Training Dataset:

		Qualitative	Quantitative		
Name	MONASH EMAIL (@student.monash.edu)	Gender	Height	Weight	Signature
Liu Chang	cliu0174	Male	175	57	Liu Chang
Chen Yang	yche0822	Male	182	65	Chen Yang
Zhao Tingting	tzha0176	Female	170	55	Zhao Tingting
Ni Tianqi	tnii0006	Male	178	58	Ni Tianqi
Yang Chen	cyan0112	Male	184	75	Yang Chen
Su xiaoyuan	xsuu0025	Female	173	68	Su xiaoyuan
Ma Yonghao	ymaa0138	Male	179	63	Ma Yonghao
Zang Hao	hzan0005	Male	176	65	Zang Hao
Huang Qiming	qhua0048	Female	171	53	Huang Qiming
Qian Yiwei	yqia0074	Male	180	60	Qian Yiwei
Wang Huai	hwan0321	Male	177	62	Wang Huai
Yan Jiale	jyan0261	Male	181	71	Yan Jiale
Liu Hao	hliu0183	Female	172	57	Liu Hao
Zhou Yuyang	yzho0299	Male	183	67	Zhou Yuyang
Ye Yuxin	yyee0057	Female	174	66	Ye Yuxin

Quantitative data:

- **Height**
- **Weight**

Qualitative data:

- **Gender**

Test Dataset:

Zhao Xin	xzha0531	Male	177	60	Zhao Xin
Yan Sijia	syan0174	Male	175	63	Yan Sijia
Wang Xiwei	xwan0472	Male	176	70	Wang Xiwei
Chen Chuxin	cche0344	Male	179	64	Chen Chuxin
Li Weizhen	wlii0178	Female	172	60	Li Weizhen
You Changjiang	cyou0030	Male	181	69	You
Chen Aowen	ache0188	Male	178	66	Chen Aowen
Yan Jin	jyan0262	Female	174	65	Yan Jin
Ye Zihan	zyee0063	Male	180	72	Ye Zihan
Liu Wenjing	wliu0089	Female	173	56	Liu Wenjing

Ten samples from Group 35's dataset.

Calculate linear regression：

```
# Train models
female_model = LinearRegression()
female_model.fit(X_female_train, y_female_train)

male_model = LinearRegression()
male_model.fit(X_male_train, y_male_train)
```

We build separate linear models of height and weight for males and females.

Calculating MAE：

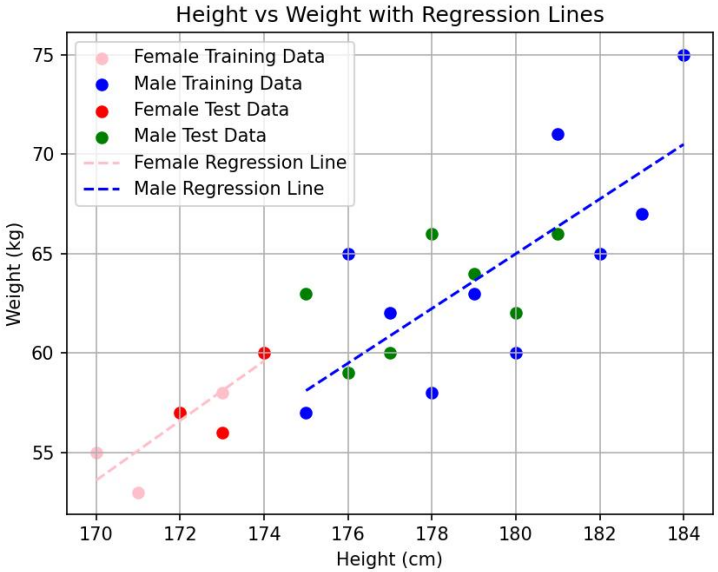
```
# Calculate MAE for training set
mae_female_train = mean_absolute_error(y_female_train, predicted_weight_female_train)  # Female MAE
mae_male_train = mean_absolute_error(y_male_train, predicted_weight_male_train)        # Male MAE
print("\nFemale Training Set - Mean Absolute Error (MAE):", mae_female_train)
print("Male Training Set - Mean Absolute Error (MAE):", mae_male_train)
# Calculate MAE for test set
mae_female_test = mean_absolute_error(y_female_test, predicted_weight_female_test)  # Female MAE
mae_male_test = mean_absolute_error(y_male_test, predicted_weight_male_test)        # Male MAE
print("\nFemale Test Set - Mean Absolute Error (MAE):", mae_female_test)
print("Male Test Set - Mean Absolute Error (MAE):", mae_male_test)
```

Visualizing Linear Regression Plots：

```
# Female regression line
all_female_heights = np.linspace(min(female_train_data['Height']), max(female_train_data['Height']), 100)
all_female_weights = female_model.predict(all_female_heights.reshape(-1, 1))
plt.plot(all_female_heights, all_female_weights, color='pink', linestyle='--', label='Female Regression Line')

# Male regression line
all_male_heights = np.linspace(min(male_train_data['Height']), max(male_train_data['Height']), 100)
all_male_weights = male_model.predict(all_male_heights.reshape(-1, 1))
plt.plot(all_male_heights, all_male_weights, color='blue', linestyle='--', label='Male Regression Line')
```

Visualizing Linear Regression Plots:



Although we incorporated gender as a factor to build separate linear models, the difference between the two models does not appear to be particularly significant.

MAE results for the training set and test set:

```
Female Training Set - Mean Absolute Error (MAE): 0.8799999999999955
Male Training Set - Mean Absolute Error (MAE): 3.1599999999999966

Female Test Set - Mean Absolute Error (MAE): 0.9666666666666686
Male Test Set - Mean Absolute Error (MAE): 1.96277056277056
```

Table 1: Contingency table of height and prediction error

	Low Error	High Error	Total
Short	6/4 ((6 - 4)^2 / 4) = 0.70	2/3 ((2 - 3)^2 / 3) = 0.80	8
Tall	2/3 ((2 - 3)^2 / 3) = 0.80	5/3 ((5 - 3)^2 / 3) = 0.92	7
Total	8	7	15

Analysis: Relationship Between Height (High-Low Grouping) and Prediction Error

$\chi 2=1.64$, $p=0.2007 \geq 0.05$ → No statistically significant difference

Conclusion : Height (high-low grouping) does not significantly affect the error.

Table 2: Contingency table of gender and prediction error

	Low Error	High Error	Total
Female	5/2 ((5 - 2)^2 / 2) = 2.04	0/2 ((0 - 2)^2 / 2) = 2.33	5
Male	3/5 ((3 - 5)^2 / 5) = 1.02	7/4 ((7 - 4)^2 / 4) = 1.17	10
Total	8	7	15

Analysis: Relationship Between Gender and Prediction Error

$\chi 2=4.05$, $p=0.0441<0.05$ → Statistically significant difference

Conclusion : Gender has a significant impact on prediction error.

Comparison of weight predictions from two models:

Height	Weight	Gender	Pred_Height	Pred_HeightGender	Error_Height	Error_HeightGender
177	60	Male	61.73	60.83	1.73	0.83
174	60	Female	58.29	59.38	1.71	0.62
179	64	Male	64.03	63.61	0.03	0.39
172	57	Female	56.0	56.6	1.0	0.4
181	66	Male	66.32	66.38	0.32	0.38
175	63	Male	59.44	58.05	3.56	4.95
178	66	Male	62.88	62.22	3.12	3.78
173	56	Female	57.15	57.99	1.15	1.99
180	62	Male	65.17	64.99	3.17	2.99
176	59	Male	60.59	59.44	1.59	0.44

Task 1: One-sample t-test on MAE < 5 (Height + Gender model)

H_0: $\mu = 5$

H_1: $\mu < 5$

n = 10, df = 9

$\bar{x}$ = 1.678, s = 1.677, SE = 0.530

t = -6.263, t(0.05, 9) = -1.833, p = 0.0001

Conclusion: Reject H_0 (MAE < 5 is significant)

In the height-and-gender model, the MAE being below 5 is considered significant, with an average MAE around 1.

Task 2: Paired t-test (Error_Height - Error_HeightGender)

H_0: $\mu_{diff} = 0$

H_1: $\mu_{diff} > 0$

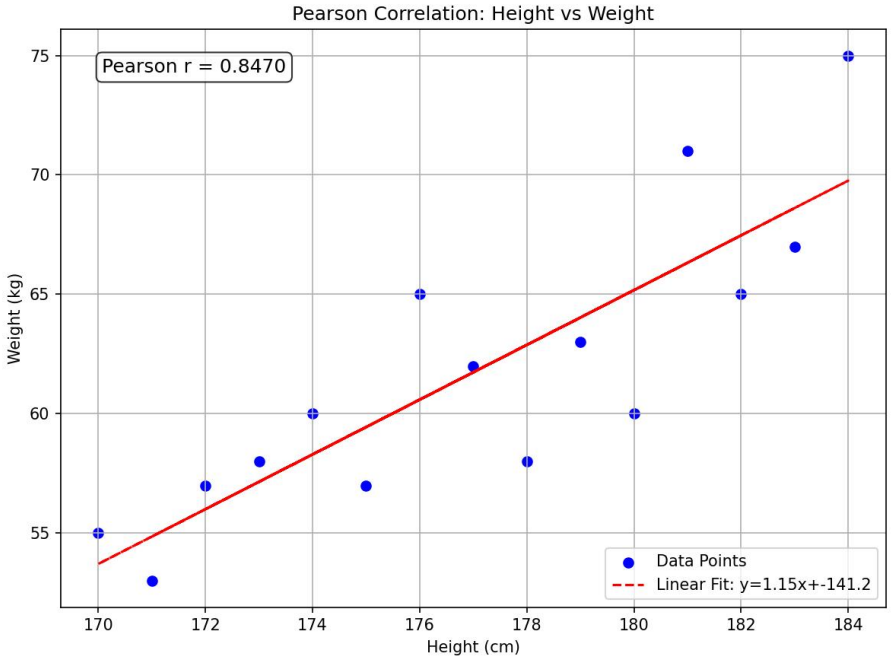
$\bar{d}$ = 0.059, s = 0.876, SE = 0.277

t = 0.212, t(0.05, 9) = 1.833, p = 0.4185

Conclusion: Retain H_0 (no significant difference)

In contrast, the height-only model does not show a significant difference in average error on the current test dataset.

Pearson Correlation: Height vs Weight



We analyze the relationship between height (cm) and weight (kg) using correlation analysis.

One statistical techniques are used: Pearson (for linear relationship) .

Technical Details

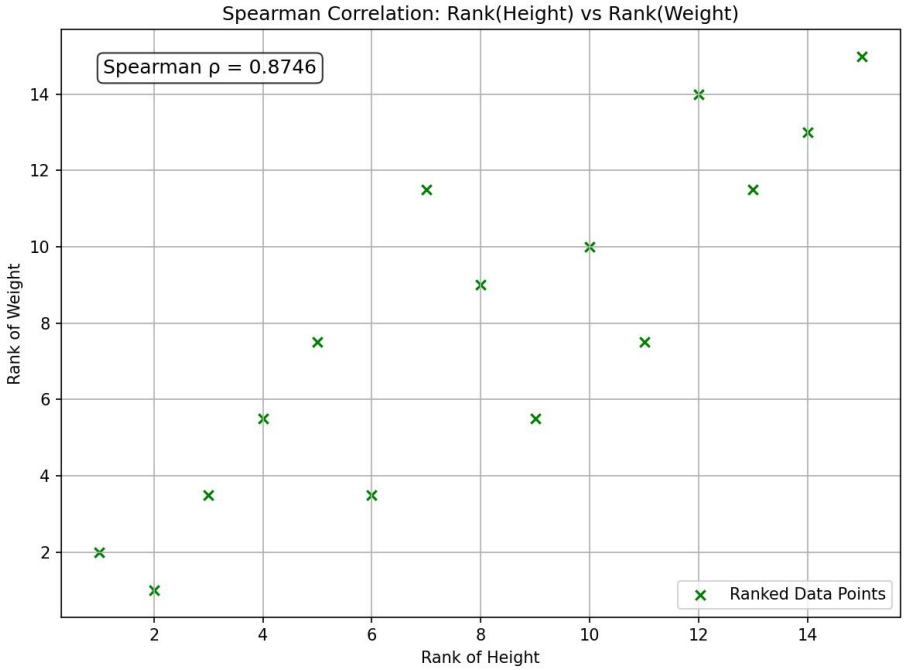
Mean Height = 177.00, Mean Weight = 61.73
Cov(Height, Weight) = 22.9286
Std Height = 4.4721, Std Weight = 6.0529
Pearson Correlation (manual) = 0.8470
Pearson Correlation (scipy) = 0.8470, p = 0.0001

Evaluation: Pearson correlation is statistically significant.

Conclusion:

Pearson correlation coefficients are positive, indicating that as height increases, weight tends to increase as well. Pearson supports a linear relationship.

Spearman Correlation: Rank(Height)vs Rank(Weight)



We analyze the relationship between height (cm) and weight (kg) using correlation analysis.

One statistical techniques are used:

Spearman (for monotonic relationship).

Technical Details

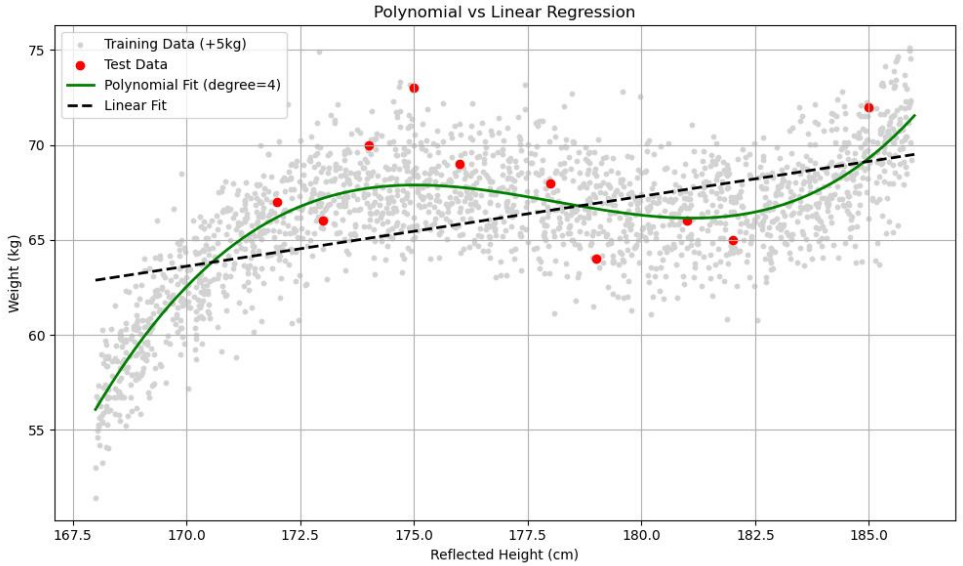
Rank(Height) = [6. 13. 1. 9. 15. 4. 10. 7. 2. 11. 8. 12. 3. 14. 5.]
Rank(Weight) = [3.5 11.5 2. 5.5 15. 5.5 10. 11.5 1. 7.5 9. 14. 3.5 13. 7.5]
Cov(Rank Height, Rank Weight) = 17.4286
Std Rank Height = 4.4721, Std Rank Weight = 4.4561
Spearman Correlation (manual) = 0.8746
Spearman Correlation (scipy) = 0.8746, p = 0.0000

Evaluation: Spearman correlation is statistically significant.

Conclusion:

Spearman correlation coefficients are positive, indicating that as height increases, weight tends to increase as well. Spearman supports a general monotonic trend.

Polynomial vs Linear Regression



Non-linear Model Information:

Model: Polynomial Regression (degree=4)
Data: 2000 augmented training samples
Fitted Function:

$$f(x) = -16362.1578 - 122.8993x + 5.1850x^2 - 0.0350x^3 + 0.0001x^4$$

Train MAE (Polynomial): 1.5999
Test MAE (Polynomial): 1.8291

Linear Model Information:

Model: Linear Regression
Same data used as above
Fitted Function: $f(x) = 1.0373 + 0.3681 \cdot x$
Train MAE (Linear): 2.2728
Test MAE (Linear): 3.1494

Explanation of the Non-linear

We use a degree-4 polynomial regression model because the relationship between height andweight is not perfectly linear-real-world data often exhibits subtle curvature and fluctuations. Ahigher-degree polynomial captures these non-linear patterns better than a straight line.

Why MAE is Lower

The non-linear model achieves a lower Mean Absolute Error (MAE) because it fits the training datamore closely while maintaining good generalization on the test set. lt adjusts for the non-lineartrends in the data that a linear model cannot capture.

Conclusion:

The non-linear model performs better than the linear model based on MAE.
Polynomial regression captures subtle nonlinear fluctuations in the data,resulting in better generalization on real test data.

MONASH University