Mid-Career Salary Analysis - NWC, Oregon/Washington

DATA 403 | Jack Boydell, Paul McSlarrow

Overview

Database Creation:

- downloading college data CSV, web scraping using Python packages
- establishing relational tables with constraints
- utilizing regular expression functions for data cleaning
- joins/subqueries to create appropriate tables

Analysis (SQL, Python, R):

- general theme of our analysis: looking at mid-career salary (10+ years experience)
- exploratory data analysis: correlation analysis, boxplot state comparison
- NWC specific, general Oregon/Washington analysis; school size differentiation
- fitting a basic regression model with mid-career salary as the target variable

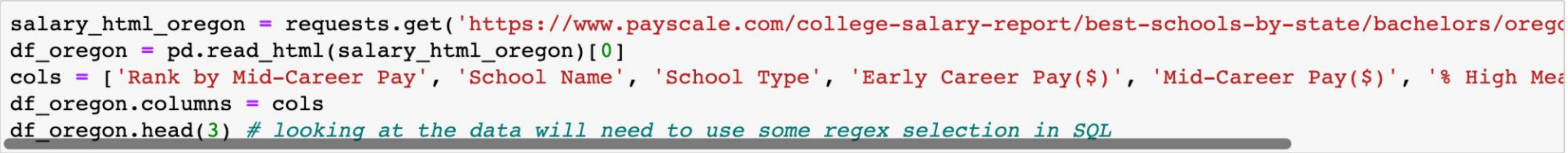
Data Collection

College Data (CSV file):

- public Kaggle dataset from 2020 with easily downloadable format

Oregon/Washington Schools (Webscraping):

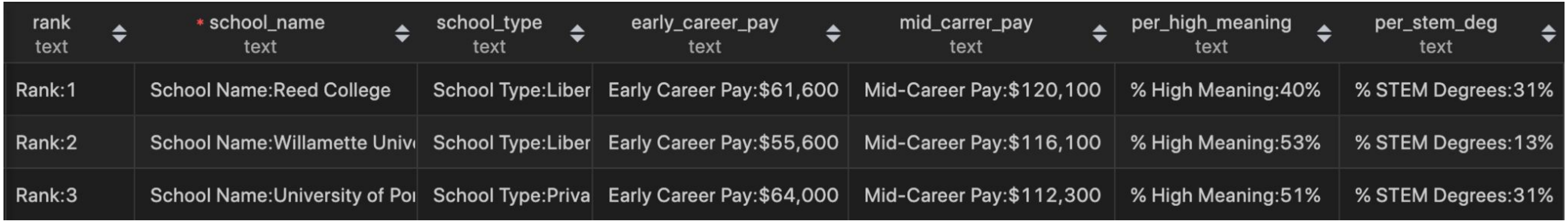
- Data comes from PayScale's College Salary Report (2021)



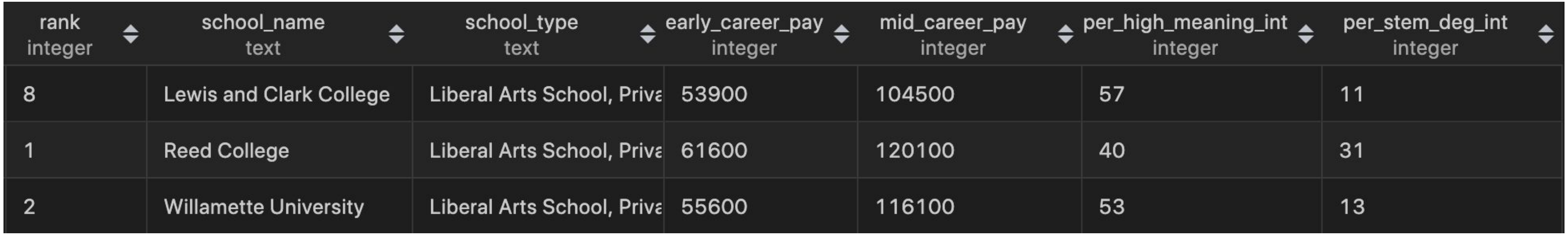
```
salary_html_oregon = requests.get('https://www.payscale.com/college-salary-report/best-schools-by-state/bachelors/orego
df_oregon = pd.read_html(salary_html_oregon)[0]
cols = ['Rank by Mid-Career Pay', 'School Name', 'School Type', 'Early Career Pay($)', 'Mid-Career Pay($)', '% High Mea
df_oregon.columns = cols
df_oregon.head(3) # looking at the data will need to use some regex selection in SQL
```

Regular Expressions

Functions: `regexp_split_to_array(text, re)`, `regexp_match(str, re)`

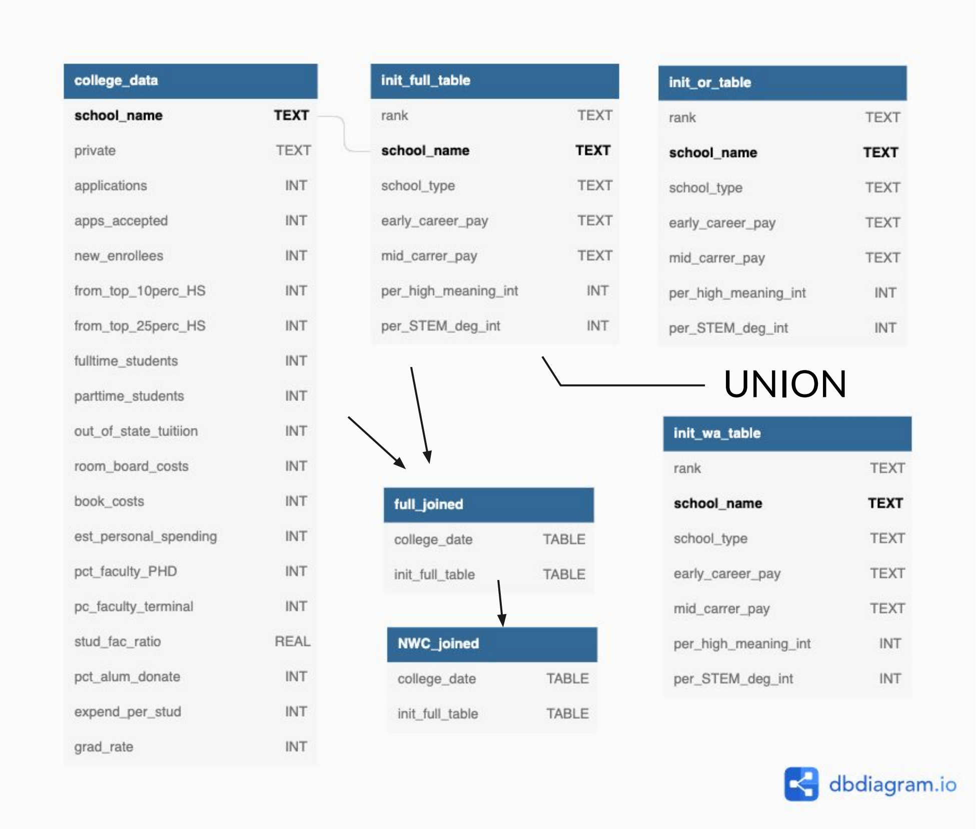


rank text	* school_name text	school_type text	early_career_pay text	mid_carrer_pay text	per_high_meaning text	per_stem_deg text
Rank:1	School Name:Reed College	School Type:Liber	Early Career Pay:$61,600	Mid-Career Pay:$120,100	% High Meaning:40%	% STEM Degrees:31%
Rank:2	School Name:Willamette Univ	School Type:Liber	Early Career Pay:$55,600	Mid-Career Pay:$116,100	% High Meaning:53%	% STEM Degrees:13%
Rank:3	School Name:University of Poi	School Type:Priva	Early Career Pay:$64,000	Mid-Career Pay:$112,300	% High Meaning:51%	% STEM Degrees:31%



rank integer	school_name text	school_type text	early_career_pay integer	mid_career_pay integer	per_high_meaning_int integer	per_stem_deg_int integer
8	Lewis and Clark College	Liberal Arts School, Priva	53900	104500	57	11
1	Reed College	Liberal Arts School, Priva	61600	120100	40	31
2	Willamette University	Liberal Arts School, Priva	55600	116100	53	13

Relational Database



college_data	
school_name	**TEXT**
private	TEXT
applications	INT
apps_accepted	INT
new_enrollees	INT
from_top_10perc_HS	INT
from_top_25perc_HS	INT
fulltime_students	INT
parttime_students	INT
out_of_state_tuitiion	INT
room_board_costs	INT
book_costs	INT
est_personal_spending	INT
pct_faculty_PHD	INT
pc_faculty_terminal	INT
stud_fac_ratio	REAL
pct_alum_donate	INT
expend_per_stud	INT
grad_rate	INT

init_full_table	
rank	TEXT
school_name	**TEXT**
school_type	TEXT
early_career_pay	TEXT
mid_carrer_pay	TEXT
per_high_meaning_int	INT
per_STEM_deg_int	INT

init_or_table	
rank	TEXT
school_name	**TEXT**
school_type	TEXT
early_career_pay	TEXT
mid_carrer_pay	TEXT
per_high_meaning_int	INT
per_STEM_deg_int	INT

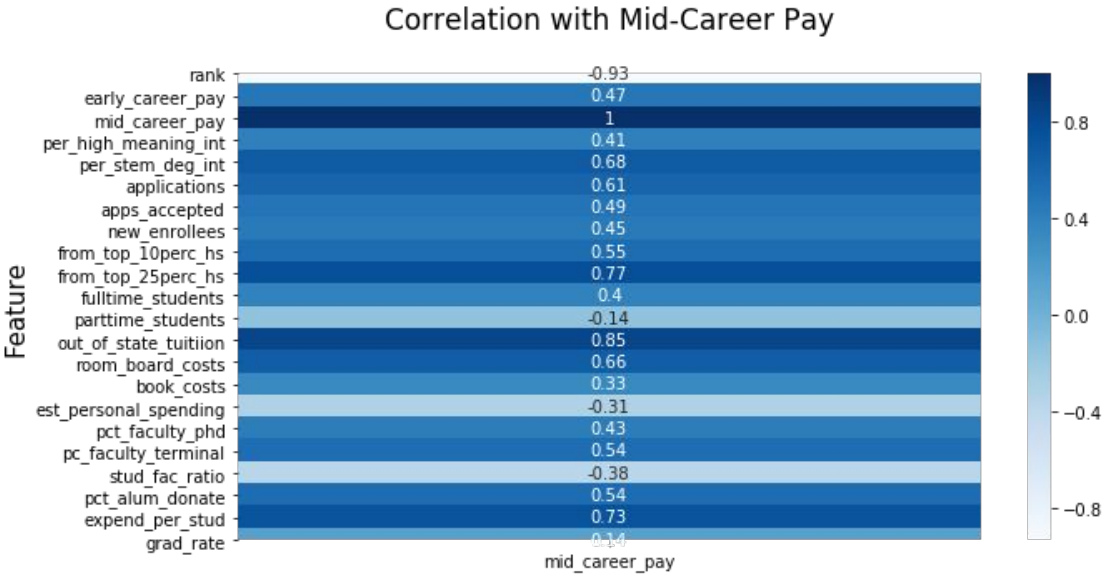
UNION

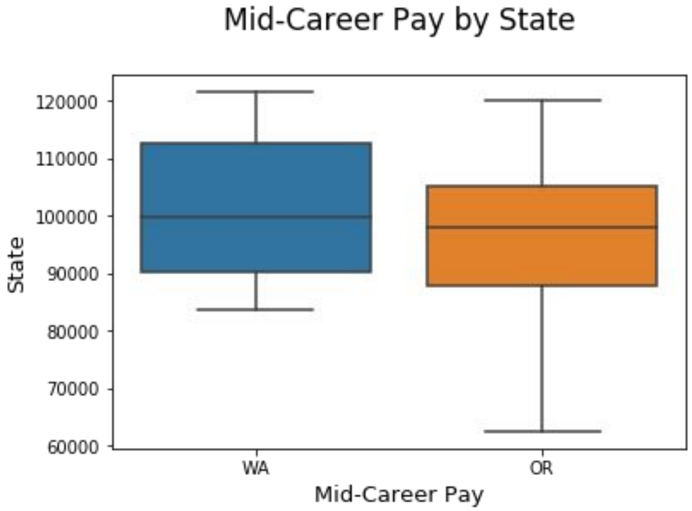
init_wa_table	
rank	TEXT
school_name	**TEXT**
school_type	TEXT
early_career_pay	TEXT
mid_carrer_pay	TEXT
per_high_meaning_int	INT
per_STEM_deg_int	INT

full_joined	
college_date	TABLE
init_full_table	TABLE

NWC_joined	
college_date	TABLE
init_full_table	TABLE

Exploratory Data Analysis - Visualization





Analysis Findings

7 of the 9 NWC schools (excluding George Fox, Whitworth) have a higher than average median mid-career salary for their respective state

Three out of the top ten schools come from the NWC – Whitman, Willamette, Puget Sound.

Willamette ranks 2nd in Oregon for median mid-career salary, 2nd in the NWC (Whitman)

We are 95% confident that the true difference in mean between the median salaries of Washington and Oregon schools (WA - OR) is between $-2220.00 and $13906.19. This suggests that there is not a significant difference between the two as $0 is contained in the interval.

[-2220.00 , 13906.19] dollars ($)

Small, Medium, Large Schools

We were able to group the colleges based on size (small: x<3000, medium: 3000>x>9999, large: x > 10000 students) and run certain aggregates across them.

We were able to learn that while large universities get more of the top 25 and top 10 percent high school students, they still tend to have the lowest average graduation rate.

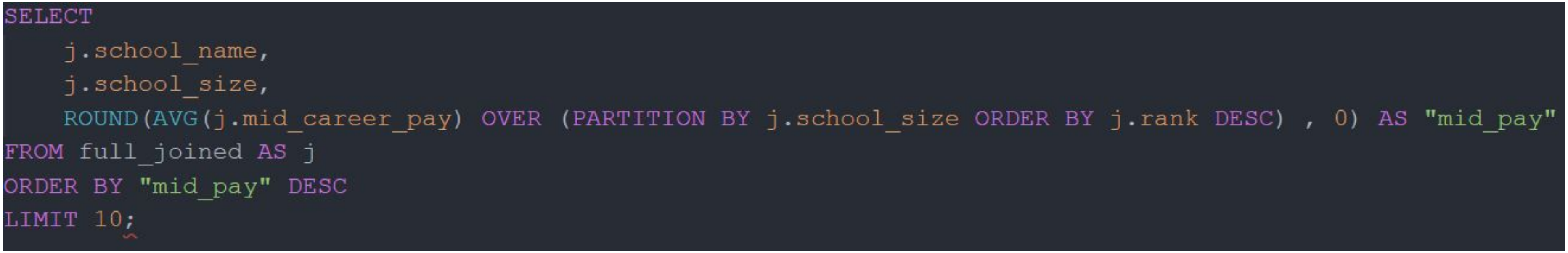
While small universities tend to have a higher average acceptance rate, they tend to have a higher average of mid-career pay (in OR and WA)

We were able to group based on the two states (OR and WA) and find that while the average percentage of PHD professors is higher in Oregon, both the graduation rate and the mid career pay is on average lower than that of Washington.

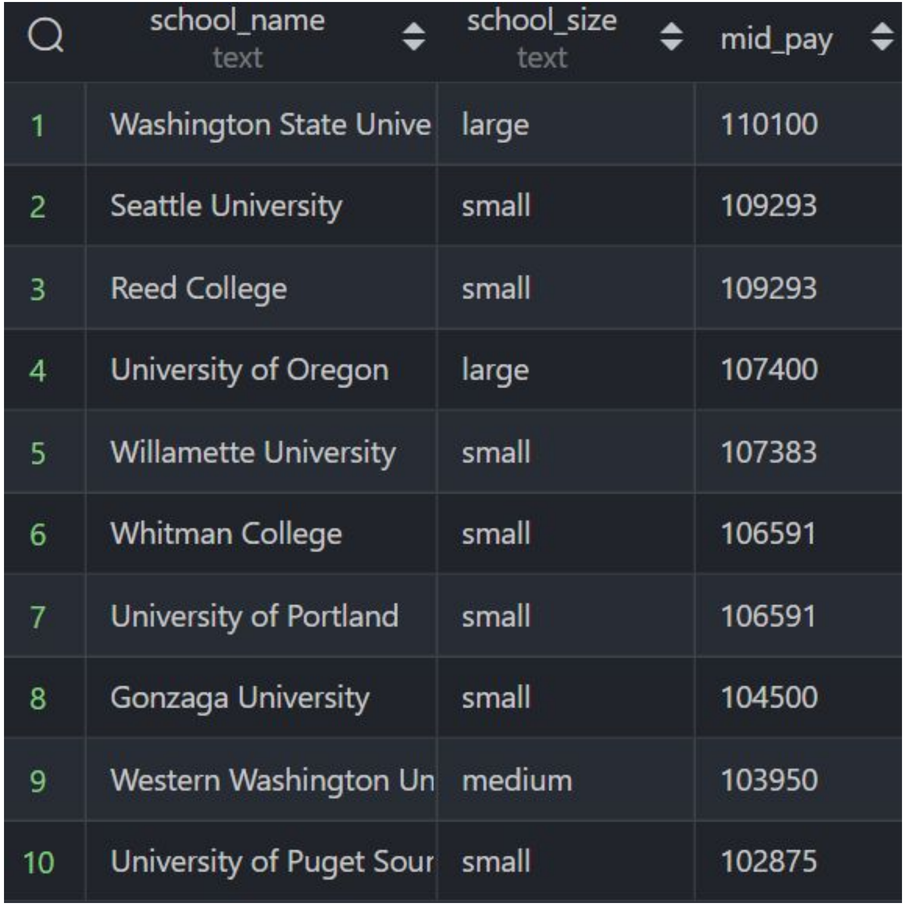
Analysis on WA and OR schools

We were able to use the partitioning method on the table on the size of the school to see the top 10 highest mid career pay. Through this analysis we were able to learn that out of the 10 schools in this list, 7 of the schools were small schools, 2 of the schools were large, and 1 of the schools was medium sized. It comes to show that regardless of the size of the university, small schools tend to have a higher average mid career pay than other school sizes.

Although, this it is important to note that this is only representative of universities in Washington and Oregon, and not statistically significant throughout all U.S. schools.



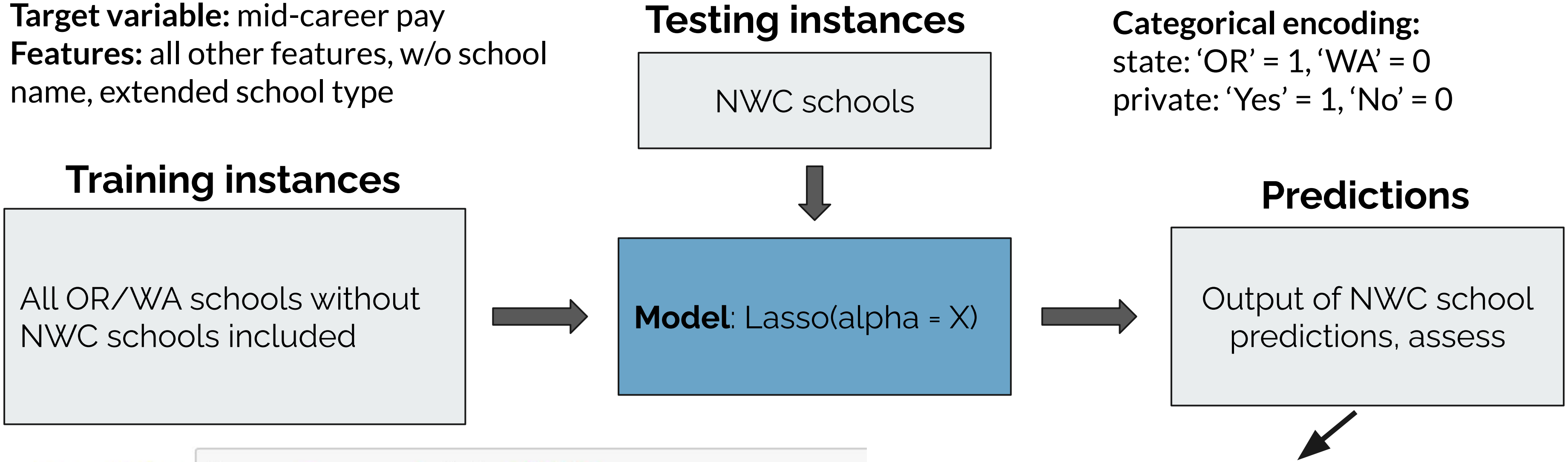
```
SELECT
    j.school_name,
    j.school_size,
    ROUND(AVG(j.mid_career_pay) OVER (PARTITION BY j.school_size ORDER BY j.rank DESC) , 0) AS "mid_pay"
FROM full_joined AS j
ORDER BY "mid_pay" DESC
LIMIT 10;
```



	school_name text	school_size text	mid_pay
1	Washington State Unive	large	110100
2	Seattle University	small	109293
3	Reed College	small	109293
4	University of Oregon	large	107400
5	Willamette University	small	107383
6	Whitman College	small	106591
7	University of Portland	small	106591
8	Gonzaga University	small	104500
9	Western Washington Un	medium	103950
10	University of Puget Sour	small	102875

Fitting a Lasso Regression Predictor

Target variable: mid-career pay
Features: all other features, w/o school name, extended school type



Testing instances

NWC schools

Categorical encoding:
state: ‘OR’ = 1, ‘WA’ = 0
private: ‘Yes’ = 1, ‘No’ = 0

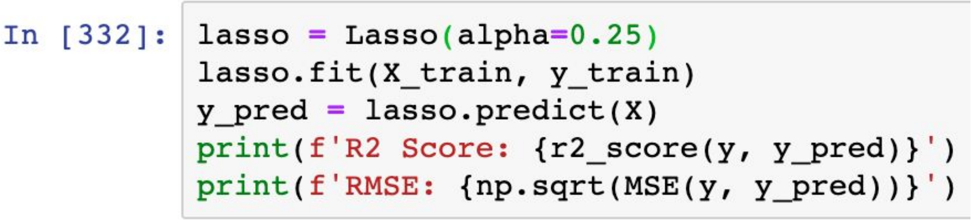
Training instances

All OR/WA schools without NWC schools included

Model: Lasso(alpha = X)

Predictions

Output of NWC school predictions, assess



```
In [332]: lasso = Lasso(alpha=0.25)
          lasso.fit(X_train, y_train)
          y_pred = lasso.predict(X)
          print(f'R2 Score: {r2_score(y, y_pred)}')
          print(f'RMSE: {np.sqrt(MSE(y, y_pred))}')
```

R^2 Score: approx. 0.72

RMSE: approx. $5200

Thank you!