Candidate Name: Christopher Mok

Please note the following:

1. Both questions use the same dataset
2. All source code must be provided with comments along with instructions on how to execute the code.
3. R is preferred, but you are free to consider other tools. Regardless of the tool you use, source code / workings must be provided and clearly documented
4. Two datasets are provided **spenddata.csv** and **testdata.csv.** You are free to decide how best to use them.
5. You do not need to augment this dataset with any external data. However, if you choose to do so, you must document the reasons.

**Question 1**

You are given a set of survey data which captures spend amount among other data points. The Marketing team is curious to understand if there are similar groups of people visiting Singapore and they have approached the data science team to help them find out.

Using the set of data provided:

* 1. Determine how many groups of tourists have visited Singapore

Using KPrototype (hybrid of K-means and K-modes), there are 5 groups of tourists identified.

Cluster 1 – Low c.40, c.47, c.65, ; High c.43. Generally average values

Cluster 2 – Low a.4, c.54, c.59, c.74, c.76; High c.79, c.81, c.86 pp.132, c.250

Cluster 3 – Low b.6, b.15, b.24, c.86, c.87, c.143, t7.149, ; High b.22, c.62, c.65, c.66, c.69, c.71, c.73, c.75, c.76

Cluster 4 – Low var3; High b.27. Generally average values

Cluster 5 – High c.54, c.57, c.74, c.82, c.137, c.139, t7.158, var2, totshopping.rep

* 1. Please explain the choice of metric used

For numerical data, Euclidean distance is used as the distance metric. This allows the algorithm to minimize the within-cluster sum of squared error such that points of nearby distances cluster together. For categorical data, Matching distance is used as the distance metric. A weighted average of the distance is the computed to cluster the data points.

* 1. What are the assumptions you made when building this solution?

K-prototype clustering is used. I assume that the variances of the distribution of each feature attribute is spherical and the sizes of cluster are even.

* 1. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.

I have considered hierarchical clustering as it does not require the assumptions to (c). However, the biggest drawback is that since hierarchical clustering takes O(n2) to run, and there are 18,379 observations in spenddata.csv, the hierarchical clustering method will not be scalable. Should the dataset (number of cross-sections) be smaller in the case of clustering companies, hierarchical clustering will be favoured.

* 1. Please explain under what conditions will the approach you choose **be not appropriate**

K-means clustering for numerical data will not work on non-spherical clusters as it aims to minimize the within-cluster sum of squared error (Euclidean distance). It also performs poorly when the clusters are not evenly sized.

* 1. Why is your approach performing well / not well?

It is not performing well as the model has too many dimensions and the features have missing values (even after dropping variables with more than 50% missing values). From the elbow plot, there is no clear elbow point for a good cluster selection.

* 1. Was any feature engineering required? If yes, what were they. If no, why?

Standard normalization is applied to numerical features to reduce biasedness arising from Euclidean distance. Missing data for numerical features are imputed with median values whereas missing data for categorical features are imputed with mode values.

**Question 2**

You are given a set of survey data which captures spend amount among other data points. However, some of the spend amount (**totshopping.rep**) is missing.

Build a model that will help impute the amount spent by a visitor based on the set of data provided.

1. Please explain the choice of metric used

I use Root Mean Squared Error (RMSE) as the metric to evaluate my model. Using the model with the lowest RMSE on the validation set, RMSE penalizes larger deviations and ensures that imputed predictions do not deviate too significantly from the truth.

1. What are the assumptions you made when building this model?

Since Ridge Regression is used, Gauss-Markov assumptions are assumed. Linear in parameters, homoskedasticity and no serial correlation (observations are independent of each other), the residuals have zero mean and are not correlated with the independent variables.

1. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.

Considered regularization models such as Ridge, Lasso, ElasticNet, Random Forest Regression. Random Forest Regression works well for non-linear relationships. However, using RMSE as a metric, found that ridge regression with alpha=1.0 works best in minimizing RMSE.

1. Please explain under what conditions will the model you choose **be not appropriate**

The model will not be appropriate if the observation has too many missing values in the variables as the missing values are imputed with median and mode of the independent variables. It will also not be appropriate if there the model lacks a linear relationship with the dependent variable.

1. How confident are you of the model’s robustness and how would you explain the model’s performance?

I am confident of the model’s robustness as on average, the predictions deviate from the actual value by 0.04 in the test set.

1. Why is your model performing well / not well?

The model performs well as L2 regularization technique is applied to ensure that the model is not overfit. The large dataset also helps to lower the variance of the model.

1. Was any feature engineering required? If yes, what were they. If no, why

Standard normalization is applied to numerical features to reduce biasedness and speed up computation. Missing data for numerical features are imputed with median values whereas missing data for categorical features are imputed with mode values. One hot encoding is applied for categorical features.