**MKT 568 -Assignment 3**

**Names: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Due: Oct 30, 11:59 PM**

**Scoring: 30 points total (15% of your grade)**

**This assignment aims to help you solidify your understanding of linear regression, segmentation and market basket analysis. With your team, discuss the following questions and answer them. In your submission, delete the question texts and *only* include your answers. Your submission must be written in an M.S. Word file with a Times New Roman, 12 Font. Any sources used in the assignment must be properly cited in the APA format (7th edition). Failing to cite your sources is considered plagiarism.**

**What you submit on the deadline includes (1) a Word document where you have responded to the questions listed in this document and (2) a Python script including the complete code used to respond to questions A and B. The Python code must be in the format of a Jupyter Notebook (.ipynb) and without errors. You will lose at least 2 points if your Python files throw any errors. So, make sure to run it once before submission and correct any errors.**

**For all questions related to QA and QB, 1/3 of the points will be dedicated to including the correct Python script and 2/3 of the points will be dedicated to providing the correct response in the submitted Word document.**

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

A car sales executive wants a method of predicting the selling price of “Toyota Corollas” when they are resold on their lot. The executive wants to understand why a particular Corolla is priced as it is and is asking for your help. You have the data on the selling price (Price) of 1436 used Corollas based on the cars’ features. The data is available on Canvas (ToyotaCorolla.xlsx). The Excel file includes the data and the data dictionary.

The sales executive asks you to tell them

* If the vehicle’s age, mileage, and manufacturing year affect sales price.
* Whether a car’s fuel type has any effect on the sales price.

Your goal is to perform ONE linear regression analysis to answer the following questions based on this data:

1. The first step in any analysis is to understand the business problem and the data. Fill in the table below. Determine whether each variable listed in the table is an independent variable or a dependent variable in the regression analysis. Then find features in the data set that correspond to each variable listed in the table and determine the type of this feature (categorical or continuous) as it is recorded in the data set. (3 points)

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |  |  |  |
| mileage |  |  |  |
| manufacturing year |  |  |  |
| sales price |  |  |  |
| fuel type |  |  |  |

1. After understanding the business and before running any analysis, you must clean the data. Write the appropriate Python code to clean the data. Include all scripts used for data cleaning in your submission (in the .ipynb file you will submit). Additionally, in the Word document submission, briefly list the steps you took to clean this dataset. (2 points)

*Note: In the assignment, you will only work with the variables in the above table, so you should only clean data captured by these variables.*

1. Run a linear regression model to predict Price based on vehicle’s age, mileage, manufacturing year, and fuel type. All necessary codes must be included in your Python file submission. After running the linear regression, interpret the performance of the model using the adjusted r-squared (what is the value of the adjusted r-squared, what does it mean and does it indicate that the linear regression performed well or not). (2 points)
2. Based on the results obtained from this linear regression, which of the independent variables had a statistically significant impact on Price? (4 points)

*Note:*

1. *Statistical significance is a statistical term, and it does NOT refer to the strength of the relationship between an input and output variable*.
2. *Interpreting the unstandardized beta for categorical inputs is different from continuous inputs. To include categorical inputs in linear regression, you must first dummy code them. For a categorical variable with N values, you will use N-1 dummy coded variables (I taught you to drop one of the dummies randomly). When you run linear regression, interpreting the effect of a dummy-coded categorical variable on the output depends on the value of the dummy-coded variable that was dropped from the analysis at the dummy coding stage. Use the internet to understand what this means (start by searching “The dummy variable trap”). Your answer to the sales executive’s question (does fuel type affect price and interpreting the unstandardized beta for this variable) depends on understanding why and which one of the dummy variables was dropped.*
3. For all independent variables that had a statistically significant influence on Price, interpret the unstandardized coefficient (i.e., how would you explain this to a non-technical executive)? (4 points)

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

In addition to assessing your understanding of clustering, the goal of this question is to get you familiar with Kaggle.com. Kaggle, a subsidiary of Google, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists, and enter competitions to solve data science challenges. You can find very useful resources on this website, and it is important for you to be familiar with this website. Thus, in this assignment, you will work with a dataset from Kaggle.com. Click on the link below to check out the web page of the data set. Since I want you to get familiar with the website, I won’t add the data file to Canvas. You must download the data from this website yourself.

<https://www.kaggle.com/arjunbhasin2013/ccdata>

Spend some time and explore this webpage to learn how the website works. Data scientists sometimes post their codes to this website or discuss the data set as well.

Most data posted on Kaggle are posted with a specific task (e.g., clustering, classification, prediction of a specific target). The person who posts the data also tells you what task can be performed on this data. The data for this assignment is collected from credit card customers. The goal is to run clustering on this data.

Descriptions of variables in the data are included on the Kaggle website (Under the “About Dataset” title). If you have any questions about this data, you can post your questions to the website (<https://www.kaggle.com/arjunbhasin2013/ccdata/discussion>), so the data owner will respond to your questions directly or contact me or the TAs.

Download this data and answer the following questions based on this data. If you cannot download it, let me know so I can help.

**Business Problem**

Bank managers have designed a new type of low-interest personal loan that customers can use for any purpose. A personal loan describes any situation in which an individual borrows money for personal needs. This product must be very popular due to its low interest rate, but the bank managers have decided to only offer this loan to highly credit-worthy customers. Being credit-worthy means that a customer can be trusted with a loan because they have a history of good financial behavior (e.g., paying back debts in time, not overusing their credit limits, making full credit card payments, being a customer for a long time, etc.). Because creditworthy customers exhibit responsible financial conduct, it is anticipated that they will repay this debt in full and avoid defaulting on the loan.

The management’s goal is to segment its customer base into a number of clusters, determine clusters that consist of customers with high levels of creditworthiness, and offer this loan to consumers in those clusters. The marketing manager asked you to perform this segmentation task.

1. The first step in clustering, after understanding the business problem, is to determine the segmentation basis—variables that will be used in the clustering analysis. This decision must be made based on the business problem and the project's goal. The goal of this project is to cluster customers based on **creditworthiness** (how much they can be trusted with a loan) and choose the cluster(s) with the highest creditworthiness levels. Examine all the variables in the Credit Card data set and select the variables that are the most appropriate for this analysis. For each variable in the data set, explain why you chose or did not choose it as the segmentation basis. In your response, create a table with three columns: (1) variable name, (2) whether you will include it in the analysis or not, and (3) your justification for this decision. (3.5 points).

*Note: Variables used in segmentation cannot be highly correlated. So, when selecting features, take care to avoid choosing variables that have strong correlations with each other.*

1. After selecting the segmentation basis, you must clean the data. Include the code required to clean the data in your Python file submission. In the Word file, briefly explain what steps you took to clean the data for the clustering analysis. (2 points)
2. After cleaning the data, you must determine the optimal number of clusters for the clustering analysis. What is the best number of clusters for this analysis, and what piece of information aided you in answering this question? Conduct the necessary analysis to identify the optimal number of clusters for this task. Include the relevant code in your submitted Python file and incorporate all appropriate charts in your response in the Word document you plan to submit. (3 points)
3. Use the variables you selected in part 1 to perform K-means and answer the following questions based on your analysis. The Python code you submit must include the code required to respond to all questions below.
   1. Describe the clusters in terms of the percentage of customers assigned to each cluster. In other words, what percentage of the customers were assigned to each cluster? (1.5 points)
   2. Report the cluster centroids averages (a table with average values of each input for each cluster, as I showed you in class). (1.5 points)
   3. Evaluate the performance of this clustering analysis. (1.5 points)

*Notes:*

*(1) This must be done using both the mathematical metrics and business meaningfulness of the clusters.*

*(2) Incorporate the code for calculating the mathematical metrics into your Python script and ensure that all results are included in the Word document to support your clustering evaluation.*

* 1. Based on your response to the previous parts, describe the type of customer that each cluster represents, then determine the most creditworthy cluster(s). (2 points)