Seminararbeit AI-assisted programming and data analysis
Data Profiling with LLM's a Case Study on HR Data

Vincent Kapp[a,1,*], Prof. Dr. Martin Spindler[b,2]

[a] *Universität Hamburg, M.Sc. BWL, Street Address, Hamburg, Postal Code*

[b] *Universität Hamburg, Statistik mit Anwendung in der Betriebswirtschaftslehre, Street Address, Hamburg, Postal Code*

Abstract

This is the abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum augue turpis, dictum non malesuada a, volutpat eget velit. Nam placerat turpis purus, eu tristique ex tincidunt et. Mauris sed augue eget turpis ultrices tincidunt. Sed et mi in leo porta egestas. Aliquam non laoreet velit. Nunc quis ex vitae eros aliquet auctor nec ac libero. Duis laoreet sapien eu mi luctus, in bibendum leo molestie. Sed hendrerit diam diam, ac dapibus nisl volutpat vitae. Aliquam bibendum varius libero, eu efficitur justo rutrum at. Sed at tempus elit.

Keywords: LLM, Data Analysis, Data Profiling

List of Figures

List of Tables

Please make sure that your manuscript follows the guidelines in the Guide for Authors of the relevant journal. It is not necessary to typeset your manuscript in exactly the same way as an article, unless you are submitting to a camera-ready copy (CRC) journal.

For detailed instructions regarding the elsevier article class, see https://www.elsevier.com/authors/policies-and-guidelines/latex-instructions

*Corresponding author

Email addresses: vincent.kapp@studium.uni-hamburg.de (Vincent Kapp), martin.spindler@uni-hamburg.de (Prof. Dr. Martin Spindler)

[1]This is the author.

[2]This is the Editor

1. Einleitung

Datengetriebene Prozesse können nicht nur Unternehmen beim Optimieren der Prozesse sondern auch bei der entwicklung neuer Strategien helfen. Aber uum festzustellen welche Daten in den Datenbanken vorliegen und wie diese zu einernander stehen braucht es Mitarbeiter*innen die einzelnd überprüfen welche Daten es gibt und in welchen zusammenhängen, ausgehend von Commen Sense, es gibt. Data Profiling stellt Unternehmen mit historisch gewachsene Datenbanksysteme vor Herausforderungen für die Dategetriebene Transformation da. Können die aktuellen entwicklunge von Large Language Modellen (LLM) dazu beitragen, autamtisiert die zusammenhängen von Spalten zu erkennen? Um diese Frage zu beantworten, widmet sich diese Seminararbeit aufbauen von der Arbeit von Trummer 2024, wie gut LLM, anhand einer Case Study, die Korrelation von Daten zu erkennen.

Philipp Bach, Victor Chernozhukov, Martin Spindler (2024). Heterogeneity in the U.S. Gender Wage Gap. Journal of the Royal Statistical Society: Series A, 187(1), 209-230, available online.

Philipp Bach, Victor Chernozhukov, Martin Spindler, Closing the U.S. gender wage gap requires understanding its heterogeneity, Working Paper, available at arXiv, 2018.

Sven Klaassen, Jan Teichert-Kluge, Philipp Bach, Victor Chernozhukov, Martin Spindler, Suhas Vijaykumar, DoubleMLDeep: Estimation of Causal Effects with Multimodal Data, available at arxiv, 2024.

(**?**)

Here are two sample references: **? ?**.

With this template using elsevier class, natbib will be used. Three bibliographic style files (*.bst) are provided and their use controled by `cite-style` option:

- `citestyle: number` (default) will use `elsarticle-num.bst` - can be used for the numbered scheme

- `citestyle: numbername` will use `elsarticle-num-names.bst` - can be used for numbered with new options of natbib.sty
- `citestyle: authoryear` will use `elsarticle-harv.bst` — can be used for author year scheme

This `citestyle` will insert the right `.bst` and set the correct `classoption` for `elsarticle` document class.

Using `natbiboptions` variable in YAML header, you can set more options for `natbib` itself . Example

```
natbiboptions: longnamesfirst,angle,semicolon
```

2. Hintergrund

If `cite-method` is set to `citeproc` in `elsevier_article()`, then pandoc is used for citations instead of `natbib`. In this case, the `csl` option is used to format the references. By default, this template will provide an appropriate style, but alternative `csl` files are available from https://www.zotero.org/styles?q=elsevier. These can be downloaded and stored locally, or the url can be used as in the example header.

2.1. Data Profiling

2.2. Korrelation ermittelnt

2.3. Sprachmodelle

2.4. NLP for Data Bases

2.5. Benchmark

2.5.1. Equations

Here is an equation:

$$f_X(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}; \alpha, \beta, x > 0.$$

Inline equations work as well: $\sum_{i=2}^{\infty}\{\alpha_i^{\beta}\}$

2.5.2. Figures and tables

?@fig-meaningless is generated using an R chunk.

#{r} #| label: fig-meaningless #| fig-cap: A meaningless scatterplot #| fig-width: 5 #| fig-height: 5 #| fig-align: center #| out-width: 50% #| echo: false plot(runif(25), runif(25))

3. Tables coming from R

Tables can also be generated using R chunks, as shown in **?@tbl-simple** example.

#{r} #| label: tbl-simple #| tbl-cap: Caption centered above table #| echo: true knitr::kable(head(mtcars)[,1:4])

3.1. Benchmark Data

3.2. Benchnmark Metrics

4. Benchmark Analysis

5. Comparing Prediction Methods

5.1. Description of Methods

5.2. Experimental Setup

5.3. Comparison Results

6. Scenario Variants

7. Results Breakdown

8. Other Correlation Metrics

9. Column Types

10. Conclusion

References