IBM Introduction to Machine Learning

Exploratory Data Analysis for Machine Learning

Summary of the data

The California Standardized Testing and Reporting dataset contains data on test performance, school characteristics and student demographic backgrounds. The data used here are from 420 districts (45 counties) in California with data available for 1998 and 1999.

	0	1	2	3	4
Observation Number	1	2	3	4	5
dist_cod	75119	61499	61549	61457	61523
county	Alameda	Butte	Butte	Butte	Butte
district	Sunol Glen Unified	Manzanita Elementary	Thermalito Union Elementary	Golden Feather Union Elementary	Palermo Union Elementary
gr_span	KK-08	KK-08	KK-08	KK-08	KK-08
enrl_tot	195	240	1550	243	1335
teachers	10.9	11.15	82.9	14	71.5
calw_pct	0.5102	15.4167	55.0323	36.4754	33.1086
meal_pct	2.0408	47.9167	76.3226	77.0492	78.427
computer	67	101	169	85	171
testscr	690.8	661.2	643.6	647.7	640.85
comp_stu	0.34359	0.420833	0.109032	0.349794	0.12809
expn_stu	6384.91	5099.38	5501.95	7101.83	5235.99
str	17.8899	21.5247	18.6972	17.3571	18.6713
avginc	22.69	9.824	8.978	8.978	9.08033
el_pct	0	4.58333	30	0	13.8577
read_scr	691.6	660.5	636.3	651.9	641.8
math_scr	690	661.9	650.9	643.5	639.9

```
data.dtypes
```

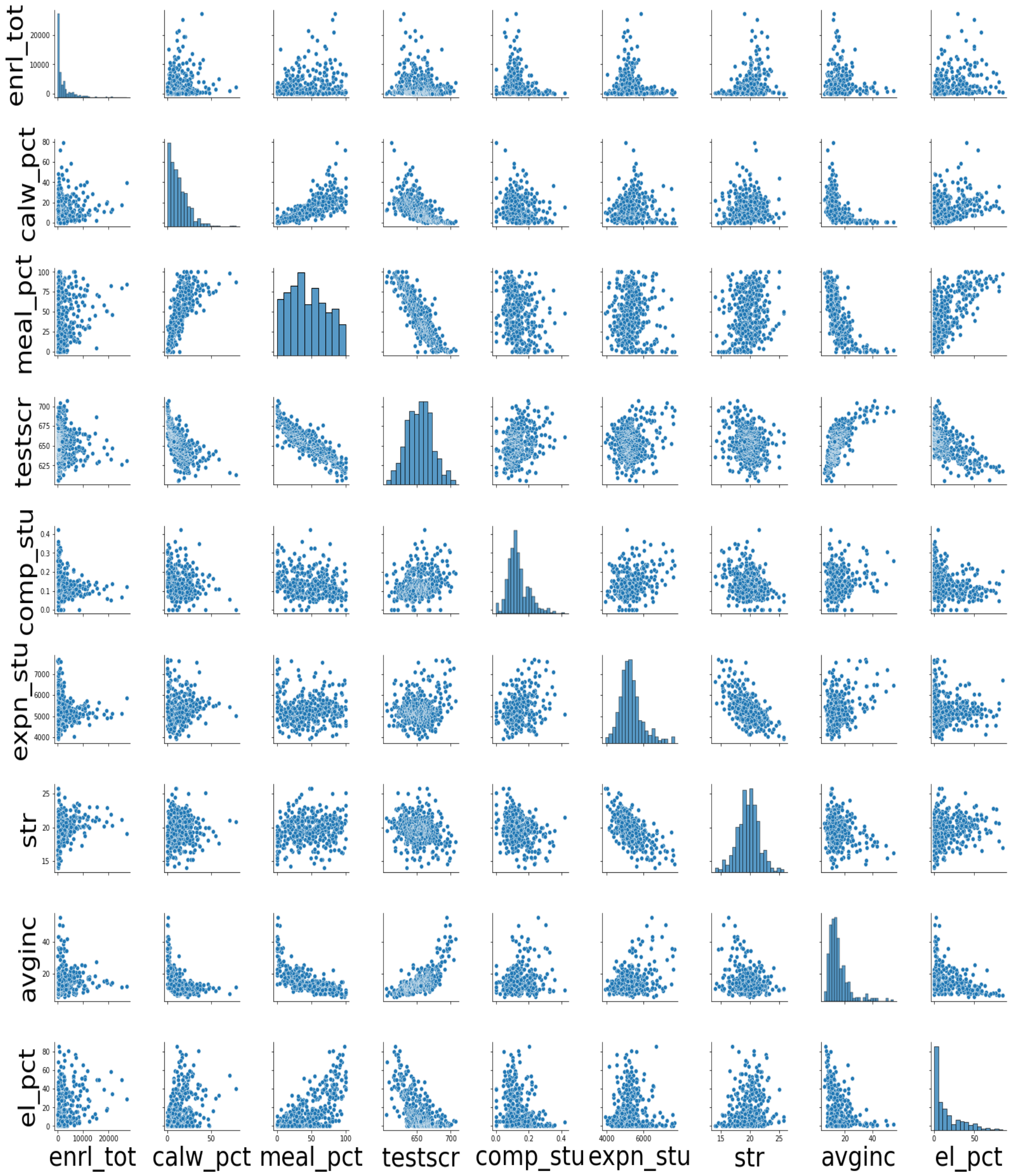
```
Observation Number      int64
dist_cod                int64
county                 object
district               object
gr_span                object
enrl_tot                int64
teachers              float64
calw_pct              float64
meal_pct              float64
computer                int64
testscr               float64
comp_stu              float64
expn_stu              float64
str                   float64
avginc                float64
el_pct                float64
read_scr              float64
math_scr              float64
dtype: object
```

School characteristics include enrollment, number of teachers, number of computers per classroom, and expenditures per student. Demographic variables include the percentage of students in the public assistance program, the percentage of students that qualify for a reduced-price lunch, and the percentage of students that are English Learners. We drop nonnumeric variables for data description. Also, data is presented by types of attributes.

	count	mean	std	min	25%	50%	75%	max
enrl_tot	420.0	2628.792857	3913.104985	81.000000	379.000000	950.500000	3008.000000	27176.000000
teachers	420.0	129.067376	187.912679	4.850000	19.662499	48.564999	146.350002	1429.000000
calw_pct	420.0	13.246042	11.454821	0.000000	4.395375	10.520450	18.981350	78.994202
meal_pct	420.0	44.705237	27.123381	0.000000	23.282200	41.750700	66.864725	100.000000
computer	420.0	303.383333	441.341298	0.000000	46.000000	117.500000	375.250000	3324.000000
testscr	420.0	654.156548	19.053348	605.550049	640.049988	654.449982	666.662506	706.750000
comp_stu	420.0	0.135927	0.064956	0.000000	0.093767	0.125464	0.164466	0.420833
expn_stu	420.0	5312.407541	633.937053	3926.069580	4906.180054	5214.516602	5601.401367	7711.506836
str	420.0	19.640425	1.891812	14.000000	18.582360	19.723208	20.871815	25.799999
avginc	420.0	15.316588	7.225890	5.335000	10.639000	13.727800	17.629001	55.327999
el_pct	420.0	15.768155	18.285927	0.000000	1.940807	8.777634	22.970003	85.539719
read_scr	420.0	654.970477	20.107980	604.500000	640.400024	655.750000	668.725006	704.000000
math_scr	420.0	653.342619	18.754202	605.400024	639.375015	652.449982	665.849991	709.500000

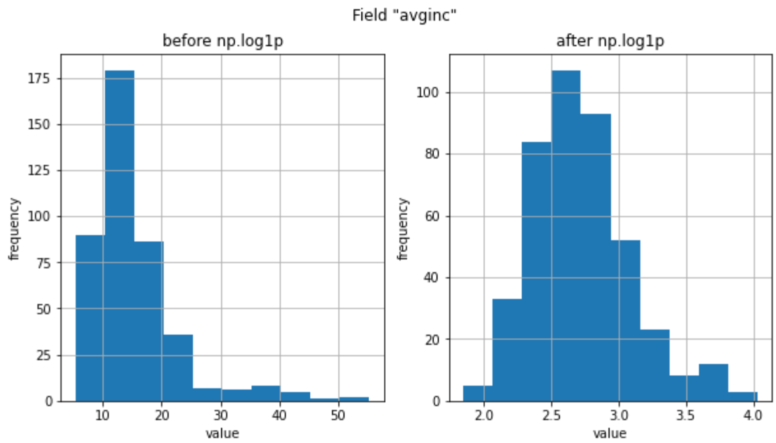
Plan for data exploration

The main objective of the analysis will be focusing on interpretation. We want to examine and determine the coefficients that do better at explaining the target variable (test scores). At a first glance we can guess that when student-teacher ratio (number of students per teacher) increases, test scores would be lower. Then we should focus on socioeconomic variables, where scores might be higher in counties with high average income. Variables like percentage of English learners (not native English speakers), computers per student or percentage of children qualifying for reduced-price lunch could be useful to determine the economic context. With this information, it would be possible to determine where to focus investment in a more efficient way in order to achieve better results.



Data cleaning and feature engineering

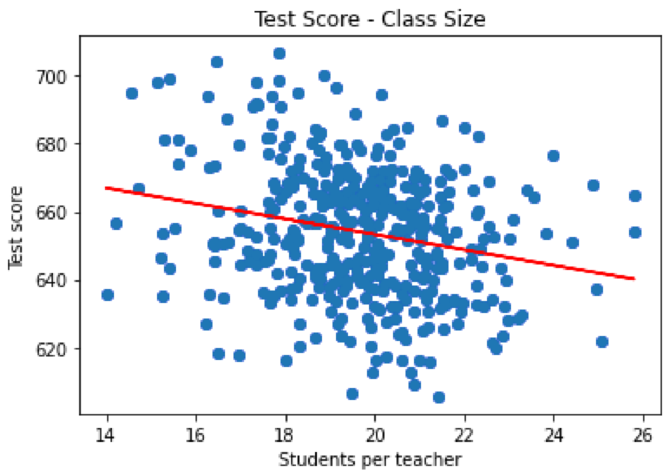
We discard absolute variables like “teachers” or “computers” and use instead “student-teacher ratio” and “computers per student”. Also, average test scores are preferred to math scores and reading scores individually, so we drop the last 2. We check if there are skew variables to make them symmetric. We define a limit of 0.75 and log transform the mean income.



	Skew
teachers	2.943065
avginc	2.223104
calw_pct	1.689099
el_pct	1.431917
expn_stu	1.071729
comp_stu	0.925678

Key findings and insights

We can see that if students per teacher increases by 1, test scores are 2.28 points lower on average. However, class size only explains 5% of test scores since R^2 is 0.05. Between test scores and income, it seems that a 2-degree polynomial fits better than a linear regression.




```
print(lr.intercept_)
print(lr.coef_)
```

```
698.9329523279787
[-2.27980829]
```

	1	avginc	avginc^2
0	1.0	3.165053	10.017561
1	1.0	2.381766	5.672809
2	1.0	2.300383	5.291760
3	1.0	2.300383	5.291760
4	1.0	2.310586	5.338809
...	...	...	...
415	1.0	3.391719	11.503759
416	1.0	3.754997	14.100005
417	1.0	3.208138	10.292152
418	1.0	2.393522	5.728948
419	1.0	2.602838	6.774765

420 rows × 3 columns

Based on these results, we find that qualifying for reduced price lunch (MEAL_PCT) and being a native English speaker (EL_PCT) are the most negative affecting features. On the other hand, as income increases (AVGINC), higher scores are predicted.

	0	1
2	meal_pct	-10.175501
7	el_pct	-3.619577
1	calw_pct	-0.890676
5	str	-0.359379
0	enrl_tot	0.001878
3	comp_stu	0.771574
4	expn_stu	0.966368
6	avginc	4.486458

Hyphotesis and significance testing

Null: $\beta_1 = 0$

Alternative: $\beta_1 \neq 0$

Null: $\beta_1 = \bar{\beta}_1$

Alternative: $\beta_1 \neq \bar{\beta}^1$

Null: $\beta_1 = -1$

Alternative: $\beta_1 \neq -1$

For the first hypothesis, we reject the null hypothesis with a 95% confidence interval (from -3.3 to -1.26) that class size (β_1) has no impact on test scores.

Suggestions for next steps

By running a correlation matrix ".corr()" we can analyze deeply the relationship between variables. Scores from reading or math could be analyzed separately and not as the mean of both (testscr) as the model deploys. Also, this dataset could be updated to date and measure the impact of better and more powerful computers in learning. Anyone can suggest or revisit the model to achieve a better explanation or a better prediction: https://github.com/estebancarboni/IBM-Introduction-to-Machine-Learning/blob/master/Exploratory%20Data%20Analysis%20Final%20Project.ipynb

	enrl_tot	calw_pct	meal_pct	testscr	comp_stu	expn_stu	str	avginc	el_pct	read_scr	math_scr
enrl_tot	1.000000	0.090161	0.129234	-0.153988	-0.212718	-0.112285	0.298481	0.028392	0.354879	-0.188399	-0.110889
calw_pct	0.090161	1.000000	0.739422	-0.626853	-0.151968	0.067889	0.018276	-0.512651	0.319576	-0.611847	-0.617691
meal_pct	0.129234	0.739422	1.000000	-0.868772	-0.203953	-0.061039	0.135203	-0.684440	0.653061	-0.878808	-0.823015
testscr	-0.153988	-0.626853	-0.868772	1.000000	0.270703	0.191273	-0.226363	0.712431	-0.644124	0.981882	0.979143
comp_stu	-0.212718	-0.151968	-0.203953	0.270703	1.000000	0.286560	-0.307070	0.194806	-0.251007	0.281158	0.248589
expn_stu	-0.112285	0.067889	-0.061039	0.191273	0.286560	1.000000	-0.619982	0.314484	-0.071396	0.217927	0.154989
str	0.298481	0.018276	0.135203	-0.226363	-0.307070	-0.619982	1.000000	-0.232194	0.187642	-0.246593	-0.195553
avginc	0.028392	-0.512651	-0.684440	0.712431	0.194806	0.314484	-0.232194	1.000000	-0.307419	0.697819	0.699398
el_pct	0.354879	0.319576	0.653061	-0.644124	-0.251007	-0.071396	0.187642	-0.307419	1.000000	-0.690286	-0.568682
read_scr	-0.188399	-0.611847	-0.878808	0.981882	0.281158	0.217927	-0.246593	0.697819	-0.690286	1.000000	0.922901
math_scr	-0.110889	-0.617691	-0.823015	0.979143	0.248589	0.154989	-0.195553	0.699398	-0.568682	0.922901	1.000000