Problem 1

The Meaning of Life

In 2021, Pew Research Center conducted a survey of 2,596 adults in the United States (report here). They asked "What aspects of your life do you currently find meaningful, fulfilling, or satisfying?" They did not report raw data, but below is a generated dataset based on reported results.

Topic	Frequency
Family	716
Friends	300
Material Well-being	265
Career	262
Challenges	285
Spirituality	250
Society	210
Health	172
Hobbies	136

(a) Compute the Relative Frequencies for each response category.

(b) Construct a bar graph of the Relative Frequencies.

(c) Interpret the data in a paragraph (2 or more sentences).

The relative frequency, $\hat{p}$, is given by

$$\hat{p} = \frac{f}{n} \tag{1}$$

where f is the frequency of the corresponding class (or topic in this study), and n is number of total observations. Using Equation (1) the relative frequencies for each topic were calculated and are displayed in Table 1, along with the frequency of each response.

Topic	Frequency	Relative Frequency
Family	716	0.276
Friends	300	0.116
Material Well-being	265	0.102
Career	262	0.101
Challenges	285	0.110
Spirituality	250	0.096
Society	210	0.081
Health	172	0.066
Hobbies	136	0.052

Table 1: Topics US adults consider meaningful, fulfilling, or satisfying.

For a visual representation of the relative frequencies see the bar graph presented in Figure 1. The bar graph makes it apparent that the relative frequency of 'Family'is over twice as high as every other topic. The next five highest topics have a similar relative frequency around 0.10.

Interpreting the data can be tricky without fully understanding how the data was collected. For example, were these responses selected from a list provided, or where they what each individual person thought of? The total number of responses equals the number of participants, so it is reasonable to assume only one response per person. However, was the response the first thing the person thought of after the question, or were they given time to contemplate to provide what they thought was the most meaningful, fulfilling, or satisfying? Assuming that the responses are what the individuals considered

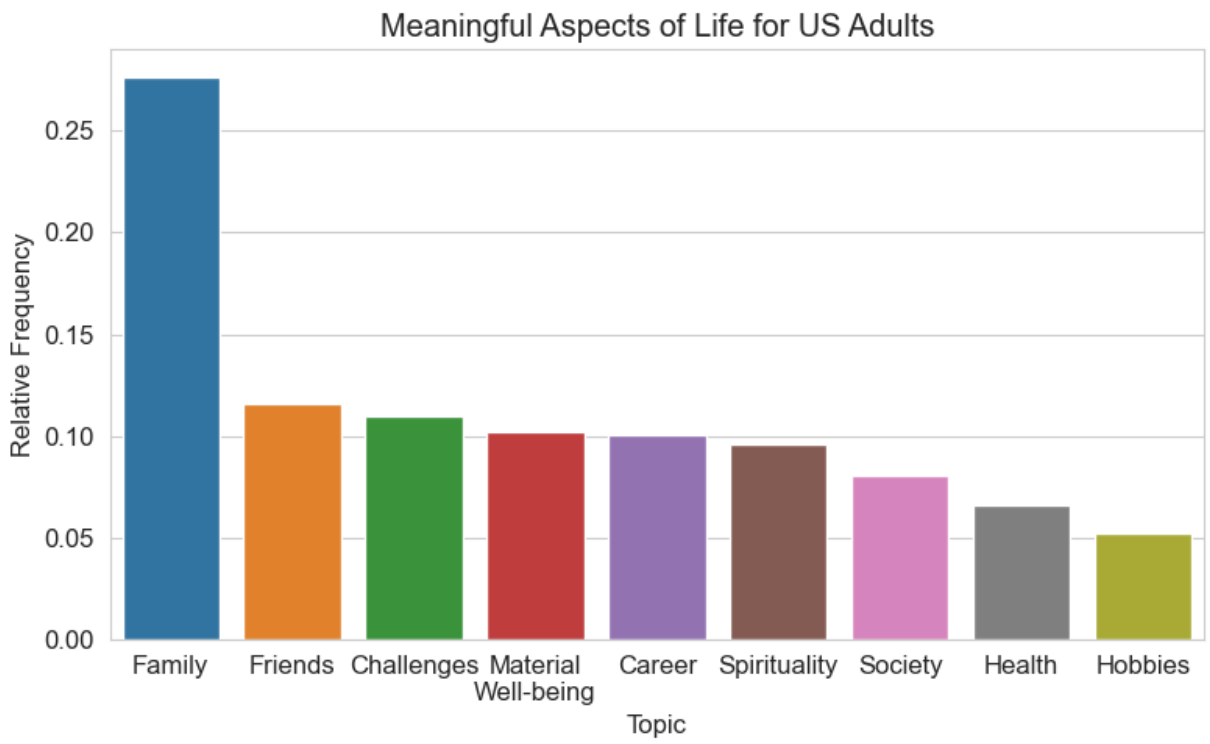

Figure 1: The relative frequencies of responses US adults gave to the question, "What aspects of your life do you currently find meaningful, fulfilling, or satisfying?"

the most important, and assuming the sample accurately represents the US adult population, then the data indicates that US adults most frequently find meaning in life from family. Considering both the family and friends topics together make up over one-third of the relative frequency, it shows that US adults find meaning in life most frequency from people close to them.

Problem 2

Board Game Weights

The file `bgg.csv` on Canvas contains a database of every board game on the popular site "Board Game Geek." These data can be used to answer the following questions.

(a) The column `averageweight` gives the average user assessment of the "weight" (i.e., complexity) of each game. Games with a value of 0 have not been rated and should not be included in any of the following analysis.
Create a table of summary statistics for the average weight of games. This table should include the Minimum, 1st Quartile, Median, Mean, 3rd Quartile, Maximum, Sample Variance, and Sample Standard Deviation (note the use of the word 'Sample' even though this is arguably a census).

(b) Create Box Plots for the average weight of games by whether it is ranked as a Family Game or not. If `Family Game Rank` is blank (coded as `NaN`), then it is not ranked as a Family Game. I recommend adding a new column using the function `np.isnan`.
Are family games more or less complex? What can you say about the relative complexity of family games and non-family games?

(c) Create a scatterplot of weight to average rating (`average`). What can you say about this relationship?

The summary statistics for the average weight of the games on "Board Game Geek" are listed in Table 2. The weight is an indication of the complexity of the game, as assessed by the users. It appears that the range scale for the weight is from one to five based on the minimum and maximum values found in the data. Half of the games have an average weight equal to or below 2.0, and three-quarters of the games are below 2.5. This indicates there are more games with lower complexity.

Statistic	Average Weight
Minimum	1.000
1st Quartile	1.341
Median	2.000
Mean	2.038
3rd Quartile	2.571
Maximum	5.000
Variance	0.650
Standard Deviation	0.806

Table 2: Summary statistics for the average weight (or user designated complexity) of games on "Board Game Geek."

Games are also given a 'Family Game Rank' if they are considered a family game. Using this, the family games and non-family games can be grouped together and compared against each other. Figure 2 shows such a comparison for the average weight of games grouped by whether or not they are considered a family game. The box plots show that the minimum for both is the same, but the maximum complexity for the family games is around 3, while it is 5 for non-family games. It also shows that on average family games are typically less complex than non-family games.

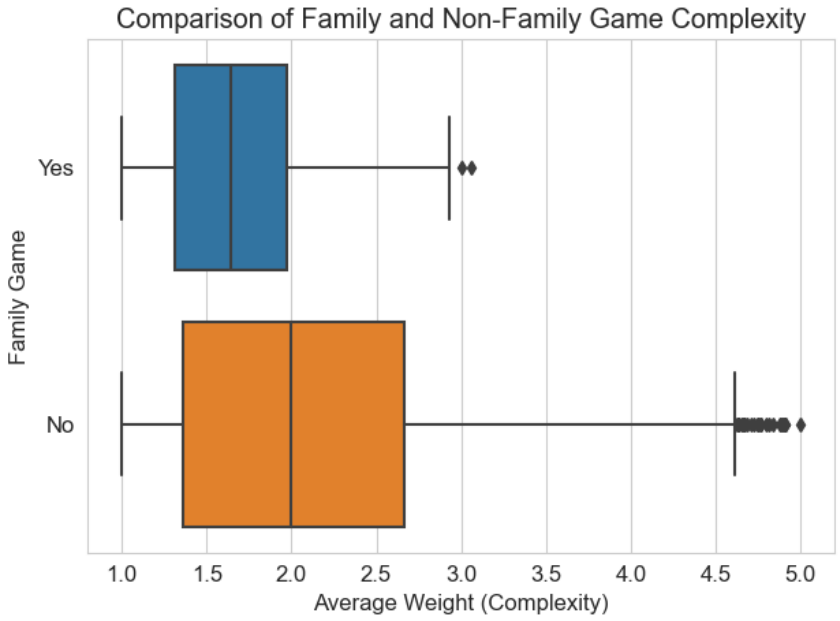

Figure 2: The complexity of family games compared to non-family games.

There is an interesting relationship between the average weight of the games and the average rating. Figure 3 shows this relationship in a scatter-plot. The plot shows a positive correlation between the two. As the average weight increases the average rating is grouped together more at the higher ratings. Notice that there are relatively few games with a complexity higher than four that have a rating lower than six.

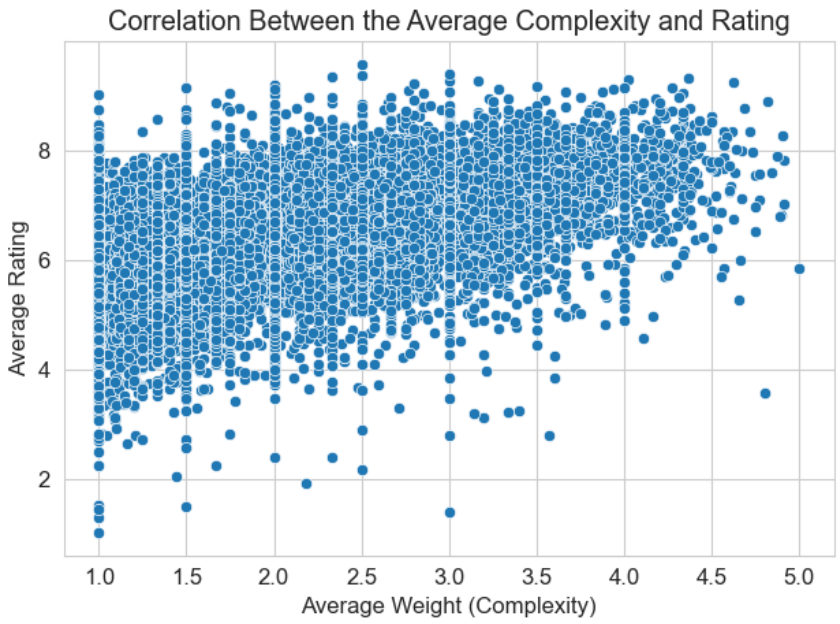

Figure 3: The correlation between the average weight and the average rating.

Problem 3

Baseball Hall-of-Famers

Baseball Hall-of-Famers (HoF) played during different eras of baseball. One common classification of eras is '19th Century' (up to the 1900 season), 'Dead Ball' (1901–1919), 'Lively Ball' (1920–1941), 'Integration' (1942–1960), 'Expansion' (1961–1976), 'Free Agency' (1977–1993), and 'Long Ball' (after 1993). For this exercise, define the era of a player based on the mid-point of their career (rounding up if necessary).

Using the file `hofbatting.csv`, containing non-pitching HoFs as of 2013, classify each player according to their era to answer the following questions.

(a) Create a Bar Graph for the number of HoFs from each era as of 2013. Interpret the data. You will need to take multiple steps to solve this problem.

- Import the CSV.
- Create a column of data to define the mid-career.
- Find a way to count the number of HoFs in each era by the mid-career column.
- Create the graphs.
- Provide a written interpretation of the data.

(b) Create a histogram showing the distribution of non-pitching HoFs' Mid-Career year.

(c) There are two major dimensions to hitting: the ability to get on base (measured by the on-base percentage `OBP`) and the ability to advance runners already on base (measured by the slugging percentage `SLG`). Create a scatterplot of OBP vs. SLG. Are there any outliers? If so, identify them by name. Is there a relationship between OBP and SLG?

(d) Consider a combined metric for hitting, the On-base Plus Slugging (OPS) statistic, which is the sum of OBP and SLG. Normalize this data (i.e., calculate the z-scores), then create a scatterplot with OPS on the y-axis and Mid-Career Year on the x-axis. Identify any outliers by name. Do you notice any patterns in the scatterplot of the data? What can you say (if anything) about the cause of any pattern?

(e) Create a Box Plot for the Home-Run Rate (HRR), defined as home-runs per at-bat (HR/AB), of HoFs during each era (i.e., you should have 7 box-plots). Also calculate descriptive statistics of HRR including Min, Q1, Median, Q3, Max, Mean, Range, and Sample Standard Deviation for each era. Provide a table of these values from the Expansion era only (to limit time spent copying and pasting from Python).

The number of non-pitching Hall-of-Famers (HOF) during each baseball era is shown in Figure 4. The order of the eras is kept in chronological order to also show the change over time. The 'Lively Ball' era has nearly twice as many non-pitching HOFs than any other era. At 21 years, it is the longest era, other than the 19th Century era. The other eras are between 15-18 years, which would still make the 'Lively Ball' era the highest even if the numbers were adjusted to be considered on a per year basis. The 'Long-Ball' era is significantly lower than all the other eras. A player is only eligible for the HOF after a certain number of years after they finish playing. Since the data was published in 2013, the actual length (adjusting for when players are eligible) of the 'Long Ball' era is likely much shorter than the other eras.

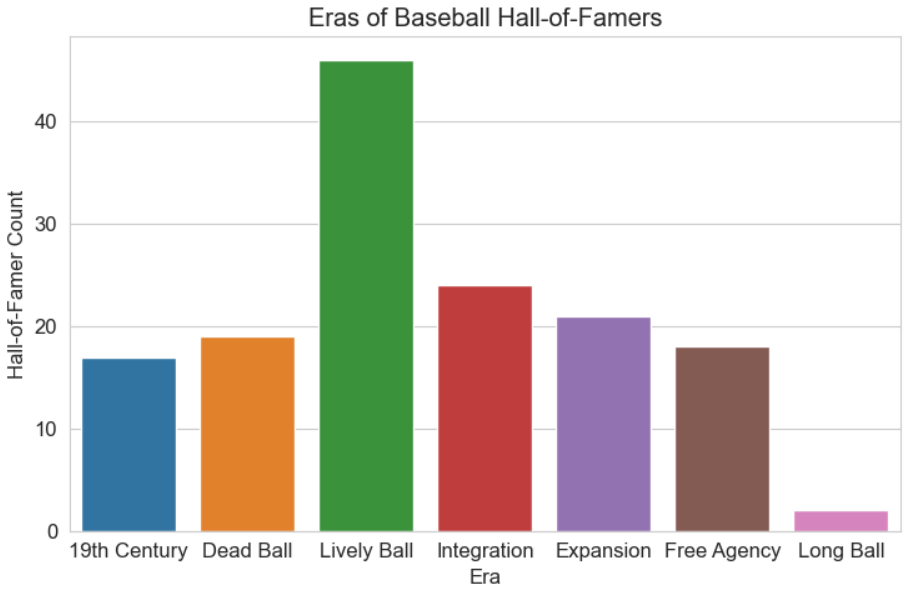

Figure 4: The number of baseball Hall-of-Famers during each era.

The players mid-career year was calculated based on the 'From' and 'To' fields in the data. While it is rare for professionals to have breaks in playing, if breaks did occur the mid-career year did not account for them. The distribution of the HOF's mid-career year is shown in the histogram in Figure 5. As expected, there is a significant increase during the years of the 'Live Ball' era. The histogram is really just breaking down the bar graph in Figure 4 into smaller bins. The bins in Figure 5 don't match exactly to the years of each era because they are all equal in size. This plot is a better representation of differences in the number of HOFs over time.

There are two major dimensions to hitting: the ability to get on base, measured by the on-base percentage (OBP), and the ability to advance runners already on base, measured by the slugging percentage (SLG). Figure 6 shows the positive correlation between these two hitting statistics. Players with higher OBP also tend to have higher SLG. There is one significant outlier that is also shown in the figure. That player is Willard Brown, who only played one year (1947) in Major League Baseball (MLB). A little extra research shows that while his stats in MLB where below average for HOFs, he played at a high level in the Negro League for several years.

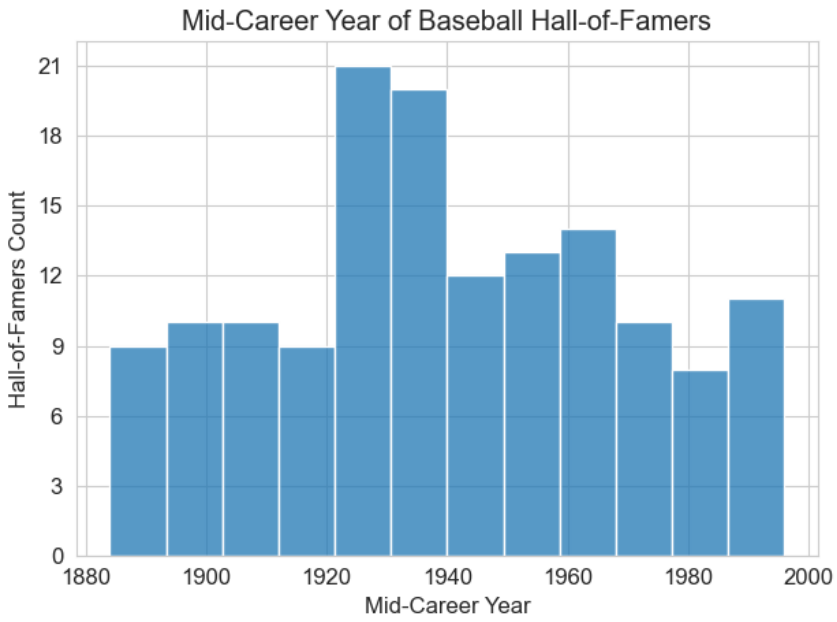

Figure 5: The distribution of the number of Hall-of-Famers based on the midpoint of their career.

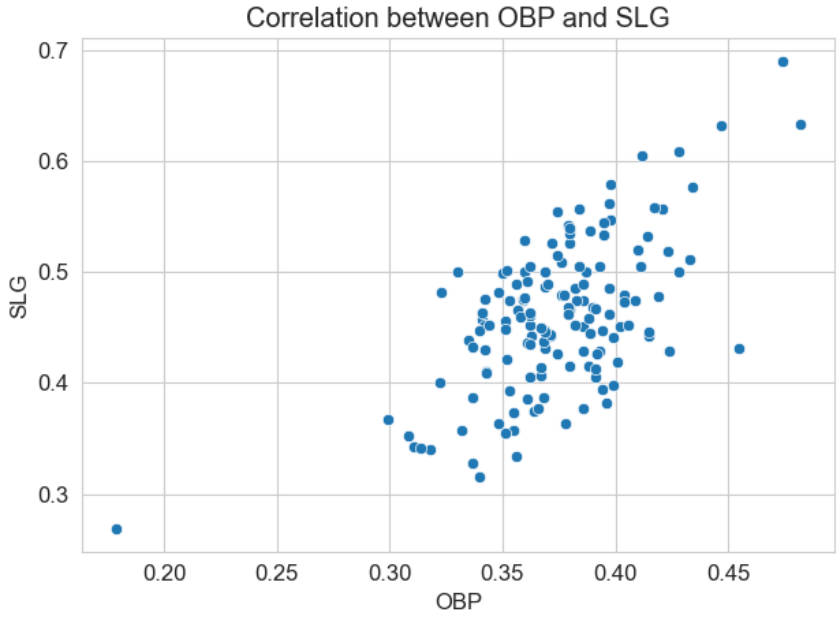

Figure 6: The positive correlation between the on-base percentage (OBP) and slugging percentage (SLG).

Another common hitting metric in baseball is the on-base plus slugging (OPS) statistic. As the name implies it is the sum of the on-base percentage and the slugging percentage. The OPS data was normalized to have a mean of zero and a standard deviation of one. The normalized OPS was plotted against the mid-career year as shown in Figure 7. The plot shows no correlation between the two. There is no noticeable increase or decrease in OPS throughout the years.

There are several outliers in the OPS data. The data point with the lowest normalized OPS in Figure 7

is Willard Brown. This is no surprise because he had the lowest OBP and SLG as shown in Figure 6. There are also two data points with a normalized OPS greater than three. They are Babe Ruth and Ted Williams. Since they are outliers on the high side there is no surprise that they are HOFs.

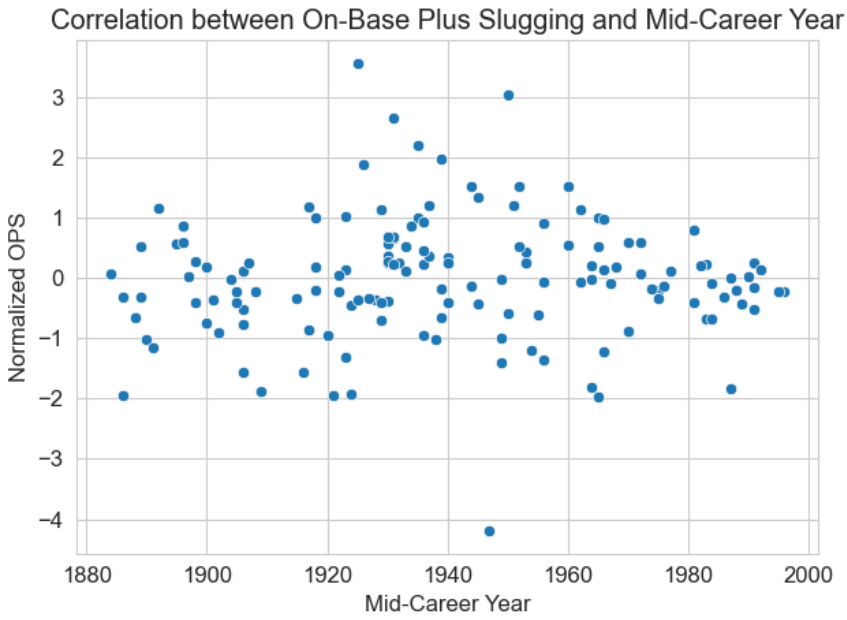

Figure 7: The OPS and the Mid-Career year do not have a correlation.

Yet another combined metric in baseball is the Home-Run Rate (HRR). The HRR is the number of home-runs per at-bat. Figure 8 shows the HRR statistics during each era. The median HRR increased significantly between the 'Dead Ball', 'Lively Ball', and 'Integration' eras. The highest median HRR was 0.041, which occurred during the 'Expansion' era. The rest of the summary statistic for that era are shown in Table 3. Note that there is a single outlier in the HRR data that is significantly higher than the rest. That HOF is Babe Ruth with a HRR of 0.085.

Statistic	HRR
Min	0.008
Q1	0.025
Median	0.041
Q3	0.058
Max	0.070
Mean	0.040
Range	0.062
Sample Standard Deviation	0.018

Table 3: Summary statistics for the home-run rate of Hall-of-Famers during the 'Expansion' era.

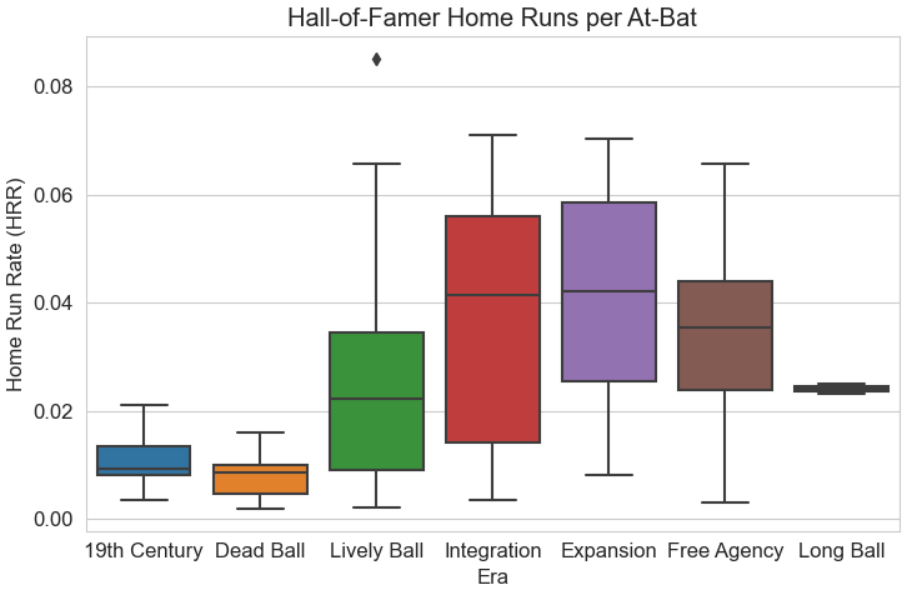

Figure 8: The Home-Run Rate (HRR) for Hall-of-Famers during each era.