1. Conceptualize a multiple linear regression model considering “weight” as a dependent variable and a set of independent variables, including binary variables for the model year. Write the regression equation in the document.

   To conceptualize a multiple linear regression model with "weight" as the dependent variable and a set of independent variables, including binary variables for the model year, we can write the regression equation as follows:

   $weight = \beta_0 + \beta_1 * foreign + \beta_2 * mpg + \beta_3 * cylinders + \beta_4 * displacement + \beta_5 * hp + \beta_6 * acceleration + \beta_7 * modyr70\text{-}82 + \varepsilon$

   Where:

   - β_0 is the intercept term.
   - $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are the coefficient of the independent variables.
   - ε is the error term.

   I selected those independent variables for my model because:

   - I chose variable foreign because the vehicle's foreign or domestic origin could impact its weight. Vehicles manufactured in different regions may have different design standards, materials, and regulations, which can affect their weight. Therefore, I assume that there could be a difference in weight between foreign and domestic vehicles, with one group tending to be heavier than the other.
   - I chose variable mpg because it is often inversely related to weight. Lighter vehicles tend to have better fuel efficiency, so I assume there might be a negative relationship between mpg and vehicle weight, where vehicles with higher gas mileage tend to be lighter.
   - I chose variable cylinders because the number of cylinders in the engine can indicate the size and power of the vehicle. Larger engines with more cylinders tend to be heavier, so I assume there might be a positive relationship between cylinder and vehicle weight, where vehicles with more cylinders (larger engines) tend to be heavier.
   - I chose variable displacement because engine displacement is a measure of the volume swept

by all the pistons in the cylinders of an engine. Larger displacement engines generally weigh more, so I assume there might be a positive relationship between displacement and vehicle weight, where vehicles with higher engine displacement tend to be heavier.

- I chose variable Hp because it is a measure of the engine's power output. Vehicles with higher horsepower often have larger, more powerful engines, which could contribute to higher weight. Therefore, I assume there might be a positive relationship between Hp and vehicle weight, where vehicles with higher horsepower tend to be heavier.
- I chose variable acceleration because it measures how quickly a vehicle can increase its speed. While there might not be a direct relationship between acceleration and weight, heavier vehicles might have lower acceleration due to their mass. Therefore, I assume there might be a negative relationship between acceleration and vehicle weight, where heavier vehicles might have slightly lower acceleration.
- I chose variable Modyr70-82 because it can reflect advancements in technology and changes in design practices over time, which could affect vehicle weight. Therefore, I assume that there could be differences in weight between vehicles from different model years.

2. Present summary statistics (min, max, mean, median, standard deviation, first quartile, and third quartile) for the variables in your model in a table and briefly comment on the summaries.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Sd
weight	1613	2209	2750	2953	3582	5140	849.86
mpg	9	17.5	23	23.79	29.80	46.6	7.997
cylinders	3	4	4	5.43	8	8	1.70
displacement	68	98	144	191.5	260.5	455	104.47
hp	46	75	92	103.5	122.8	230	38.42
acceleration	8	13.88	15.5	15.62	17.23	24.8	2.82
modelyr	70	73	76	76.03	79	82	3.72
foreign	0	0	0	0.39	1	1	0.49

From the observation, we can see that the mean value is close to the median value of cylinder and modelyr variables, which suggests that the data is symmetrically distributed around the mean. This also indicates a normal distribution or approximately normal distribution. On the other hand, we can see that the mean of weight and mpg variables are significantly different from the median, which suggests skewness in the data.

3. Calculate a pairwise correlation matrix for your model variables (while calculating the correlation matrix, only include the dependent variable and the non-binary independent variables that you selected for the model). Present the results in a table and comment on the correlation matrix

	weight	mpg	cylinders	displacement	hp	acceleration	modelyr

weight	1.0	-0.83	0.90	0.93	0.87	-0.43	-0.30
mpg	-0.83	1.0	-0.78	-0.80	-0.78	0.44	0.58
cylinders	0.90	-0.78	1.0	0.95	0.84	-0.51	-0.35
displacement	0.93	-0.80	0.95	1.0	0.90	-0.55	-0.36
hp	0.87	-0.78	0.84	0.90	1.0	-0.69	-0.41
acceleration	-0.43	0.44	-0.51	-0.55	-0.69	1.0	0.23
modelyr	-0.30	0.58	-0.35	-0.36	-0.41	0.27	1.0

From the correlation matrix table, we can see that weight has a strong positive relationship with cylinders, displacement, and hp. On the other hand, it has a strong negative relationship with mpg and acceleration. This matches my assumption from question 1.

4. Using R, generate your model and follow these steps:

- Calculate VIF for the independent variables of your model and show the values in a table:

	VIF
mpg	5.60
cylinders	11.28
displacement	20.34
hp	9.70
acceleration	2.12
foreign	2.24
modyr70-82	inf

After running the multiple regression model the first time and then calculating the VIF, we got the VIF value of modyr70-82 as infinite. This indicates perfect multicollinearity between that variable and the other independent variables in the model, Also, during the first testing, the VIF values of cylinders and displacements were greater than 10, which implies that they may need further examination, so we kept them and remove modyr70-82 from the model and re-run the model.

After running the model the second time, the VIF values of cylinders and displacements were still greater than 10, so we decided to remove them and re-run the model.

- Show the final model results in a table:

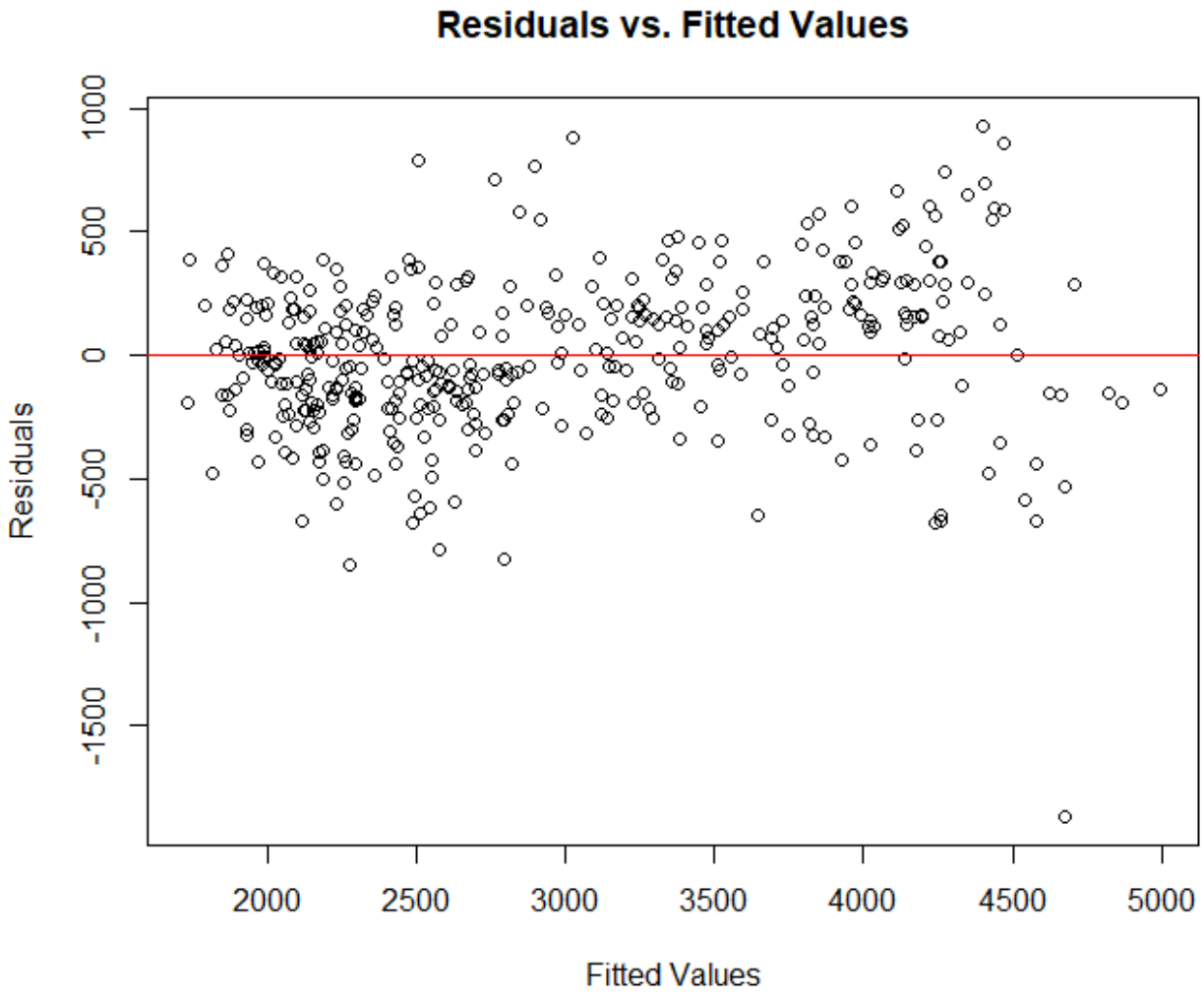
	Estimate	Std.Error	t	p
Intercept	707.39**	238.87	2.96	0.00325
mpg	-26.93***	3.45	-7.81	4.98e-14
hp	17.07***	0.85	20.18	< 2e-16
acceleration	78.26***	8.01	9.77	< 2e-16
foreign	-263.81***	40.02	-6.59	1.38e-10

The coefficient -26.93 means that while holding all other factors fixed if the vehicle increases its gas mileage by one, the weight of the car will decrease by 26.93, which translates to an approximate 269.3% decrease in weight.

The coefficient 17.07 means that while holding all other factors fixed if the vehicle increases its horsepower by one, the weight of the car will increase by 17.07, which translates to an approximate 170.7% increase in weight.

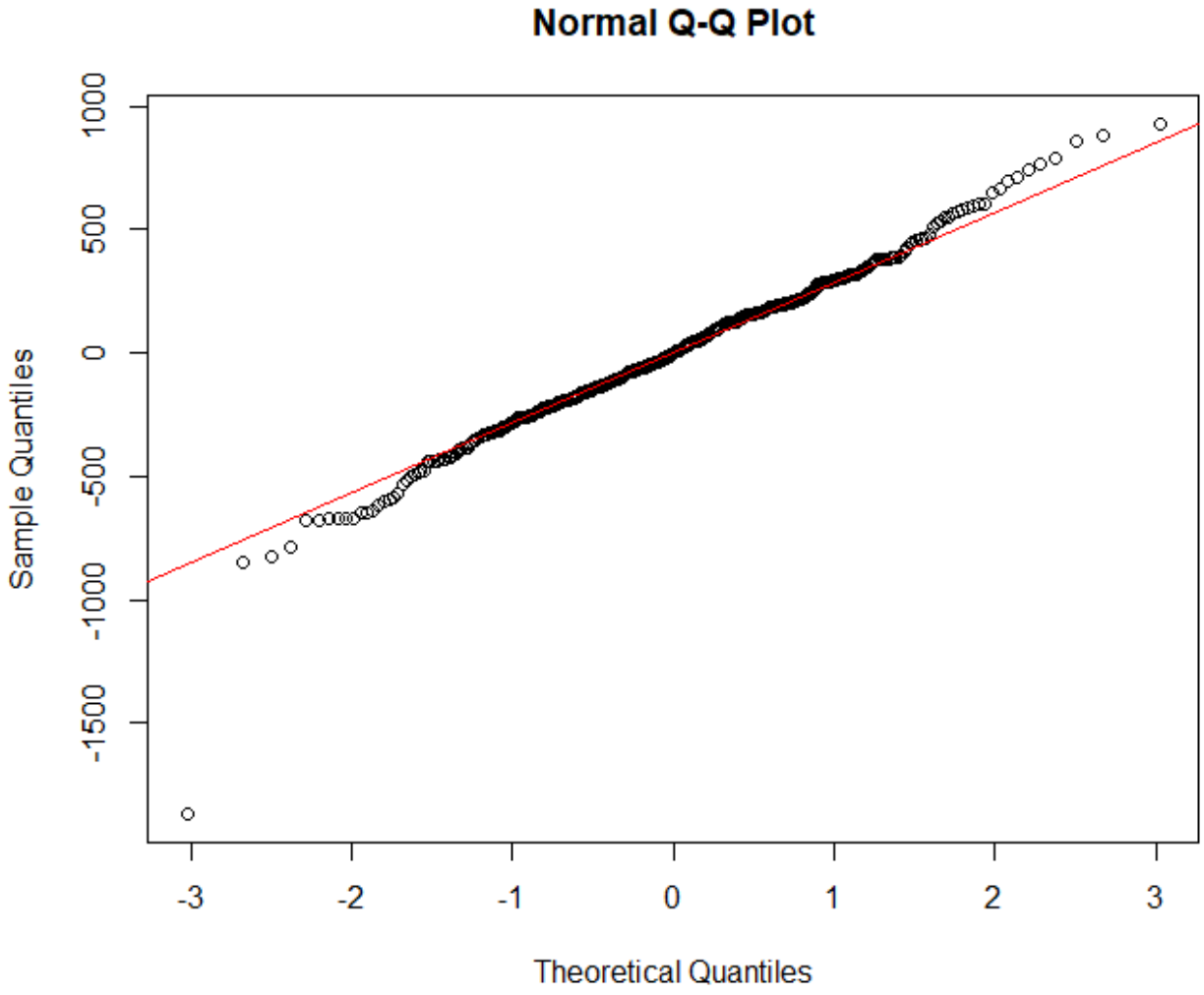
5. Check the following assumptions of the linear regression model

- Linearity:



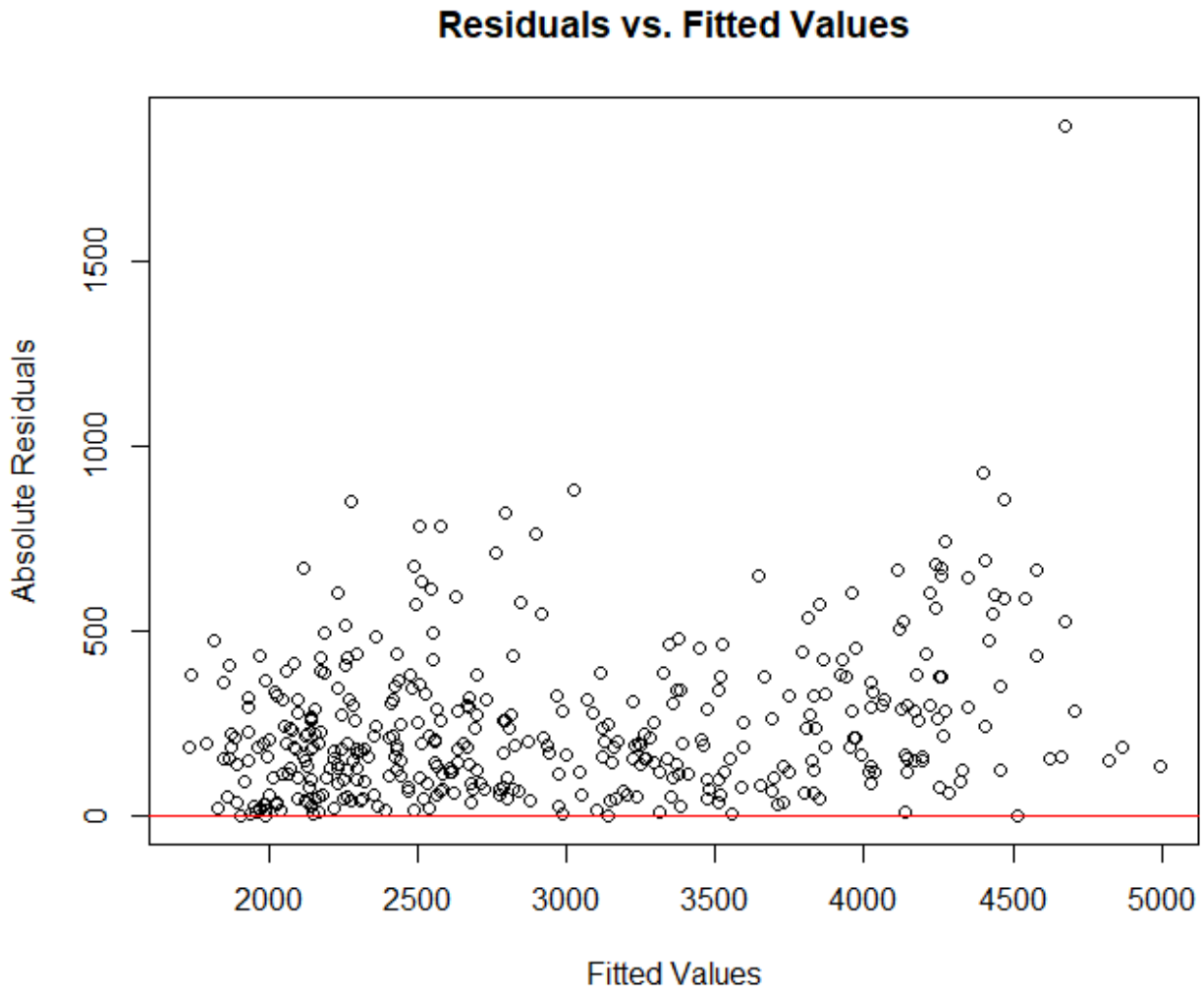
From the plot, you can see the points are randomly scattered around the red reference line (representing perfect linearity), which implies that the assumption of linearity is met.

- Normality in Errors



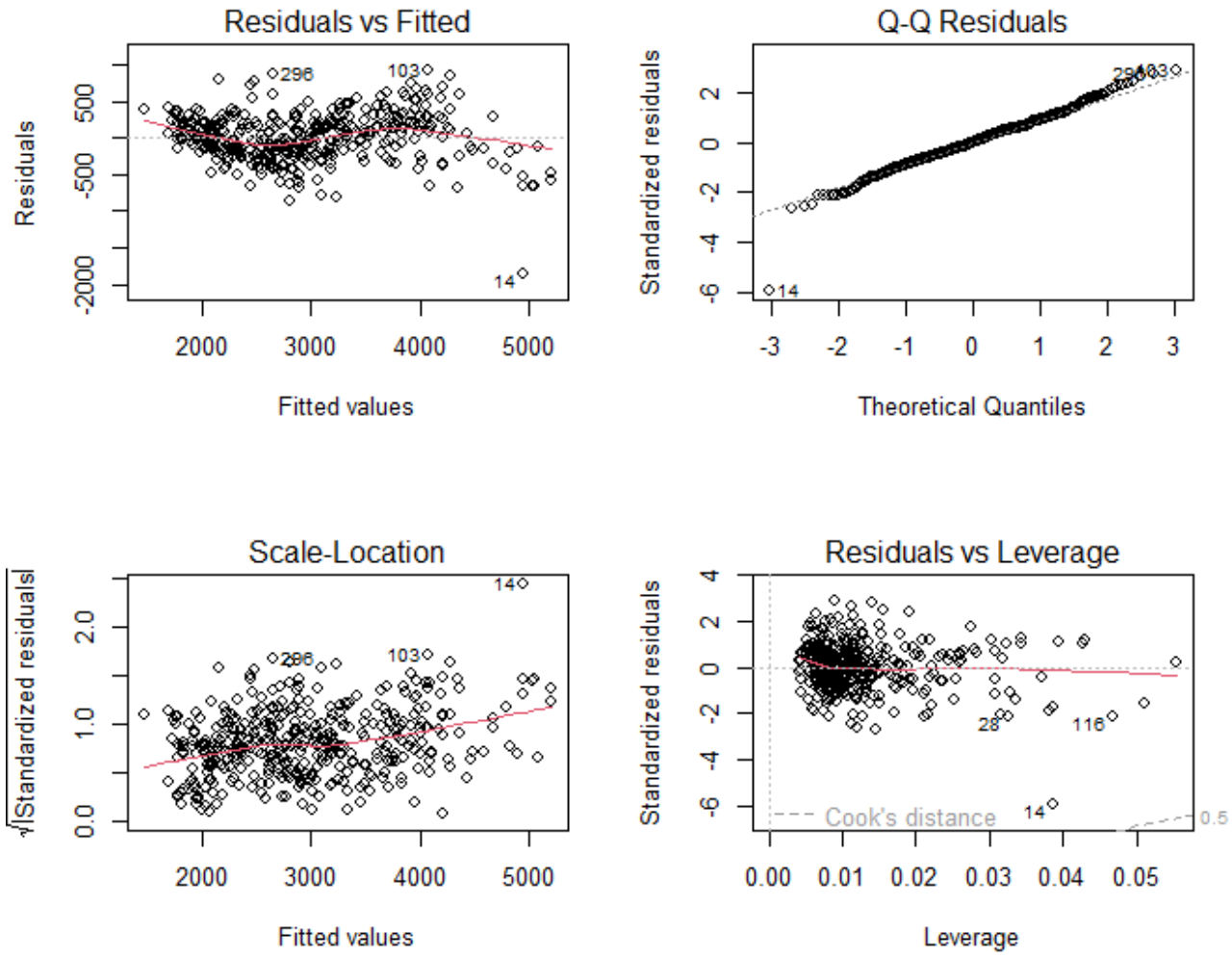
From the plot, you can see the points fall approximately along the red diagonal line. This shows that the residuals are normally distributed.

- Homoskedasticity



From the plot, you can see the spread of residuals appears relatively constant across all values of the predicted values. This shows that the assumption of homoskedasticity is met.

- Unusual and influential observations:



From these plots, you can see some data points deviate significantly from the overall pattern of the data. They may represent extreme values or errors in the data collection process.

6. Do American cars run less mileage per gallon than foreign cars? Use an appropriate statistical test (make sure to write each step and the relevant values in detail in the document)

Since foreign is a binary variable (1 for foreign, 0 for domestic), a negative coefficient would indicate that foreign cars have higher mileage per gallon than domestic cars. In our case, we got the coefficient for the foreign variable as -263.8045, which proves that foreign cars have higher mileage per gallon than American cars. In addition, we need to extract the p-value from the model to check the significance of the coefficient associated with the foreign variable. In our case, the p-value is 1.382677e-10, which is less than the alpha of 0.01. This means we reject the null hypothesis and we have the confidence that the difference in mileage per gallon between American and foreign cars is statistically significant.

7. Do the model year binary variables jointly have explanatory power? Use an appropriate statistical test (make sure to write each step and the relevant values in detail in the document)

To test whether the model year binary variables jointly have explanatory power in the regression model, I use the variance (ANOVA) test. This test compares the fit of the full model (including all predictor variables, including the model year binary variables) to a reduced model (excluding these binary variables). After doing this test, I extracted the p-value from the ANOVA and got the p-value as 2.908971e-05. Since the p-value is much smaller than the typical significance level of 0.05, we reject the null hypothesis. This shows the full model, which includes the model year binary variable fits the data substantially better than the reduced model, which does not include these variables. It also implies that the model year binary variables jointly have explanatory power in predicting the weight of cars.

8. Calculate the model’s goodness of fit and comment.

To calculate the model's goodness of fit, I calculate and get the result of R-squared as 0.86, adjusted R-squared as 0.859, RMSE as 317.28, and RSE as 319.21. I think my linear regression model is a reasonably good fit for the data. The high R-squared and adjusted R-squared values suggest that a significant portion of the variance in the dependent variable is explained by the independent variables. Additionally, the relatively low RMSE and RSE values indicate that the model's predictions are close to the actual observed values