CodeCoven

Anna Radmilovich
Brooke Long
Heather Aubry
McKenna O'Bryant
Isabella Malek

What various factors affect student's average scores?

VARIABLES

Target Variable: Student average test scores
Dummy Variable: Whether or not you have a part time job

Factors We Are Looking At:

★ Job Status
★ Absence Days
★ Amount of Weekly Study Hours

DATA CLEANING PROCESS

Original Data:

- Dropped Individual class columns.
- Combined individual class columns into a new column 'average_scores'
- Dropped Columns: Email, Gender, first_name, last_name.

	id	first_name	last_name	email	gender	part_time_job	absence_days	extracurricular_activities	weekly_self_study_hours
0	1	Paul	Casey	paul.casey.1@gslingacademy.com	male	False	3	False	27
1	2	Danielle	Sandoval	danielle.sandoval.2@gslingacademy.com	female	False	2	False	47
2	3	Tina	Andrews	tina.andrews.3@gslingacademy.com	female	False	9	True	13
3	4	Tara	Clark	tara.clark.4@gslingacademy.com	female	False	5	False	3
4	5	Anthony	Campos	anthony.campos.5@gslingacademy.com	male	False	5	False	10
...	...	...	...	...	...	...	...	...	...
1995	1996	Alan	Reynolds	alan.reynolds.1996@gslingacademy.com	male	False	2	False	30
1996	1997	Thomas	Gilbert	thomas.gilbert.1997@gslingacademy.com	male	False	2	False	20
1997	1998	Madison	Cross	madison.cross.1998@gslingacademy.com	female	False	5	False	14
1998	1999	Brittany	Compton	brittany.compton.1999@gslingacademy.com	female	True	10	True	5
1999	2000	Natalie	Smith	natalie.smith.2000@gslingacademy.com	female	False	5	False	27

2000 rows × 19 columns

math_score	history_score	physics_score	chemistry_score	biology_score	english_score	geography_score
73	81	93	97	63	80	87
90	86	96	100	90	88	90
81	97	95	96	65	77	94
71	74	88	80	89	63	86
84	77	65	65	80	74	76
...	...	...	...	...	...	...
83	77	84	73	75	84	82
89	65	73	80	87	67	73
97	85	63	93	68	94	78
51	96	72	89	95	88	75
82	99	91	69	83	93	100

DATA CLEANING PROCESS

	id	absence_days	weekly_self_study_hours	career_aspiration	@dropdown	average_scores	part_time_binary	ec_activities_binary
0	1	3	27	Lawyer	NaN	82	0	0
1	2	2	47	Doctor	NaN	91	0	0
2	3	9	13	Government Officer	NaN	86	0	1
3	4	5	3	Artist	NaN	78	0	0
4	5	5	10	Unknown	NaN	74	0	0
...	...	...	...	...	...	...	...	...
1995	1996	2	30	Construction Engineer	NaN	79	0	0
1996	1997	2	20	Software Engineer	NaN	76	0	0
1997	1998	5	14	Software Engineer	NaN	82	0	0
1998	1999	10	5	Business Owner	NaN	80	1	1
1999	2000	5	27	Accountant	NaN	88	0	0

Cleaned Data:

- Added in Dummy Variables derived from columns 'part_time_job' and 'extracurricular_activities'.
- Drop columns converted to binary.

DECIDING ON A VARIABLE

Analyzing R Values:

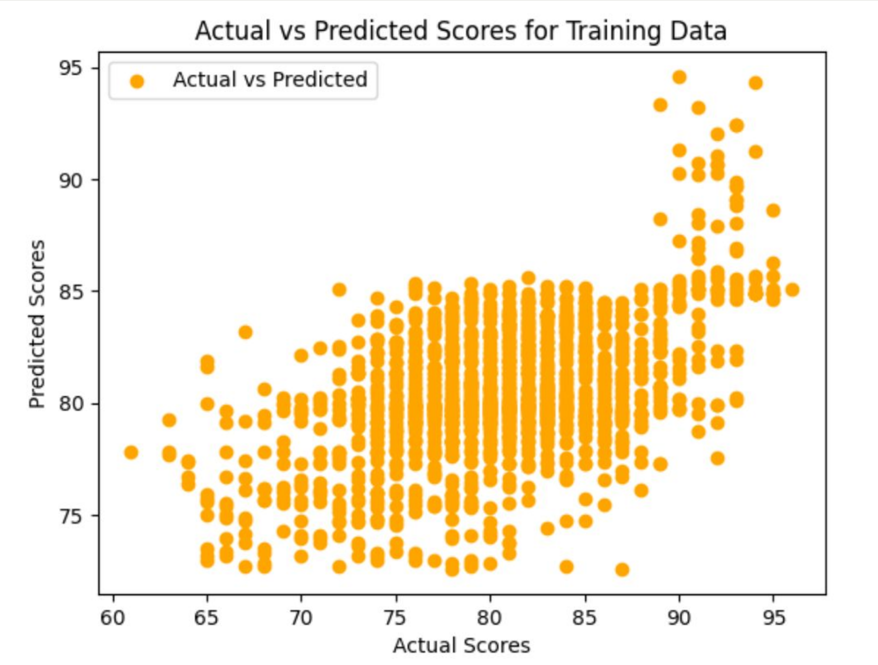
```
absence_correlation = st.pearsonr(final_df['absence_days'], final_df['average_scores'])
part_time_correlation = st.pearsonr(final_df['part_time_binary'], final_df['average_scores'])
ec_activities_correlation = st.pearsonr(final_df['ec_activities_binary'], final_df['average_scores'])
self_study_correlation = st.pearsonr(final_df['weekly_self_study_hours'], final_df['average_scores'])




print(f'{absence_correlation} \n {part_time_correlation} \n {ec_activities_correlation} \n {self_study_correlati
```

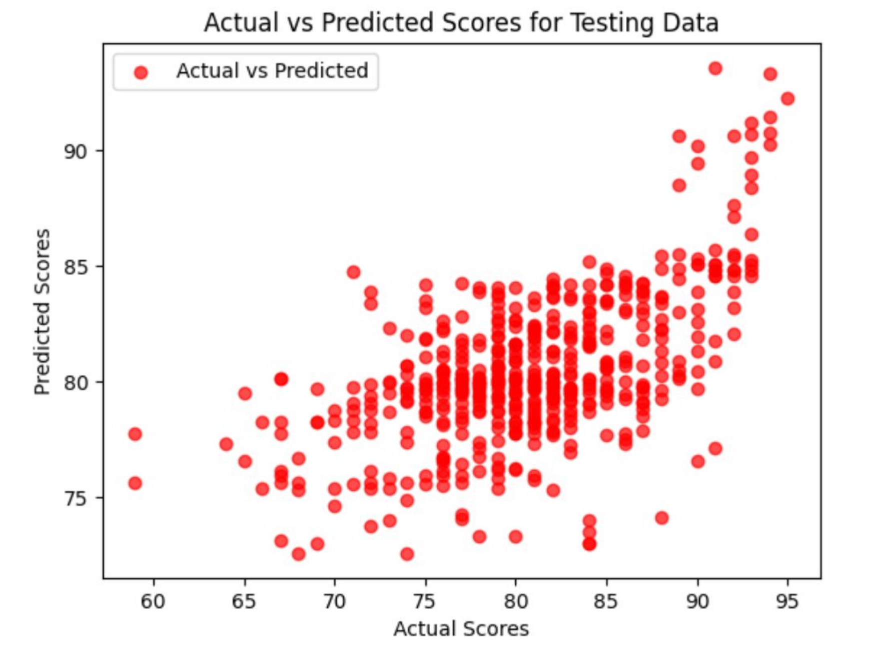
```
PearsonRResult(statistic=-0.23212551390350877, pvalue=7.101916903918721e-26)
 PearsonRResult(statistic=-0.19038102328458462, pvalue=8.893604670772564e-18)
 PearsonRResult(statistic=-0.03236683632613075, pvalue=0.1479072643171271)
 PearsonRResult(statistic=0.5008872480744411, pvalue=1.6725877881290696e-127)
```

Training Model



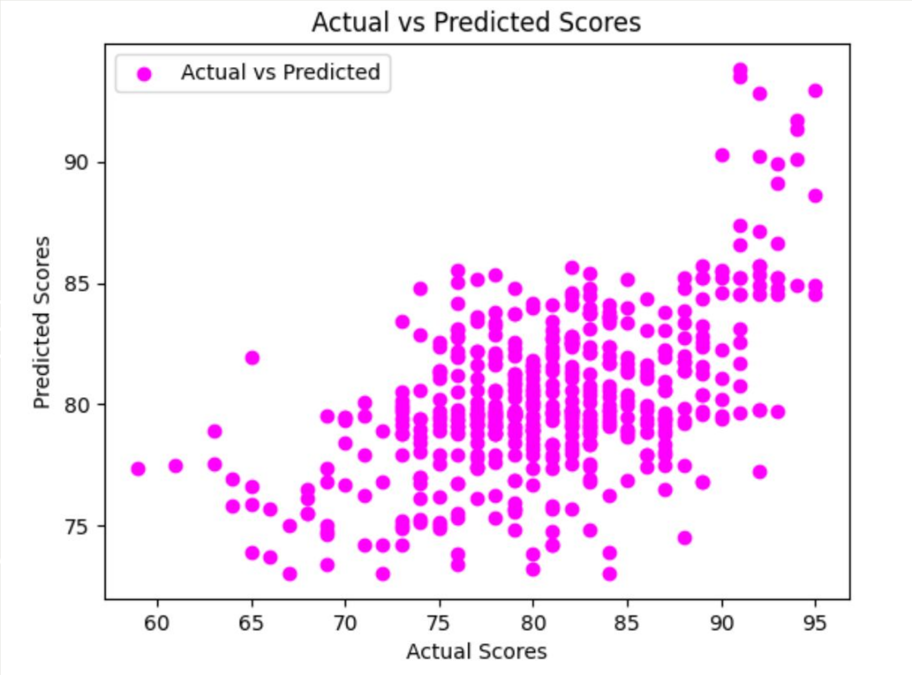
- ★ Contains 75% of the csv set and is trained to predict how scores are impacted by three features including: Job, Amount of Absence Days, and Weekly Study Hours.
- ★ Degree = 2; allows for R^2 to be represented
- ★ Not a linear relationship; specifically middle cluster lessens the effect of external factors impacting scores
- ★ As the predicted scores increase so does the likelihood of the external factors impacting the predicted score

Testing Model



- ★ Contains 25% of the data and evaluates the performance of our model.
- ★ The training model provides an accurate prediction with the tested data;
  - ○ similar placement in the middle where clustering appeared on the training model,
  - ○ increased effect of external factors as scores increased
- ★ Less data = Less clustering

Training vs. Testing Model



- ★ Combination of both testing and training models.
- ★ Shows us that the model is a correct representation of our data as all the models are distributed similarly.
- ★ Tells us that students predicted scores often don't align with their actual scores.
- ★ Impact could include the various combinations of the three features,:
  - More absent day could increase Self-study
  - Part time work could increase absence
  - Part time work could decrease self study hours

SUMMARY

- ★ Overall our data shows us that actual student scores and predicted scores are consistent, and there's is initially a minimal relationship between them.
- ★ Besides the outliers, the combination of higher absence days and minimal weekly study hours is more likely to produce lower predicted scores.
- ★ Predicted scores are more accurate as scores are higher(about 85 - 95), but are still lower than the actual, meaning these external factors are more likely to not produce high scoring tests.