Regression Analysis

Group13: 李心揚 宋宇然 白宗翰 葉威霆

CONTENT

- DATA SUMMARY
- EDA
- MODEL DEVELOPMENT
- Implications and Recommendations
- Extra info : ANN

DATA SUMMARY(Training Set)

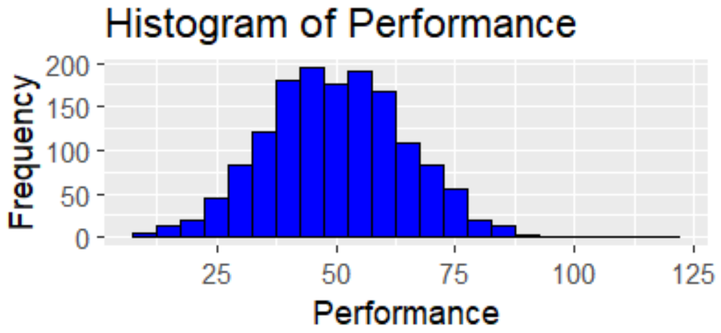
Variables	Min	1st Qu.	Median	Mean	3rd Qu.	Max
performance	10	40	50	49.91	60	120
Study_hrs	0	5	8	8.08	11	21
Sleep_hrs	0	5	8	8.08	11	22
High_math	1	50	58	57.86	66	99
High_hist	1	49	55	55.01	61	94
High_eng	1	53	60	60.36	68	99

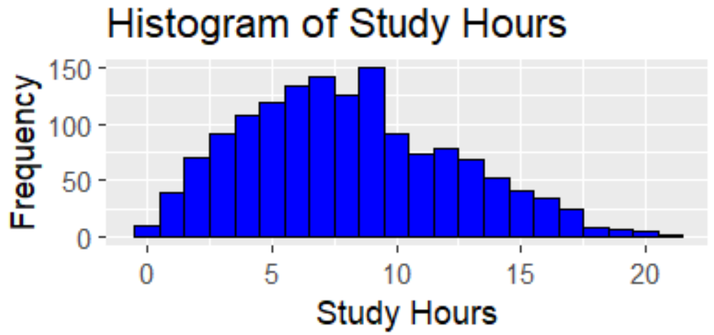
Variables	F	M
gender	715	761

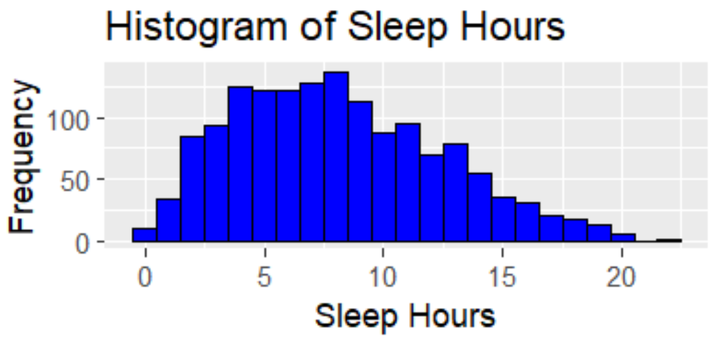
Variables	Rural	Urban
environment	359	1117

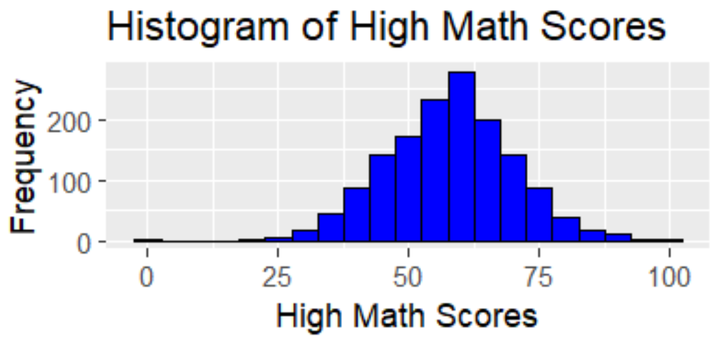
Variables	H	L	M
stress	250	701	525

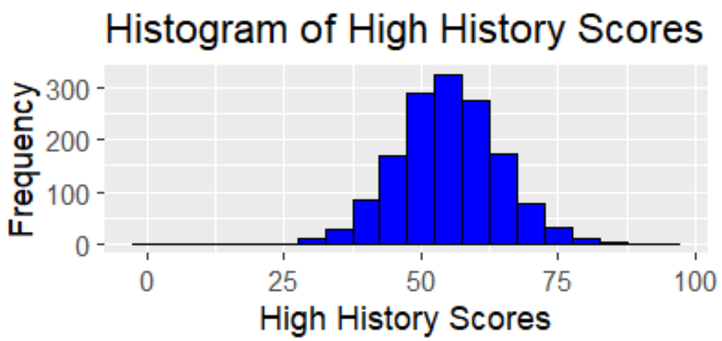
Histograms

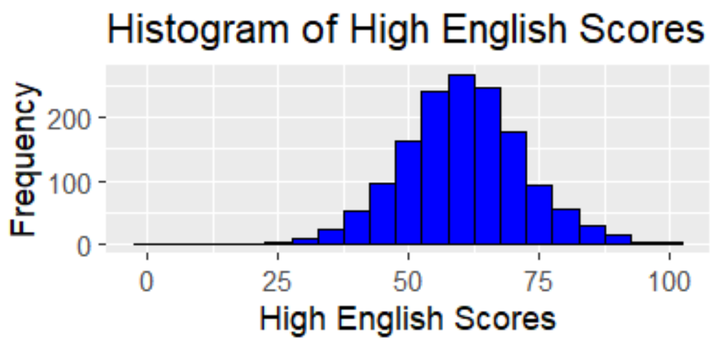

Histogram of Performance
Frequency
0
50
100
150
200
25
50
75
100
125
Performance

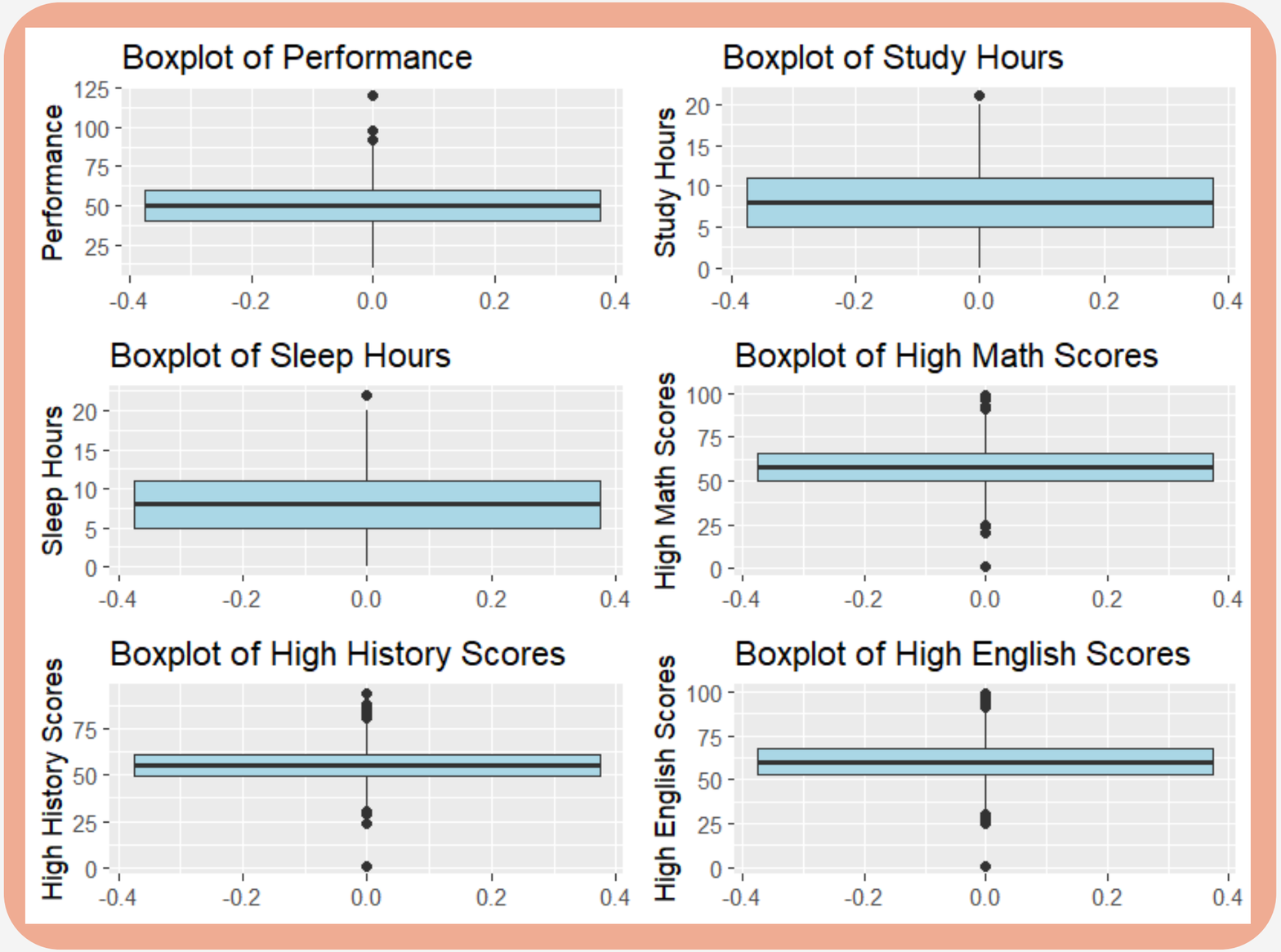

Histogram of Study Hours
Frequency
0
50
100
150
0
5
10
15
20
Study Hours


Histogram of Sleep Hours
Frequency
0
50
100
0
5
10
15
20
Sleep Hours


Histogram of High Math Scores
Frequency
0
100
200
0
25
50
75
100
High Math Scores


Histogram of High History Scores
Frequency
0
100
200
300
0
25
50
75
100
High History Scores


Histogram of High English Scores
Frequency
0
100
200
0
25
50
75
100
High English Scores


Boxplot of Performance
Performance
125
100
75
50
25
-0.4
-0.2
0.0
0.2
0.4
Boxplot of Study Hours
Study Hours
20
15
10
5
0
Boxplot of Sleep Hours
Sleep Hours
Boxplot of High Math Scores
High Math Scores
Boxplot of High History Scores
High History Scores
Boxplot of High English Scores
High English Scores

Boxplot

DATA SUMMARY(Testing Set)

Variables	Min	1st Qu.	Median	Mean	3rd Qu.	Max
performance	14	41	50	50.4	60	92
Study_hrs	0	5	8	8.043	11	19
Sleep_hrs	0	5	7	7.853	10	20
High_math	30	50	58	58.23	65.25	98
High_hist	30	50	55	55.4	61	87
High_eng	25	53.75	61	60	68	98

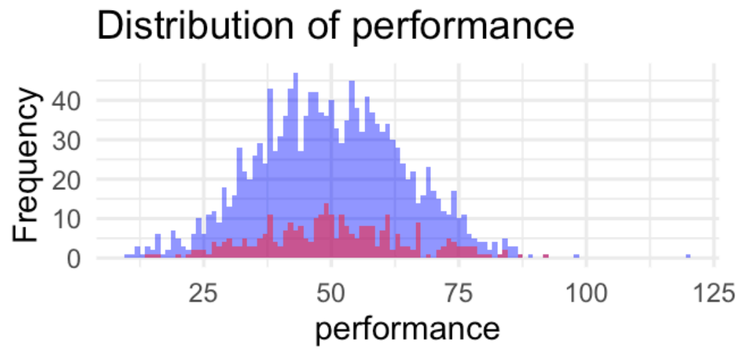
Variables	F	M
gender	150	150

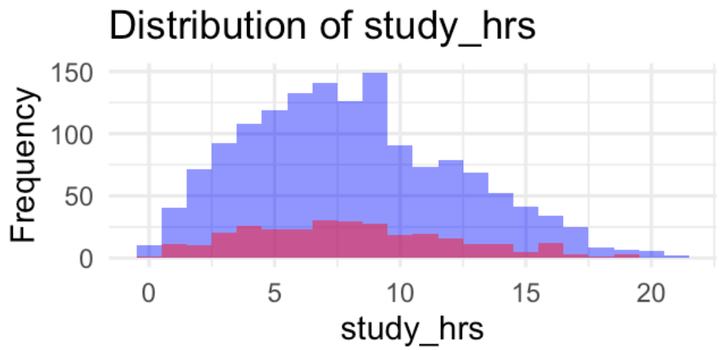
Variables	Rural	Urban
environment	53	247

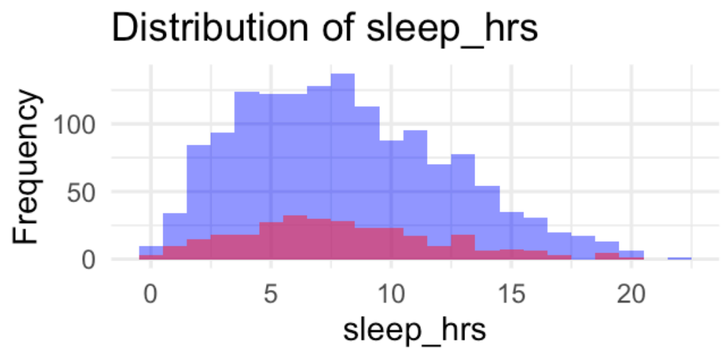
Variables	H	L	M
stress	100	100	100

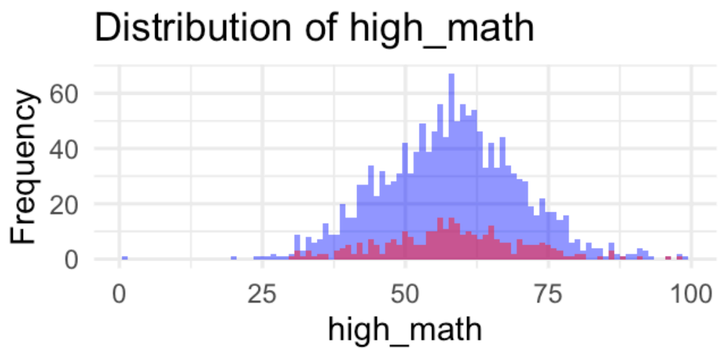
Histograms

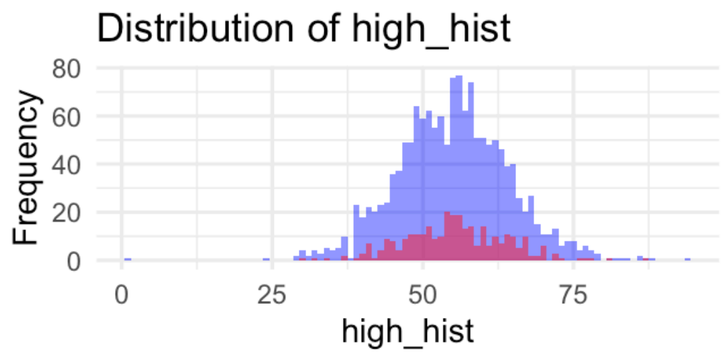
(Adding Testing Set)

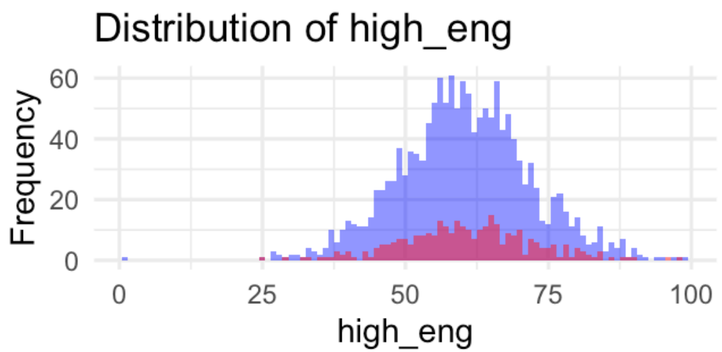






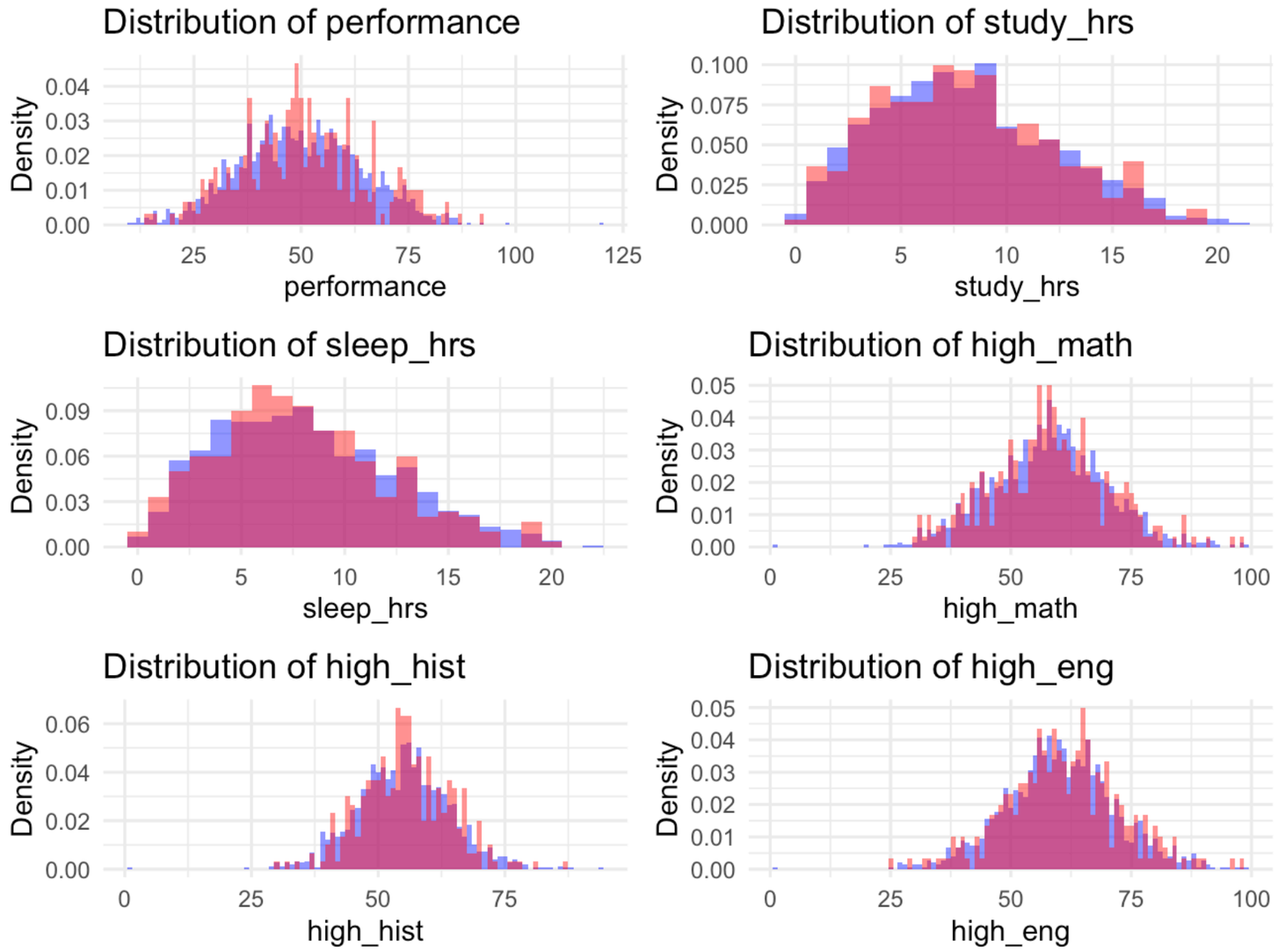




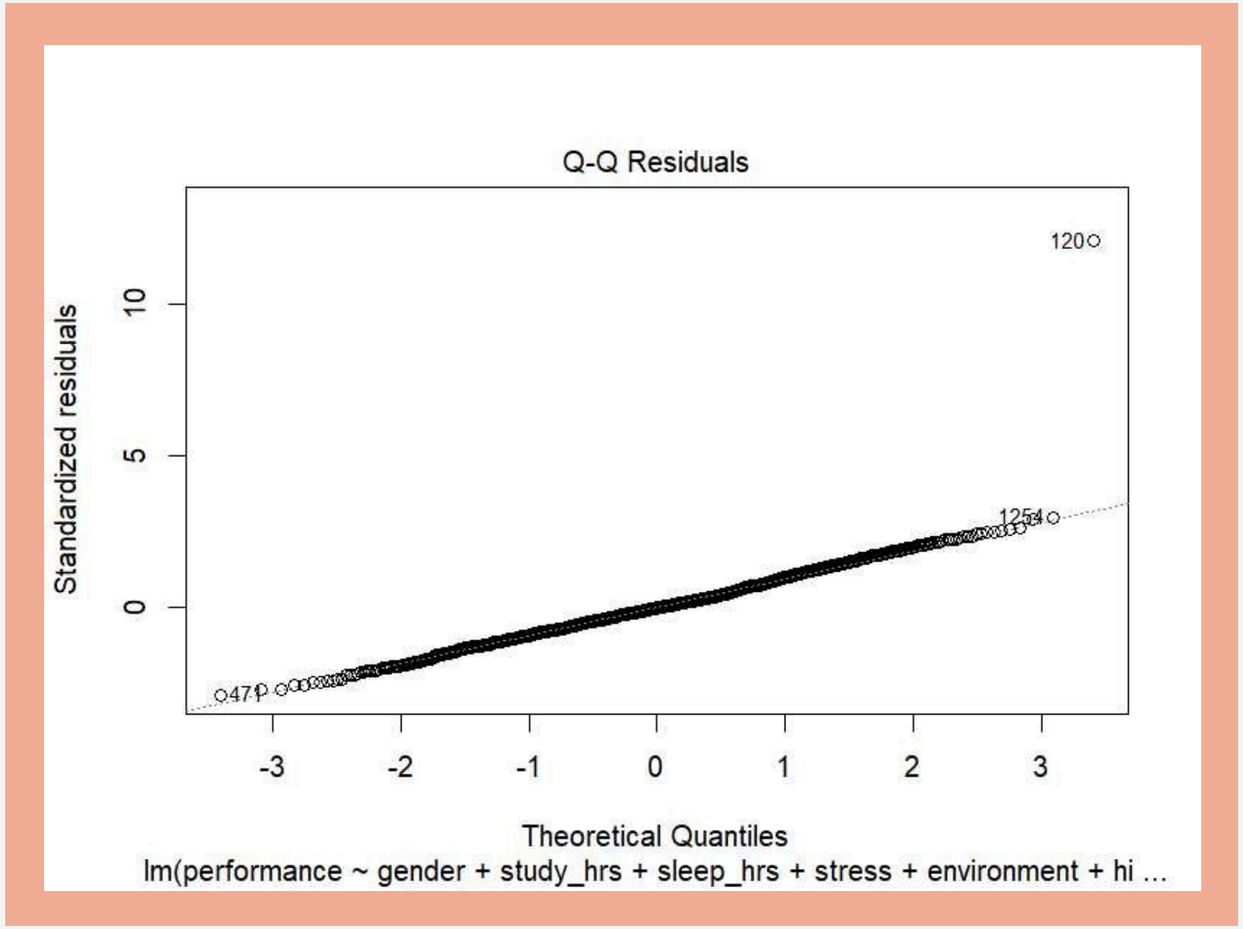


Histogram for Density Distribution

(Adding Testing Set)



QQ-Plot


Q-Q Residuals
Standardized residuals
10
5
0
120
1254
47
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(performance ~ gender + study_hrs + sleep_hrs + stress + environment + hi ...

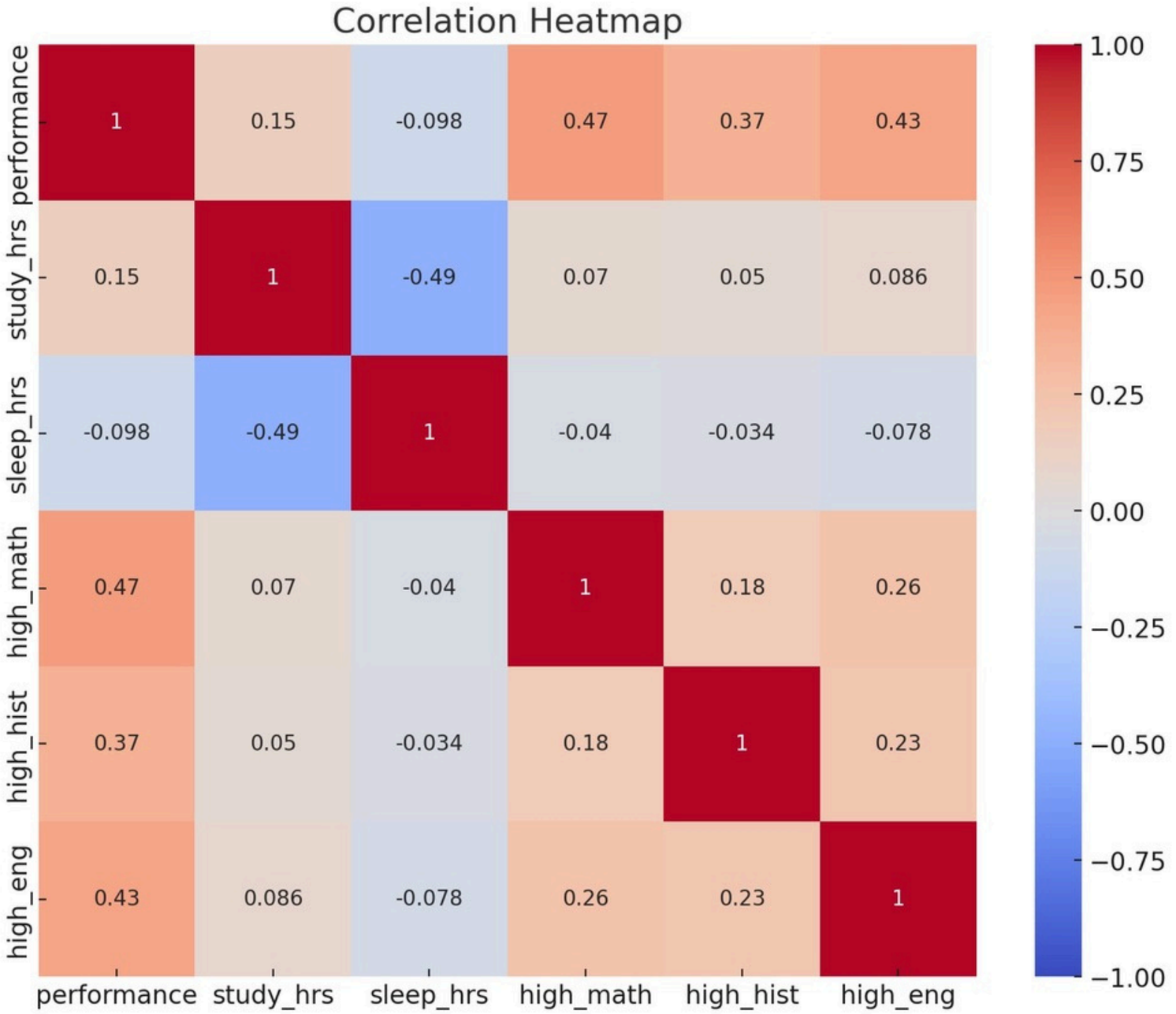
GVIF

(Generalized Variance Inflation Factor)

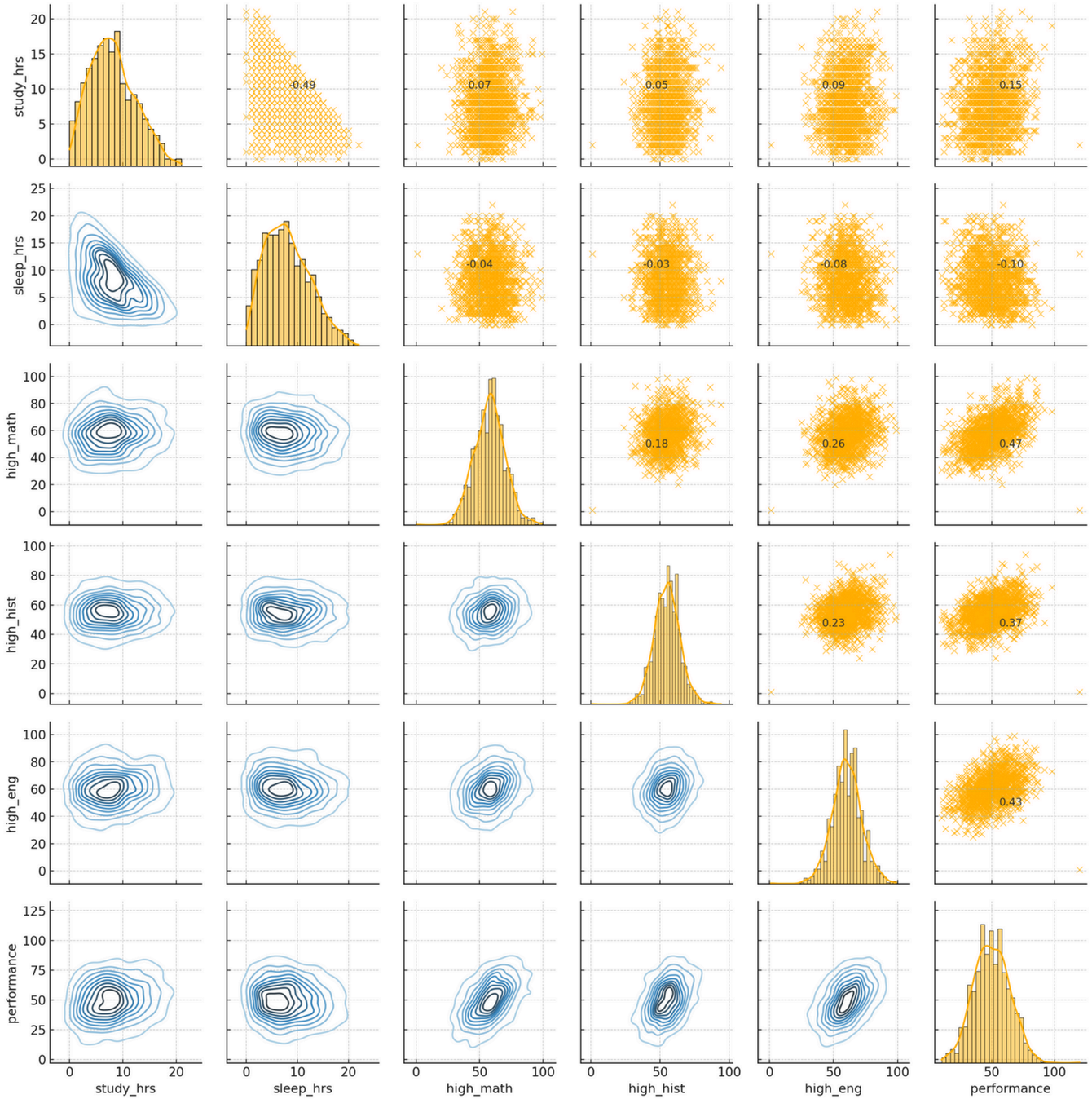
GVIF是普通VIF的推廣形式。對於數值型變量，GVIF等同於VIF；
對於因子變量，GVIF考慮了因子變量的多重共線性。

	GVIF		GVIF		GVIF
High_math	1.108	**Sleep_hrs**	1.347	**Gender**	1.022
High_hist	1.083	**Study_hrs**	1.330	**Stress**	1.168
High_eng	1.151			**Environment**	1.093

Correlation Matrix



Correlation Matrix



Multiple Linear Regression

Data Preprocessing

One-Hot Encoding

'gender', 'stress', 'environment'

Standardization 標準化

'highmath', 'higheng', 'highhist', 'studyhrs', 'sleephrs'

Feature Interaction 特徵交互

```
from sklearn.preprocessing import PolynomialFeatures
poly =
PolynomialFeatures(interaction_only=True,include_bias=False)
```

Remove Outlier 移除離群值

三個標準差以上

Backward selection 後向選取

也有嘗試Forward selection 和 RFE，但都沒有比較好

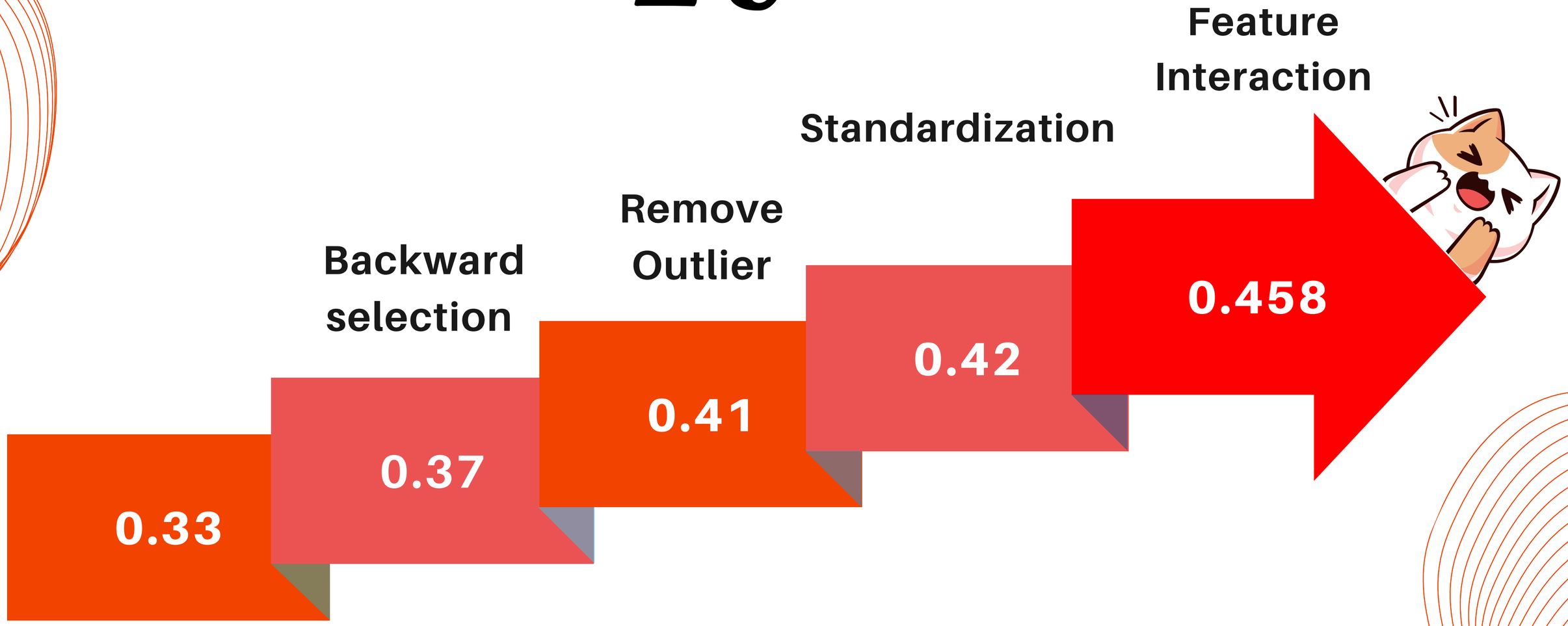
Before Backward selection

- study_hrs
- sleep_hrs
- high_math
- high_hist
- high_eng
- gender_M
- stress_L
- stress_M
- environment_Urban
- study_hrs sleep_hrs
- study_hrs high_math
- study_hrs high_hist
- study_hrs high_eng
- study_hrs gender_M
- study_hrs stress_L
- study_hrs stress_M
- study_hrs environment_Urban
- sleep_hrs high_math
- sleep_hrs high_hist
- sleep_hrs high_eng
- sleep_hrs gender_M
- sleep_hrs stress_L
- sleep_hrs stress_M
- sleep_hrs environment_Urban
- high_math high_hist
- high_math high_eng
- high_math gender_M
- high_math stress_L
- high_math stress_M
- high_math environment_Urban
- high_hist high_eng
- high_hist gender_M
- high_hist stress_L
- high_hist stress_M
- high_hist environment_Urban
- high_eng gender_M
- high_eng stress_L
- high_eng stress_M
- high_eng environment_Urban
- gender_M stress_L
- gender_M stress_M
- gender_M environment_Urban
- stress_L stress_M
- stress_L environment_Urban
- stress_M environment_Urban

After Backward selection

	p	VIF
study_hrs	(0)	(1.849)
high_math	(0)	(1.095)
high_hist	(0)	(1.091)
high_eng	(0)	(1.131)
stress_L	(0.038)	(2.04)
study_hrs,stress_L	(0)	(1.83)
high_hist,high_eng	(0.006)	(1.04)
gender_M,stress_L	(0.001)	(2.48)
gender_M,environment_Urban	(0.002)	(1.99)
stress_M,environment_Urban	(0.033)	(1.44)

R^2


0.33
Backward selection
0.37
Remove Outlier
0.41
Standardization
0.42
Feature Interaction
0.458

OLS Regression Results

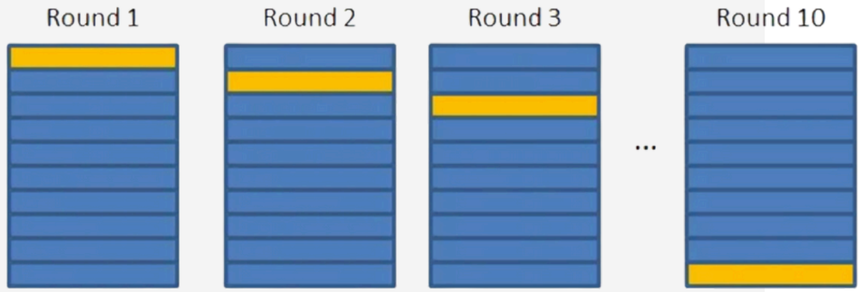
Dep. Variable:	performance	R-squared:	0.453
Model:	OLS	Adj. R-squared:	0.449
Method:	Least Squares	F-statistic:	96.80
Date:	Wed, 29 May 2024	Prob (F-statistic):	1.80e-145
Time:	23:25:50	Log-Likelihood:	-4446.3
No. Observations:	1178	AIC:	8915.
Df Residuals:	1167	BIC:	8970.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	49.7975	0.309	161.247	0.000	49.192	50.403
study_hrs	2.5772	0.419	6.147	0.000	1.755	3.400
high_math	5.2432	0.322	16.269	0.000	4.611	5.876
high_hist	3.8009	0.328	11.574	0.000	3.157	4.445
high_eng	4.3450	0.330	13.163	0.000	3.697	4.993
stress_L	-0.9937	0.478	-2.079	0.038	-1.932	-0.056
study_hrs stress_L	-1.5669	0.420	-3.731	0.000	-2.391	-0.743
high_hist high_eng	-1.0876	0.396	-2.748	0.006	-1.864	-0.311
gender_M stress_L	1.4032	0.434	3.231	0.001	0.551	2.255
gender_M environment_Urban	-1.1203	0.361	-3.106	0.002	-1.828	-0.413
stress_M environment_Urban	0.8625	0.403	2.138	0.033	0.071	1.654

Omnibus:	1.133	Durbin-Watson:	2.055
Prob(Omnibus):	0.567	Jarque-Bera (JB):	1.200
Skew:	0.049	Prob(JB):	0.549
Kurtosis:	2.879	Cond. No.	2.90

Performance =

$$
\begin{aligned}
&49.79 \\
&+ 2.57 \times \text{Study hours} \\
&+ 5.24 \times \text{high math} \\
&+ 3.8 \times \text{high hist} \\
&+ 4.35 \times \text{high English} \\
&- 0.99 \times \text{stress} \\
&- 1.56 \times (\text{study hours} \times \text{stress}) \\
&- 0.87 \times (\text{high hist} \times \text{high english}) \\
&+ 1.4 \times (\text{gender} \times \text{stress}) \\
&- 1.12 \times (\text{gender} \times \text{environment}) \\
&+ 0.86 \times (\text{stress} \times \text{environment})
\end{aligned}
$$



MLR Model Evaluation

Test Set Evaluation

- R-square: 0.4583
- Mean Absolute Error (MAE): 8.2116
- Root Mean Square Error (RMSE): 10.508

Training Set Evaluation

- R-square: 0.4533
- Mean Absolute Error (MAE): 8.3811
- Root Mean Square Error (RMSE): 10.543

Cross-Validation Results

- R-square: 0.4383
- Mean Absolute Error (MAE): 8.4667
- Root Mean Square Error (RMSE): 10.635

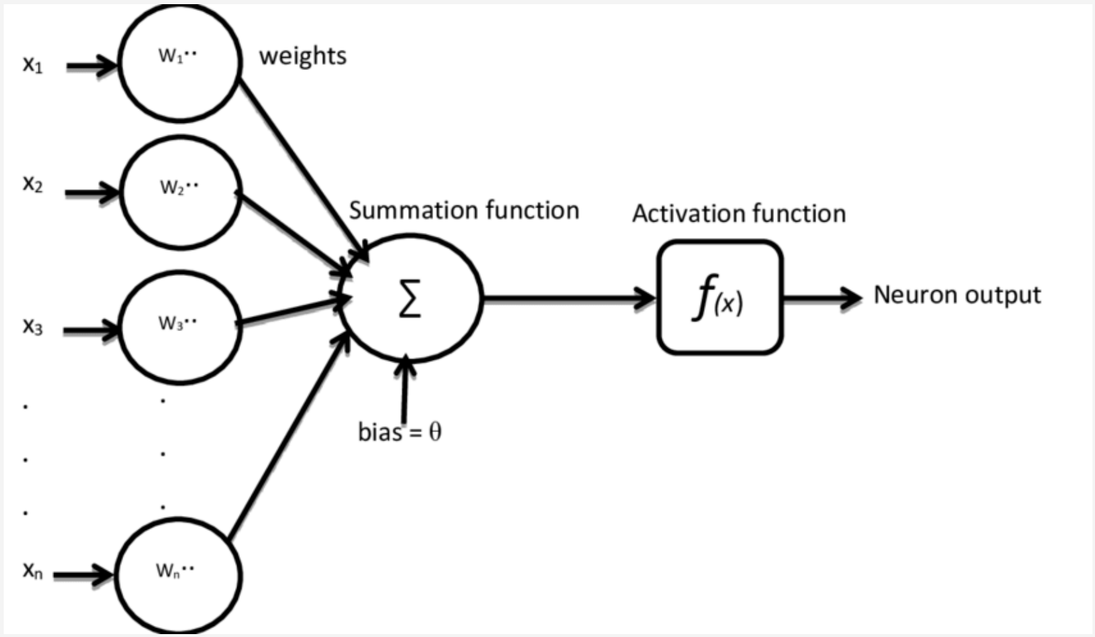
Implications and Recommendations

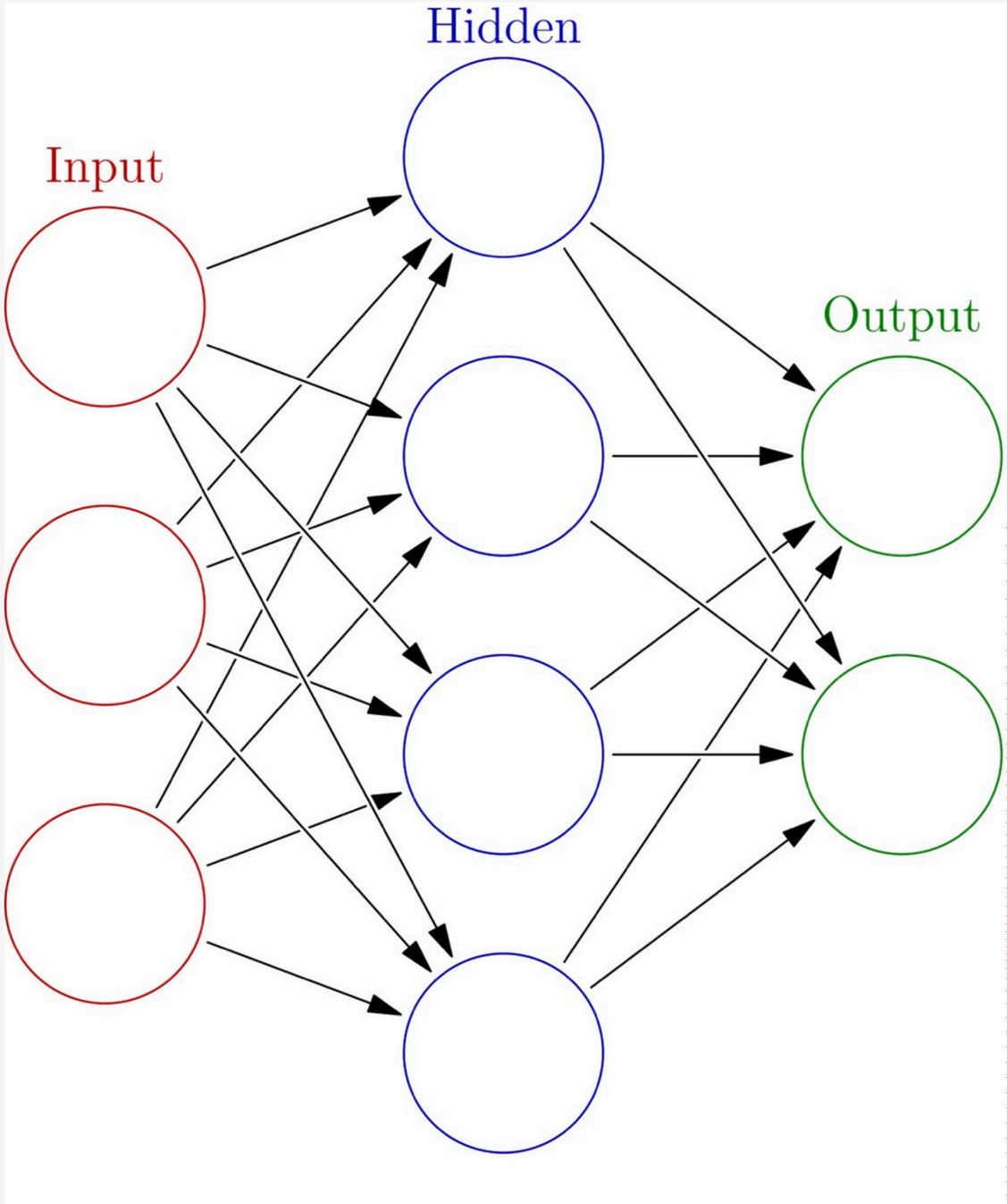
- **Study Hours and High School Scores:** Increased study hours and higher high school scores (particularly in math and English) are positively correlated with better freshman performance. Students should be encouraged to maintain consistent study habits and strive for strong high school performance.

- **Stress Levels:** Stress has a notable impact on performance. Strategies to manage and reduce stress could potentially improve academic outcomes.

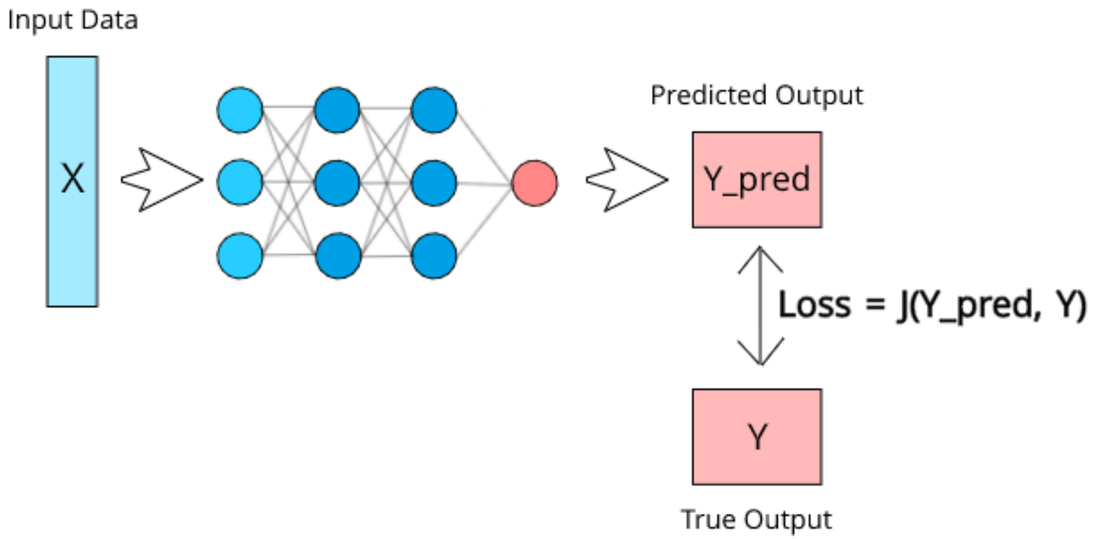
Extra Info :
Artificial Neural Network

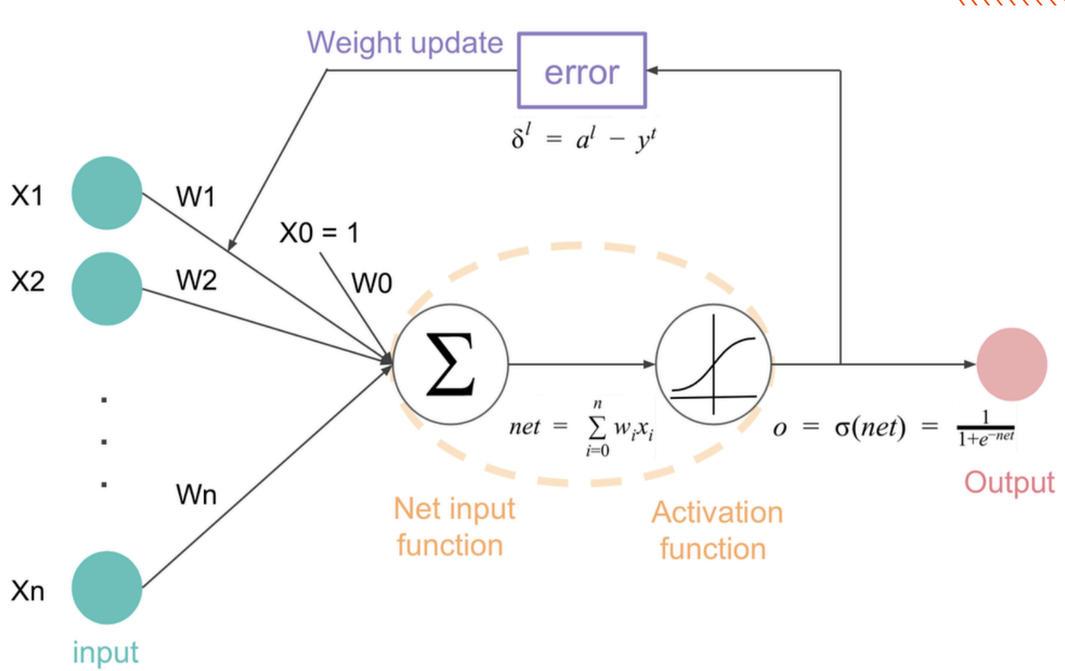
ANN MODEL


x1
x2
x3
xn
w1
w2
w3
wn
weights
Summation function
Activation function
Σ
f(x)
Neuron output
bias = θ

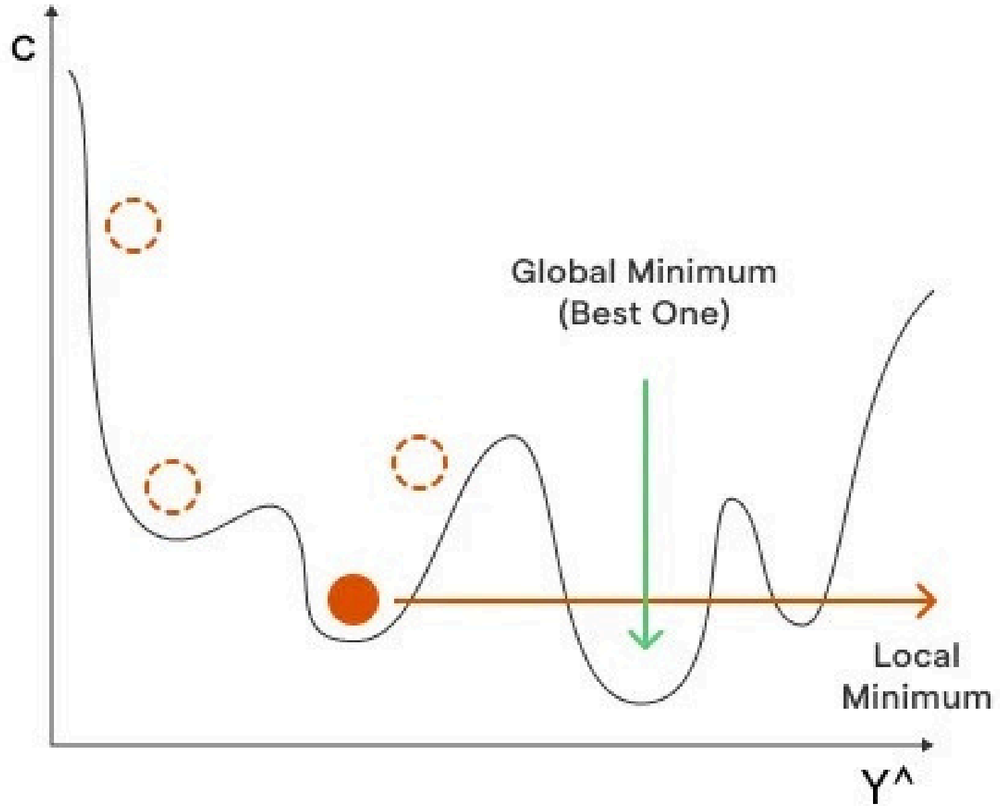

Hidden
Input
Output

Loss function





Gradient Descent




```
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 414us/step - loss: 2687.2400
Epoch 2/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 546us/step - loss: 2642.8262
Epoch 3/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 409us/step - loss: 2573.1497
Epoch 4/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 355us/step - loss: 2559.1226
Epoch 5/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 378us/step - loss: 2504.2283
Epoch 6/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 396us/step - loss: 2373.4856
Epoch 7/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 457us/step - loss: 2300.7292
Epoch 8/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 458us/step - loss: 2192.6270
Epoch 9/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 524us/step - loss: 2004.6803
Epoch 10/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 351us/step - loss: 1914.0354
Epoch 11/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 431us/step - loss: 1760.8042
Epoch 12/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - loss: 1610.3127
Epoch 13/100
45/45 ━━━━━━━━━━━━━━━━━━━━ 0s 345us/step - loss: 1491.8370
```

	MLR	ANN
R-SQUARED	0.4583	0.4626
ROOT MEAN SQUARE ERRO (RMSE)	8.2116	10.3470
MEAN ABSOLUTE ERROR (MAE)	10.508	8.007

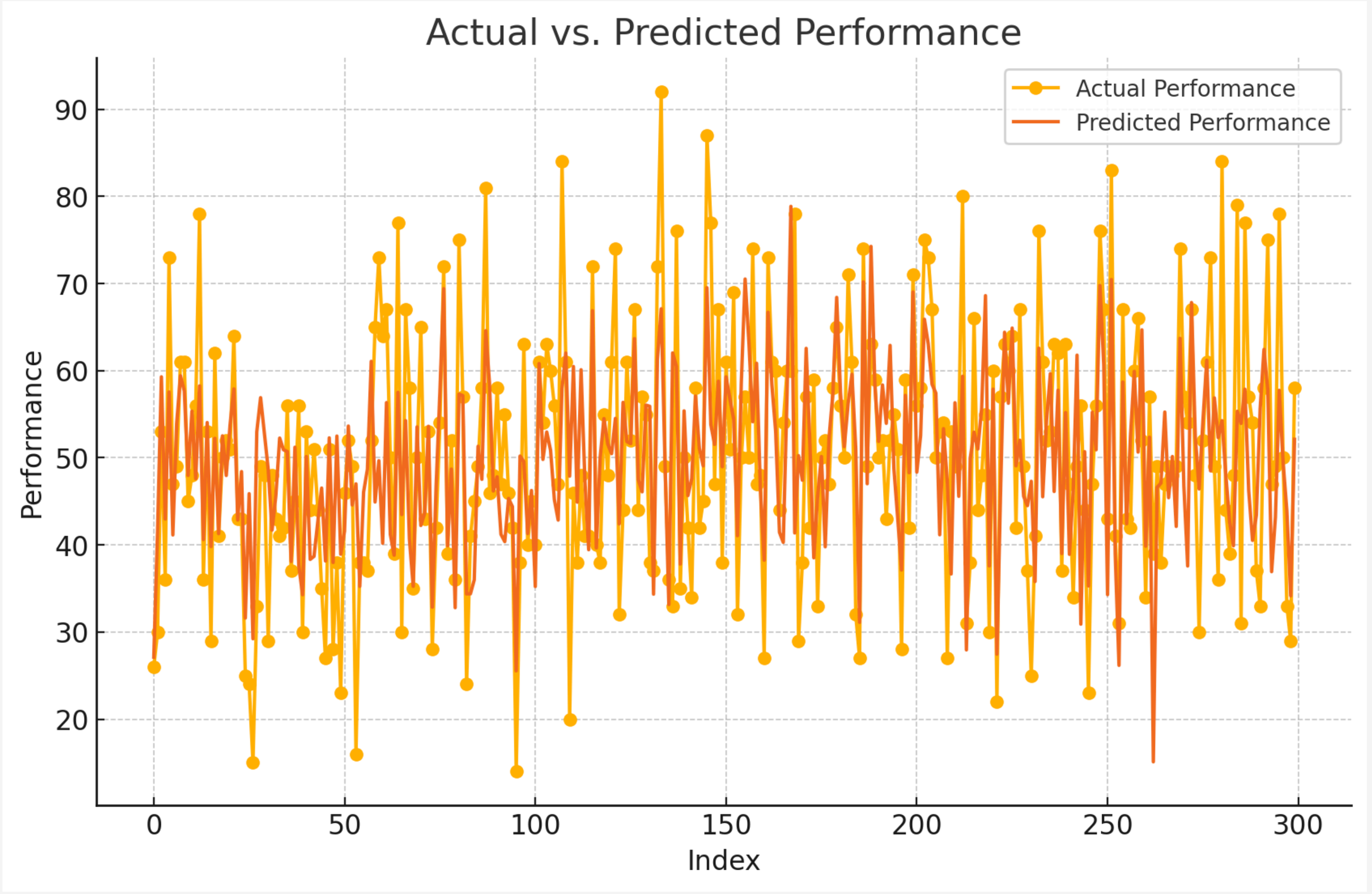
CODE

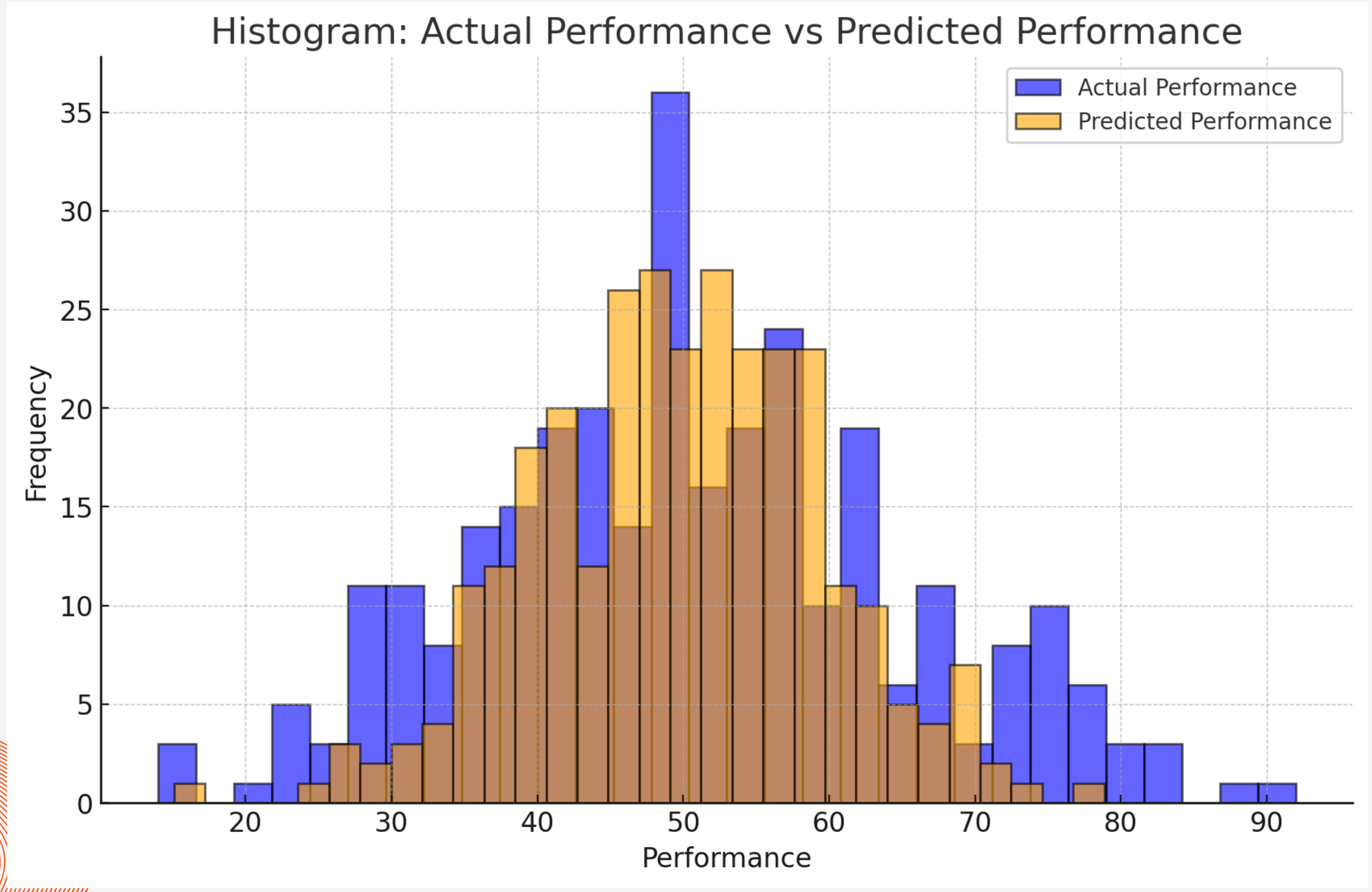
```
data = pd.read_csv(data_path)


categorical_features = ['stress']
numeric_features = ['study_hrs', 'sleep_hrs', 'high_math', 'high_hist', 'high_eng']


preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
        ('cat', OneHotEncoder(), categorical_features)
    ])

X = data.drop('performance', axis=1)
y = data['performance']
X = preprocessor.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=0)


model = Sequential([
    Dense(10, activation='linear', input_dim=X_train.shape[1]),
    Dense(1)
])


model.compile(optimizer='adam', loss='mean_squared_error')


model.fit(X_train, y_train, epochs=100, batch_size=25, verbose=1)
```

Predicted vs. Actual

Predicted vs. Actual

A	B
performance	predicted_perf
26	27.1166334
30	41.7634552
53	59.2916233
36	42.9727047
73	57.5489143
47	41.1400406
49	53.9419719
61	59.4093872
61	57.659693
45	47.9853039
48	55.3457448
56	47.6034405
78	58.2507603
36	40.6145719
53	54.0450943
29	39.8189923
62	52.2251317
41	41.2938456
50	52.5171539
52	47.9681001


Actual vs. Predicted Performance
Actual Performance
Predicted Performance
Performance
Index
90
80
70
60
50
40
30
20
0
50
100
150
200
250
300


Histogram: Actual Performance vs Predicted Performance
Actual Performance
Predicted Performance
Frequency
Performance
0
5
10
15
20
25
30
35
20
30
40
50
60
70
80
90

THANK YOU FOR LISTENING