Overview

You will take over the role of *Market and Insights Analyst* at the consulting services department of a multinational professional services firm. As part of this role, you are asked to work across the following three (3) different client engagement projects.

Project 1

For this project you will analyse data from a survey in which 200 respondents were asked to rate the importance of a number of store attributes when choosing where to buy office equipment. The file `office.csv` contains data for the project. For each respondent, we have the following variables:

Variable	**Description**
`respondent_id`	An identifier for our observations
`variety_of_choice`	Importance of this attribute on a 0-10 scale
`electronics`	Importance of this attribute on a 0-10 scale
`furniture`	Importance of this attribute on a 0-10 scale
`quality_of_service`	Importance of this attribute on a 0-10 scale
`low_prices`	Importance of this attribute on a 0-10 scale
`return_policy`	Importance of this attribute on a 0-10 scale
`professional`	Whether the respondent is a professional or not (e.g., student)
`income`	Gross annual income expressed in thousands of pound sterling
`age`	Respondents' age in years

Task

1. Read and inspect the data set. Provide a descriptive analysis for each of the variables in the data set.

2. Make a new data object (e.g., a `data.frame or tibble`) for clustering that includes only the attitudinal variables from the original data set. Then normalise (use z-score standardisation) all variables in this new data object. Which variable has the smallest minimum value and which variable has the largest maximum value in the normalized data set?

3. Run the hierarchical clustering algorithm using `method = "ward.D2"` on the normalised data and use `set.seed(123)`for reproducibility. Plot the dendogram.

4. Suppose that after looking at the dendrogram and discussing with the marketing department, you decide to proceed with a 6-cluster solution. Divide the data points into 6 clusters. How many observations are assigned to each cluster?

5. Use the normalised data to calculate the means for each of the attitudinal variables per cluster. Use the `flexclust` package to generate a segment profile plot. Comment on whether any cluster memberships

have changed, if any. Check the concordance between the `hclust` and `as.kcca` procedures.

6 Describe the 6-cluster solution using the cluster numbers corresponding to the hierarchical clustering procedure.

7 Comment on why you may decide to NOT proceed with this 6-cluster solution.

8 Generate a 5-cluster solution. How many observations are assigned to each cluster?

9 Repeat the steps performed previously to describe the clusters for the 5-cluster solution (i.e., calculate cluster means and segmentation plot). Describe the 5-cluster solution using the cluster numbers corresponding to the hierarchical clustering procedure. Give “expressive” labels to the clusters.

10 Comment on why you may find this 5-cluster solution better than the previous 6-cluster solution.

11 Use all the variables not included in the clustering procedure to evaluate whether the 5-cluster solution is meaningful. Generate ideas on how to target each segment (at least one idea per segment).

12 Run the k-means clustering algorithm on the normalised data, creating 5 clusters. Use `iter.max = 1000` and `nstart = 100` and `set.seed(123)` for reproducibility. How many observations are assigned to each cluster?

13 Check the concordance between the `hclust` and `kmeans` procedures. What is the Hit Rate?

Project 2

For this project you will model website user conversion. You will be working on a dataset with more than 20 thousand unique users of a website based in four countries. The file `ecommerce.csv` contains data for the project. For each user, we have the following variables:

Variable	**Description**
`country`	The country the user accessed the site from (France, Germany, Ireland, or UK)
`source`	The source through which the user accessed the site (ads, search, or direct link)
`total_pages_visited`	The number of pages visited by the user
`visit_duration`	The amount of time the user spent in the site (in seconds)
`discount`	Whether the user was offered a discount (10% off first order; yes, no)
`conversion`	Whether the user converted, or made a purchase (yes, no)

Task

1 Read and inspect the data set. Provide a descriptive analysis for each of the variables in the data set.

2 Build a simple logistic regression model of `conversion` on `discount`. Call this model `m1`. Comment on the coefficient estimate of `discountyes`. What is the sign of the coefficient? Is the effect statistically significant?

3 Calculate the odds ratio for `discountyes`. What does this mean?

4 Calculate the 95% confidence interval for the odds ratio for `discountyes`. What does this mean?

5 Generate a double-decker mosaic plot (using the `ggmosaic` package) to visualise the count of the combinations of the following variables: `discount` (on x-axis), `conversion` (as fill colour), and `source` (as facets). Use the plot to describe whether (and how) the effect of `discount` on `conversion` is different for the three `source` channels.

6 Build a logistic regression model that predicts `conversion` from `discount` and `source`. Call this model `m2`. Comment on the coefficient estimates of `sourcedirect` and `sourcesearch`.

7 Calculate the odds ratios for `sourcedirect` and `sourcesearch`. What do these mean?

8 Build a logistic regression model that predicts `conversion` from `discount` and `source` and also includes their interaction. Call this model `m3`. Comment on the coefficient estimates of the interaction terms.

9 Calculate the 95% confidence intervals for the odds ratios for the interaction terms. What do these mean?

10 Build a logistic regression model that predicts `conversion` from all available variables in the data set. This model should also include an interaction of the `discount` and `source` variables. Call this model `m4`. Which variables are significant at the 95% level?

11 Calculate the correlation between the two numerical variables in the data set (`total_pages_visited` and `visit_duration`). Comment on the result. How may this affect `m4`?

12 Build another logistic regression model from `m4` by removing the `visit_duration` variable. Call this model `m5`. How has the effect of `total_pages_visited` changed compared to `m4`?

13 Make a plot that visualises the odds ratios (as points) of the variables in `m5` as well as their confidence intervals (as error bars).

14 Use model `m5` to predict the conversion probabilities for each user in the data set. Store these probabilities in the data set, in a variable called `base_prob`. What is the mean value of `base_prob`?

15 Calculate an indicator variable for whether individuals will convert or not, based on their predicted probabilities from the previous task, using a threshold value of 0.5. Call this variable `pred_conversion`. How many users so we predict to convert?

16 What is the accuracy or hit rate?

17 What is the area under the curve?

18 Predict new probabilities under a hypothetical scenario that the values variable `total_pages_visited` were increased by one unit (i.e., one page) for all users. Store these probabilities in the data set, in a variable called `new_prob`. What is the mean value of `new_prob`?

19 Calculate the lift metric for the hypothetical scenario from the previous task (i.e., Task 18).

Project 3

For this project you will run a Choice-Based Conjoint study in the Cloud Services Platform market (e.g. Amazon Web Services, Google Cloud, Microsoft Azure). The client wants to make some product design decisions such as core feature-sets, pricing, and tiers of service to optimise revenue or new sign-ups.

You will work with the `cloud.csv` file. The file contains data on choices made by 200 respondents. Each respondent evaluated 15 choice sets. Thus, the file contains data on 200 × 15 = 3000 choice sets. Each choice set had three alternatives. A respondent's task was to choose one alternative from a choice set. The following table describes the variables in the dataset:

Variable	Description
`respondent_id`	Identifier for each respondent (1 to 200)
`choiseset_id`	Identifier for each choice set for each respondent (1 to 15)
`alternative_id`	Identifier for each alternative in a choice set (1 to 3)
`choice_id`	Identifier for each choice set in the entire study (1 to 3000)
`cloud_storage`	Attribute cloud storage with three levels: 30GB / 2000GB / 5000GB
`customer_support`	Attribute customer support with two levels: Yes / No
`cloud_services`	Attribute cloud services with three levels: Email / Email + Video / Email + Video + Productivity
`price`	Attribute price with three levels: £6 per month / £12 per month / £18 per month
`choice`	Shows which alternative was chosen in each choice set (Dummy coded: 1 if alternative was chosen; 0 otherwise)

Task

1. Read and inspect the data set. Provide a descriptive analysis for each of the variables in the data set. Make sure you provide an analysis that is meaningful for each variable type (e.g., factors, identifiers).

2. Convert the attribute variables `cloud_storage` and `price` so that the factor reference levels are the levels representing the smallest values (i.e., 30GB for `cloud_storage` and p6 for `price`). Why there is no need to perform this step on the rest of the attribute variables?

3. Create a new variable in the data set that turns `price` into numeric class (do not overwrite `price`). Call this new variable `price_n`. What is the mean of variable `price_n`?

4 There are 3000 choice sets in the data set. Therefore, there were 3000 choices made. Out of these 3000 choices, how many times did respondents choose a 30GB cloud storage? What is the percentage of respondents who chose email only as cloud service?

5 Use the `dfidx()` function from the `dfidx` package to create a specially formatted data object that will be used in the process of estimating a multinomial conjoint model. In the argument `idx`, use a `list` of the two indexes (`choice_id` and `respondent_id`) that define unique observations. Also use `alternative_id` as the variable defining the levels of the alternatives. Call this data object `m_data`. How many variables (i.e., columns) does `m_data` have?

6 Use `m_data` to build a multinomial logit model that predicts `choice` from `cloud_storage`, `customer_support`, `cloud_services`, and `price`. Make sure that you tell the `mlogit()` function to exclude the intercept term. Call this model `model1`. Use `set.seed(123)` right before running the command that builds the model. Comment on the coefficient estimates of `cloud_storage5000gb` and `pricep12`.

7 Now follow the same process as in Task 6 to build a multinomial logit model that uses `price_n` instead of `price`. Call this model `model2`. Again use `set.seed(123)` right before running the command that builds the model. Comment on the coefficient estimate of `price_n`. What does this mean?

8 Use a likelihood ratio test to test the `model2` against `model1`. What is the outcome of the test? Are `model2` and `model1` significantly different? Which model we should choose between the two and for what reason(s)?

9 Use `model2` to predict the choice probabilities for different alternatives in the data. What is the predicted probability of choosing the third alternative in the first choice set?

10 Use the predicted probabilities from Task 9 to compute the predicted alternatives using the maximum choice probabilities. Which is the predicted alternative in the third choice set?

11 Then we can extract the selected alternatives from the original data. Which is the selected alternative in the fifteenth choice set?

12 Compute the confusion matrix for `model2`. What is the accuracy (or hit rate) of `model2`? How does `model2` compare to the baseline method (i.e., making random predictions)?

13 Now let us see how we can use the `model2` parameters to predict market shares under hypothetical market scenarios for an arbitrary set of products. First, build a custom function to predict market share for an arbitrary set of alternatives available in a data set `d`. You can find the commands for building the custom function in the “Multinomial Choice Modelling Practical”. Call the custom function `predict.share`.

14 Create a data object (i.e., `data.frame` or `tibble`) with the following hypothetical market consisting of five alternatives:

`cloud_storage`	`customer_support`	`cloud_services`	`price_n`
30gb	no	email	6
30gb	no	email, video	12
30gb	yes	email	12
5000gb	yes	email	18
5000gb	no	email, video, productivity	18

Call this data object `d_base`.

15 Run the customer function `predict.share` using `model2` and `d_base` as input arguments. What is the predicted market share for alternative four of this hypothetical market?

16 Now consider a modification on the previous hypothetical market, in which the level of the `cloud_services` attribute changes for the fifth alternative to “email, video”. What is the predicted market share for alternative four of this new hypothetical market?

17 Which alternative was affected the most from this modification of the hypothetical market, and by how much (in percentage terms)?

18 Use the `model2` coefficients to calculate how much a consumer would be willing to pay (in £ per month) for customer support.

19 Use the `model2` coefficients to calculate how much a consumer would be willing to pay (in £ per month) for an upgrade from 30GB to 2000GB cloud storage.

20 Use the `model2` coefficients to calculate how much a consumer would be willing to pay (in £ per month) for an upgrade from 2000GB to 5000GB cloud storage.

END OF DOCUMENT