Case Study: Analyzing MLB Player Data (Tables in Excel)

Executive Summary: [¼ page description of the project highlights. What are the key results?]

We will follow part of the data science process and focus on importing data and do some first exploratory data analysis.

Table of Contents

1. Frame the problem

[What it the question you can answer? Why are they important?]

Choosing the right players to from a baseball team is important for the team's success. We will use player data from the Major League Baseball (MLB) organization [1] to analyze the current players and teams.

Here are some questions we can investigate:

- Do some teams have more younger players?
- Does the weight of a player have an impact on their performance?

Found patterns can be used to inform selecting new players for a team. (Note: I have no idea about baseball!)

2. Collect the needed data

[Data source, data quality and reliability.]

Data Sources

I found these potential data sources using a quick web search:

- http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_MLB_HeightsWeights
- http://mlb.mlb.com/stats/sortable.jsp#elem=%5Bobject+Object%5D&tab_level=child&click_text=Sortable+Player+hitting&game_type='R'&season=2018&season_type=ANY&league_code='MLB'&sectionType=sp&statType=hitting&page=1&ts=1547740208757

Import Data

Data comes in different formats

- Text: CSV, fixed-width text, XML
- Application specific: Excel, SAS, SPSS, Stata, etc.
- Web: HTML, XHTML
- Relational database

You can try:

- Copy & paste into Excel (use `CRTL + SHIFT` for marking the table)
- Excel provides: `Data > From Web`
- Use other tools. For example, Firefox: Install `Add-ons > Extensions > Table to Excel` (use red icon next to URL in Browser)

3. Prepare and explore the data

[Clean and connect the data.]

Format Data as Table

Place cursor into a table cell and use `Insert > Table` or select data for the whole table with `CTRL + Shift + Arrow keys` and use `Home > Format as Table`.

Use `Design` (after clicking inside the table) to set the name. There is also the `Formulas > Name Manager` to manage cell and table names.

Tables provide some automatic functionality (sorting, summaries context menu, creation of new columns automatically, structured references).

To remove the table functionality, click inside the table and select `Design > Convert to Range`.

To remove the formatting, use `Home > Clear > Clear Formats`.

Split Data in Columns

Add empty columns to the right.

`Data > Text to Columns`

- Player names are in one column separated by “_”. We cannot sort them alphabetically by last name. Clean the player name and separate them into the columns first and last name. Careful, there are some players with middle names or “Jr.” Use sorting to find them.

Clean Text Values

To clean text you can used the functions in `Formulas > Text > …`

- Use the function substitute to clean the position name (replace “_” by “ ”).

Function result depend on the original data. If you want to replace the original value then you have to Convert function results into fixed values using `Copy` (`CTRL + C`) and then `Home > Paste > Value`

Check and Fix Data Quality

It is important to make sure that all data contains valid values (e.g., no player has a weight of 0). You can use

- sorting or
- `Data > Data Validation`

Empty cells often mean missing values.

- Sorting always puts empty cells to the end of the table.
- You can also find cells with `Home > Find & Select > Go to Special` and select Blanks. Selected cells can be replaced with entering a value and `CTRL + Enter`

Sorting, Filtering, Slicers and Conditional Formatting

Data tables have filters and support sorting for each column (in the column header). For example, you can filter by team or position.

`Insert > Slicer` for more convenient filtering. **Note:** Slicers and other objects (e.g., charts) may get hidden when filtering hides rows (only in older versions of Excel). To avoid this right-click on the slicer and select `Size and Properties > Don’t move or size with cell`

Tables can be selected using `CTRL + A` `(A = “all”)` and provide sorting and filtering and a context menu (when selected, not available in all Excel versions). To select columns in a Table, go to the column header and click when you see a down arrow.

- Use the context menu for data tables (right click) to add averages.
- Add conditional formatting. Go to `Home > Conditional Formatting` for more choices.

Hint: Always look at the new menus that appear when you select a table or a chart.

Referencing Cells and Formulas

Column/row references use letters and numbers and are relative to the current position. For example, `=A1`

Absolute references use $ always reference the same cell and not adapt when copied. For example, `=$A$1`.

Data tables use structured references that look like `=Tablename[@[Columnname]]` (references one value)

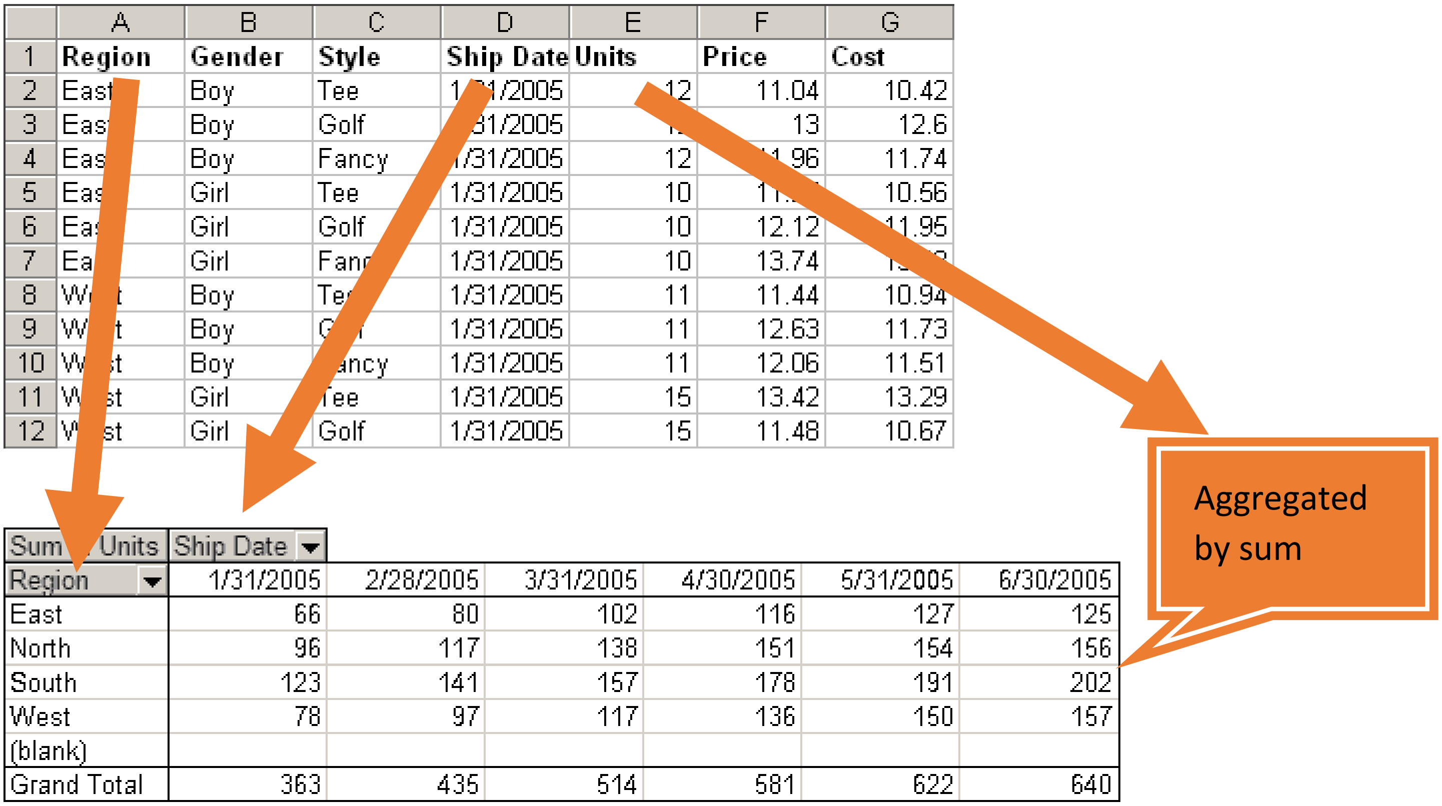
or `=Tablename[Columnname]` (references whole column)

Formulas use references.

- Use `Formulas > More > Statistical > CountIF` to calculate the proportion (percentage) of players younger than 25 years.
- Are some players under/overweight? Calculate the Body Mass Index (look up formula and add a column).

Pivot Tables

Reorganize and summarize selected columns to create a report. Example (from Wikipedia):



	A	B	C	D	E	F	G
1	**Region**	**Gender**	**Style**	**Ship Date**	**Units**	**Price**	**Cost**
2	East	Boy	Tee	1/31/2005	12	11.04	10.42
3	East	Boy	Golf	1/31/2005	[illegible]	13	12.6
4	East	Boy	Fancy	1/31/2005	12	11.96	11.74
5	East	Girl	Tee	1/31/2005	10	[illegible]	10.56
6	East	Girl	Golf	1/31/2005	10	12.12	11.95
7	East	Girl	Fancy	1/31/2005	10	13.74	[illegible]
8	West	Boy	Tee	1/31/2005	11	11.44	10.94
9	West	Boy	Golf	1/31/2005	11	12.63	11.73
10	West	Boy	Fancy	1/31/2005	11	12.06	11.51
11	West	Girl	Tee	1/31/2005	15	13.42	13.29
12	West	Girl	Golf	1/31/2005	15	11.48	10.67

Sum of Units	Ship Date ▼					
Region ▼	1/31/2005	2/28/2005	3/31/2005	4/30/2005	5/31/2005	6/30/2005
East	66	80	102	116	127	125
North	96	117	138	151	154	156
South	123	141	157	178	191	202
West	78	97	117	136	150	157
(blank)						
Grand Total	363	435	514	581	622	640

Highlight the data table and use `Insert > Pivot Table` and then define columns, rows and the aggregation function.

- How many players are there for each position?
- What is the average weight/age/height in each team? Are they different depending on the position?
- Use conditional formatting (check conditional formatting rules), sorting, slicer (team, age, etc.) to create a dashboard.

Visualization

Appropriate visualizations can be created by selecting a column and using `insert > recommended charts`

- Create a histogram for height, weight, and age.
- Create a scatter plot for weight and age.
- Add a bar chart for the Pivot table showing the number of players for each position.

4. Models and Algorithms

[What type of problem do we have? Forecast values, make yes/no decisions, compare data?]

We will learn about models and algorithms later in this course.

5. Communicate the results and/or implement a data-driven product

[Prepare report, visualizations, real-time dash boards. Implement decision support tools or incorporate algorithms in apps and web sites.]

Discuss the patterns that you have found and how they can be used to make decisions about new players.

6. Evaluate the value of the project

[Does the project answer the initial questions?]

7. References

[1] The Official Site of Major League Baseball, https://www.mlb.com/, Visited: February 2021.