**Output of Questions**

Q1 :

**Output :**

Please enter the car's fuel efficiency (miles/gallon):50

Please enter the size of the fuel tank (in gallons): 30

What would you like to do:

1. See current fuel level

2. Drive

3. Add gas

4. Exit

1

You have 0.00 gallons of gas left.

What would you like to do:

1. See current fuel level

2. Drive

3. Add gas

4. Exit

2

How many miles do you want to drive? 3

You don't have enough fuel to drive 3.0 miles. You can drive another 0.0 miles on this gas.

What would you like to do:

1. See current fuel level

2. Drive

3. Add gas

4. Exit

28

Invalid choice. Please enter a number between 1 and 4.

What would you like to do:

1. See current fuel level

2. Drive

3. Add gas

4. Exit

2

How many miles do you want to drive? 30

You don't have enough fuel to drive 30.0 miles. You can drive another 0.0 miles on this gas.

What would you like to do:

1. See current fuel level

2. Drive

3. Add gas

4. Exit

5

Invalid choice. Please enter a number between 1 and 4.

What would you like to do:

1. See current fuel level

2. Drive

3. Add gas

4. Exit

4

GoodBye

**Q2 :**

Q1 : There may be duplicate records in the data. Remove them. How many records do you have now?

A black text on a white background

Description automatically generated with low confidence

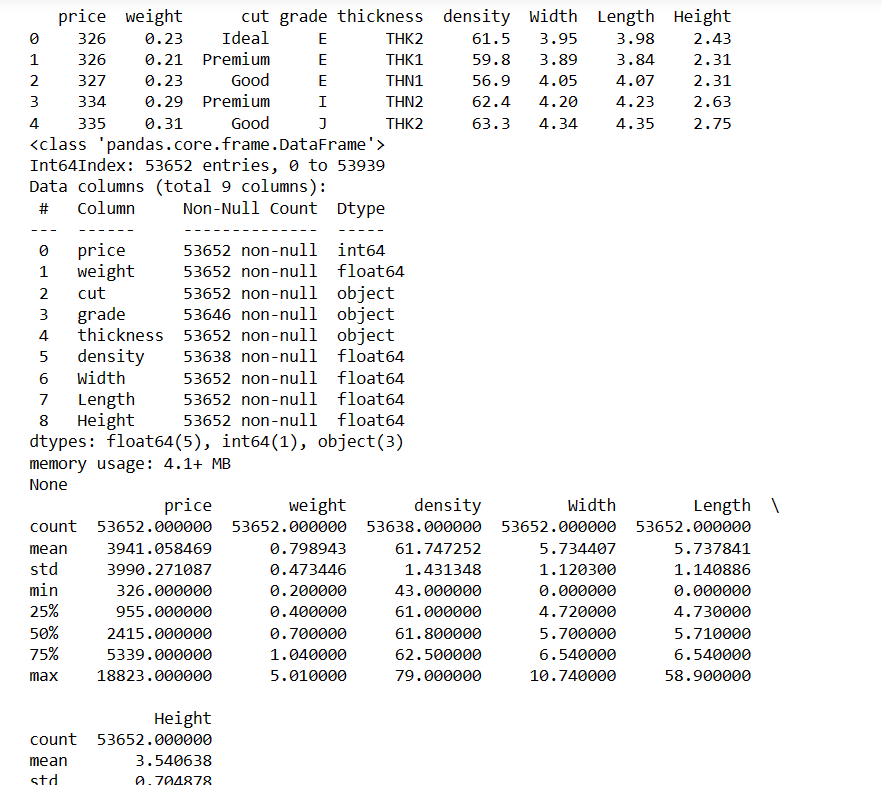
Q2 : Draw a Histogram of the Price variable. Is it a bell curve? If not, what is it?

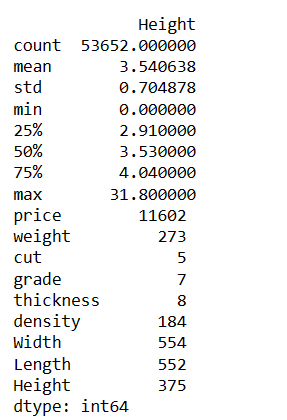
A picture containing text, screenshot, diagram, plot

Description automatically generated

**The histogram of the Price variable reveals that the distribution is positively right-skewed with a long tail to the right. This means that there are some very high-priced aluminum sheets in the dataset.**

Q3 : Do some basic data exploration (e.g. using commands as head( ), info( ), describe( ), nunique( ), etc). Which variables will you NOT select?



****

**Answer :**

Based on the data exploration, the variables that could potentially be excluded from the model are:

From the output, we can see that the grade and thickness variables are strings(objects) and have a large number of unique values compared to the other variables.

Thickness: it has only 8 unique values, and it's not clear how it relates to the other variables. Grade: it has only 7 unique values, and it's not clear how it relates to the other variables. We will choose not to select these variables for our analysis.

Q4 : Are there any outliers in the data? What about missing values? If any of either, treat them.

**A picture containing screenshot, diagram, rectangle, line

Description automatically generated**

**A picture containing text, screenshot, font

Description automatically generated**

**A screenshot of a computer

Description automatically generated with medium confidence**

Q5 : Partition the data into a training set (with 70% of the observations), and testing set (with 30% of the observations) using the random state of 12345 for cross validation.

**Output :**

X\_train shape: (37542, 25)

y\_train shape: (37542,)

X\_test shape: (16090, 25)

y\_test shape: (16090,)

Q6 : On the partitioned data, build the best KNN model. Show the accuracy numbers. (Hint: What is the best value of k? How do you decide the ‘best k’?)

**Output :**

Best k: 4

Accuracy: 0.8428772886647922

**A picture containing line, plot, diagram, text

Description automatically generated**

Train set accuracy: 0.9952053699856162

Test set accuracy: 0.14195152268489744

Q7 : On the partitioned data, build the best logistic regression model. Show the accuracy numbers

Output :

Train set accuracy: 0.09762399445953865

Test set accuracy: 0.09683032939714108

Q8 : Based on the results of k-nearest neighbor, and logistic regression, what is the best model to classify the data? Provide explanation to support your argument.

Answer :

Based on the results, KNN model performs significantly better than logistic regression. The accuracy of the KNN model on the training set is very high, at 0.995, indicating that it is overfitting the data. However, on the test set, it still performs better than logistic regression with an accuracy of 0.142.

On the other hand, the logistic regression model performs very poorly on both the training and test sets, with an accuracy of only 0.098.

Therefore, based on these results, the KNN model is the better choice for classifying the data. However, it is important to note that further analysis and fine-tuning of the models may be necessary to achieve even better results.