	86.7%	7.4%	5.9%	0.0%
1959	84.8%	8.8%	6.5%	0.0%
1960	82.3%	8.9%	8.9%	0.0%

The dataset above shows the initial dataset used by the author, although we did not make any changes to the dataset, we had to change the percentage value to decimal system for more clear representation of data. The dataset contains the year column and has each ethnicity column and the percentage of players from each ethnicity group that played each year from 1947 to 2016.

Data source: **https://sabr.org/bioproj/topic/baseball-demographics-1947-2016/**

Redesigned graph:



The visualization above is the new redesigned graph of representation of different ethnicities that started playing baseball after the breakout of Jackie Robinson who was an African American player. Before his breakout the sport was mainly dominated by the white ethnicity players, his debut encouraged African American players and other ethnicity groups to pursue the sport.

We maintained the same original thought that the author did while he did the visualization but just represented it in a new version and better way adding interactive way to view. When we view the graph, we can hover over any point and it will show us the year we hovered over and each ethnicity group's percentage of players in that year and also for better understanding we can zoom in the graph and view it for much better understanding and clarity. One more added feature of dygraph is that we are able to scroll the entire x-axis and the visualization adjusts itself dynamically.

The primary advantage of this dygraph is the abundance of additional viewing options it offers, including pan, mouseover, zoom, etc. In actuality, Google created dygraphs, a tool that is currently being used in various projects by this business.

This charting tool is perfect for use when working on complicated projects because it can analyze a lot of data without suffering any performance issues.

Redesign Results and Analysis:

As we can see from the above visualization, the African American ethnicity group percentage grew slowly from 1947 after Jackie Robinson's debut and it is steady until 1970, after which for the period between 1972-1996, there was no significant change (no significant rise/decline). However, after 1996, we are able to see decline in the African American ethnicity group. As per our analysis, we are assuming that maybe there was some other sport that might have become famous after 1996, due to which the players were drawn/inclined towards it and started moving into that particular sport and hence the decline.

Also, an additional reason we think for this decline is ethnicity domination. As we can see from the visualization, the white ethnicity group has majority players involved. It might be the case that the major/important positions in the sport were occupied by white ethnicity group due to which the African American ethnicity group started leaving the sport or was not recognized.

Apart from these, we can see the Latino ethnicity group has observed an increase in the number of players. This can become another topic of analysis as to how and what can be the reason for their growth. While for the Asian ethnicity group, we are able to observe a very minimal growth. The group started involvement in the sport in 1964. And even after that there was not noticeable participation from Asian ethnicity group.

Challenges faced during project redesign:

Our first job was to select a dataset or visualization that might be altered to provide more insightful results. Finding ones with a credible dataset, metadata, and the potential for better improvements proved to be difficult, despite the fact that we can locate bad visualizations with ease. Making certain the data we analyzed was error-free was the first significant challenge we encountered.

Choosing the right type of data visualizations to provide was one of the challenging tasks. It was difficult to produce graphics that were both insightful and simple to understand.

The dataset that the author used for visualization does not have all ethnicities combined under one column which makes it difficult to compare all of them together in the graph, it also made it difficult to choose visualizations like shiny, Plotly etc so it made a challenging task to search for R programming visualization which can take this kind of data and compare all of them correctly.

The visualization that we opted for is the dygraphs. The challenge was that we were not familiar with dygraph plotting and had to learn about it and how it is going to take the data and if it's going to represent the data exactly how we want it to. Cannot deny the fact that it was a good learning opportunity.

We used dygraphs through R Studio. The parameter in the original graph used was percentage for each ethnicity group players. While plotting with dygraph we found it difficult to use percentage in the programming, so we made it into a simple system by converting the percentage to decimal parameter.

The data visualization assignments and lectures greatly aided us in generating a variety of visualization ideas, selecting the best visualization from the data, and gaining understanding of the R packages and tools needed to put the ideas into practice.

Conclusion:

We redesigned one bad graph for this midterm project that represented percentage of baseball players by ethnicity between years 1967-2016. Due to the structure of the dataset, it was difficult to implement it using R Shiny or Plotly. However, we came across dygraphs that function similar to plotly and are interactive. This gave us an opportunity to study dygraphs, a concept which definitely helped us grow our knowledge base and this added a new skill to our bucket. This midterm project has helped us to interpret more clearly the difference between good and bad visualizations. It has also helped us understand new concepts in detail. We will dive in more into these concepts in our future projects and personal projects as well.

References:

1. Mark Armour and Daniel R.Levitt, “Baseball Demographics, 1947-2016”. [Online]. Available: https://sabr.org/bioproj/topic/baseball-demographics-1947-2016/. [Accessed October 5,2022].
2. “dygraphs for R”. [Online]. Available: https://rstudio.github.io/dygraphs/index.html. [Accessed October 5,2022].
3. “Interactive time series with dygraphs”. [Online]. Available: https://www.rstudio.com/blog/interactive-time-series-with-dygraphs/. [Accessed October 5,2022].