PhD stipends

This project aims to analyze the data on doctoral scholarships available at Kaggle and build a model to predict remuneration. A future student could use this predictor to estimate the approximate salary in the chosen university and area of knowledge. This source presents data from 2012 to 2020.

primary predictor variables available on dataset

Universities
Departments
Program tradition
Living wage ratio

target variable available on dataset

Overall pay

secondary predictor variables created based on dataset

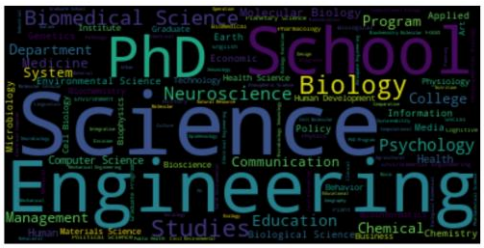
State
Area of knowledge

ETL phase

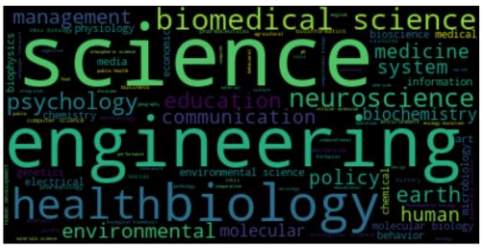
1	Unnecessary columns cleaning Univesity null entries cleaning
2	Universities text entries grouped by similarity 'University' -> 'State' dictionary creation Non-US universities cleaning 'State' column creation based on dictionary
3	Design of a classification function that links each department to one area of knowledge Area of knowledge column creation
4	Overall Pay column processing: treating non-numeric data and conversion to numerical type

Exploratory data analysis phase results

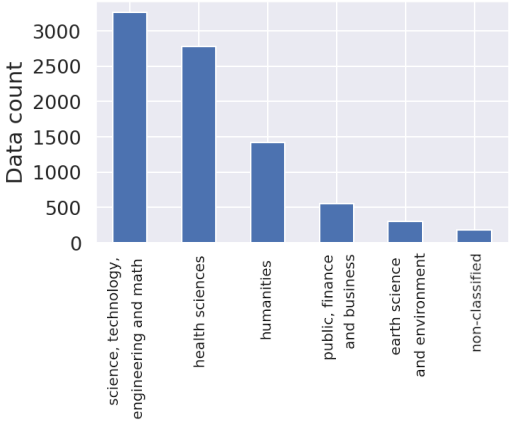
Departements Wordcloud
Before text cleaning



After text cleaning:
Phd, School, Program, and other unnecessary words were removed



Data count after knowledge area classification



Next steps:

- Filter Data by year
- Present the Overall Pay depending on the state for each year [2012:2020]
- Compare the Overall Pay with living cost ratio
- Analyze the correlation matrix to see if the predictor variables are not collinear
- Begin to train and test models (KNN, decision tree, random forest, and others)
- Best model selection and predictor deployment on the internet

https://github.com/amandaventurac/PhD_salaries