1. What and how have I learned?

1.1. Starting with an overview

I started my learning journey in Data Science during the winter semester break 2020. Together with a peer of mine, we structured the learning and pathway for the semester. It was a good starting point for the topic.

1.2. Refreshing my skills from prior learnings

I checked my old literature from statistic classes from my former studies of psychology. Modules like *The Science of Methods I-III* and *Psychological Diagnostics* covered topics like probability, descriptive statistics inferential statistics, and hypothesis testing.

- In the module *The Science of Methods I*,I learned content like measurement, uni- and bivariate descriptive statistics, basics of inferential statistics, t-tests, simple analysis of variance, and the consolidation and practical application of statistical methods.
- The content covered by the module *The Science of Methods II* was theoretical derivation and testing of hypotheses, multi-factorial analysis of variance with and without repeated measures, and non-parametric evaluation procedures.
- In the module Psychological *Diagnostics*, I covered topics like interrelationships between statistical procedures including their repetition (ANOVA and regression), the extension of the range of methods to multivariate procedures, MANOVA, multiple regression, non-linear regression, structural equation models, and multilevel models.

1.3. Creating my learning documentation and checking my learning progress

During my time at CODE, I learned that having all notes in one place works best for me. Therefore, whenever new content came across my way, I documented it in my notion-learning book. I planned via my learning with a 'Learning plan' and a checklist 'qualification objectives for module SE 25' and documented my tutorials in a 'cheat sheet'. Throughout the entire learning journey, I took notes from all different learning sources, and it was a process of collecting and adding new information about a topic I have not understood before. That helped me to grow my knowledge base.

1.4. Starting the own project

I got caught by the topic 'data science' and I started my project earlier than planned. With only a few outstanding tasks I could solve all analyses that I planned. To double-check my work, I watched the Pandas tutorial and the Seaborn tutorials. Based on my judgment I have covered all qualification goals of module SE 25 about handling data, data exploration, and effective communication.

2. What did I set out to learn and to build?

2.1. Why did I decide to take this module?

I chose the module 'data science' to learn software libraries written for the Python programming language for data manipulation and analysis. Also, the topic of data visualizations with software libraries is an area of interest for me. Therefore, I started watching tutorials on Pandas and Seaborn. Before finishing them I also started guided exercises to test my new knowledge. Some of the concepts I could already use, others I looked up during the process. Working with a jupyter notebook was new to me and after a brief moment of 'oh-my-god-how-does-it-work', I fell in love with it. Today, I use it regularly and enjoy it a lot.

2.2. In which part was I successful?

Did I learn what I set out to learn? I enjoy the topic 'data science' and I love finding out new elements that I do not know. Working with Pandas, NumPy and Seaborn feels like a new standard and it feels good.

After my midsemester check-in at the beginning of April, I completely trashed my work, started from scratch and dove into the world of machine learning. I had never looked into the topic before and watched a DataCamp course on it to get a first understanding of the concept. In the process, I learned the theory about the machine learning library Extreme Gradient Boosting (XGBoost). I understood how this algorithm uses, adapts, and combines many base learners to make a final prediction based on the whole algorithm. To reduce the variance of each base learner, I studied the meta-algorithm 'boosting'. I finetuned the hyperparameters (learning rate, max_depth, and min_child_weight) using GridSearch. I learned how to evaluate a model based on accuracy considerations and cross-validation and applied this knowledge. I also considered topics like overfitting and underfitting. The most exciting part for me was when I was able to derive insights from the ML model. SHAP values highlight the contribution that each predictor variable has to the dependent variable -that is a concept that I find very compelling. With their help, I was able to identify the variables with high relevance for the model prediction.

What was fascinating for me was that I could check the findings using descriptive and inferential statistics. With this step, a refinement took place because I could keep or discard identified variables, and I was sure in my decision because of the statistical check.

Bottom line, I have never learned and applied so many new skills within a month. I am thrilled with the possibilities that ML models offer in testing connections. I am curious to explore this topic further, for example, how predictions from a machine learning model differ or are similar to predictions from statistical analysis of variance. That is an area where I could relate my knowledge from my psychology degree well to my current studies.

2.3. Where were you less successful?

As I am learning new software libraries, I sometimes do not know how to solve things, but I got used to checking stack overflow and the library descriptions for help. There is one topic I keep getting confused about, even though I have used it so many times: what visualization options does matplot offer, and what seaborn. In particular, the representation of many plots in an overview is implemented differently in both libraries. I find this irritating. Another topic that I could not solve satisfactorily concerns the import of functions from other Jupyter notebooks. My original solution with the nbimporter library, which solved everything via 'import nbimporter // from other_notebook import function name', worked smoothly for me. Some of my fellow students, to whom I gave the files for testing, got an error message. I also read that the developer of the nbimporter library no longer supports it. Therefore, to make sure that the functions work, I integrated them into the main document - which is not the cleanest solution. How an import from one notebook to another has to be structured so that it always works smoothly is something I will research further.

3. Which skill level do you think that you have achieved, and why?

As of now, my current knowledge of descriptive statistics covers the skills needed for profoundly analyzing data. Handling data with Python libraries works well and communicating results effectively I feel well prepared, and my visualization skills are very proficient. To complete the assignment, I additionally learned about a topic that was completely new to me and very successfully integrated the new knowledge into my work. This becomes particularly evident in the three-stage analysis process that I implemented in my work. Therefore, I think it is appropriate to aim for a level 2.