Exam projects from Advanced Applied Statistics 2021-2022

This is a list of projects from the 2021 and 2022 editions of *Advanced Applied Statistics and Multivariate Calculus* in the Data Science master. This list is provided just for inspiration. It's *not* a list of topics to choose from! Also, note that **the course literature and coverage of methods taught has changed** since 2022!

Topic	Dataset	Statistical methods
Statistical Analysis of a Chess Game Dataset	Chess games (Kaggle)	Bayesian models
Modeling Time Series Data using Bayesian Approaches in Recurrent Neural Networks	4 datasets from Kaggle (Temperatures, Food prices, Mineral prices, Stock prices)	Bayesian models
Correlations between regional foods and ingredients	Self-scraped dataset of recipes	Clustering, dimensionality reduction
Applied statistics on ten social dimensions	Reddit posts annotated for social dimensions (Kaggle)	Clustering, dimensionality reduction, Bayesian models
Statistical exploration of salary in the IT sector	Salary data (Kaggle)	Bayesian models
Prevention of overfitting	3 datasets from Kaggle (simulated data, human activities and postural transitions, League of Legends)	Regularisation, Bayesian models
Classifying Pneumonia from Chest X-Ray Images using Linear Discriminant, Principal Component and k-Means Clustering Analysis	Published dataset from a research paper	Dimensionality reduction, clustering
Comparative Study on Performance of Markov Chain Monte Carlo (MCMC) and Automatic Differentiation Variational Inference (ADVI)	Self-generated synthetic data	Bayesian models
Understanding the Certain Power of Uncertainty in Bayesian Neural Networks	MNIST (handwritten digit recognition)	Bayesian models
Geospatial Clustering Analysis on Crimes in Denmark	Crime data from statbank.dk	Clustering
Comparative performance study on Varying Bayesian Neural Networks Approaches	Occupancy detection (UCI); Heart disease, Ozone levels (Kaggle)	Bayesian methods
Dirichlet Process Mixture Model to Compress Convolutional Neural Network	Pretrained neural network weights	Weight clustering
Data-driven approaches to estimate market value of football players	Soccer player's values and statistics (Kaggle)	PCA, Lasso/ridge regression, Random forests
Capital region real estate price market analysis	Self-scraped data from boliga.dk	Bayesian regression, PCA, clustering
Modelling team performance in major league baseball	Datasets from seanlahman.com, baseball-reference.com, fangraphs.com	Linear regression, PCA, sampling

Structural complexity of Romanic and Germanic names	Dataset of Facebook names; Europarl	Information theory, clustering
Impact of demographic factors on income	US census income dataset	Linear regression, PCA, clustering
Statistical analysis of political affiliation test	DR political candidate test	PCA, clustering, propensity score matching
Applied statistical testing: An introduction	various	Hypothesis testing
Assessing the agreement between common feature selection methods on feature importance for different regression tasks	US house prices; Austing house prices; AirBnB price; Bike sharing; Car prices	Ridge and Lasso regression; Regression trees; Random forests
Clustering countries for humanitarian aid	Health and socioeconomic statistics (Kaggle)	Clustering
Finding important factors of customer churn using regularization and dimensionality reduction	IBM Telco customer churn dataset	Logistic regression, regularisation, clustering, PCA
Analyzing statistical methods to predict wind energy generation	Weather data (UK met office), Windmill power data (Scottish Electrical Network)	Linear regression, SVM, Multi-layer perceptron
Uncorking the secrets of wine clustering	Wine dataset (UCI ML repository)	Clustering, EM, mixture models
Evaluating fidelity of synthetic image edits: An analysis of k-means and t-SNE	Self-generated images from denoising diffusion model	Clustering, t-SNE