TU Dortmund University
Department of Statistics

Prof. Dr. Andreas Groll

Application Report Master Data Science

Summersemester 2023

Dear Applicants,

many thanks for your interest in our study program M.Sc. Data Science. In order to successfully study at our department, we have to ensure that you bring all necessary skills with you. One of the most important skills as a Data Scientist is the ability to understand, solve and present the results for a given data task. Therefore, we ask you to solve the following data analysis task. We require you to write a report of exactly 10 pages presenting the results of your analysis. **It is mandatory to add the report in pdf-format to your uni-assist application.** The report can be written in English or in German language.

Only applicants whose reports are evaluated by us as at least sufficient can be considered in the further application process. Therefore, please take the writing of the report seriously; you should plan for a working time of approximately 20 hours. Please use our self-disclosure document to inform yourself about our additional enrollment criteria. You should only start this task after being sure that your Bachelor degree qualifies for our Master. In case of doubt contact our study advisory service (study-advisory@statistik.tu-dortmund.de).

You can find the data for the application period **summer semester 2023** under this link:
`https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Studium/Studiengaenge-Infos/animal_speeds.csv`.
The data comes from the animal kingdom, covering both the weight and the highest movement speed of several different animal species. We are interested in how the average weight of an animal species influences the highspeed of that species. The intuition here is that larger animals are generally faster than smaller animals, however, the heaviest animals are not the fastest ones.

In the data set, you find information on 159 animal species, including the average weight kilogram, the highspeed in kilometers per hour and the movement type (climbing, flying, running or swimming). The data is collected from various internet sources, since both the average weight and the highspeed of an animal species are not easy to measure, and hence, are not publically available for all species. For example, the average weights for male and female animals often differ. Moreover, there is a selection bias in the data, since highspeeds are often only known for the fastest animals.

a) Perform a detailed descriptive analysis of the data set. Use appropriate statistical measures to describe the variables in the data set. Make sure to include at least one statistical graphic. The analysis should show in a descriptive way how the average weight influences the highspeed.

b) Perform a linear regression to analyze the relationship between the variables weight and highspeed taking into account the covariable movement type. Think about useful data transformations. Interpret the coefficients of the linear model(s).

You are free in the choice of methods, as long as the task is answered appropriately. Your report should contain an introduction and a short description of the task and the data set. Moreover, you have to describe all statistical methods you used. Please do so in a mathematical way: If possible, give mathematical definitions of the methods. The report should end with a summary and a discussion of the results. Be sure to do proper literature work: If you use a method, cite a book in which it is explained. When you make a statement, support it with a proof or a literature citation.

You are already familiar with writing statistical reports? That's pretty good, feel free to just get started. For all the others, we recommend to read the guidelines on the following pages.

We wish you best luck with the report and with your application!

Important notes on writing statistical reports

On the following pages, we will give you a lot of advice on how to write a statistical report. There are some things you should stick to, some other aspects are optional. Much depends on the task at hand and your approach to solving it.

- **Layout**: The final layout is your own choice, however, some rules of thumb: Line spacing 1.5 times, font size neither too small nor too large (e.g. 12p for Times New Roman), uniform font throughout the report, sufficient margins.
- **Structure**: The following list is a basic framework for the layout of a statistical report. It gives relevant information about the required sections and some basic information about their contents. An individual report can deriviate from this structure, e.g. by introducing additional subsections, but at its core, the report should follow that structure.

  i) Title page (1st page): It should include your name, the name of the project, the current date and it should have a nice layout. This is the first page we see from your report, so the first impression should be a good one. Does not count for the site limit.

  ii) Table of contents (2nd page): List all sections and subsections, including appendix and bibliography with the respective page numbers. This should be done as simply and clearly as possible. Most programs can automatically create a neat table of contents, you should use such functionality. Does not count for the site limit.

  iii) Introduction / motivation (approximately 1 page): Give a short motivation for the given problem. Why should one be interested in solving it? Do not write: *I am doing it because it is the application report*, but give a real-world motivation. Briefly describe the exact problem, your approach to solve it and present the main result: this is not an adventure story that has to be thrilling, on the contrary, you should state the final result in the introduction. At the end of the introduction, give an overview on the structure of the report.

  iv) Detailed description of the problem (approximately 0.5 to 1 pages): What is the exact task? Do not copy-paste from the task definition, and try not to repeat what you have already written in the introduction. Be as exact as possible, give some *research questions* you want to answer in this report. Also, have a first look into the data, from a technical point of view: Where does the data come from? How were they collected? What are the variables, what is the scale level of the individual variables? Are there any missing values? Do not make a long list (a table in the appendix is sometimes a good alternative), and keep the task in mind. Do not give unnecessary information that is irrelevant to the task.

  v) Methods (approx. 3–4 pages): Give a sound description of the (statistical) methods used in the report. You can expect that we know all basic mathematical stuff, like, what is a sum, a logarithm, an arithmetic mean. Any advanced statistical method you use, like, for example, a histogram, a t-test or a decision tree has to be described and defined. Make sure the description is mathematically sound, you can define most methods in a pure mathematical way. Try to do so. In addition, explain how the methods works. You can also mention advantages and disadvantages of the method. Add references to appropriate literature, you did not came up with the method yourself. This section can be further divided into individual subsections of the described methods.

  vi) Evaluation (approx. 3–4 pages): In the evaluation, answer the questions given on the first page using appropriate statistical methods. This evaluation can be further divided into individual parts for the individual subtasks. You should make one subsection for the descriptive

analysis. Give us insight into the data. Add some statistical numbers, for example, describing the location and dispersion of the data. And, this is a must: Add a statistical graphic. Please choose a graphic that is appropriate for the task, make sure that it looks nice, that it shows the relevant information, that it does not contain any unnecessary information and it must be self-explanatory. In the second subsection, answer the second task. Justify the choice of your method. Apply the method and show the result. Another graphic might be necessary here, depending on the used method. Be neutral, don't interpret the results yet. Let the numbers speak for themselves.

vii) Summary (approximately 1 page): Brief recapitulation of the project's research question and of the most important results. Try to answer every research question in the summary. Discussion and interpretation of the result in context of the real world context. Open question, further topics that could or should be analyzed. The summary should be readable and understandable without reading the remainder of the report.

viii) Bibliography (1 page or more): Please list all the literature you used. A citation must include (if applicable): the names of the authors, the title of the publication, the year of its publication, the name of the journal, the publisher and page numbers. For web pages, the date of the query must be included. All papers listed in the bibliography must be cited in the text and vice versa. Do not forget to cite the software (and all additional packages / libraries) you used. Does not count for the site limit.

ix) Appendix (as long as necessary, as short as possible): Additional tables and figures that did not fit into the main report. General rule: try not to use an appendix. If you need an appendix, keep it as short as possible, at most: 3 pages. Does not count for the site limit.

- **Tables and figures**: If you decide to add tables and graphics (you have to add at least one graphic), please stick to the following guidelines:
  - All figures and tables must have either a sub-heading or a heading. Please be consistent (i.e. always a heading for figures, always a sub-heading for tables).
  - Figures and tables are numbered consecutively, even figures and tables in the appendix.
  - Each figure and table must be referred to at least once in the text (this also applies to those in the appendix). The corresponding number of the figure or table is used for this.
  - An illustration and a table typically appear on the page or following page of the first response in the text, but never before.
  - Figures and tables should be self-explanatory. This includes that all axes are labeled. Additional relevant information can, for example, be presented in a legend.
- **Citations**: There are several different ways on how to cite literature within your report. Neither of them is wrong, as long as you consistently stick to one of them. We recommend to use the author (year) or the (author, year) style, it adds much more readability compared to just numbering your literature with [1], [2], [3] and so on. If there are two authors, both should be indicated, if there are more than two authors the abbreviated form Author 1 et al. (year) or (Author 1 et al., year) should be used. If there are several author from the same year, they are additionally differentiated by letters after the letters after the year numbers: (2005a), (2005b). When citing from books, it is often useful to specify page numbers.
- **Mathematical formulas**
  - Mathematical formulas and symbols (whether indented or in body text) must be set off from the rest of the typeface

  - Once chosen, designations should be retained throughout the report
  - Mathematical formulas can also be numbered so that they can be referred to in the text

- **Further advice**
  - Keep the thread of the report in mind. Do not write unnecessary things. For example, do not introduce methods that you will not use in the end. At every point in the report, it should be clear why you are explaining it and what your goal is. You have one goal in mind: to answer your research questions. The rest of the report is for that purpose only.
  - Linguistically, precise and factual/scientific formulations should be chosen. Do not use passive wording. Avoid: Experiential style, colloquial language, filler words, too many repeated words and long nested sentences. Grammar and spelling should be error-free.
  - Quotations must be identified as such (indentation of the text, quotation marks, exact indication of the book/article and page number) and must not consist of longer text passages. **Copying of text passages is not allowed! If longer passages are copied or even quotations are not marked as such, the report has to be evaluated as plagiarism and counts as failed.**
  - Be consistent with yourself. If there are two ways to write a word (dataset and data set), stick to one of the two ways. In many places, decisions must be made about the appearance of the report. Decide on the variant you like better and stay consistent with that decision. There is no one perfect report, each of us writes differently.