Assignment I

Yasemin Sarpkaya 29172
Ege Kaan Özalp 28989

QUESTION 3

Part a

53,43720311

-3,480152213

2,950337894

-1,9640234

Rating=53.4372−3.4802×Fat+2.9503×Fiber−1.9640×Sugars

Part b:

When the independent variable Fiber increases the dependent variable also increases. Indicating a positive relation between Fiber and Rating. Rest of the independent variables have a negative relationship with the dependent variable. For instance, when other variables are constant for each additional unit of fat, the rating will decrease by 3.48 points. So when Fat or Sugars increase, Ratings decrease. To conclude, cereals with higher fat and sugar content are more likely to have lower ratings while creals with higher fiber content are more likely to have higher ratings based on the multiple linear regression model we have found in part a.

Part c:

- **Total Sum of Squares (SST):** 14996.80
- **Regression Sum of Squares (SSR):** = 12995,83
- **Residual Sum of Squares (SSE):** 2000.97
- **R-squared:** 0.866
- **Correlation Coefficient:** 0.93
- **Mean Absolute Error (MAE):** 4.18
- **Root Mean Squared Error (RMSE):** 5.098
- **Relative Absolute Error (RAE):** 0.3823
- **Root Relative Squared Error (RRSE):** 0.3653

Part d:

PHYTON code results

```
Intercept: 53.437203112654856
Coefficient for Fat: -3.4801522125761855
Coefficient for Fiber: 2.950337893751642
Coefficient for Sugars: -1.9640233997901102
Rating = 53.4372 + (-3.4802) * Fat + (2.9503) * Fiber + (-1.9640) * Sugars
```

What we have found at part a:

Rating=53.4372−3.4802×Fat+2.9503×Fiber−1.9640×Sugars

```
{'SST': 14996.800399790134, 'SSR': 12995.687584213987, 'SSE':
2000.8734936350077, 'R_squared': 0.8665799743748668, 'Correlation
Coefficient': 0.9309027738988791, 'MAE': 4.183154831168831, 'RMSE':
5.097584726829745, 'Relative Absolute Error': 0.3822806928219616, 'Root
Relative Squared Error': 0.36526706069002884}
```

Part e:

When we enter the relative values to the equation model we have found

Rating=53.4372−3.4802×1 + 2.9503×13 −1.9640×0

Rating = 88.31

Part f:

In forward selection we start with an empty model and iteratively add features to increase model accuracy.

Selecting based on MSE

```
Best Feature Subset: ('calories', 'protein', 'fat', 'sodium', 'fiber', 'carbo', 'sugars', 'potass', 'vitamins', 'weight', 'cups')
Lowest MSE: 7.761705534786187e-14
```

Selecting based on adjusted R2

```
Best Feature Subset: ('calories', 'protein', 'fat', 'sodium', 'fiber', 'carbo', 'sugars', 'potass', 'vitamins')
Highest Adjusted R-squared: 0.9999999999999987
```

In backward elimination we start with all features and eliminate features iteratively until we end up with the features that lead to most effective predictions.

Eliminating features based on MSE

```
Best Feature Subset: ['calories', 'protein', 'fat', 'sodium', 'fiber', 'carbo', 'sugars', 'potass', 'vitamins', 'cups']
Lowest MSE: 7.830877612528858e-14
```

Eliminating features based on adjusted R2

```
Best Feature Subset: ['calories', 'protein', 'fat', 'sodium', 'fiber', 'carbo', 'sugars', 'potass', 'vitamins']
Highest Adjusted R-squared: 0.9999999999999987
```

While using MSE the model found all features useful and added them all to the model. This might cause the model to overfit for future instances. We got R^2 close to 1.0 in both forward selection and backward elimination, indicating that the accuracy is almost as good as MSE. This means we got almost a perfect fit with only 9 of the features, instead of including all features to the model. This prevents overfitting while providing accurate predictions.

Part g:

	Model	RMSE	MAE	MAPE
0	OLS	5.097585	4.183151	10.430002
1	Ridge	5.121186	4.201541	10.518289
2	Lasso	5.133491	4.193542	10.499429

When we compared the results in terms of RMSE, we see that OLS has the lowest value among the three. This means that OLS is better at minimizing the error compared to Ridge and Lasso. For MAE we observed that the values are quite similar to each other, but OLS performs slightly better. For MAPE, OLS performs better once again. Since Lasso and Ridge methods use regularization to avoid overfitting, OLS might seem to be performing better.