PhD Stipends in the United States: To Maximize Overall Pay While Studying Engineering, Head Towards Stanford University!

1. Original Dataframe

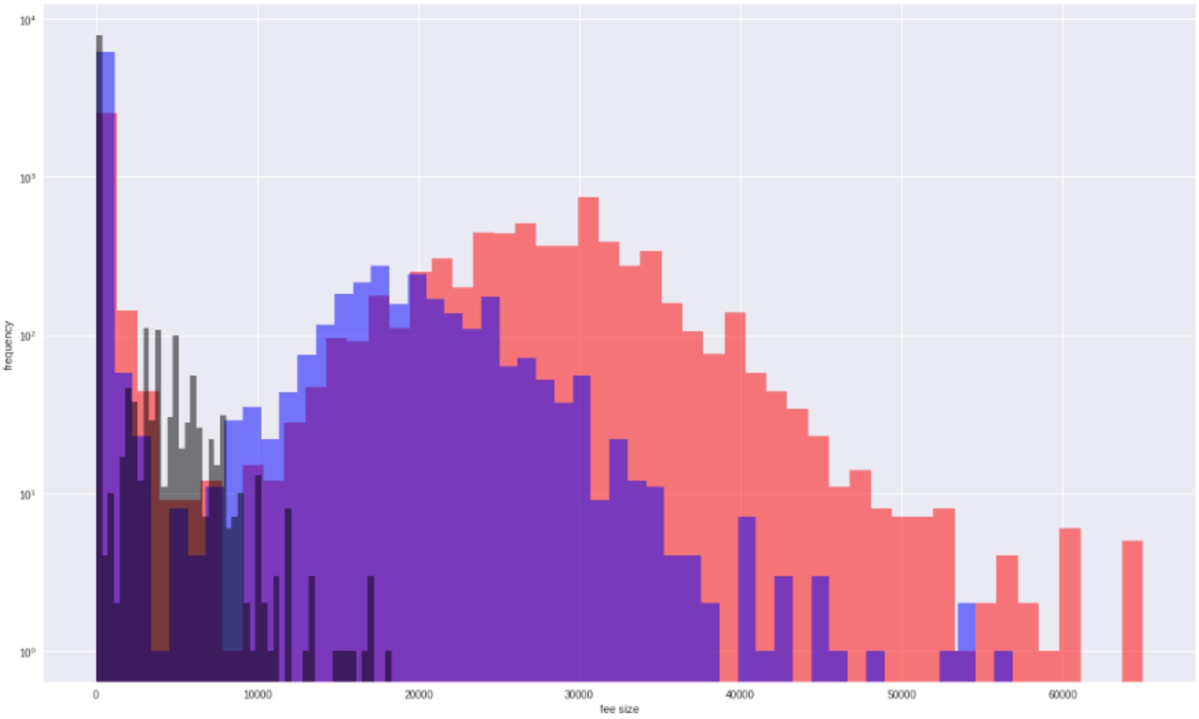
On July 13, I downloaded the PhD stipend dataset (8707 observations, 11 columns each) from kaggle. The more I cleaned and investigated, the more conspiracy theories I started to generate about what aspects of the mess must for sure have been introduced on purpose. The dataset is heavily biased towards responders from the United States. If PhD students indicated being paid on a per 9 months scheme, an additional field appeared where gross pay for the 3 summer months could be entered.

2. Data Cleaning and Outliers

Since neither an overall pay nor fees of 1 Million USD seem credible to me, I dropped all observations containing either of:

- 12 Months gross pay > 65’000 USD
- 9 Months gross pay > 60’000 USD
- 3 Months gross pay > 20’000 USD
- fees > 58’000 USD.

➔ *Bimodal distribution of the remaining 12M (red) / 9M (blue) / 3M (black) values:*



I further dropped all observations containing living wage ratios that were either negative (I don't believe that there are areas giving you money for free), or equalled zero (-> zero divisions).

If any of

- university
- department
- program year
- academic year
- overall pay
- low wage ratio

was missing, I deleted the entire row as well, since these observations are of not much value.

Finally, I removed 121 remaining duplicates.

3. Gross Pay (GP) and Fees – Missing Value Imputation

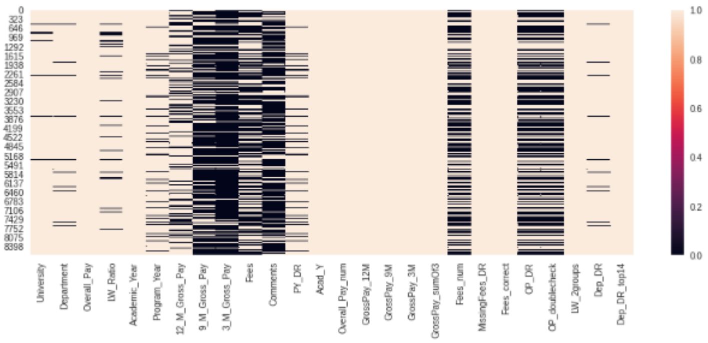
I interpreted missing values in the three gross pay variables as true zeros and recoded them as such.

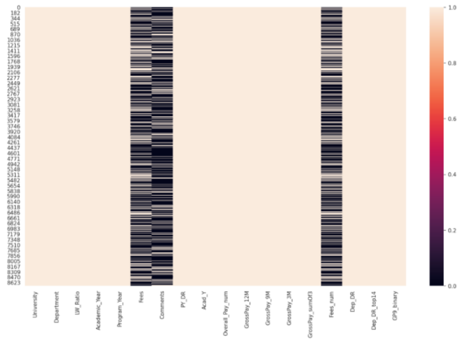
I added up the three gross pay variables for each observation and noticed a perfect match between all available Fee values and (GP3 + GP9 + GP12 – Overall Pay). I thus recalculated the missing fees using this formula. The resulting distribution of the missing fees very closely matched the distribution of the available fees.

The interesting part about this is, that the survey did not contain a field for overall pay, this was calculated from the indicated GP's and fees, instead. Some funny guy must on purpose have removed the fee values for more than half of the observations.

Left: missing values (black) in original dataframe

Right: dataframe after cleaning, before imputation of missing values for fees. Besides fees, comments was the only other column that contained missing values (interpretation: no comment).





4 Slicing the Cleaned Dataframe to Obtain Different Subframes & Department Recoding

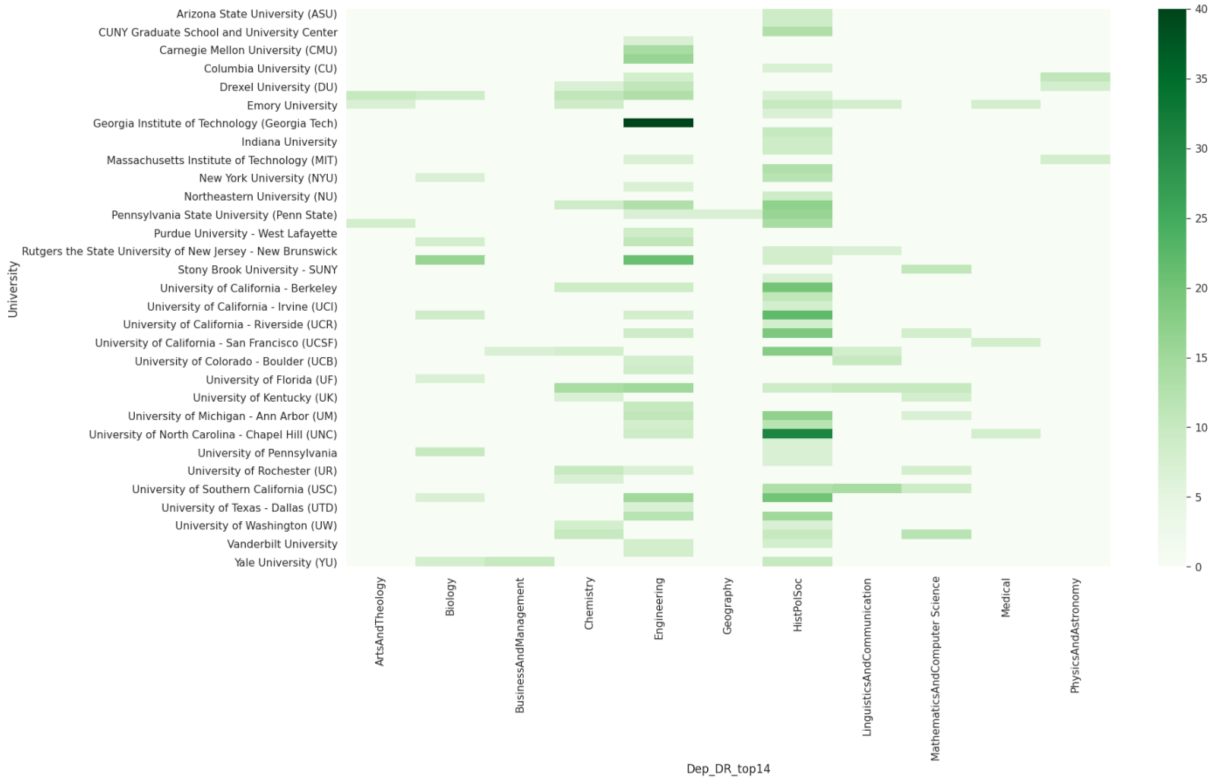
I decided to restrict the analysis to

- universities that occur at least in 20 unique observations (6604 rows remained) and
- university x department combinations that were reported at least 6 times (→ 5429 rows).

I did not use the original version of department labels, but recoded them into 13 categories that appeared reasonable to me: Engineering / Mathematics & Computer Science / Physics & Astronomy / Chemistry / Biology / Medical / Geography / Education / History & Politics & Sociology & Psychology (please forgive me!) / Linguistics & Communication / Business & Management / Arts & Theology. I stopped manual recoding after some years of work ;) and smashed not yet recoded values into a common category called "others".

I also created another subframe where I dropped all observations of "others" (→ 3054 rows).

Heatmap of the crosstab for (university x department) of the 3054 row version:

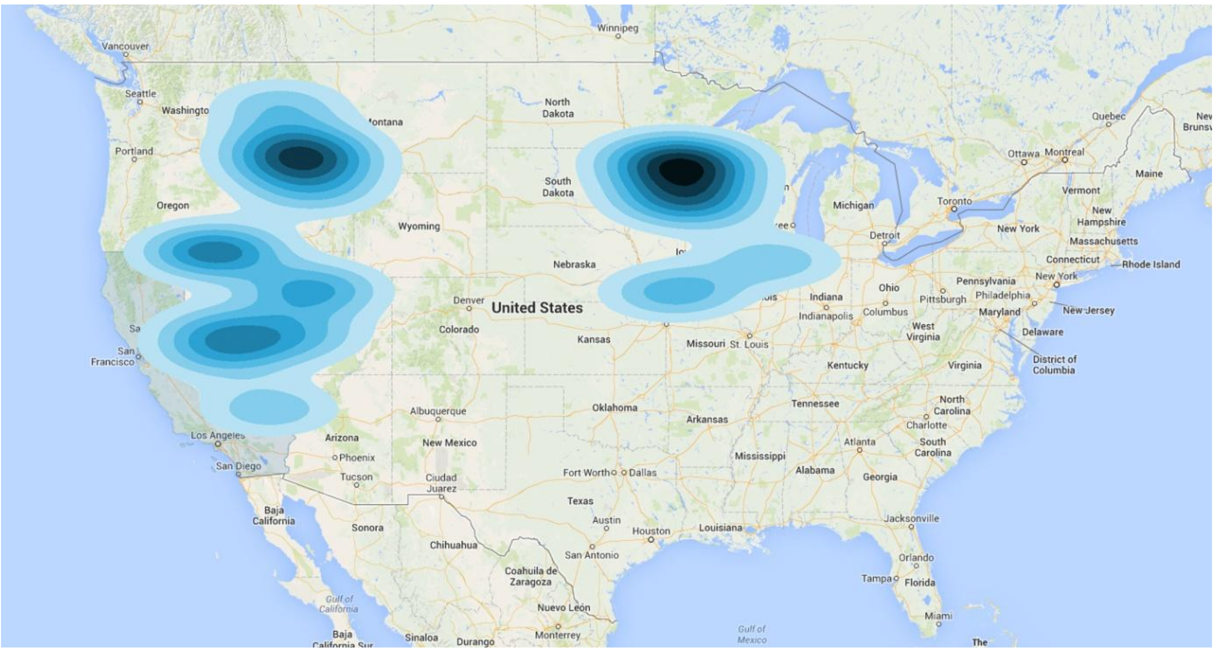


5 Which Are the Most Frequent Universities in the Clean Dataset?

All of the 20 most frequent universities are within the United States.

I looked up the GPS coordinates for these 20 universities and mapped them to the dataset using one dictionary each for longitude and latitude.

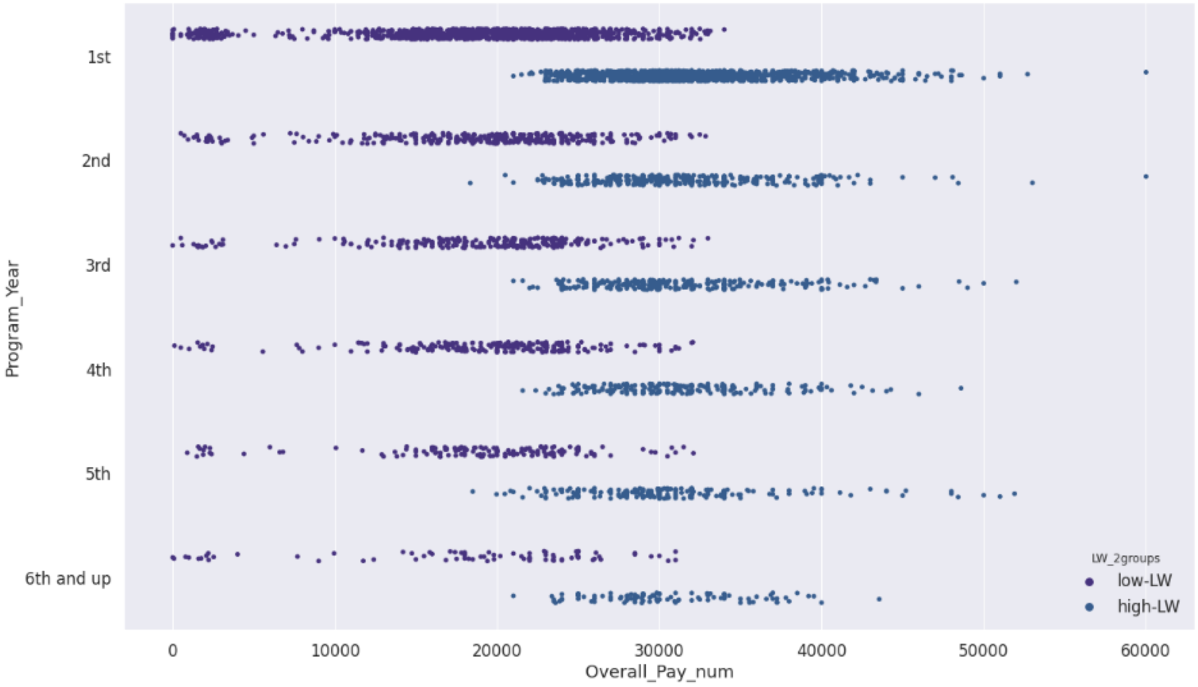
The most frequent universities are clustered in specific subregions of the United States:



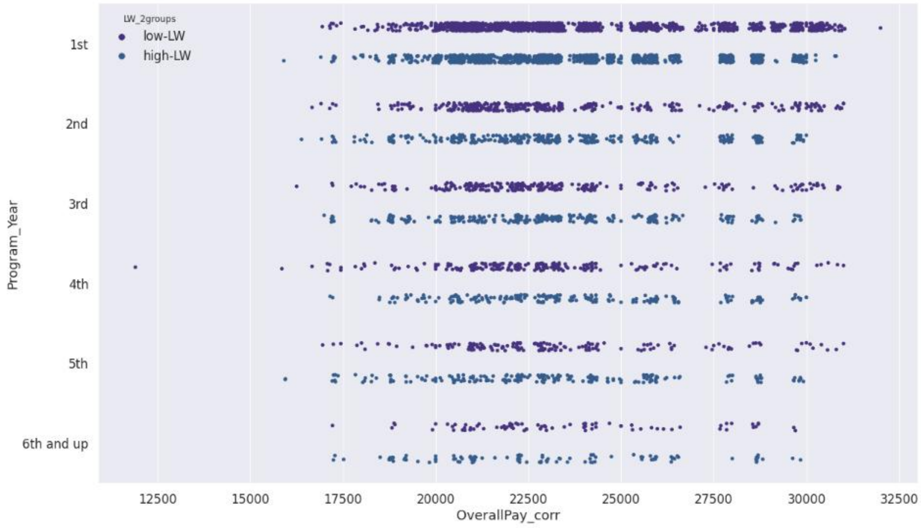
6 Adjusting Overall Pay & Fees by Living Wage Ratio (LWR)

I really wonder where the LW Ratio comes from, since it was not directly included in any of the survey fields (IP address of computer?). To make overall pay and fees comparable between different observations, I divided these by the living wage ratio to create LWR-adjusted versions of the two.

Overall Pay by Program Year is higher in observations with high LWR (blue) than in those with low LWR (dark blue):



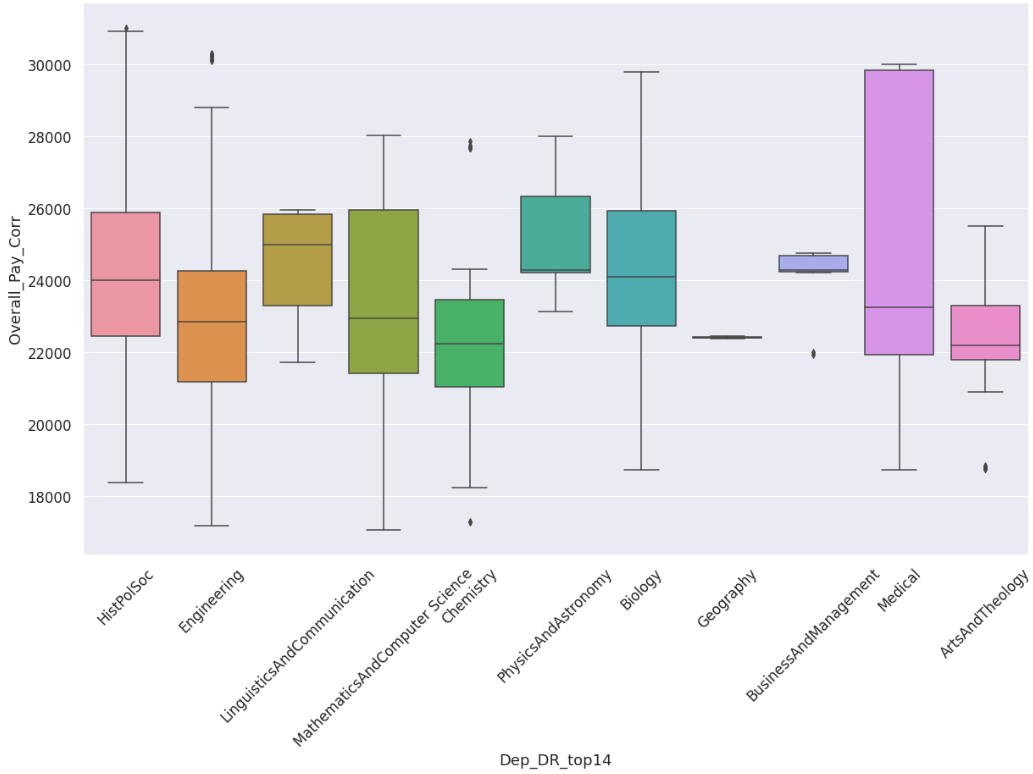
After adjusting for LWR, however, the world looks pretty fair:



I would have liked to compare the distribution of LWR in my dataset to the one of the entire population of the United States, but I spent some days in the mountains, and also my kiddies claimed some of my holiday week's time. ☺ Unfortunately, I could not even find a precise definition of living wage RATIO (which x is divided by which y? …). I would intuitively associate lower LWR with poorer regions and higher LWR with richer ones. In any case, LWR follows a heavily left-skewed distribution.

7 Getting Down to the Nitty-Gritty: Comparing Overall Pay Between Departments

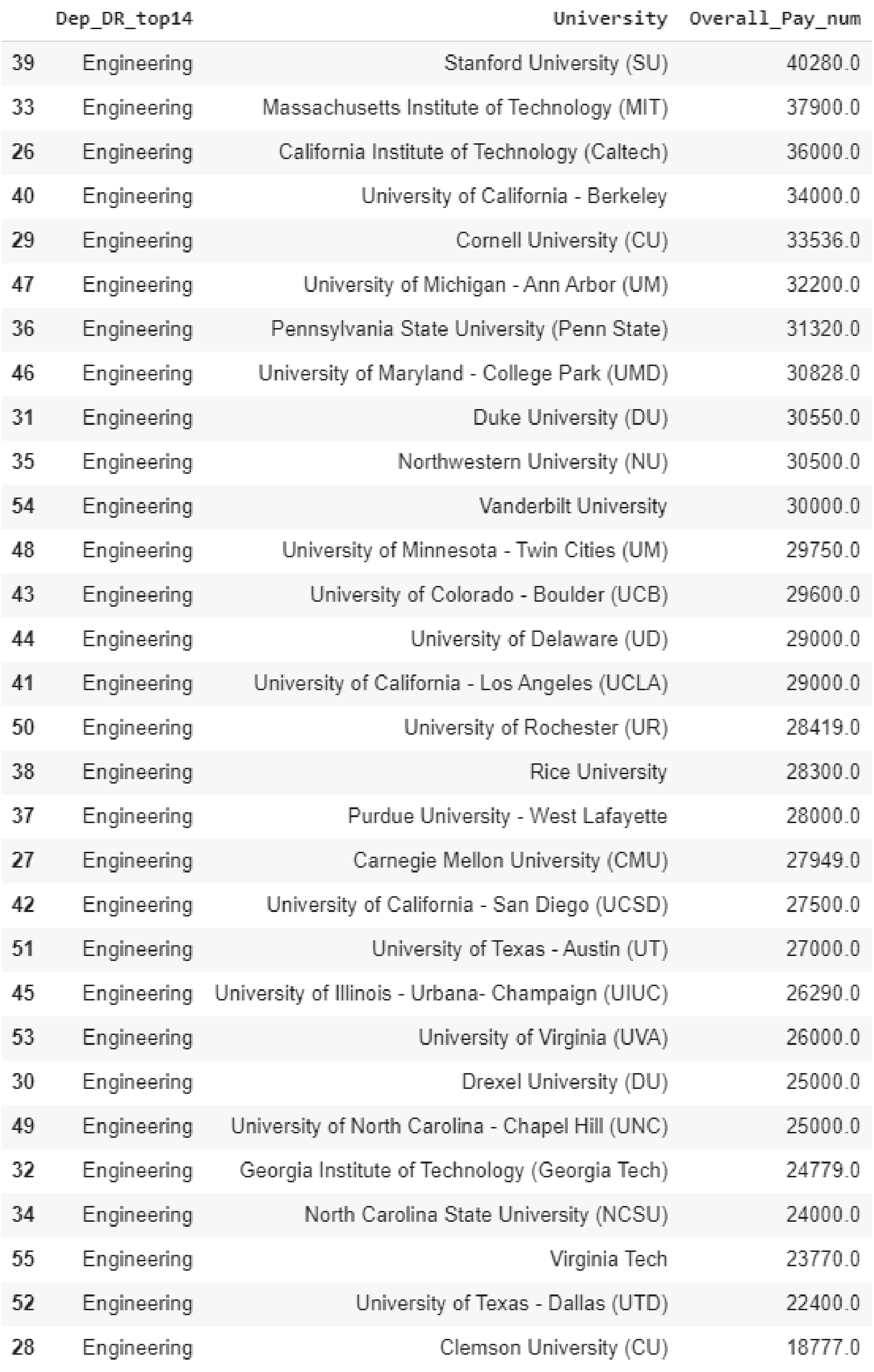
Overall Pay values adjusted for LWR are mostly in the range of USD 17'000 – 30'000.



Unfortunately, my Jupyter notebook is too chaotic for me to extract the means for the various departments in a reasonable amount of time.

8 In Which University do Engineering Students Get the Highest Stipends (unadjusted)?

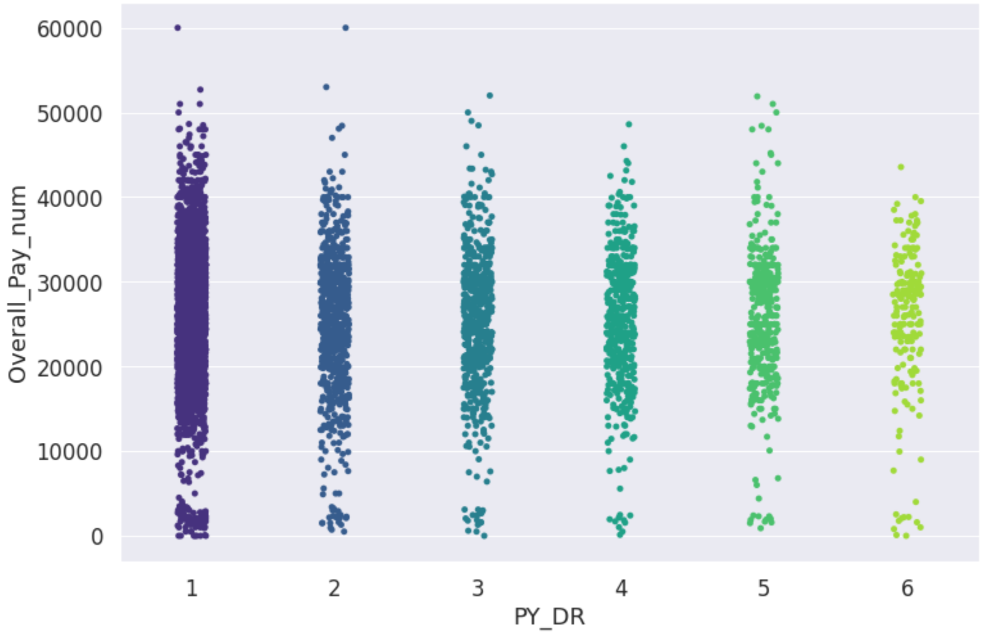
Unlike in question 7, overall pay was not adjusted for LWR here. Stanford University is on rank 1, followed by MIT and Caltech. Engineering students at Stanford get more than twice the amount of those studying at Clemson university (last place):



	Dep_DR_top14	University	Overall_Pay_num
39	Engineering	Stanford University (SU)	40280.0
33	Engineering	Massachusetts Institute of Technology (MIT)	37900.0
26	Engineering	California Institute of Technology (Caltech)	36000.0
40	Engineering	University of California - Berkeley	34000.0
29	Engineering	Cornell University (CU)	33536.0
47	Engineering	University of Michigan - Ann Arbor (UM)	32200.0
36	Engineering	Pennsylvania State University (Penn State)	31320.0
46	Engineering	University of Maryland - College Park (UMD)	30828.0
31	Engineering	Duke University (DU)	30550.0
35	Engineering	Northwestern University (NU)	30500.0
54	Engineering	Vanderbilt University	30000.0
48	Engineering	University of Minnesota - Twin Cities (UM)	29750.0
43	Engineering	University of Colorado - Boulder (UCB)	29600.0
44	Engineering	University of Delaware (UD)	29000.0
41	Engineering	University of California - Los Angeles (UCLA)	29000.0
50	Engineering	University of Rochester (UR)	28419.0
38	Engineering	Rice University	28300.0
37	Engineering	Purdue University - West Lafayette	28000.0
27	Engineering	Carnegie Mellon University (CMU)	27949.0
42	Engineering	University of California - San Diego (UCSD)	27500.0
51	Engineering	University of Texas - Austin (UT)	27000.0
45	Engineering	University of Illinois - Urbana- Champaign (UIUC)	26290.0
53	Engineering	University of Virginia (UVA)	26000.0
30	Engineering	Drexel University (DU)	25000.0
49	Engineering	University of North Carolina - Chapel Hill (UNC)	25000.0
32	Engineering	Georgia Institute of Technology (Georgia Tech)	24779.0
34	Engineering	North Carolina State University (NCSU)	24000.0
55	Engineering	Virginia Tech	23770.0
52	Engineering	University of Texas - Dallas (UTD)	22400.0
28	Engineering	Clemson University (CU)	18777.0

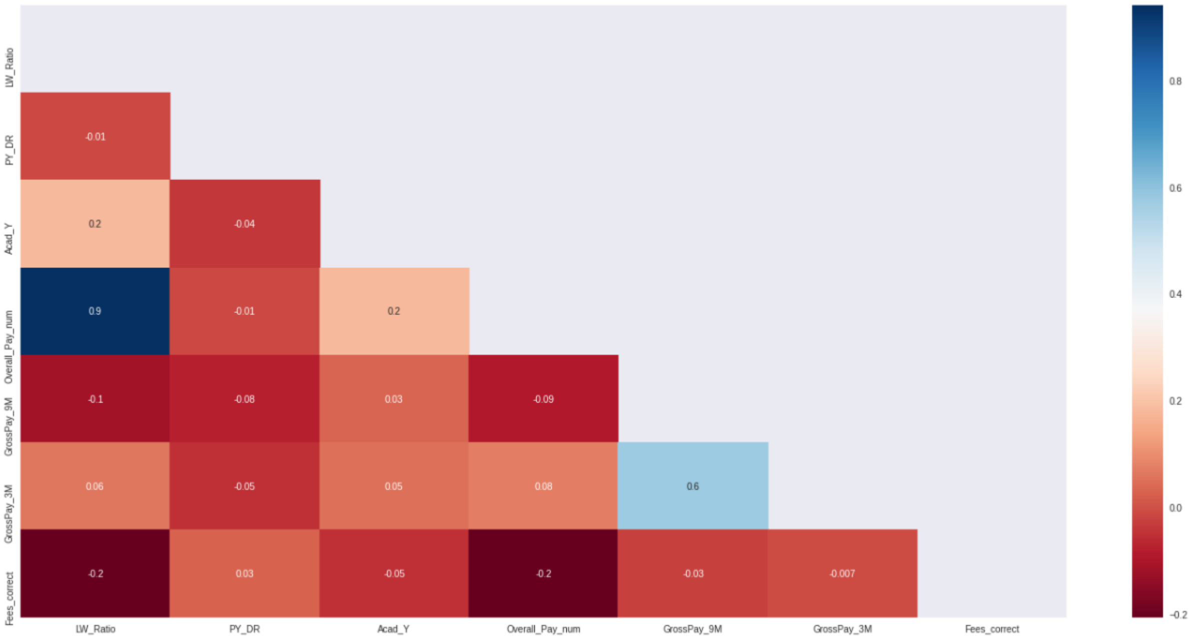
9 Overall Pay by Program Year

Independent of program year, there are always two clusters of PhD students: those receiving (almost) nothing (overall pay, unadjusted in the range of 0 - 5’000 USD per year, fees very close to total gross pay) and those receiving stipends that exceed their fees by roughly 10’000 - 40’000 USD per year:



10 Correlations

One of the strongest correlations is between Overall Pay, unadjusted, and LWR (r=0.9, dark blue):

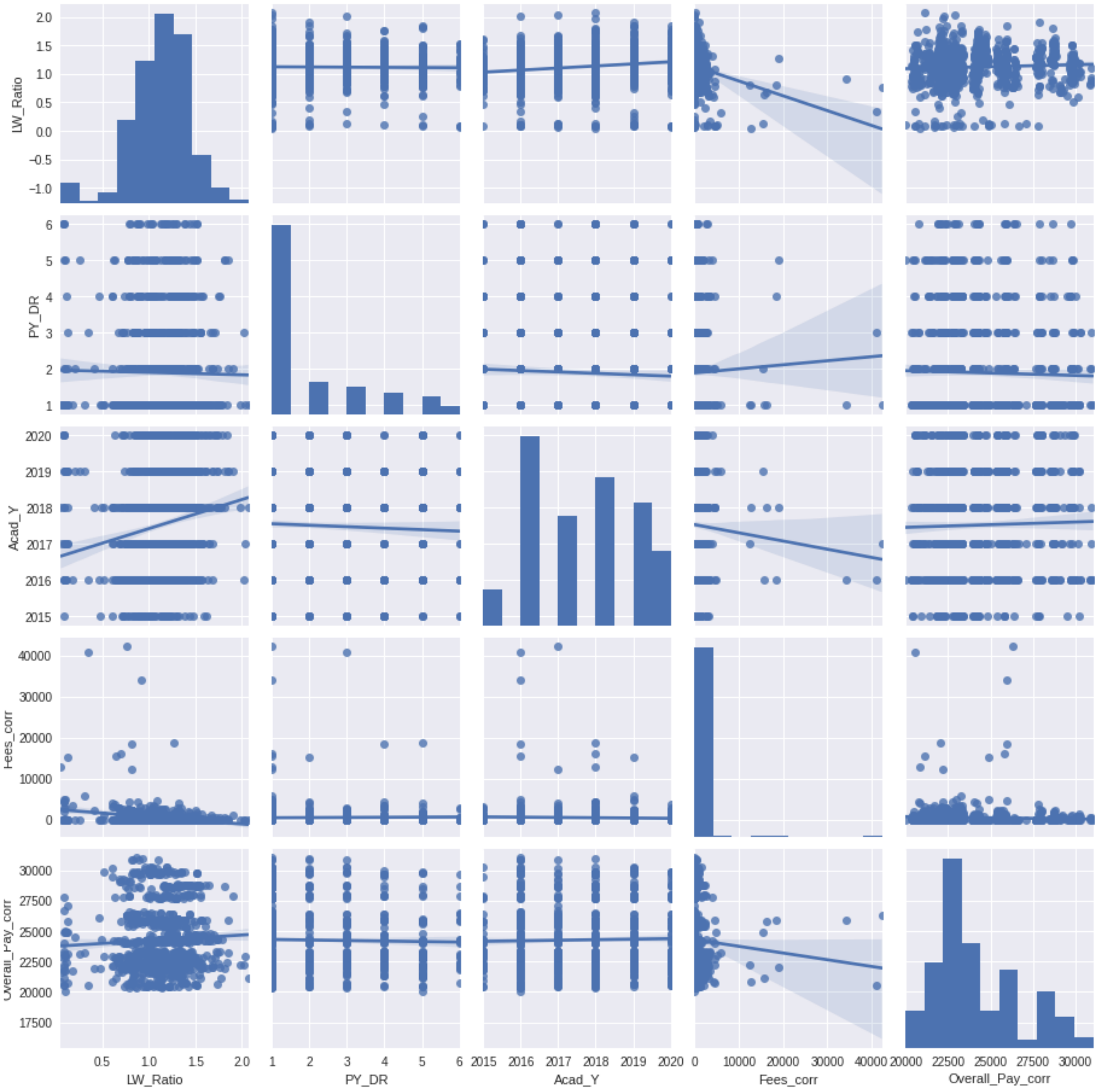


After adjusting fees and overall pay values for LWR, two correlations are particularly striking:

- LWR increases with Academic Year (2015 - 2020) (positive correlation)
- Fees, adjusted for LWR, decrease with increasing LWR.

I restricted this analysis to the years between 2015 – 2020, since much less observations are available for the years before 2015. Including all years may introduce several kinds of biases.

Pairplots of the five main numerical features in the PhD Stipend dataset revealing correlations:



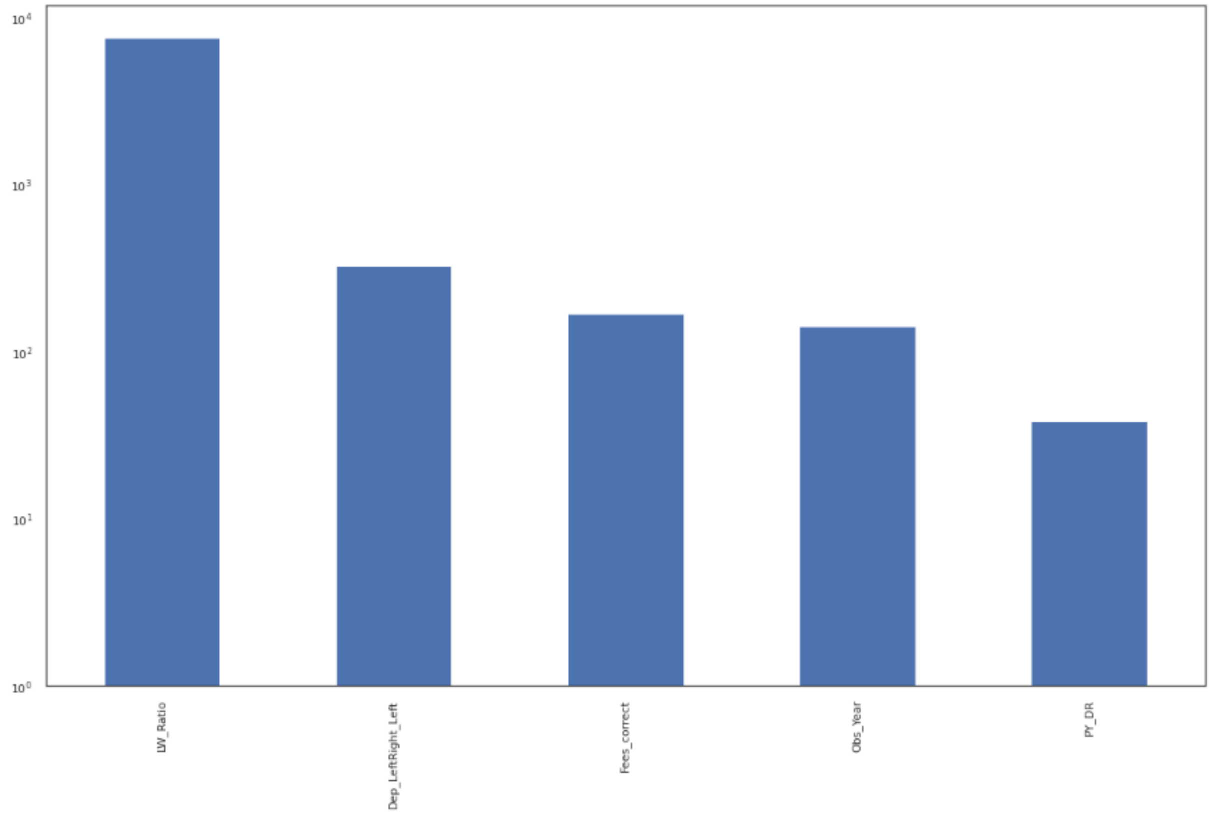
11 Prediction of Overall Pay (Unadjusted) – Linear Regression Using Scikit-Learn & Feature-Engine

I tried to predict the unadjusted overall pay using not only four numerical features (LWR, fees, academic year and program year), but also a simplified version of department categories.

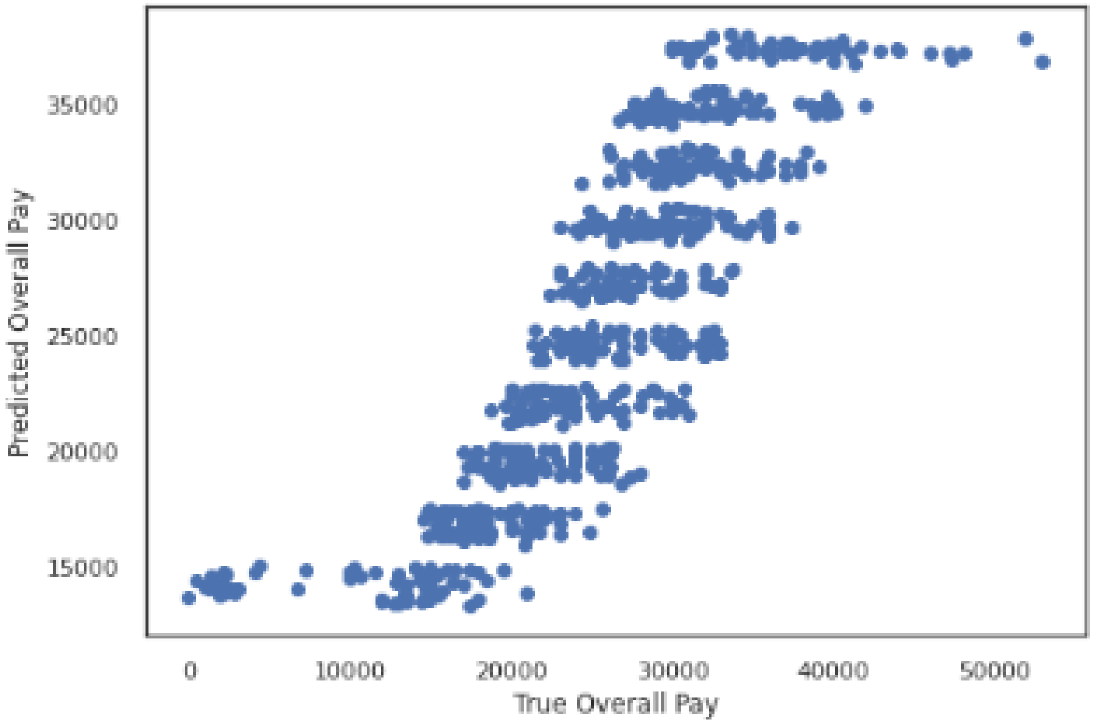
All departments particularly associated with activation of the left brain hemisphere were recoded into zeros, whereas those associated with activation of the right brain hemisphere were recoded into ones (1: History-Politology-Sociology-Psychology, Arts & Theology, Linguistics & Communication, Education). Limitation: this may not be the most scientific way to engineer a binary feature out of the original department variable.

I fitted a linear regressor to a test set (30% of all observations) using an equal frequency discretizer followed by ordinal categorical encoding for the two skewed variables fees and LWR (and one hot encoding for the categorical department feature). I obtained an r^2 for the test set of 0.73.

LW ratio is the strongest predictor of overall pay (unadjusted for LWR), followed by department (left vs. brain hemisphere), fees (unadjusted for LWR), academic year (2015-2020 only) and program year:



True overall pay values in the intermediate range of 15’000 – 40’000 USD are the easiest to predict using linear regression. True values below 10’000 USD are overestimated, whereas those above 45’000 USD are underestimated by the linear regressor:


Predicted Overall Pay
True Overall Pay