# Baseball Statistics
# Group 10 - Final Report

Petar Cvetkovic (01563662) and Christoph Spreitzer (01467712)

Universität Wien – 052215 Visualization and Visual Data Analysis (VU)

**Project:** http://www.unet.univie.ac.at/~cvetkovicp97/VIS/project.html

**Public Tableau – Dashboard 1:** https://public.tableau.com/views/BaseballStats-Player/Dashboard2?:embed=y&:display_count=yes
**Public Tableau – Dashboard 2:** https://public.tableau.com/views/BaseballStatistics-Teams/Dashboard1?:embed=y&:display_count=yes

**Abstract.** A user-friendly visualization of "The Lahman baseball database, 1871-present". This dataset will give the interested user the opportunity to explore and gain insight into the baseball statistics in an interactive and interesting way.

**Keywords:** baseball, statistics, homeruns, hits, doubles, triples, teams, players, popular, Babe Ruth, base, bat, catcher, play, run, strike out, visualization, tableau.

## 1 Motivation

### 1.1 General information

Baseball is a widely popular sport in America. Players take turns batting and fielding. Players of the batting team take turns hitting against the pitcher of the fielding team, which tries to prevent runs by getting hitters out. The teams switch between batting and fielding whenever the fielding team records three outs. One turn batting for both teams, beginning with the visiting team, constitutes an inning. A game is composed of nine innings, and the team with the greater number of runs at the end of the game wins. [1]

Sports statistics help fans understand the game better, predict the winner and compare old players with their favorite player today. Baseball fans, as well as baseball novices often search for data in order to achieve those things, but can't understand the datasets.

With that in mind, our project has the objective to provide the user with a complete overview of the dataset, as well as the ability to get detailed insight about specific baseball teams and/or players.

### 1.2 Tasks

The project should provide an easy to understand visualization of baseball statistics, portraying the most important information of the data, giving an overview of all the players based on many different criteria (e.g. country, hits, runs, batting hand…) and giving the user the ability to play around with that criteria in order to get valuable insight from it.

Key criteria for the task:
- user friendliness
- interactivity
- providing great insight

Use Cases:

- User is able to see the name, year of birth, year of debut, city of birth as well as the games played for each team for every player.
- User can search a certain player by name and get the information for that exact player (find out information about a specific player).
- User is able to see the average at bats, salary and win/lose ratio for a player as well for more players together.

- User is able to select two or more teams and years and compare their stats (wins, games played, ranks) [see how various aspects changed over time]
- User can see the statistics for batting (filter/select statistics for hits, homeruns, doubles, triples…) for all players as well as their batting hand (right, left, both) and the amount of hits he hit. [find out who had the most hits, homeruns, and with which batting hand]
- User can find out about salaries pro and furthermore click on a certain state and get the statistics for batting for players from that country. (user can find out which country has the highest salaries)
- User can select a player from the batting statistics graph and get specific insight for that player.
- A user is able to see how the weight and height impact the player performance.

### 1.3 Users

Our target audience is anyone interested in understanding baseball statistics and/or finding more about a certain player, no matter whether for personal or professional purposes. Factors such as age, gender and nationality don't matter, as long as they have a basic understanding of graphs.

Examples:

- An average baseball fan who wants to find out the salary for his favorite player, and which teams he played for. He has a solid understanding of graphs and knows how baseball works. All he needs to do is write the name of the player in the search bar and read the information from the graphs.

- Journalist who has an assignment to find the left-handed player with most homeruns from the most paid state in America. Has a good understanding of graphs, but has never watched a baseball game. All he needs to do is select the most paid state from the map, filter the batting statistics to “homeruns” and read the graph.

### 1.4 Data

We selected the project from suggested project topics, and worked with the given dataset:

- The Lahman baseball database, 1871-present [2]

This dataset contains pitching, hitting, and fielding statistics for Major League Baseball from 1871. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875.

## 2 Related Work

### 2.1 Related visualizations

A related visualization is one from Ryan Sleeper [3]. He explains how to improve the understanding of baseball statistics. He explains the baseball glossary and how certain stats are calculated.

Another related visualization is “Most Popular Athletes by State” [4]. It is actually a blog that is reporting facts and trends in the world of sports and entertainment. Even though we stumbled upon their website only after M3, they have a very simple, user-friendly approach to presenting data.

### 2.2 Previous visualization ideas we incorporated

While researching ways to make our visualization easy to understand, we came across transfermarkt.at [5] which had a very interesting solution to our problem. We incorporated the fundamental idea with the table containing all player names from that visualization, as well as the income over time for a specific player, in order to make our project more convenient and user friendly.

### 2.3 References to tools used

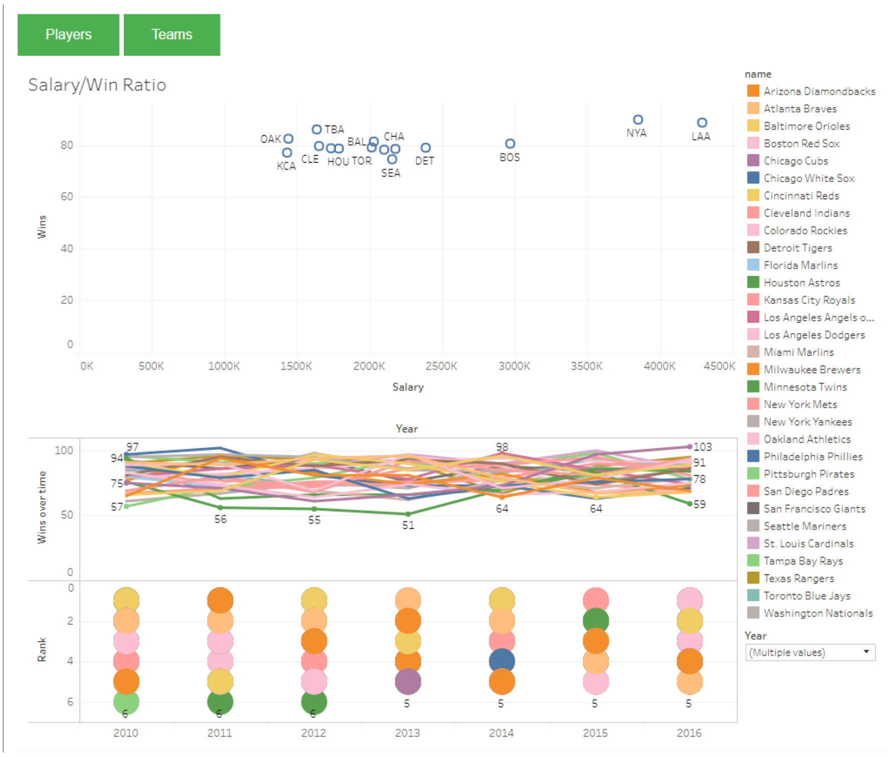
The only tool we used is Tableau [6], since we could incorporate things we learned about it in class.

## 3 Approach

### 3.1 Description of visualization design

After a lot of trial and error, listening to feedback, adaptation and improvement, our Project has reached its final state. It consists of two Tableau dashboards (Teams & Players) and buttons to switch between them.

- Dashboard “Teams” consists of two graphs that function together. (Picture 1)



*Picture 1*

The top graph is a scatterplot which shows the ratio of salary and wins. In other words, it shows the efficiency of the team depending on their salary. The idea of this scatterplot is to show the teams of the baseball major league over time and find out more about winning behavior in dependency to the salary. The more to the left teams go, the less they are paid. The higher the teams are, more wins they have. Y-axis represents wins, and X-axis represents the salary.

The bottom graph consists of two parts. One is a line chart, the other one a bubble chart. On the right, there is a color legend, and the bottom right a “year” drop down list, for the user to select a year or multiple years.
The line-chart represents wins over time. Y-axis shows the wins, and the X-axis years.
The bubble chart represents teams ranking over the years. Y-axis is inverted, because the rank 1 is higher than rank 5. X-axis represents years.

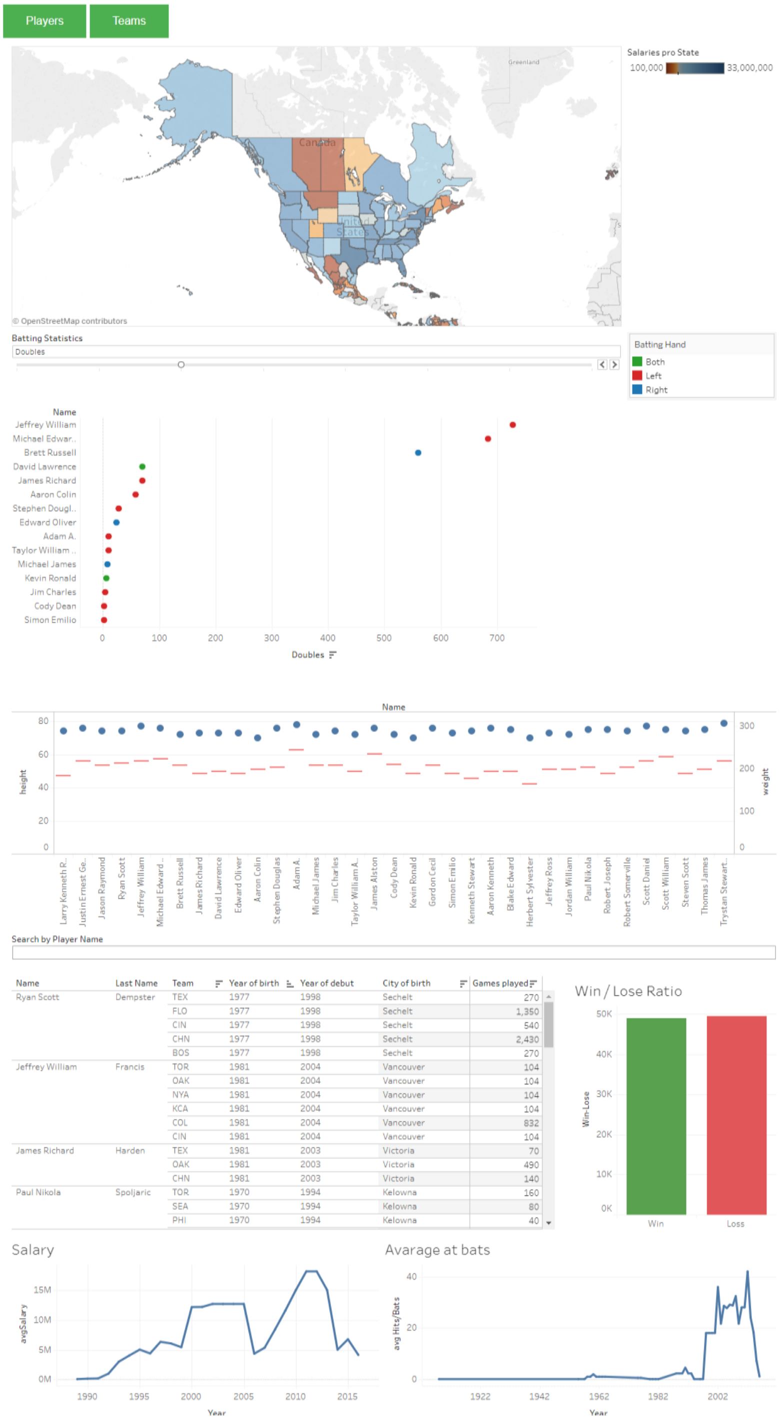
- Dashboard “Players” consists of seven visible graphs. (Picture 2) Visible, because the parameter “batting statistics” provides adequate ways of filtering (hiding and showing graphs). Only one graph from the batting statistics (runs, homeruns, doubles, triples…) can be shown on the dashboard. This dashboard has an overview of the whole dataset, as well as views that are details, showing statistics for a single player.

The geo-map shows salaries per state. Color legend next to it shows that depending on the color of the state, it has a higher or lower salary. There is also a tooltip that shows the exact salary. This map is also used as a filter for the other graphs. For example, if a user clicks on Arizona, the bar-chart will change and show the “hits” for players originated from Arizona.

The filter parameter “batting statistics” allows the user to choose a certain batting statistic which will sort all the players in the descending order for the chosen stat. For example, if a user clicks on “Hits” he will see a bar chart consisting of all players in a descending order by the amount of hits they made.
The first player will have the most amount of hits, the second will have the 2nd most and so on. The colors of the circles represent the batting hand of the player. This is very helpful if the user is only interested in players that bat with a certain hand.
Color legend next to it shows which color represents which hand. X-axis shows the number of chosen statistic, and Y-axis represents the players name.

The chart in the middle is showing the height/width ratio of the players. Which allows the user to see how the height/width ratio impacts player performance. Circles on the Y-axis represent the width, and the lines represent the height. X-axis holds the player names.

Next, we created a table-chart with all the players listed, with parameters name, last name, team, birth year, debut and city of birth as well as a search parameter for the names. If a user types a name in the search bar, he can see the average bats, salary, height/weight ratio and his win/lose rate. The table has a tooltip, showing how many games a player has played for each team.
The bar chart “win/lose ratio” represents total wins and losses. Salary and Average at bats are simple line-charts which represent players salary and at bats over time.



| Name | Last Name | Team | Year of birth | Year of debut | City of birth | Games played |
|---|---|---|---|---|---|---|
| Ryan Scott | Dempster | TEX | 1977 | 1998 | Sechelt | 270 |
| | | FLO | 1977 | 1998 | Sechelt | 1,350 |
| | | CIN | 1977 | 1998 | Sechelt | 540 |
| | | CHN | 1977 | 1998 | Sechelt | 2,430 |
| | | BOS | 1977 | 1998 | Sechelt | 270 |
| Jeffrey William | Francis | TOR | 1981 | 2004 | Vancouver | 104 |
| | | OAK | 1981 | 2004 | Vancouver | 104 |
| | | NYA | 1981 | 2004 | Vancouver | 104 |
| | | KCA | 1981 | 2004 | Vancouver | 104 |
| | | COL | 1981 | 2004 | Vancouver | 832 |
| | | CIN | 1981 | 2004 | Vancouver | 104 |
| James Richard | Harden | TEX | 1981 | 2003 | Victoria | 70 |
| | | OAK | 1981 | 2003 | Victoria | 490 |
| | | CHN | 1981 | 2003 | Victoria | 140 |
| Paul Nikola | Spoljaric | TOR | 1970 | 1994 | Kelowna | 160 |
| | | SEA | 1970 | 1994 | Kelowna | 80 |
| | | PHI | 1970 | 1994 | Kelowna | 40 |

*Picture 2*

### 3.2 Reasons for design choices

Reasons for our design choices come from class and feedback. For example, we started with two dashboards one under another. Which we merged together after the feedback, and added another dashboard that extends the use-case which was also required from us. We also reasoned for certain charts because we learned that it should be that way in class. We also designed them how we did, keeping the lessons in mind.

To be more specific, we chose the geo-map because it looks nice and serves as a great starting point for the user. It is a great way to represent a location and show the average salaries per state. The batting-statistics is a bar chart because it is easy to compare players with one another. It provides a nice overview for a lot of values.
The height/width chart is like that because of the feedback we received. It is a nice way to compare players width and height and see how it impacts their performance.
We went with the table-chart and a search parameter because popular websites use that style and it’s proven to be user friendly. The line-charts which represent salary and average bats over time are simple and easily understandable. Win/lose bar chart looks like that for the same reason.
The scatterplot presenting the salary/win ratio in team dashboard looks like that because it looks nice and presents the data in a very good way. Lastly, we decided for the double graph because it shows important information and that is connected.

## 4 Implementation

### 4.1 Implementation Tools

Our project is available online on our homepage. To create our homepage, we used HTML5 and CSS. We used Tableau for creating the visualization from the dataset. Once finished, we uploaded it to Tableau Public and embedded it on our homepage.

### 4.2 Challenges and Implementation Problems

- Our number one major challenge was incorporating as much data and interactions, while keeping the loading time short. It takes a lot of time for Tableau to execute a query and it was hard keeping the waiting times low while implementing everything we have planned.

- A concrete example of an implementation problem we had is that we couldn’t implement a graph for defensive player positions because it required merging another .csv file with full-outer join. Since we already had around five .csv files that we worked with, once we added the sixth one, it took Tableau thirty minutes to let us change something, and then another thirty minutes to accept the change. Since we wanted that graph to be interactive with other graphs from the dashboard, we weren’t able to solve this problem and ended up not using the graph for defensive positions.

- Another problem was showing the color legend (for batting hand) only once on the dashboard (for multiple sheets) and keeping it consistent. Since we work with a parameter that hides sheets, the color legend disappeared every time the graph changed even though the color legend applied to all graphs. We solved this problem by taking a screenshot of the color legend and importing it into the dashboard.

- We had a very interesting idea which was showing a mapped baseball field with the player positions. (Picture 3). Our plan was to map a baseball field and add a tooltip for the defense positions showing how many times a player played as that position. However, when it came to the implementation we weren’t able to do it. We tried a lot of different approaches, but none of them worked because our dataset didn’t have suitable values.

*Picture 3*

- One of the biggest challenges was coming up with mockups and a user-friendly design that efficiently shows huge amount of data. We spent a lot of time researching and trying to come up with creative solutions to this challenge. It was a long journey, but we managed to make the project with trial and error and help from professors.

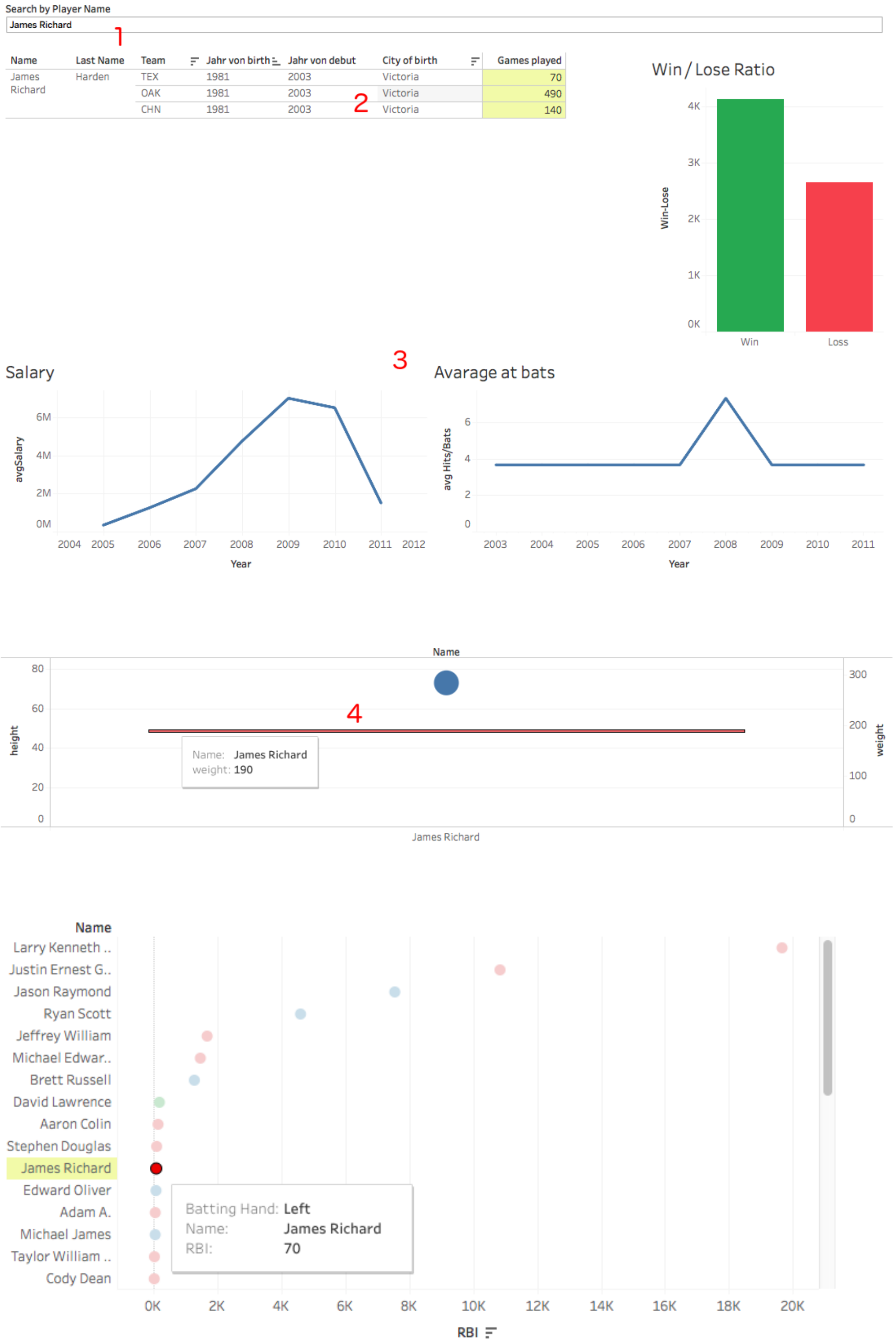
- Another challenge was starting from scratch after M2. Since our first project didn’t have a big enough scope, we had to reset everything for M3. We were battling with a short time frame and we had to find a suitable topic, come up with tasks, use cases, draw mockups, and implement a Lo-Fi Prototype all at once. We pushed through it with longevity and perseverance, since we are sedulous people.

## 5 Results

### 5.1 Scenario 1

Marcus wants to know something about a single player, because he saw a documentation about good players around the turn of the millennium. With our homepage he can do that easily: (picture 4)

1) Search the right name in the search parameter, in our example click with the left mouse in the field and type in “James Richard” and press Enter.
2) Point with the mouse at the table and click in the row “name”.
3) You can see two line-charts about the Salary over the years and the average at bats over the years. You can also see a bar chart with the Win/Lose ratio.
4) If you scroll up you can see players weight and height, as well as his batting statistics.



Search by Player Name

James Richard

| Name | Last Name | Team | Jahr von birth | Jahr von debut | City of birth | Games played |
|---|---|---|---|---|---|---|
| James Richard | Harden | TEX | 1981 | 2003 | Victoria | 70 |
| | | OAK | 1981 | 2003 | Victoria | 490 |
| | | CHN | 1981 | 2003 | Victoria | 140 |

Name: James Richard
weight: 190

Batting Hand: Left
Name: James Richard
RBI: 70

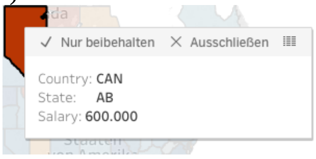
*Picture 4*

## 5.2 Scenario 2

Daniel is interested in the best batter from different states in the dependency of the salary, because in some states the best batter gets paid more, even if he is less efficient. (picture 5)

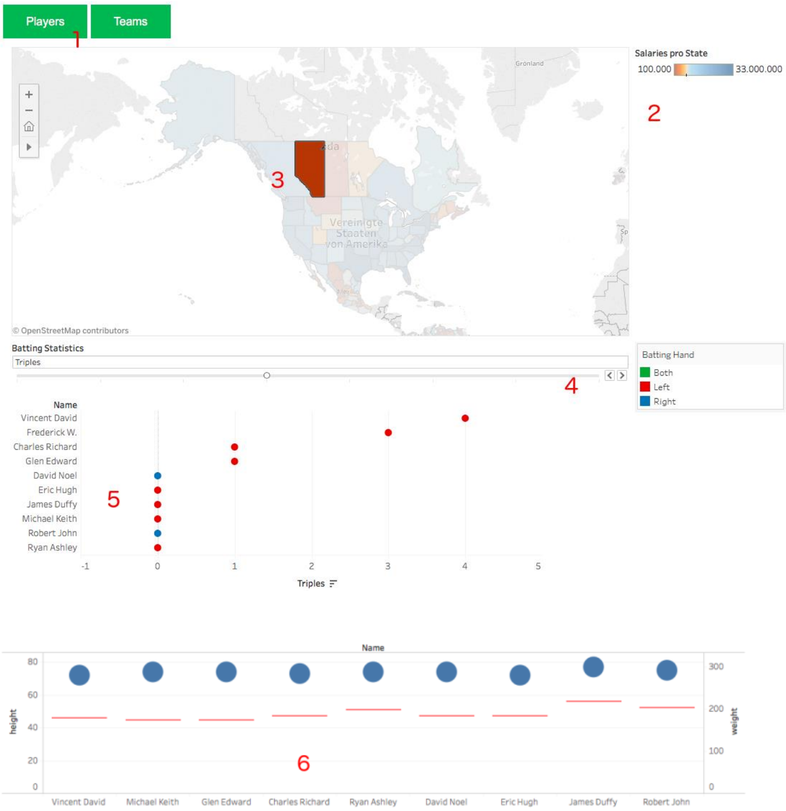
To see that, user can visit our page and do the following:

1)Choose the player tab to get on the right dashboard for player information.
2)Look at the salary legend to understand the map.
3)Drive over the area and search for the lowest salary, and click on it.



Country: CAN
State: AB
Salary: 600.000

4)He can then see different Batting Statistics by dragging on the menu arrows.
5) He can see the players sorted in the descending order based on the chosen batting statistic for the chosen state, as well as their batting hand.
6) He can see how the weight and height impact the player performance.

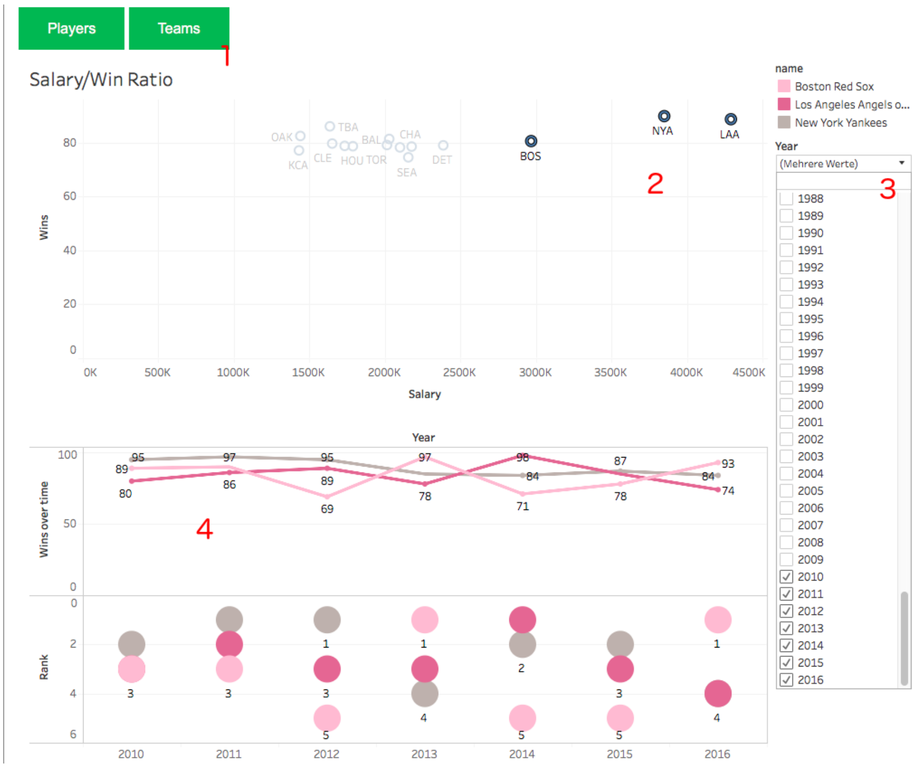


*Picture 5*

### 5.3 Scenario 3

Mike wants to know something about his favorite baseball team in the major league. He always hears that his favorite team is only in the midfield. So, he decided to argument about his team with a good visualization. With our visualization he can show that his team has a consistently good rank for the team income.  He can do so in the following way: (picture 6)

1) Choose the teams tab.
2) There are two possibilities to choose the teams the first one is to hold the “cmd (ctrl)” key, and click with the left mouse on the teams. The second one is to hold the left mouse button and drag over the teams you want to choose.
3) Select years you want to see statistics for.
4) There is one line-chart with the wins over time, and one with the rank.



*Picture 6*

### 5.4 Performance and Feedback

The performance of our project was perfect for M3, but after M3 we combined our two dashboards together which resulted in a higher execution time. We changed one use case, so we could delete one of the joins in our csv-data. After that, the combined dashboard had a passing execution time. We used the deleted csv-data in our new dashboard “Teams” in the rank chart. So now the performance of  our project is decent, however there is a noticeable loading time that occurs when the dashboard “Players” is loaded for the first time.

For our evaluation we created two simple tasks, for each one of the dashboards which our colleagues had to execute.

The main idea of the tasks was to test the usability and complexity of our visualizations. The first task was to search a baseball player by name, and say his salary and batting hand. For this task, the students went to the search field and searched for the name. After that they read the “salary” line-chart, scrolled up and found the player batting hand. It didn’t take long for them to do it, and they had fun doing it. However, one student had a problem seeing the charts because he forgot to select the name after searching for it.

The second task was to find out the ranking of a baseball team in one specific year. For this task the users navigated to dashboard “teams” and completed it with ease.

The feedback was good. They said they could see themselves using this tool if they were baseball fans, and that it would be interesting to see this visualization for other sports as well.

# 6 Discussion

## 6.1 Strengths and weaknesses of the approach and implementation

Based on our feedback from the students and our own experience, we created the strengths and weaknesses of our implementation:

Strengths:

- Simplified huge data set
- User-friendly
- Adaptive
- Combined filters
- Easy-to-Understand Charts
- Interactive

Weaknesses:

- High execution times
- Decent performance
- No available data for old players

To summarize: The strength of our project lies in the interactivity and information value focusing on user friendliness and a carefully chosen set of visualization graphs that are nice for the viewers eye. We have made sure that our project provides both: complete overview of the data, as well as the detailed views and specifics for certain data, meaning it fits the needs of most users.

The weakness of our project lies in the execution time. We have not come up with a way to make the execution times lower, except for extracting the data before posting it online. Another bad thing is that for some users there aren’t any data available, so sometimes, certain graphs will be empty.

## 6.2 Lessons learned

We first took the topic “baby names”, because it sounded interesting. After the feedback from M2, we realized we had to change the topic, since the scope of the dataset we were working with wasn’t big enough.
We learned that it is hard to think about visualizations with only a few variables available, and that some visualizations that make sense for us don’t necessarily make sense for users.

After that we switched to “baseball statistics”. This dataset was huge, we had over fifteen csv-files with a lot of information. We spent a lot of time analyzing the data and making a new concept for our visualization project.

We visualized our charts with Tableau and during the time working, learned a lot about Tableau and earned some valuable experience. There are a lot of things which Tableau makes possible to do, but there are also things that Tableau isn't so good at. So, what we learned is that, even if it is easier to work with a visualization tool like Tableau, D3 is more powerful.

After the feedback for M3, we had to come up with an idea about a new use case. We learned how to bounce of each other's ideas, implement mockups and come up with creative solutions. Our new use case was a dashboard for teams, where a user can compare them with one another.

We learned how to make meaningful graphs, how to work with .csv files, how to improve the project based on feedback from users and how to use the theory and apply it in a project.

## 7 Task Separation

| Christoph | Petar | Date |
|---|---|---|
| Tableau:<br>Adding color legend, fixing table names | Tableau:<br>Fixing win/lose chart, fixing scrollable graphs, improving dashboards. | 15.01.18 |
| Tableau:<br>Improving chart height/weight.<br>Implementing "salary/win ratio" graph. | Tableau:<br>Coming up with idea for use case.<br>Implementing chart "Geomap salary", working on dashboard 2 overview & interactions. | 16.01.18 |
| Tableau:<br>Creating dashboard teams. | Tableau:<br>Merging dashboard 1 & 2, creating interactions, creating dashboard players. | 17.01.18 |
| Written report:<br>Results, Discussion | Written report:<br>Motivation, Related Work | 18.01.18<br>19.01.18 |
| Written report:<br>Task Separation, References, format written report | Written report:<br>Approach, Implementation, grammar correction.<br>Website: implementing Tableau dashboards online | 20.01.18<br>21.01.18 |

**Table 1.** Task separation

### References

1. Bjarkman, Peter C. (2004). Diamonds Around the Globe: The Encyclopedia of International Baseball. Greenwood
2. http://seanlahman.com/baseball-archive/statistics/
3. https://public.tableau.com/en-us/s/blog/2014/03/using-tableau-improve-understanding-baseball-statistics
4. https://www.besttickets.com/blog/most-popular-athletes/
5. https://www.transfermarkt.at/
6. https://www.tableau.com/