ETC1010-5510: Introduction to Data Analysis

Assignment 2

Instructions to Students

This is an individual assignment and you must work on it on your own. Collaboration on the assignment constitute collusion. For more on collusion and misconduct please see this webpage. (https://connect.monash.edu/s/article/FAQ-2144)

This assignment gets you to use R to complete tasks and for you to explain your results in sentences.

Please make sure you read the hints throughout the assignment to help guide you on the tasks.

The points allocated for each of the elements in the assignment are marked in the questions and in certain cases.

Marking + Grades

This assignment is out of 65 and will be worth **15%** of your total grade. **Due on: Tuesday, 15 October 2024, 4:30 PM (Melbourne time)**.

There are two sections: Part A and Part B. Part A is out of 26. Part B is out of 39.

For this assignment, you will need to upload the following into Moodle:

- The rendered html file **saved as a pdf**. The assignment will be only marked if the pdf is uploaded in Moodle. **The submitted assignment pdf file must have all the code and output visible.**
- **At a minimum, your assignment should be able to be "knitted"** using the `Knit` button for your Rmarkdown document so that you can produce a html file that you will save as pdf file and upload it into Moodle.

If you want to look at what the assignment looks like as you progress , remember that you can set the R chunk options to `eval = FALSE` like so to ensure that you can knit the file:

If you use `eval = FALSE` or `echo = FALSE` , please remember to ensure that you have set to `eval = TRUE` and `echo = TRUE` when you submit the assignment, to ensure all your R codes run.

IMPORTANT: You must use R code to answer all the questions in the report.

Due Date

This assignment is due on Tuesday, 15 October 2024, 4:30 PM. You will submit the knitted html file **saved as a pdf** via Moodle. Please make sure you add your name on the YAML part of the Rmd file before you knit it and save it as pdf.

The answers to some questions are subjective, so in a few sentences explain what you think. **DO NOT** use dot points.

There is less scaffolding in this assignment than in Assignment 1. You should write whatever code you think helps answer the question.

You need to keep your code neat - you will be graded on the presentation quality of your code.

Libraries

You will need to use:

- RColorBrewer
- ggdendro

Part A: A little bit of visualisation

In Part A we will look at some visualisations using the Harry Potter data from the lectures.

The dataset `hp_chars.csv` contains an excerpt list of student characters from the novels. The variables in the dataset are as follows.

- `name` - name of a student character.
- `schoolyear` - the year in which the student began at Hogwarts.
- `house` - the house in which the sorting hat placed said student.

The dataset `hp_edges.csv` contains data on which characters in the first dataset (recorded under `name` in `hp_chars.csv`) conversed with other characters in that list, and in which books. In both datasets, the list is incomplete. Not all character conversations are recorded, and not all students are represented. The variables in this dataset are as follows.

- `name1` - name of a student wizarding character.
- `name2` - name of a student wizarding character whom the character in `name1` spoke to.
- `book` - In which book the conversation took place.

Read in the data

Part A.1

Question 1

Using the `hp_chars` data, present a table that shows which houses have the highest proportion of male students. **(3pts)**

Question 2

Represent this information visually using barcharts. Load the `RColorBrewer` package and use the `Set1` colour scheme. *Hint: use the `position="fill"` option.* **(2pts)**

Question 3

In one or two sentences, describe what this graph tells us. **(2 pts)**

Question 4

Using the `Dark2` color scheme, visually represent the counts by gender using the `dodge` option. **(2pts)**.

Question 5

In one or two sentences, describe what this graph tells us. **(2 pts)**

Question 6

Which graph do you think best displays the gender composition of the houses? Give your reasoning in two or three sentences. **(3 pts)**

Part A.2

The dataset `hp_edges.csv` records times in the books where the character in `name1` spoke to characters in column `name2`, along with which book this conversation occurred.

Question 1

Combine the two datasets into a single dataframe named `merge_hp`. Make sure that this dataframe contains all rows from the dataframe `hp_edges` and any row that matches the key in the dataframe `hp_chars`. **(4 pts)**

Question 2

Display the first 4 rows of the `merge_hp` dataframe. **(1 pt)**

Question 3

Using the merge_hp dataset, create a tibble that shows which characters speak with the highest number of unique wizarding characters throughout each of the books. **(5 pts)**

Question 4

Show the first 4 rows of this tibble and use this to identify which character speaks with the highest number of unique wizarding characters in books 2 and 3? **(2 pts)**

Part B Cluster analysis

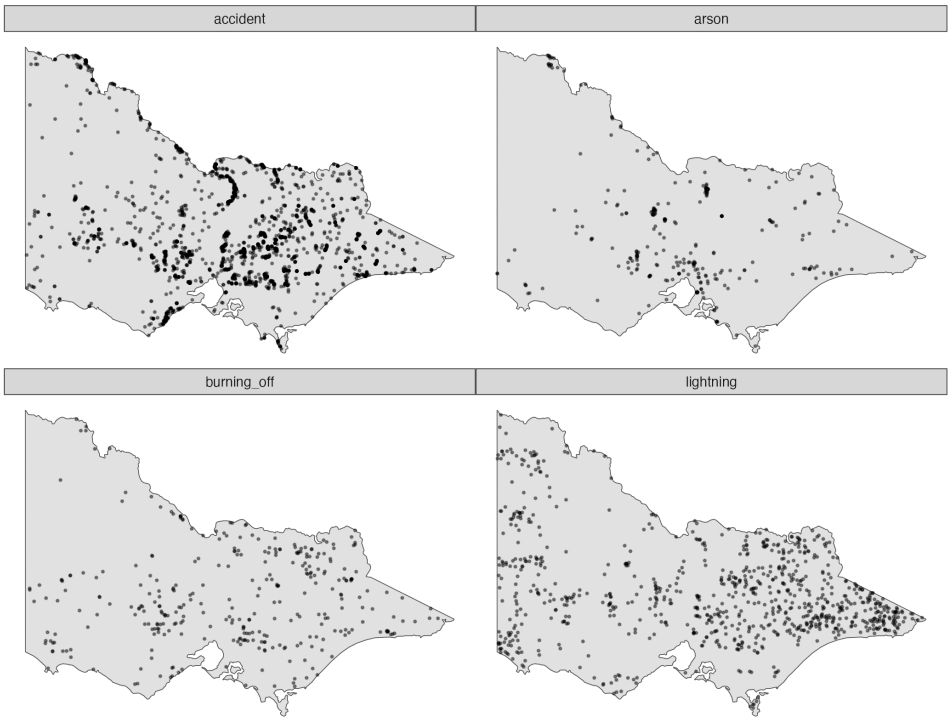
In this section, we will use a Victoria bushfire dataset compiled by Patrick Li in 2020 from various open sources on the web. The dataset includes 4049 historical bushfire events in Victoria from 2015 to 2018, recorded by the Country Fire Authority (CFA). It contains 53 variables such as date, forest coverage, rainfall, solar exposure, temperature, wind speed, and distances to the nearest CFA station, recreation site, and road. The dataset can be accessed at this URL (https://raw.githubusercontent.com/numbats/ida2024s2/master/data/vic_bushfire.csv).

Here is the data dictionary for the dataset:

Covariate name	description	Units
lon	Longitude	degrees
lat	Latitude	degrees
FOR_TYPE	Forest type. Eg. Acacia, Callitris, Casuarina, etc.	
rf	Rainfall on that day	mm
rf7	Average rainfall in the past 7 days	mm
arf14	Average rainfall in the past 14 days	mm
arf28	Average rainfall in the past 28 days	mm
arf60	Average rainfall in the past 60 days	mm
arf90	Average rainfall in the past 90 days	mm
arf180	Average rainfall in the past 180 days	mm
arf360	Average rainfall in the past 360 days	mm
arf720	Average rainfall in the past 720 days	mm
se	Global solar exposure on that day	MJ/m^2
ase7	Average global solar exposure in past 7 days	MJ/m^2
ase14	Average global solar exposure in past 14 days	MJ/m^2
ase28	Average global solar exposure in past 28 days	MJ/m^2
ase60	Average global solar exposure in past 60 days	MJ/m^2
ase90	Average global solar exposure in past 90 days	MJ/m^2
ase180	Average global solar exposure in past 180 days	MJ/m^2
ase360	Average global solar exposure in past 360 days	MJ/m^2
ase720	Average global solar exposure in past 720 days	MJ/m^2
maxt	Maximum temperature on that day	Celsius degree
amaxt7	Average maximum temperature in the past 7 days	Celsius degree
amaxt14	Average maximum temperature in the past 14 days	Celsius degree
amaxt28	Average maximum temperature in the past 28 days	Celsius degree
amaxt60	Average maximum temperature in the past 60 days	Celsius degree

Covariate name	description	Units
amaxt90	Average maximum temperature in the past 90 days	Celsius degree
amaxt180	Average maximum temperature in the past 180 days	Celsius degree
amaxt360	Average maximum temperature in the past 360 days	Celsius degree
amaxt720	Average maximum temperature in the past 720 days	Celsius degree
mint	Minimum temperature on that day	Celsius degree
amint7	Average minimum temperature in the past 7 days	Celsius degree
amint14	Average minimum temperature in the past 14 days	Celsius degree
amint28	Average minimum temperature in the past 28 days	Celsius degree
amint60	Average minimum temperature in the past 60 days	Celsius degree
amint90	Average minimum temperature in the past 90 days	Celsius degree
amint180	Average minimum temperature in the past 180 days	Celsius degree
amint360	Average minimum temperature in the past 360 days	Celsius degree
amint720	Average minimum temperature in the past 720 days	Celsius degree
ws	Average wind speed on that day	m/s
aws_m0	Average wind speed on that month	m/s
aws_m1	Average wind speed in last month	m/s
aws_m3	Average wind speed in last 3 months	m/s
aws_m6	Average wind speed in last 6 months	m/s
aws_m12	Average wind speed in last 12 months	m/s
aws_m24	Average wind speed in last 24 months	m/s
dist_cfa	Distance to the nearest CFA station	m
dist_camp	Distance to the nearest recreation site	m
dist_road	Distance to the nearest road	m

The following overview presents all bushfires mapped across Victoria, categorized by their cause. There are four types of causes: accident, arson, burning off, and lightning. Accidents and arson are classified as human-ignited fires, burning off refers to planned fires set by CFA, and lightning represents fires ignited by lightning strikes.



Question B.1

a. Read the CSV file “vic_bushfire.csv” using URL from GitHub. **(1 pt)**

b. Display the first 5 rows of the `vic_bushfire` tibble. **(1 pt)**

We want to identify the different characteristics of these bushfires using cluster analysis.

c. Considering that we will be using Euclidean distance, should any variables be dropped from the data before conducting the cluster analysis? Should the data be standardized? Briefly explain your reasoning. **(3 pts)**

Question B.2

We will conduct hierarchical cluster analysis on this data.

a. In one sentence, explain what the Agglomerative clustering method does. **(1 pt)**

b.

- Conduct a cluster analysis using Euclidean distance and the `ward.D2` method. **(2 pts)**
- Produce a dendrogram *(Hint: disable the labels by setting `labels = FALSE`)* **(2 pts)**
- Identify the number of clusters by drawing a horizontal dashed line (`geom_hline`) at the point where you intend to cut the hierarchical tree. **(1 pt)**

HINT: use `select_if` and `is.numeric` to ensure that only numeric data is used.

c. Using the dendrogram, explain the rationale behind your choice of the number of clusters in a short paragraph. **(2 pts)**

d.

- Cut the tree into 6 clusters, and draw a `lat` vs `lon` plot faceted by `cause` and cluster label. **(2 pts)**
- Describe any notable patterns or observations you see within the clusters, focusing on at least three key points. **(3 pts)**

Some code is provided below to help you.

```
vic_bushfire_6_hclust <- ___

plot_vic_map() +
  geom_point(data = vic_bushfire_6_hclust,
             aes(___), alpha = 0.4, size = 0.3) +
  ggthemes::theme_map() +
  facet_grid(___ ~ ___) +
  theme(legend.position = "none")
```

Question B.3

We will now conduct non-hierarchical cluster analysis on this data.

a. Apply K-means clustering using the number of clusters you determined in B.2.b. **(2 pts)**

b. Why can't we obtain a dendrogram for this method? **(1 pt)**

c.

- Create a line plot with variable names on the x-axis and their corresponding means on the y-axis. Use different colors for lines to represent each cluster label. **(2 pts)**
- Compare the means and explain your findings. **(2 pts)**

Some code is provided below to help you.

```
vic_bushfire_kmeans$___ %>%
  as_tibble() %>%
  mutate(cluster = ___) %>%
  pivot____(___) %>%
  ggplot() +
  geom____(aes(___, ___, group = ___, col = ___)) +
  theme_light() +
  scale_color_brewer(palette = "Dark2") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  xlab(___) +
  ylab(___)
```

d.

- Compare the cluster labels from hierarchical clustering and K-means clustering using a two-way table (`table`). You might need to rename some cluster labels to facilitate a clearer comparison. **(2 pt)**
- Describe your observations. **(2 pt)**

e. Name one advantage and one disadvantage of k-means clustering **(2 pts)**

f.

- Using the seed `2024` , and re-run the k-means clustering with 6 clusters. Draw a `lat` vs `lon` plot faceted by `cause` and cluster label. **(2 pts)**
- Compare this plot to the one in B.2.d. Does k-means clustering more effectively separate groups of bushfires in terms of spatial distribution? **(2 pts)**

Some code is provided below to help you.

```
set.seed(2024)
vic_bushfire_6_kmeans <- ___

plot_vic_map() +
  geom_point(data = vic_bushfire_6_kmeans,
             aes(___), alpha = 0.4, size = 0.3) +
  ggthemes::theme_map() +
  facet_grid(___) +
  theme(legend.position = "none")
```

g. Do the clusters you determined in B.2.b and those from B.3.a align with the known groupings based on the causes of the fires? Use relevant plots or tables to support your analysis. **(2 pts)**

h. Name at least two cluster statistics that can be used to help determine the optimal number of clusters for k-means or hierarchical clustering, and include an interpretation of each statistic. Then, calculate one cluster statistic for the k-means solution in B.3.a. **(2 pts)**