QUESTION 1

10 points Save Answer

a. (2 marks) Use your Project 1 as an example to explain that *Exploratory Data Analysis* is an iterative process.

b. (2 marks) Give two types of plots that are suitable for illustrating the covariation between a continuous variable and a categorical variable. Explain your answers.

c. (2 marks) Assuming that you have a data frame `df` which has a continuous variable `var1` and a categorical variable `var2`, write R code using the *ggplot2* library to illustrate your answers for the two types of plots in Part *b* above.

d. (2 marks) Briefly explain one similarity and one difference between `geom_hex` and a `geom_point` from the *ggplot2* library.

e. (2 marks) Explain when it would be suitable to use `geom_boxplot` for a variable `var` of a data frame. Briefly describe what the outputs are from `boxplot.stats(var)$stats`.

100%

QUESTION 2

10 points Save Answer

a. (4 marks) Given below is a data frame `df` showing the measurements of *temperature* (in degree C), *wind speed* (in km/h), and *wind direction* of various locations in the Perth area on a specific day.

Location	Temperature	Wind.Speed	Wind.Dir
Bickley	19.1	19	SW
Garden Island	18.8	41	SSW
Mandurah	19.7	22	SW
Perth	19.8	19	SSW
Perth Airport	21.1	35	SSW
Rottnest Island	18.3	43	SSW
Swan Valley	21.2	19	SSW
Swanbourne	18.7	26	SSW

i. (2 marks) Briefly explain a plot that is suitable for visualising variable `Temperature` versus variable `Wind.Speed`. Write R code using the *ggplot2* library to illustrate your answer.

ii. (2 marks) Briefly explain a plot that is suitable for visualising variable `Temperature` versus variable `Wind.Dir`. Write R code using the *ggplot2* library to illustrate your answer.

b. (3 marks) Referring to the data frame `df` in part *a*, suppose that we want to convert the `Wind.Speed` column to the following levels to form a new categorical column called `Wind`:

- `Gentle` if the wind speed is less than 20 km/h;
- `Medium` if the wind speed is between 20 km/h (inclusive) and 40 km/h (non-inclusive);
- `Strong` if the wind speed is at least 40 km/h.

Write R code to add this `Wind` column to the data frame.

c. (3 marks) Explain what *z-normalisation* is. Is it suitable for detecting outliers? Explain your answer.

QUESTION 3

10 points | Save Answer

Given below are the first few rows of two data frames, `ins.df` and `person.df`:

`ins.df`

Name	Type	Premium	Discount
AAA Ins	contents	700	5
AAA Ins	building	1200	2
Epic Insurance	vehicle	500	3
Epic Insurance	building	2000	3
Epic Insurance	contents	1100	3
QBA	vehicle	450	2
RAB Ins	contents	680	1
RAB Ins	building	1150	2

`person.df`

Person	Ins.name	Type	Years
Jack	QBA	vehicle	5
Jack	AAA Ins	contents	10
Jill	RAB Ins	contents	2
Jill	RAB Ins	building	6

The data frame `ins.df` contains the following columns: `Name`, which stores the names of various insurance companies; `Type`, which stores the types of insurance covered by the insurance companies; `Premium`, which stores the annual premiums (in dollars) that they charge; and `Discount`, which stores the discount percentage on the premiums if their customers have been insured with them for more than 5 years. For example, if a customer has been insured with `AAA Ins` for her/his household contents for 6 years, then s/he only needs to pay 95% × 700 = 665 dollars.

The data frame person.df has the following columns: `Person`, which contains the names of various persons; `Ins.name` and `Type` contain the insurance company and type of insurance for each person; and `Years` stores the number of years the person has been with the insurance company.

a. (3 marks) Write R code to get the total number of insurance companies that offer insurance on both `building` and `contents`.

b. (2 marks) Write R code for the ***left join*** *and* ***semi join*** operations on `ins.df` and `person.df` using
   i. the `Type` columns from the two data frames;
   ii. `Name` and `Ins.name` as the two matching columns from the two data frames.

c. (5 marks) Write R code to appropriately combine the two data frames to output the total annual insurance premium that `Jack` has to pay. Note: Your code must include the use of the *pipe* operator.

QUESTION 4

10 points Save Answer

Given below is a small data frame `car.df` recording 10 observations. Each observation contains the `Colour` (`"red"` or `"yellow"`), `Origin` (`"domestic"` or `"imported"`), `Price` (in 10^3 dollars), and `Type` (`"sedan"` or `"SUV"`) of a vehicle.

```
library(tibble)
car.df <- tribble(
    ~Colour, ~Origin, ~Price, ~Type,
    "red", "domestic", 38, "sedan",
    "red", "domestic", 40, "sedan",
    "red", "domestic", 48, "SUV",
    "red", "domestic", 45, "sedan",
    "red", "imported", 78, "SUV",
    "red", "imported", 52, "sedan",
    "yellow", "domestic", 50, "SUV",
    "yellow", "domestic", 51, "SUV",
    "yellow", "imported", 53, "sedan",
    "yellow", "imported", 75, "SUV"
)
```

a. (1 mark) Briefly define what a typical Null model would be for this data frame, where the response variable that we want to predict is the `Type` column. Write R code to output the predicted probability from your Null model.

b. (5 marks) Using only the `Colour` and `Origin` columns as the input feature variables and the `Type` column as the output response variable:

   i. (3 marks) Briefly describe how the *Naïve Bayes* algorithm predicts the `Type` variable using these two input feature variables. Write R code (including the library/libraries) to show how you would train this *Naïve Bayes* classifier.

   ii. (2 marks) Construct a small test set `testdf` containing the following two instances:
   - the first test instance has `Colour` being `"red"` and `Origin` being `"imported"`
   - the second test instance has `Colour` being `"yellow"` but `Origin` is unknown.

   Write R code to show how you would use your trained *Naïve Bayes* classifier from Part *i* above to predict the types of vehicle for the two instances in `testdf`.

c. (4 marks) Using only the `Origin` and `Price` columns as the input variables together with the `Type` column as the output response variable:

   i. (2 marks) Write R code (including the needed library/libraries) to show how to train a *Decision Tree* classifier and how to display the binary tree. (No need to include the diagram)

   ii. (2 marks) Based on the output binary tree from your code, briefly explain how the *Decision Tree* algorithm works and how it would predict the `Type` of a domestic vehicle that costs $55,000.

QUESTION 5

10 points Save Answer

a. (4 marks) Explain how the *k-Nearest Neighbours* classification algorithm works for the `car.df` data frame shown in the Question 4. For each of the two input variables `Colour` and `Price`, what will be a suitable distance function? Explain your answers.

b. (2 marks) Briefly explain two differences between the the *k-Nearest Neighbours* algorithm and the *Decision Tree* algorithm for classification problems. DT need training but knn dont need it. DT is making judgement based on data features but KNN is making judgement based on close data points

c. (2 marks) Briefly explain how you would use the *Receiver Operating Characteristic* (ROC) curves to compare the performance of two binary classifiers.

d. (2 marks) Briefly explain how the *hierarchical clustering* algorithm works.

QUESTION 6

10 points Save Answer

Given below are the first 8 observations of a fictitious data frame `df` which has 3 input variables `X`, `Y`, and `Z` and an output variable `Class`. The `Class` column is a categorical variable having 4 levels: `apple`, `mandarin`, `orange`, and `pear`.

X	Y	Z	Class
-0.6210	-0.0303	-2.5022	apple
0.0396	0.3315	-4.9666	mandarin
3.6174	2.8361	-2.2986	mandarin
0.6410	1.5397	-3.4728	orange
0.7586	1.6012	-4.0678	apple
3.9301	1.1660	-3.2180	pear
1.4218	0.1662	-4.0260	apple
-2.0301	3.6804	-3.7289	apple

a. (1 mark) Write R code to replace the `Class` column to an integer column as follows:
   - group `apple` and `pear` into one level and give them the value 1
   - group `mandarin` and `orange` into another level of value 0

b. (4 marks) Write an R function to perform *z-normalisation* on the input variables.
   i. (3 marks) Your function should be named `normalise` and it should take in two arguments: a data frame and an integer, which denotes the column index of the data frame where the normalisation will be applied.

   ii. (1 mark) Call your `normalise` function three times using a *for* loop to normalise the `X`, `Y`, and `Z` columns of `df`. The normalised columns should be saved back to the data frame.

c. (5 marks) Write R code to do the following:
   i. (1 mark) Split `df` into a training set and a test set using the 90/10 split ratio.

   ii. (1 mark) Train a *Logistic Regression* classifier on the data frame `df` using all the three input variables to predict the `Class` variable.

   iii. (2 marks) Perform predictions on the training set and the test set using the trained model from Part *ii* above.

   iv. (1 mark) Compute the accuracies of the predictions of the model on the training set and the test set.