Create a Tableau Story: Write Up

This is the final project of the Udacity Data Analysis Nanodegree. The task was to use Tableau to create an explanatory data visualisation from a dataset that communicates a clear finding or that highlights relationships or patterns in a dataset.

Summary

I chose Option 1 – the Baseball dataset. This data set contains 1,157 baseball players including their handedness (right or left handed), height (in inches), weight (in pounds), batting average, and home runs. The data set is a csv file, which I downloaded from:

https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud507/baseball_data.csv&sa=D&ust=1544457037299000

The initial visualisation is presented as a data story in Tableau with 4 pages plus a cover and conclusion slide. There is an initial exploratory section which provides context but the main part of the visualisation is explanatory and allows the user to click through analyses of player characteristics to view how these affect the performance of the players. The initial visualisation can be found here:

https://public.tableau.com/profile/paula.jasper#!/vizhome/Baseballproject-initial/Story1?publish=yes

And the final version following the implementation of feedback contains 5 pages plus a cover and conclusion page and can be found here:

https://public.tableau.com/profile/paula.jasper#!/vizhome/Baseballproject-finalversion/Story1?publish=yes

Design

Initial Design

Firstly I carried out some online research on Baseball as I am from the UK where it is not a national sport so I didn’t have much knowledge about it. After I had gained more understanding of what Batting Averages and Home Runs are, and also what effects handedness and size are thought to have on these statistics I felt more prepared to plan my visualisation.

I then sketched out some designs, initially focussing on explaining what these concepts were to a British audience. However, I then decided that it was taking too long to get to the actual analysis so I scrapped this idea and decided to launch into visualising the distributions of the variables using bar charts. Although this was partly exploratory work, it helped me better understand the data and also started to provide some insights into the data, and the limitations of it. I then sketched out bar charts for categorical variables, scatter plots for continuous variables and also line graphs for trends.

Following this, I used Tableau to create the visualisations, making the following design decisions for clarity, and following best practice guidelines.

- Renamed fields for clarity –avg to Batting Average, HR to Home Runs and handedness to Both, Left and Right.
- Created new binned fields for Weight and Height to help with the visualisations.
- Used 1 colour for all my exploratory distributions and univariate analysis.
- Used a muted blue for the univariate analysis rather than a bright colour in order not to distract from my findings.
- Used high data / ink ratio by not including unnecessary axis labels or borders.
- Ensured all graphics had a title and an explanation.
- Used the blue and orange palette for multivariate analysis as this is considered the best for colour blindness. I avoided green for this reason.
- Used the same colour encodings for all my multivariate analysis (handedness page).
- Included a legend for the multivariate graphs.
- Used simple plots:
  - Simple bar graphs for univariate analysis - distributions of continuous and categorical variables and for looking at the BA HR performance of players which are also continuous.
  - Box and whisker plots for bivariate analysis of a categorical and continuous variable – to show handedness against the performance variables.
  - Colour encoded scatter plot for multivariate analysis, using the colour encoding to distinguish the categorical variable handedness over the continuous performance metrics. For consistency, these were the same colours used for the box and whisker plots.
  - Used line plots for looking at height and weight trends for the performance variables and these were all in separate plots.
- Decided to allow the user to interact with the visualisations showing the performance of all the players. They could scroll down to see the highest and lowest performers. I originally created fixed charts of just the top and bottom performers but then thought it was more interesting for the user to interact with the graphs.
- I provided filters on the handedness multivariate charts (box and scatter) to allow the user to select which hand and scroll through the performance stats on the box plots in order to focus on a subset of the data, e.g. excluding outliers

Next I sought feedback for my initial visualisation – see section after design for full feedback.

Post feedback changes

- Included a photo of a batter on the introduction page to add interest, but also to add context that the data was concerned with batter performance rather than performance of all positions.
- Added ‘count’ to the y axis of the first graph to aid understanding of the graphs.
- Labelling the graphs better and provided better explanatory captions.
- Highlighted areas of interest on the graphs in orange (again best 2nd colour for colour blindness) in response to the interest shown in the data by the people who provided feedback.

- Highlighted the ‘zero’ HR and BA data and stated the question - as this was picked up as an area of interest and confusion by the people who provided feedback.
- Included definitions of HR and BA for a British audience as the people who provided feedback questioned what these terms meant.
- Added search buttons to the Performance page, and changed top caption to explain objective of search boxes – this required learning about parameters and combining these with calculated fields.
- Added better explanations of the handedness plots – particularly the Box and Whisker plots. I also referred to a further conclusion on the end slide in order to help the flow of the story.
- Highlighted names of top performers in the bottom caption – in line with my 3rd piece of feedback who thought he would rather recruit a player with a moderate HR and BA statistic.
- Smoothed out the line graphs for height and weight graphs in order to see trends better. This allowed me to provide better insights.
- Separated height and weight graphs into two pages.
- Added a better conclusion which addressed the interest on ‘zero’ statistics (i.e. where HR and BA was 0). I couldn’t not provide a definitive answer but I provided my best assumption as to why there were zero statistics.

Feedback

Feedback 1

- What is the y axis on the first page? It would be better if there was a label.
- What is a Batting Average? Why is it a decimal?
- Loads of people have a Batting Average of 0 – weird!! Doesn’t make sense to me.
- Why has someone got a BA and HR statistic of 0? Is he a pitcher? You need to provide an explanation for this!
- I like the slider on the handedness page. That is interesting.
- I **really** like the handedness scatter plot. I like the interactivity and that you can see the name of the players there.
- Interesting that taller people are less effective.

Feedback 2

- Why do some people have a Batting Average of 0? This needs to be explained in the conclusion.
- Also why do some people have 0 Home Runs?
- Why do people have high BAs but low HRs?
- I like that these graphs are interactive.

Feedback 3

- I know that there are multiple positions in baseball but this seems to be just looking at batters. You need to make this clear that this work is on the performance of batters – maybe have the batter photo at the start rather than the end to highlight that it is about batters.

- Make the labels more specific. E.g. not just height or weight but height or weight of the batter or player.
- The labels are actually the wrong way round for height and weight.
- What does Batting Average mean?
- Change ‘number of records’ (on the distributions hover over) to ‘number of players’ to make it clearer.
- Add some better ‘handedness’ conclusions. I don’t think your conclusions really reflect what I am seeing here.
- Separate pages for height and weight would make graphs clearer.
- There are outliers in the height and weight data which are hindering seeing the trends. Investigate these and maybe try excluding these or smoothing the lines then you might find that there are trends.
- Could you do a search box on the player bar graphs? I would like to compare the performance stats and if I could search for the Batting Average of the player who has the highest number of Home Runs that would be really interesting.
- Draw some better handedness conclusions. Both handers seem to be more reliable.
- I would rather recruit Jim Rice as he is good at both (BA and HR) rather than being really good at one or the other.
- It would be useful to see all the points in the data – i.e. first base, second base, etc … not just HRs – but I know this is a limitation of the data.
- Look at the height peak of 67”. If this was excluded we could see the trend better.
- There are not enough samples to do this analysis well. It might be that the other peak of 78” is also not useful and distorts the results. It would be interesting to look into the peaks and troughs.
- Also, the same with the weight analysis – explain outliers and if you ignore them you could better see trends in the data. Look to see what the real optimum weight is.
- There aren’t enough samples as the peak is just one person with a high number of HRs.
- You could conclude that there isn’t enough data to really analyse height and weight but there could well be clear upward and downward trends.
- The graphs look great. I really like them. Well done!

Resources

I used the following websites to get a better general understanding of baseball to provide context and pique my interest.

https://en.wikipedia.org/wiki/Batting_average
en.wikipedia.org/wiki/Home_run

https://sporteology.net/top-10-most-popular-sports-in-america/

https://sporteology.net/top-ten-popular-sports-uk/

https://www.azsnakepit.com/2010/7/5/1550963/baseball-players-does-size-matter

https://www.livestrong.com/article/540491-does-height-affect-ability-in-sports/

https://www.gamesensesports.com/knowledge/2017/3/17/righties-vs-lefties-the-importance-of-handedness-training-in-baseball-hitting

https://www.researchgate.net/publication/12815793_Correlations_for_weight_height_and_two_measures_of_batting_performance

https://www.livescience.com/2665-baseball-rigged-lefties.html

https://www.thebaseballindex.com/tall-baseball-pitchers-better-short-pitchers/

And the following to help me with my visualisations

How to implement a search box:

https://www.analytics-tuts.com/search-box-in-tableau/

How to smooth the line graph:

https://community.tableau.com/thread/214271

Highlighting points:

https://onlinehelp.tableau.com/current/pro/desktop/en-us/actions_highlight_highlighter.htm

Submissions

- Data_Story_Writeup.pdf
- Baseball_data_final_export.csv – this is the file exported from my final Tableau Data Story including the extra fields automatically created and the renaming of columns. The original data file was downloaded from https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud507/baseball_data.csv&sa=D&ust=1544457037299000