BASEBALL DATA ANALYSIS

Tableau initial: https://public.tableau.com/profile/kartik.garg7913#!/vizhome/Kartikinitialproject8baseball_0/Story1?publish=yes

Tableau final link: https://public.tableau.com/profile/kartik.garg7913#!/vizhome/Kartikproject8baseball/Baseball_Story

• SUMMARY

Data set contained 1,157 baseball players including their handedness (Left, Right, Both), weight (in ounces), height (in inches), batting average and homeruns. *Before doing the visualization, I first cleaned the data in python Jupyter notebook which resulted in reduction of rows to 891.* I choose to plot the performance of baseball players based on their weight, height, handedness, performance (Batting avg., Home runs) and individuality.

• Major Inferences:

Major conclusion from the data are as follows:

1. Right handed players were in majority.
2. Average homeruns of players increased with increase in height while average batting average decreased.
3. After all the visualizations it was clear that the left handed players were the best performers.
4. Majority of players had batting average between 0.20 and 0.28.
5. Players were segregated into 4 parts based on their performance and one could see what category a particular player lies in.
6. Reggie Jackson was the best performer with batting average 0.26 and home runs = 563.
7. Dean Chance was the worst performer with batting average 0.07 and 0 home runs.

• DESIGN

1. Data Cleaning :-

Explanation: Removing all the rows with batting average and homerun = 0 as they seemed to be missing values. This cleaning was done in Jupyter notebook: python. The python file used for cleaning has been added in the zip folder along with original and finally cleaned excel file. The steps have been properly defined in that .ipynb file.

NOTE: After a feedback from reviewer I came to know that there are two names which are repeated twice although they are different individuals. I have now named them as "Bobby Mitchell1" and "Bobby Mitchell2" and same has been done for other two people. This change has also been made to the baseball_clean_data file and was performed in .ipynb file.

I have also changed he names of short forms, mentioned units for required measures and changed the story name.

2. Visualization :-

a. I choose circle chart to show the count of handedness as there were only 3 categories and were easy to differentiate between overall counts. Then later on I changed this chart to pie chart because it seemed more fitting for percentage representation.

b. I then plotted circle graph to represent No. of records vs Height and Weight. The size of circle was directly proportional to the no. of records of that particular Height or Weight but upon feedback I realised I should better use density graph as size of graph and height for same measure could be misleading and confusing.

c. Thirdly, I decided to choose line graphs to represent how the average performance of players depended on their heights and weights. I also added colour measures and trend lines for more clarification. Furthermore I added width to these lines based on number of records which I added to the mark size.

d. Another useful insight that I decided to plot was how handedness affected the performance. A bar graph was used for the same. Again no. of records was used in the mark size which affected the size of the bars.

e. Then a circle graph was used to represent the relation between average home runs and no of records vs. batting average. This graph was easy to read and understand. The *No. of records* measure was added to size and colour mark to give more clarity.

f. From previous graph I also created a further insight by adding handedness measure to the filter section and colour mark. A really interesting insight was plotted. One could easily compare that for particular batting average, what were the average home runs scored and that too handedness wise. In the size mark, no of records measure was added, to give required size to all the circles.

g. I created a simple plot for individuals performance so that is can be combined with the upcoming scatter plot in the dashboard for detailed individual performance. This plot had facility to filter individuals based on the category they lie.

h. I used a scatter plot and created 4 quadrants based on average home runs scored and batting average to categorize players in 4 levels - good, consistent, inconsistent and worst. Furthermore, I used different shapes and colour measure for each quadrant.

i. I created two dashboards: one representing *Insights Based on handedness* and another representing all *detailed analysis of performance of individuals*. The graphs were made responsive and hovering, selecting one section would highlight and filter graphs on that same dashboard.

3. Other visual practices:

a. I ensured at every step that proper formatting of sheets and labels is done and suitable titles are mentioned everywhere.

b. It was carefully seen that appropriate and same colour scheme was followed for graphs having similar dimensions and measures. Ex. for handedness (can be seen in handedness dashboard).

c. Parameter, calculated fields and filters were used wherever seemed necessary. One of the most interesting one was the calculated field: Quadrant identifier.

d. Dashboards were made responsive and actions were also added like highlight on selecting and linking sheets on that same dashboard.

e. Proper comments have been given to each sheet in story, with all possible insights from the graphs on those sheets.

- **FEEDBACK**

Feedback 1: Circle graph didn't seem suitable for representation of percentage of no. of records by handedness.

Action taken: The graph was changed to pie graph as its area representation clearly depicted percentage wise no. of records by handedness.

Feedback 2: Line graph representing avg. performances vs height could be made more meaningful if no. of records could be added as a parameter and for weight graph it is better to use a bin than weights directly.

*Action taken***:** No. of records was added to colour and size marks in the line graph to give width and dense colour at points where no. of records are more. Thus providing better view to the graph. Weight graph was plotted by using bins and thus this plot become more meaningful.

Feedback 3: It would be better if the circle graph representing batting average vs home runs and no. of records can also be seen from handedness perspective.

Action Taken: In order to add handedness perspective, the circles were further divided handedness wise by dragging handedness dimension in the filter section. The completely changed the perspective of viewing it and provided many more insights. Handedness was further added to colour mark to distinguish right, left and both handed from one another.

Feedback 4: I was asked if I could distinguish between good and bad players somehow through visualisation as it would be interesting observation.

Action Taken: Through brainstorming, searching and after some help from one friend, I designed an overall performance scatter graph which segregated good, consistent, inconsistent and worst players into quadrants. Every individual was represented in graph. Different shapes and colours were used for different quadrants using colour and size marks. The graph seemed awesome and a filter was also added name wise to select particular player and see his performance.

Feedback 5: Dashboard which represented effect of handed on performance was not too productive, neither the second dashboard represented performance clearly.

Action Taken: The circle graph which was created for performance, handedness wise was added to the dashboard1 to give it more meaning. The dashboard2 which represented performance analysis, was also added with the scatter plot representing quadrants and individual performance graph too. Thus provided deep insights like individual performance vs whole etc.

Feedback 6: There are duplicate names in the dataset although they are different individuals, due to which visualisations are bit skewed.

Action Taken: There were total of 4 duplicate names, two of each kind. The names were changed to "Bobby Mitchell1" and "Bobby Mitchell2" and same has been done for other two people. This change has also been made to the baseball_clean_data file and was performed in .ipynb file using "duplicated" and "loc" functions.

Feedback 7: The short forms should be avoided like Home runs should be used instead of HR and units should be mentioned for required measures.

Action Taken: The weight was given pounds as it seemed fitting according to the values and similarly for height inches was chosen as it appeared suitable. Short forms like HR has been changed to Home Runs.

• RESOURCES

- https://stackoverflow.com
- Udacity tutorials
- https://public.tableau.com/en-us/s/gallery
- Github.com