Data Analysis Skills – Group Project 2 Datasets

Datasets

There are 28 datasets from which your group has been allocated one to work on this assignment. The dataset number for your group is the same number as your group (e.g. group 10 will use dataset 10). Please ensure you are working with the correct dataset and are answering your assigned question of interest. The datasets are available for you to download from Moodle in "Week 9 – Group Project 2"

Datasets 1- 5

Datasets 1-5 come from the FIES (Family Income and Expenditure Survey) recorded in the Philippines. The survey, which is undertaken every three years, is aimed at providing data on family income and expenditure. Each of the 5 datasets obtained from this survey are from different regions across the Philippines. You will have access to the following variables, recorded by household (Note: head of the household refers to the person "in charge" of the house):

- `Total.Household.Income` – Annual household income (in Philippine peso)
- `Region` – The region of the Philippines which you have data for
- `Total.Food.Expenditure` – Annual expenditure by the household on food (in Philippine peso)
- `Household.Head.Sex` – Head of the households sex
- `Household.Head.Age` – Head of the households age (in years)
- `Type.of.Household` – Relationship between the group of people living in the house
- `Total.Number.of.Family.members` – Number of people living in the house
- `House.Floor.Area` – Floor area of the house (in m^2)
- `House.Age` – Age of the building (in years)
- `Number.of.bedrooms` – Number of bedrooms in the house
- `Electricity` – Does the house have electricity? (1=Yes, 0=No)

Task

Imagine you have been asked by the government to investigate the following question of interest:

- Which household related variables influence the number of people living in a household?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

Datasets 6 – 10

Datasets 6-10 come from the IMDB film database. The database contains a variety of information on all films that have been released. You will work with a subset of this database and will have access to the following variables, recorded by film:

- `film.id` – The unique identifier for the film
- `year` – Year of release of the film in cinemas
- `length` – Duration (in minutes)
- `budget` – Budget for the films production (in $1000000s)
- `votes` – Number of positive votes received by viewers
- `genre` – Genre of the film
- `rating` – IMDB rating from 0-10

Task

Imagine you have been asked by a film producer to investigate the following question of interest:

- Which properties of films influence whether they are rated by IMDB as greater than 7 or not?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

Datasets 10 – 15

Datasets 10-15 come from the Coffee Quality Database (CQD). The database contains information from the Coffee Quality Institute which is a non-profit organisation working internationally to improve the quality of coffee and the lives of the people who produce it. Each of the 5 datasets contain information on features of coffee and its production, including an overall score of quality. You will have access to the following variables, recorded by batch.

- `country_of_origin` – Country where the coffee bean originates from.
- `aroma` – Aroma grade (ranging from 0-10)
- `flavor` – Flavour grade (ranging from 0-10)
- `acidity` – Acidity grade (ranging from 0-10)
- `category_two_defects` – Count of category 2 type defects in the batch of coffee beans tested.
- `altitiude_mean_meters` – Mean altitude of the growers farm (in metres)
- `harvested` – Year the batch was harvested
- `Qualityclass` – Quality score for the batch (Good - ≥ 82.5, Poor - <82.5). Note: 82.5 was selected as the cut off as this is the median score for all the batches tested.

Task

Imagine you have been asked by a local coffee farmer to investigate the following question of interest:

- What influence do different features of coffee have on whether the quality of a batch of coffee is classified as good or poor?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

Datasets 16-20

Datasets 16-20 come from the Dallas animal shelter. Each of the 5 datasets contain a variety of information relating to each animal admitted to the shelter. You will have access to the following variables, recorded by animal admission:

- Animal_type – The type of animal admitted to the shelter
- Month – Month the animal was admitted, recorded numerically with January=1
- Year. – Year the animal was admitted to the shelter.
- Intake_type – Reason for the animal being admitted to the shelter
- Outcome_type – Final outcome for the admitted animal
- Chip_Status – Did the animal have a microchip with owner information?
- Time_at_Shelter – Days spent at the shelter between being admitted and the final outcome.

Task

Imagine you have been asked by the shelter management to investigate the following questions of interest:

- Which factors influence the number of days an animal spends in the shelter before their final outcome is decided?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

Datasets 21-25

Datasets 21-25 come from IKEA Saudi Arabia. Each of the 5 datasets contain information on items of furniture available for purchase. You will have access to the following variables, recorded by furniture item:

- `item_id` – Unique item ID for item of furniture
- `category` – The furniture category the item belongs to
- `price` – The current price in Saudi Riyals (as recorded on 20/04/2020)
- `sellable_online` – Is the item available to purchase online?
- `other_colors` – Is the item available in other colours
- `depth` – Depth of the item in cm
- `height` – Height of the item in cm
- `width` – width of the item in cm

Task

Imagine you have been asked by the company to investigate the following questions of interest:

- Which properties of furniture influence whether they cost more than 1000 Saudi Riyals?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

Datasets 26-28

Datasets 26-28 comes from the WineEnthusiast. Each of the 3 datasets contains ratings on a variety of wines on a score from 1-100 (though the only wines included in this data are rated 80 or higher). You will have access to the following variables, recorded by title of the review:

- `country` – Country of origin
- `points` – The number of points awarded for the wine on a scale of 1-100 (although reviews are only posted for ratings 80 and above)
- `price` – The cost for a bottle of wine
- `province` – The province or state the wine is from
- `title` – The title of the wine review
- `variety` – The type of grape
- `winery` – The winery that made the wine

Task

Imagine you have been asked by the wine reviewer to investigate the following questions of interest:

- Which properties of wine influence whether the number of points awarded is greater than 90?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.