Comparing US Wheat Yield with Middle-Aged Marriages in Salzburg

Data Experiment for Data Stewardship

Helmuth Breitenfellner

19.04.2019

Abstract

This paper analyses the correlation between yield of wheat in the United States on one side, and the number of marriages by middle-aged (35-44) men and women in Salzburg.

The source of the project can be found on GitHub: https://github.com/helmuthb/dmp-exercise1. All software and the report is licensed under MIT license. The data is licensed as by the data providers.

- CC0: US Wheat Production Timetable (USDA 2019)
- CC BY 4.0: Age of partners at marriage (Salzburg 2019)

This data experiment is the result of exercise 1 of the lecture "Data Stewardship".

Data Gathering

The data for (Salzburg 2019) is available for download from the Austrian Data portal (https://www.data.gv.at), and the data for (USDA 2019) from the US Department of Agriculture, Economic Research Service (https://www.ers.usda.gov/).

They have been downloaded and are available as part of the repository in the folder `data/source`.

Data Processing

The data from Salzburg is available in CSV or JSON format. For the purpose of this experiment the CSV version has been taken.

The data from the USDA is available as an Excel sheet.

For processing of the data the language *R* is used, which allows both reading of various data sources, statistical analysis (like cross correlation), and plotting of data for visual analysis and comparison.

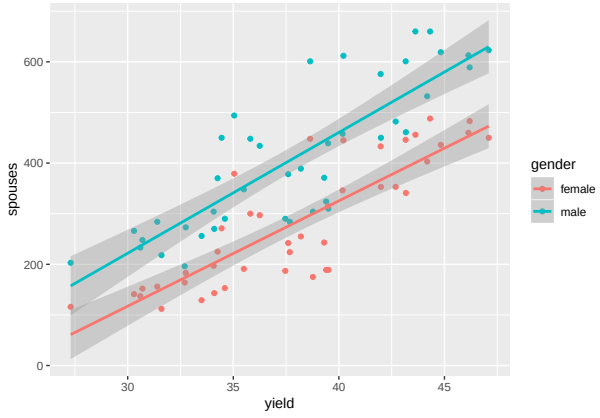
The data files are read in, using `read.csv2` for (Salzburg 2019) and `read_excel` for (USDA 2019).

For (Salzburg 2019) the age classes are transformed into numerical values (the lowest age in each class), and the two columns `wives` and `husbands` are rearranged as two lines with one column `spouses` and `gender` ("female" or "male"). The age range is then filter for the range of interest (35-44 years) and added up.

For (USDA 2019) only a range of the second sheet is read (the region corresponding to wheat). The field `year` is adjusted to numerical values. Originally it contained always two consecutive years, like 1972/1973, to indicate the harvest season. After the adjustment it contains only the second year.

The two data sets were then merged by `year` and stored as a CSV file in `data/processed/us-wheat-sbg-marriages.csv`.

Data Visualization



The previous plot shows wheat yield in comparison to weddings in Salzburg. The corresponding linear regression line has been added for enhanced visualization.

More specifically, looking at the correlation coefficient between husbands in the middle age range (35 - 44

years) and wheat yield gives a coefficient of 0.848 for women and 0.836 for men.

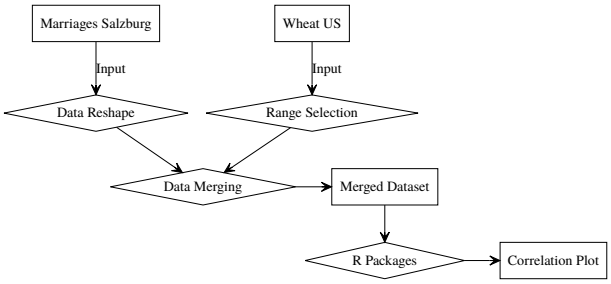
This indicates a high level of correlation.

Technical Description of Experiment

The following *R* packages have been used in this data experiment:

- `readxl` - for reading Excel sheet
- `dplyr` - for data manipulation
- `ggplot2` - for graphical plots
- `DiagrammeR` - for UML diagrams
- `DiagrammeRsvg` - for UML diagrams
- `rsvg` - for including UML diagrams

The following chart shows the steps of the experiment:



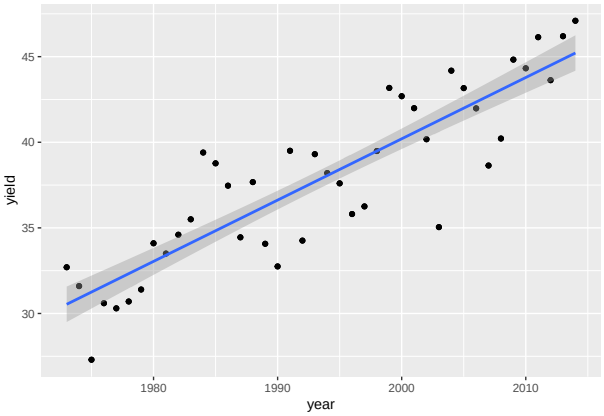
A `Dockerfile` is provided together with the required *R* packages and the LaTeXlibraries used for creating this report. The Docker image is created on Docker Hub and can be used pre-compiled.

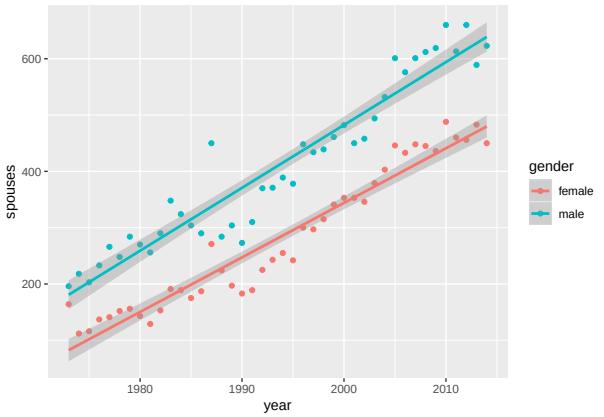
Further Analysis

Clearly one can hardly explain this correlation between marriages in Salzburg and wheat yield in the United States. So what is the reason for this correlation then?

When looking through the data set more in detail, it becomes obvious that the actual correlation is in the year. Both the yield of wheat in the United States is increasing, as is the number of marriages of middle-aged persons in Salzburg.

These correlations are even stronger, the corresponding coefficients are 0.966 for women, 0.956 for men, and 0.875 for wheat yield.





References

Salzburg, Landesstatistik. 2019. Land Salzburg. 2019. https://www.data.gv.at/katalog/dataset/096e7096-8e0a-40a1-9da3-5cceee97d53a.

USDA. 2019. “Wheat Data: Yearbook Tables.” Economic Research Service, United States Department of Agriculture. 2019. http://www.ers.usda.gov/data-products/wheat-data.aspx.