BatStat - Final Report

Christian Gottsnahm* a01404728
Jakob Schafellner† a01404727
David Schiester‡ a01404729

1 Motivation

1.1 Problem description

Nowadays, statistics play an important role, but due to the constantly growing amount of information, it is becoming increasingly difficult to filter out the interesting data. In the field of professional sports, there are several programs, but in school sports, that is not the case. We try to fill this gap by providing an interactive dashboard to help sorting out the unnecessary data.
Specifically, we want to enable ambitious, young athletes to make a successful career by choosing the best or most promising university. Of course, many other criteria along sports play an important role in the choice of a suitable university, which we can only partially incorporate. But for finding the best career opportunities, our tool allows for a precise and easy-to-use comparison of diverse colleges to evaluate those that trained the most successful athletes.

1.2 Tasks

The application should provide a readable and easily understandable overview of all relevant universities. In order to perform a suitable search, the player position must be specified. This will automatically display the visualization most appropriate for the users needs. By selecting specific areas in the shown diagrams, the colleges can be constrained using interactive filters. This results in our main task, which is to compare many colleges based on selected categories and to find the best ones for your personal interests.
The second part of the main task uses the results of the first step and offers the possibility to continue and specify the search. By comparing two colleges based on much more detailed information, it should be easy to determine a final winner.

1.3 Users

Our main target group are young baseball players looking for a suitable university for further education. By using the general comparison, hundreds of universities can quickly be limited to a handful of eligible institutions. This selection takes place on the basis of simple, free choose able criteria such as geographical location, success of the graduates in their careers and some other parameters. With the smaller selection of fitting schools, you now can use the specific mode to compare two universities based on more detailed information which leads to quickly evaluating his favourite.
In addition to the primary target group, there are still two secondary ones, firstly the sports agents, who, with the help of our visualization, can easily check the best universities in their area for suitable players. And secondly, enthusiasts of sports betting, who want to bet on professional games as well as games in the college league. Our tool also enables them to gain a better understanding as well as a deeper insight into the college sport, which should benefit them with their betting decisions.

*e-mail: a01404728@unet.univie.ac.at
†e-mail: a01404727@unet.univie.ac.at
‡e-mail: a01404729@unet.univie.ac.at

1.4 Data

As data source we use the database published by Lahman, which contains statistics for baseball. This database turned out to be extremely huge and contains exact information for all players, the played matches inclusive obtained points as well as the Colleges players attended, among many other things. As usual for databases, this data is stored divided into several tables. The tables relevant for our visualization are briefly explained below:

- Master/Player - player names, DOB, debut/final game
- Batting - batting statistics
- Pitching - pitching statistics
- Fielding - fielding statistics
- Salaries - player salary data
- Schools - list of colleges
- CollegePlaying - connects players to colleges

These tables form the basic data for our project. Using custom selects and joins with respect to the given relations between the tables we cleared or rather filtered the data. Furthermore, we introduced new fields, which were calculated or derived from existing attributes, to provide additional options for filtering and prepare the data for the visualization.

2 Related Work

Despite the huge popularity of baseball worldwide, especially in the US, and the extensive statistics, there are not many scientific papers discussing visualization. In addition, a large part of this research deals with the simulation or animation of moves in order to analyze them more accurately. Although these works provide interesting results, unfortunately these are hardly relevant for our project. In addition to the attempts of game reconstruction, a few have also dealt with the visualization of the more general statistics of this sport. But most of them are academic projects which have not been published. In the following the sources are briefly summarized:
The work [1] is based on the same database we use for our project. The introduction gives a good overview of the available tables as well as their connection to each other, which allowed us to gain a deeper understanding of the database we are using. In addition, the different problems and the provided solution via diagrams yield some exciting ideas for our project, including the correlation between salary and wins.
In [2] they present the change in baseball over time as clearly as possible in order to examine the impact of individual exceptional players on the average numbers or a specific era. The approach for filtering the displayed data using radio buttons, to enable the access to a lot of information is quite interesting. However, visualizing only one specific property makes a comprehensive comparison very difficult or even impossible. In the end, this work gave us the idea to create two different dashboards. One for pitching and one for batting, in order to process the data individual and show it as structured and well-arranged as possible.

Furthermore, we found some more interesting sources during our research. Although they did not have a significant impact on our approach, they should be briefly mentioned for the sake of completeness. In the book [3] a lot of relationships in baseball data are highlighted by graphics using the statistical program R. In [4], [5] and [6] further interesting approaches for visualizing baseball statistics are presented.

3 Approach

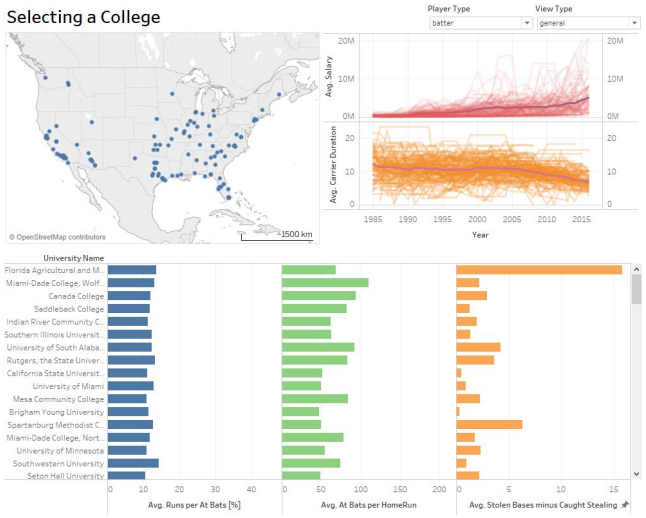

Figure 1: Full Dashboard of the general view type

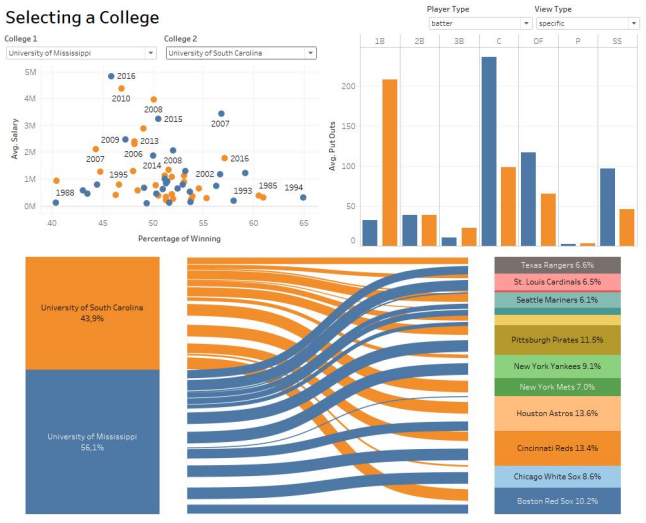

Figure 2: Full Dashboard of the specific view type

3.1 Description of the visualization

3.1.1 Parameter

We decided to use two different parameters in Tableau to display specific diagrams for each parameter. The first parameter is the view type parameter. It offers two options:

- ”general”: Shows a general overview of the statistics which will be described in more detail below.
- ”specific”: Matches two universities, giving you a closer look at two colleges to choose from.

The second parameter is the player type Parameter. This also offers two possibilities:

- ”pitcher”: Different statistics are displayed concerning the pitcher.
- ”defense/batter”: Different statistics regarding the batter are displayed here.

For each parameter, there must always be one option selected to obtain a working visualization.

Figure 3: Parameter player type and view type

3.1.2 Map

To get an overview of the existing colleges, we have integrated a map. On this all registered universities are indicated by a blue dot on the map. When hovering a point with the mouse, a tooltip appears which contains more information about the university. The university name, the city and the state are reproduced textually.

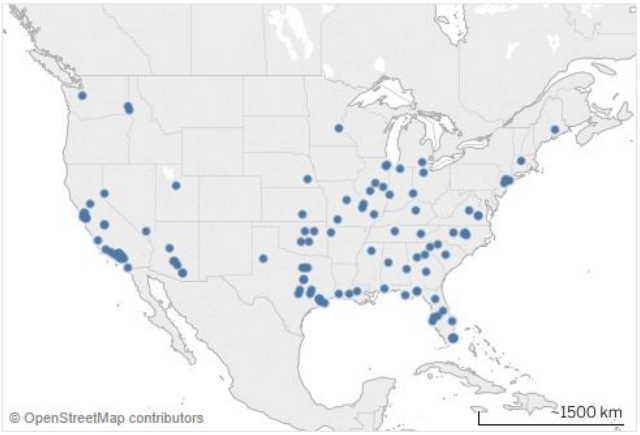

Figure 4: Map of america with all universities

3.1.3 Multiplot line chart

In this line chart, two attributes are displayed over time. In the upper one you can read off the average salary of each selected university in each year. The lower line chart is the average career duration of each player in a university. Likewise, we provided a tooltip for both diagrams which shows the exact numbers.

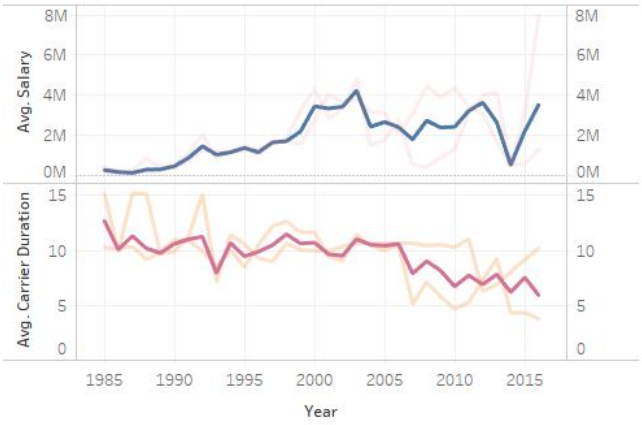

Figure 5: Line chart with multiple plots concerning the pitcher

3.1.4 Multiplot Pitcher bar chart

The bar chart with the selected player type ”pitcher”, displays three different charts. Here you will find on one axis the university name and on the other axis three subdivisions. On the left the avg. earned runs, in the middle the avg. opponent’s batting average and on the right the avg. outs pitches. The tooltip always shows the exact number of all three diagrams.

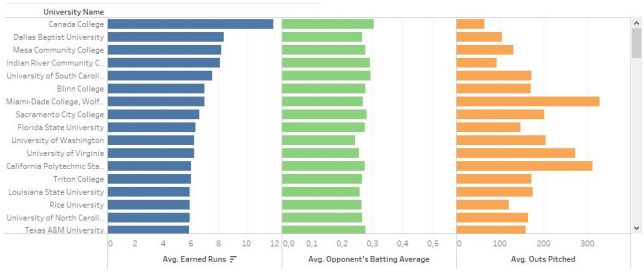

Figure 6: Bar chart with multiple plots concerning the pitcher

3.1.5 Parameter University

The parameter is only available when using the ”specific” view type. Through this parameter you can face exactly two universities.

Figure 7: Parameters for the two universities you want to compare

3.1.6 Scatter plot

This diagram is displayed when using the ”specific” view type. Here, the average salary is correlated to the percentage wins of a university.

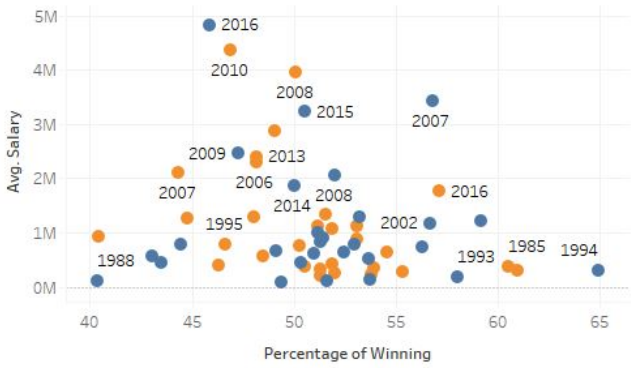

Figure 8: The Scatter plot for the salary / win correlation

3.1.7 Multiplot defence bar chart

This chart shows the average putouts on the respective positions. All positions where such a putout is possible are represented. Only the two chosen universities are compared, as it is only displayed in the view type ”specific”.

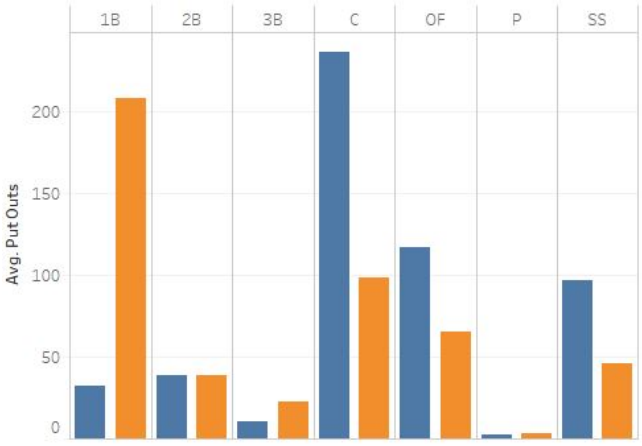

Figure 9: Bar chart with multiple plots concerning the defence

3.1.8 Sankey chart

The sankey chart shows the two selected universities and to which team how many players go after they graduate. The thicker the line in the middle is the more players go to this team. Furthermore, the percentage of graduated players is displayed at the two universities and at the teams.

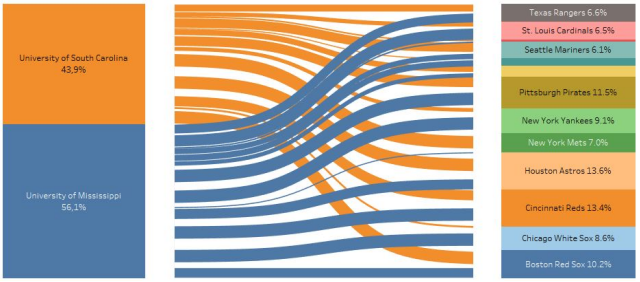

Figure 10: Sankey chart with two universities and to which team the graduated students will go

3.2 Design choices

Since Milestone 3, we have once again fundamentally rethought our design decisions and tried to implement the given feedback as well as possible. In the following, our design decisions are explained in detail:
The quality of our data source unfortunately turned out to be not as high-grade as expected. The database contains some blotted and sometimes even incorrect records, these had to be filtered out. In addition, universities with only a few associated data records were sorted out. Although this violates the recommendation to use as much information as possible, the low variance in these data does not lead to meaningful statistical visualizations and only increases the risk of false conclusions.
As already described in the previous reports, we decided to use different views, which can be controlled by parameters. On the one hand there is a separation between general and specific comparison, on the other hand a division by player position. These options offer the advantage of adapting the diagrams according to the situation and thus presenting only the most relevant information.
On the recommendation of the feedback for our Milestone 3, we decided to remove the additional information regarding salary from the map. The payment is now displayed along the time axis in a separate diagram. These changes improved the readability and additionally increased the information content by the time dependency of the paid fees. Furthermore, the criticized pie chart was replaced by a new chart, at least in terms of space. The most important information of the pie chart was transferred to a bar chart, which was then combined with the existing bar chart. The presentation of several facts with a common axis in only one diagram saves space and offers a good overview as well as a lot of parameters for the filtering.
The specific dashboard has been completely redesigned, first and foremost we reduced from the original comparison of some colleges to the comparison of two schools. This allowed us many new options in terms of visualization. Again, our first instinct was to use a bar chart, but we could not the same approach as in the overview. Due to the limited number of universities we could use the proximity effect to improve the readability of our bar chart. Since we were told while our presentation that many young players probably aspire to a certain team, we decided to implement a heat flow chart otherwise known as Sankey-diagram. A quick glance at this chart is enough to determine which teams have recruited the graduates of the chosen universities. Finally, we wanted to give an information about the expected payment. In order to also provide data about the chances of winning major league games and to detect if there is a correlation with the salary, we use a scatterplot. Last but not least, we came up with the idea to encode the universities by color and prevent the diagrams from displaying the now unnecessary axes in order to clean up the display.

4 Implementation

4.1 Platform

It took us a while to decide which platform we would use to implement the project. First, we agreed to use one of the visualization methods presented in the course. Thus, the decision was between Tableau and D3. Each of those tools has its advantages and disadvantages. After a long discussion, we chose Tableau because it offers a simple and intuitive interface. This allow us to quickly create prototypes to better understand the data itself. This was also a reason why we decided against D3. Since we were using a large dataset, we were initially unable to evaluate much from the data. In Tableau, we got a good overview based on simply generated charts. Furthermore, it is easy to integrate the dashboard into our website via a Tableau Public Account.

4.2 Challenges

4.2.1 Waiting period for data retrieval

Due to the very large dataset, the first test charts in Tableau were extremely computationally intensive and resulted in a long waiting time for displaying the charts. To make our dashboard more user-friendly, we had to make changes. For this purpose, all contained tables were analyzed for their contents and attributes, furthermore unnecessary tables were filtered out. Our focus has been on information about universities and their related tables to establish correlation between education and career success. So, tables that were not required were already excluded from the data sources in advance. In the end, only a handful of tables remained, but their relations with each other led to a huge amount of data. With the help of custom SQL-queries, the data volume could be further reduced by excluding all superfluous attributes. In addition, the attributes were assigned descriptive names. As a result of this change, the data retrieval has been greatly reduced and thus the waiting period could be shortened immensely or even no waiting time can occur.

4.2.2 Overlaying in Tableau

It took some time until a possibility was found to ensure the change of charts at the same position. We absolutely wanted to implement this feature, as it saves a lot of space. In previous tests, we could change a chart with a few simple clicks when changing a parameter. However, this feature was removed or changed in an update of Tableau and we could not realize this function so easily anymore. Initially, we overlaid several charts, but we came across a problem. When switching diagrams via parameter controls, the hidden diagram appears to be transparent, but remains in the foreground, thus preventing any interaction with the underlying, activated diagram. After some research on the internet, we found a method that was very similar to the old one. Using this method, we could display different diagrams when changing a parameter.

4.2.3 Initial-view of Dashboard

At the start of the Dashboard no interactive filters are set, so the displayed charts are quite cluttered and almost unreadable. With the help of predefined filters, the vast amount of information could be filtered by default, but this could lead to a decline in the intuitive handling.

4.2.4 Tableau-Actions

The possibilities of actions in Tableau are very limited. There is only just filtering and highlighting. The interactive filters of Tableau were a big problem for us. If multiple charts are based on the same attributes and interactive filters are implemented, the filtering via interactions seems to be random and the filters will not work properly. Despite the very limited possibilities, the project could be realized with more complex solutions. With the use of D3 this problem could have been avoided from the beginning.

5 Results

5.1 Scenario of use

Imagine a young successful baseball player looking for a suitable university. First, he indicates his position, Pitcher, to get the appropriate view. Grown up in the cold north, he now would prefer a warmer area. As a high-performance athlete too humid climates are out of question, so his decision falls to southern California:

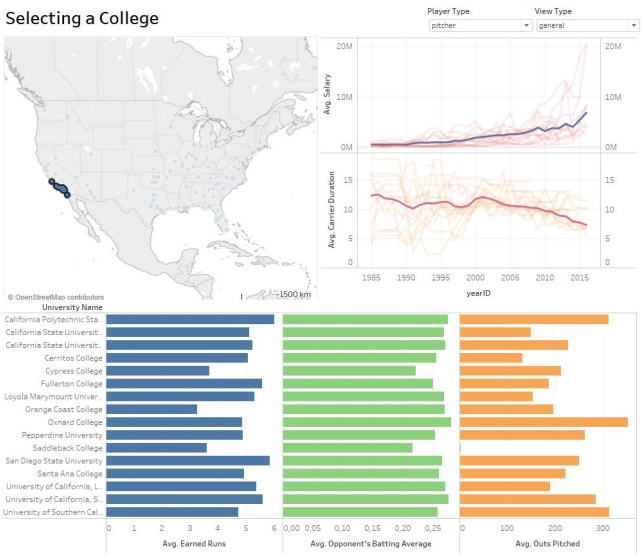

Figure 11: General overview with selected location

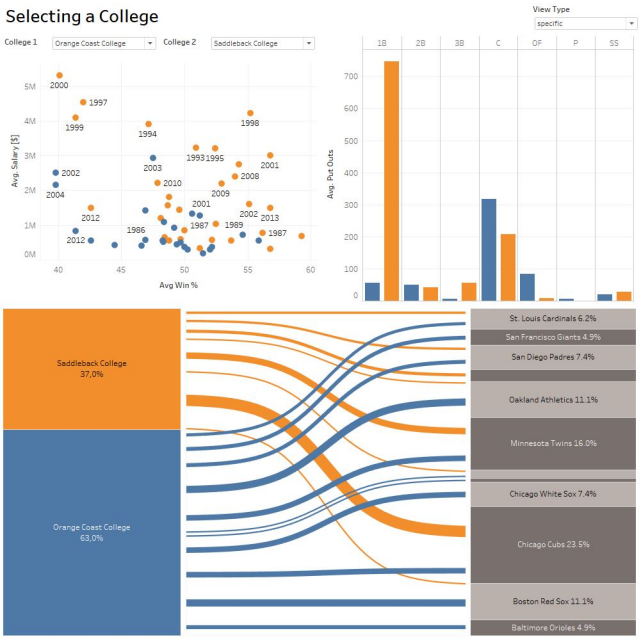

Figure 13: Comparison of Orange Coast- and Saddleback-College

Having chosen the desired location, he now limits his selection to the best three colleges. In his opinion the most important stat are the earned runs, as they are often used to rate a pitcher. The lower the value the better the player, so he sorts by descending values and selects the top three schools:

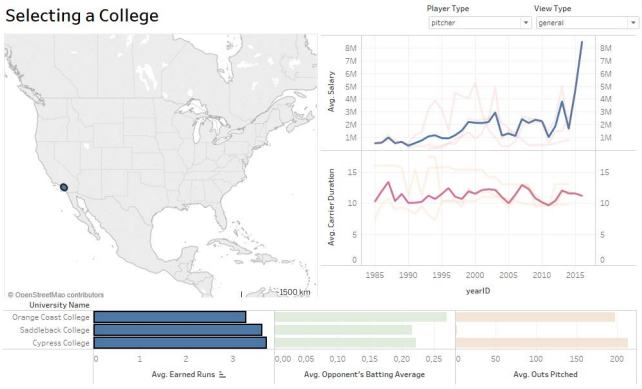

Figure 12: General overview with selected colleges

To ensure the common salaries for graduates from this colleges match his expectations, he crosschecks his selections via the displayed line chart. Knowing the chances for high fees he continues his search. Already impressed by the reputation he favours the Orange Coast College, but to make sure its the right choice, he decides to use the specific comparison for further information:

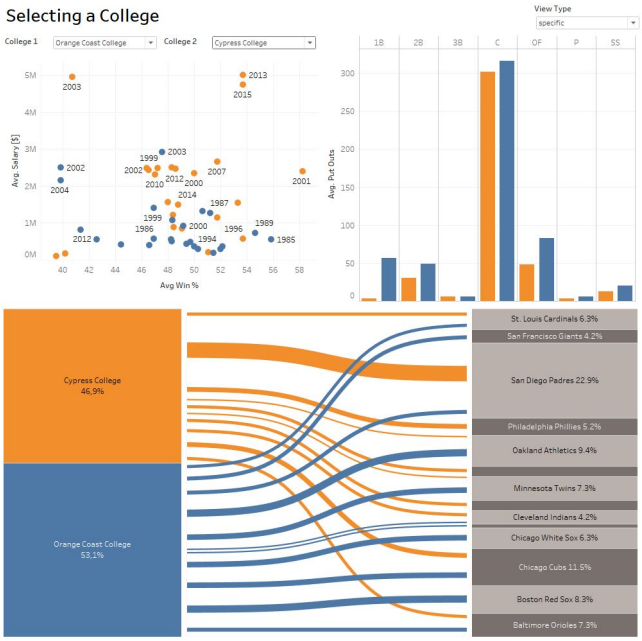

Figure 14: Comparison of Orange Coast- and Cypress-College

Comparing his favourite to the other two options, he realises the expected salaries are higher even if the Orange Coast College seems to train the better players. In addition, he detects that no graduates from Cypress College have ever played for his favourite team, the Boston Red Sox. Shocked by his finding he excludes Cypress College from his selection and decides to visit the remaining schools for his final decision.

5.2 Performance

Overall, we are reasonably happy with the performance of our dashboard. You can scroll and zoom on the charts without any further delays. The only factor is in filtering the data maybe have a slightly increased load time.

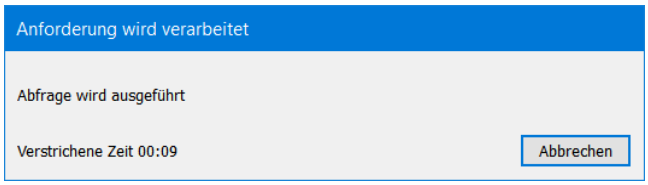

Figure 15: Load time with 8 to 15 seconds

Because the database queries and the database itself is very large. Using D3 might perhaps fix such load time, as you have more data processing options in D3 than in Tableau. However, at the beginning of the project, we decided to do it in Tableau. This eliminates this suggestion for improvement. When all data has been loaded after filtering, it is possible to work smoothly with the dashboard again. All types of implemented highlighting are done very quickly.
Another possible reason for the poorer performance could be the use of two dashboards in one. Maybe it would have been better to divide the dashboard into a ”general”-dashboard and a ”specific”-dashboard. Since this would require less data to be loaded per dashboard.
Due to the long wait, we thought about shrinking the dataset a bit to increase the performance. But then we refrained from doing so, because we did not know exactly how much time it would cost and if it would bring a noticeable performance improvement.
We basically implemented the suggestions for improvement from the last presentation. However, we have incorporated new features based on these implementations, resulting in higher load times. Nevertheless, this feedback was very helpful to get a more intuitive dashboard. Where also the essential data were displayed, which are important for the user.

5.3 Usability studies

For the verification of a user-friendly interface, we conducted a small study with some people. They should only test the usability without any further understanding of the data or baseball rules. The subjects had no experience in visualization or design choices. Every single person was briefly explained the application and afterwards some tasks were put together to prove the comprehensibility of the surface. Among other things, they were asked to select individual colleges. Due to the filtering of the data after selecting some colleges, most people were overwhelmed. Because they generally did not know how to cancel or undo the filter in Tableau, for example. Since no person has ever worked with Tableau, this was to be expected. In the following they received a brief explanation in Tableau and then everything went great. Almost all the test persons forgot to change the parameter Player Type. This may be related to the lack of baseball experience, so they are advised to change this parameter for testing purposes. The change to the ”specific” ”View Type” worked very well because they wanted to have more detailed data about the selected colleges. For the people this was a logical conclusion. For the persons the application was quite good and they said that for college seeking students in terms of baseball it could be very helpful. The selected data that is displayed are very interesting for subject-related people. However, it has been criticized that the map for selecting the colleges is a bit too small and therefore the choice is more difficult. On these points, however, we found no possible change as we want that the chosen colleges can be changed at any time, and the bigger we make the map, the smaller the diagrams become. That’s why we opted for the slightly smaller map.
Based on the results obtained by the test persons, some small changes were made to make the user interface more user-friendly. Some criticized points have not been changed, as they will not occur for persons with knowledge of the sport.

5.4 Functionality studies

After the usability study, another study was done to prove the functionality of the visualization. A person was interviewed who knows this topic and can test it for functionality. It was more about using the data and the existing built-in visualization capabilities. These include the individual diagrams as well as the filters and highlights and how they interact with each other. This allowed the full functionality to be tested and if the used data makes any sense at all. Therefore, we exactly explained what we would like to achieve with this visualization and what our target audience is. During the study we came to some criticisms and to praise. In general, the visualization was titled as comfortable and helpful to use. The criticisms included:

- Line chart: An overload of data occurs at the average career duration as well as the average salary. This makes it very difficult to make a meaningful decision from this diagram. However, choosing fewer universities can solve this problem. Only when the visualization starts can the user feel overwhelmed.
- Scatterplot: Since the choice of the two universities is freely available to anyone, this can lead to the fact that no beautiful correlation of the data can be achieved. When two universities are compared, where one does very well and the other does very badly. Again, this issue can be ruled out by better-chosen universities.

The points of positive feedback were:

- Map: The map visualizes the individual locations of the universities very well, so that it can perceive each person well visually. Tableau’s standard features such as zoom and any kind of selection work great.
- Bar chart: The bar charts are kept very simple, both with the player type ”pitcher” and ”batter”. As a result, the data provided are very well reproduced and it can be sorted nicely according to the properties shown.
- Sankey chart: Because this chart is represented by the view type specific, there can be no excessive overloading of the data. In the particular type of chart, you can see from the two selected universities, the percentage of graduates of the colleges and to which team they go after graduation. The function ”edge bundling” is very well used, making it very legible and there are less or no errors. Only for teams that are less sought after, the name is no longer displayed, but this is bypassed by a tooltip. Furthermore, the line strength from the universities to the teams is very crucial and it is easy for every user to see if many students are going to a team or not.

6 Discussion

6.1 Strengths and weaknesses of implementation

Since every visualization has its strengths and weaknesses, ours also has its positive and negative sides. Based on our studies with the most diverse persons some weaknesses could be found and solved. However, we were unable to eliminate them all.
One of the strengths of our visualization is the large amount of data that has been used. However, it is not sufficient to use a large amount of data, but this data must also be combined correctly to be able to achieve a helpful presentation. Furthermore, the individual diagrams harmonize very well with each other, so that we make it possible for the selected user group to transfer a lot of information to the user in a small space.
The studies also made it clear that baseball fans are very good with

this tool and find it very informative to compare the different universities.
The filters used in Tableau work well and give the user even more visual feedback. Based on this feedback, they can better address universities in the specific view.
Apart from the strengths of the visualization, there are also some weaknesses that we will explain below.
One of the biggest problems is the too long loading time when using a filter or changing a parameter.
Our custom SQL-queries are too costly, related to the size of the dataset. Any change in the visualization, be it through a filter or changing a parameter, will result in a long wait time.
Furthermore, there are occasional problems when using filters on multiple charts at once. This leads to errors in the data. Likewise, records with null entries can come up.
The line chart that shows the average career duration and the average salary is a bit confusing. Since a variety of lines are displayed, which the user can not initially assign as useful.
Some planned properties could not be implemented directly in Tableau so we had to do the visualization without them. Since these are not feasible in Tableau or they increase the load time even more.

6.2 Lessons learned

The lessons we learned are, among other things, that a good visualization does not have to look just beautiful. But also, the functions that are offered must be present and they must harmonize perfectly with the visualization.
One of our basic problems was that we encountered many difficulties with our prototypes. Since the data set was very large and a meaningful filtering of the data was very difficult to accomplish. Furthermore, it was a good experience for all of us because we could learn more about one of America's favorite sport, baseball. Maybe it would have been better to choose a topic where at least one of the colleagues already has experience. This might have eliminated some of the problems. However, finding an interesting topic is not that easy and no good data source was found to meet the requirements for the project. During the project, we also came across much more interesting topics. Therefore, a more intensive employment in finding a topic would have been helpful.

7 Task separation

7.1 Gottsnahm

- summarised motivation
- analysed related work
- improved dashboard
- explained design choices
- described scenario of use

7.2 Schafellner

- implemented and finalised dashboard
- discussed performance and lessons learned

7.3 Schiester

- updated homepage
- descripted visualisation
- documented implementation
- recapped results

References

[1] Jason Yeung. Visualizing 146 years of baseball in a single qlik app. https://www.linkedin.com/pulse/visualizing-146-years-baseball-single-qlik-app-jason-yeung, September 2017. Last accessed 21.01.2018.

[2] Ian Knight. Baseball stats through the ages: An interactive visualization. http://studentwork.prattinfoschool.nyc/blog/coursework/information-visualization/baseball-stats-ages-interactive-visualization, July 2016. Last accessed 21.01.2018.

[3] Jim Albert. *Visualizing Baseball*. Chapman and Hall/CRC, August 2017.

[4] Ryan Sleeper. Using tableau to improve the understanding of baseball statistics. https://public.tableau.com/en-us/s/blog/2014/03/using-tableau-improve-understanding-baseball-statistics, March 2014. Last accessed 21.01.2018.

[5] Steve Jones. Mlb statistics visualization. http://stevejones.io/portfolio/mlb-stat-vis/, 2017. Last accessed 21.01.2018.

[6] Ken Cherven. Visual-baseball project. http://visual-baseball.com/wordpress/, 2016. Last accessed 21.01.2018.