**Question A. Automobile Selling Price (15 points total)**

1.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Independent or Dependent variable? | What is the name of the variable that measures this concept in the ToyotaCorrolla dataset? | Is the feature you named in the previous column measured as a categorical or continuous variable in the dataset? |
| vehicle’s age | Independent | Vehicle\_Age | Continuous |
| mileage | Independent | KM | Continuous |
| manufacturing year | Independent | Mfg\_Year | Continuous |
| sales price | Dependent | Price | Continuous |
| fuel type | Independent | Fuel\_Type | Categorical |

2. To prepare the dataset for analysis, I performed the following data cleaning steps:

1. **Selected Relevant Columns**: I focused on the variables needed for the analysis: Price, Vehicle\_Age, KM, Mfg\_Year, and Fuel\_Type.
2. **Checked for Missing Values**: I examined each selected column for missing values and confirmed that none of them contained missing entries.
3. **Converted Categorical Variable**: Since Fuel\_Type is a categorical variable, I converted it into dummy variables (Fuel\_Type\_Diesel and Fuel\_Type\_Petrol), with CNG as the reference category. This encoding is necessary to include Fuel\_Type in the regression analysis.
4. **Verified Data Types**: I checked the data types of each column to ensure they were appropriate for analysis.

These cleaning steps ensure the dataset is ready for linear regression analysis.

3.

* **Adjusted R-squared**: The adjusted R-squared value is **0.808**.
* **What it Means**: This value means that approximately 80.8% of the variance in the car's selling price (Price) can be explained by the model, which includes the variables Vehicle\_Age, KM, Mfg\_Year, and Fuel\_Type.
* **Model Performance**: An adjusted R-squared of 0.808 is relatively high, suggesting that the model fits the data well and that these variables are strong predictors of Price.

**Answer for the Question**

Based on the adjusted R-squared, the linear regression model performed well. With an adjusted R-squared of 0.808, the model explains around 80.8% of the variation in the selling price of Toyota Corollas, which is a strong indication that the chosen predictors have a significant impact on price.

**Question 4: Identify Statistically Significant Variables**

In regression analysis, statistical significance is determined by the **p-value** of each variable:

* A **p-value** less than 0.05 indicates that the variable has a statistically significant impact on the dependent variable (in this case, Price).

Here’s the p-value information from the summary:

1. **Vehicle\_Age**: p-value = 0.103 (not statistically significant)
2. **KM (Mileage)**: p-value = 0.021 (statistically significant)
3. **Mfg\_Year (Manufacturing Year)**: p-value = 0.000 (statistically significant)
4. **Fuel\_Type\_Diesel**: p-value = 0.019 (statistically significant)
5. **Fuel\_Type\_Petrol**: p-value = 0.614 (not statistically significant)

**Answer for Question 4**

The variables that have a statistically significant impact on Price are:

* KM (Mileage)
* Mfg\_Year (Manufacturing Year)
* Fuel\_Type\_Diesel

These variables have p-values less than 0.05, indicating they significantly influence the selling price of Toyota Corollas.

**Question 5: Interpret the Unstandardized Coefficients for Statistically Significant Variables**

Now, let’s interpret the unstandardized coefficients (also called beta values) for the significant variables in simple terms. Here’s how each significant variable affects Price:

1. **KM (Mileage)**: Coefficient = -0.0195
   * **Interpretation**: For each additional kilometer on the odometer, the selling price of the car decreases by approximately 0.0195 euros, holding other factors constant. This reflects the expected decrease in price as mileage increases.
2. **Mfg\_Year (Manufacturing Year)**: Coefficient = 1585.4440
   * **Interpretation**: For each additional year (i.e., a more recent manufacturing year), the selling price of the car increases by approximately 1585.44 euros, holding other factors constant. This shows that newer cars tend to be more expensive.
3. **Fuel\_Type\_Diesel**: Coefficient = 958.9086
   * **Interpretation**: If the car’s fuel type is Diesel, its price is expected to be about 958.91 euros higher compared to the reference category, CNG (which was dropped during dummy encoding). This suggests that Diesel cars are more expensive than CNG cars, all else being equal.

**Final Answer for Question 5**

To summarize for a non-technical executive:

* **Higher mileage** slightly reduces the selling price of a car.
* **Newer manufacturing years** increase the car's selling price.
* **Diesel fuel type** tends to increase the car's price compared to CNG.

These interpretations provide insights into how each significant variable affects the resale price of Toyota Corollas.

**Question B. Bank Customer Segmentation (15 points)**

|  |  |  |
| --- | --- | --- |
| Variable | Include/Exclude | Reason |
| CUST\_ID | Exclude | Identifier; does not contribute to clustering. |
| BALANCE | Include | Reflects the outstanding balance, which may indicate credit usage. |
| BALANCE\_FREQUENCY | Include | Shows how often the balance is updated, indirectly reflecting account activity. |
| PURCHASES | Include | Total purchase amount; relevant for spending behavior. |
| ONEOFF\_PURCHASES | Include | Reflects high-value purchases; useful for spending pattern analysis. |
| INSTALLMENTS\_PURCHASES | Include | Indicates use of installment purchases, relevant for payment behavior. |
| CASH\_ADVANCE | Include | Cash advances may indicate financial instability or need for liquidity. |
| PURCHASES\_FREQUENCY | Include | Shows frequency of purchases, relevant for spending behavior. |
| ONEOFF\_PURCHASES\_FREQUENCY | Include | Frequency of large one-time purchases; indicates purchasing behavior. |
| PURCHASES\_INSTALLMENTS\_FREQUENCY | Include | Frequency of installment purchases, relevant for payment behavior. |
| CASH\_ADVANCE\_FREQUENCY | Include | Frequency of cash advances, indicating reliance on credit. |
| CASH\_ADVANCE\_TRX | Include | Number of cash advance transactions, also indicates credit reliance. |
| PURCHASES\_TRX | Include | Number of purchase transactions, relevant for spending activity. |
| CREDIT\_LIMIT | Include | Credit limit may indicate customer creditworthiness. |
| PAYMENTS | Include | Total payments made, relevant for repayment behavior. |
| MINIMUM\_PAYMENTS | Include | Minimum payments show customer’s payment pattern and ability to pay. |
| PRC\_FULL\_PAYMENT | Include | Percentage of full payments, useful for assessing payment reliability. |
| TENURE | Include | Tenure may reflect loyalty and stability as a customer. |

**Answer to Question 2: Data Cleaning Steps**

To prepare the dataset for clustering, the following data cleaning steps were taken:

1. **Selected Relevant Columns**: Only the columns relevant to creditworthiness were selected, such as balance, purchases, payment behavior, and credit usage.
2. **Checked for Missing Values**: We identified that CREDIT\_LIMIT had 1 missing value and MINIMUM\_PAYMENTS had 313 missing values.
3. **Imputed Missing Values**: The missing values in CREDIT\_LIMIT and MINIMUM\_PAYMENTS were filled with the mean values of their respective columns. This imputation method was chosen to retain these variables without introducing significant bias, as they contain valuable information about creditworthiness.

These steps ensure that the dataset is complete and ready for clustering analysis.

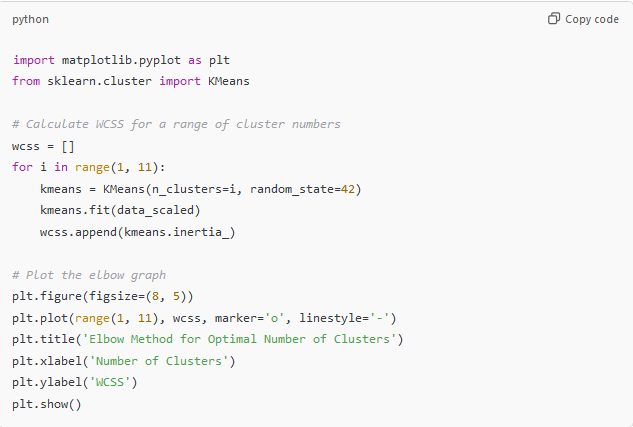
**Answer to Question 3: Determining the Optimal Number of Clusters**

To determine the optimal number of clusters, we used the **elbow method**, which evaluates the Within-Cluster Sum of Squares (WCSS) for different cluster counts. The elbow point in the plot occurs where the reduction in WCSS slows down, indicating that adding more clusters beyond this point does not significantly improve the model's fit.

**Steps and Analysis**:

1. We plotted the WCSS values for cluster counts from 1 to 10.
2. The elbow in the plot was observed at **3 clusters**, which indicates that 3 is an optimal choice for balancing simplicity and data segmentation quality.

**Conclusion**: Based on the elbow method, **3 clusters** were chosen as the best number for clustering. This choice is supported by the elbow point, which shows that adding more than 3 clusters yields diminishing returns in reducing WCSS.



**Answer for Question 4: K-Means Clustering Analysis**

Now we have all the information needed to complete Question 4. Here are the organized answers to each part.

**Part I: Percentage of Customers in Each Cluster**

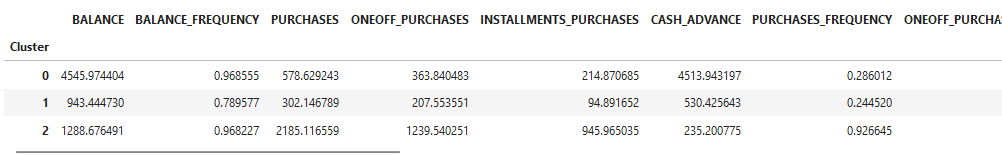
The distribution of customers across the clusters is as follows:

* **Cluster 1**: 50.94%
* **Cluster 2**: 35.20%
* **Cluster 0**: 13.87%

This distribution indicates that Cluster 1 contains the largest portion of customers, followed by Cluster 2, with Cluster 0 having the smallest share.

**Part II: Cluster Centroid Averages**

The table below shows the centroid (mean) values of each variable for each cluster, giving insight into the typical customer profile within each cluster.



Each cluster’s centroid values provide a summary of the typical behavior of customers within that group.

**Part III: Clustering Performance Evaluation**

1. **Mathematical Metric - Silhouette Score**:
   * The Silhouette Score for this clustering is **0.29**. This score, while not very high, suggests moderate cluster separation. A score closer to 1 would indicate more distinct and well-defined clusters, so a score of 0.29 implies some overlap among clusters.
2. **Business Meaningfulness**:
   * **Cluster 0** (13.87%): High-balance customers who frequently rely on cash advances. This could represent a financially constrained group that heavily utilizes credit.
   * **Cluster 1** (50.94%): Low-balance, low-activity customers. They engage minimally with their accounts, possibly indicating conservative spending behavior or limited credit usage.
   * **Cluster 2** (35.20%): Moderate-balance customers with high purchase activity and minimal cash advance reliance. This group likely represents active and responsible credit users.

Despite the moderate silhouette score, the clusters appear meaningful in a business context. They provide actionable insights into different customer types, enabling targeted strategies based on credit usage and spending behavior.

**Part IV: Customer Types and Most Creditworthy Cluster(s)**

Based on the centroid analysis, here’s a description of each cluster:

1. **Cluster 0**:
   * **Customer Type**: High-balance customers who rely on cash advances.
   * **Creditworthiness**: Likely moderate, as frequent cash advances may indicate financial strain or reliance on credit.
2. **Cluster 1**:
   * **Customer Type**: Low-balance, low-activity customers.
   * **Creditworthiness**: Likely low, as minimal account activity and low balances suggest limited credit engagement.
3. **Cluster 2**:
   * **Customer Type**: Moderate-balanced, high-purchase-frequency customers with low reliance on cash advances.
   * **Creditworthiness**: Likely high, as high purchase activity with minimal cash advance usage indicates responsible credit behavior.

**Most Creditworthy Cluster**: **Cluster 2** appears to be the most creditworthy segment due to their frequent purchases and low reliance on cash advances, suggesting stable and responsible credit usage