7CS039 Assessment

Statistics for AI & Data Science

Assessment Overview

Your assessment takes the form of an exploratory data analysis using `R`. You must include screenshots of your `R` code and you must include charts and graphics produced in `R` as appropriate. Your assessment should be at least five pages in length, including images, but it should not be more than ten pages. The assessment should be typed and Harvard style referencing should be used where appropriate.

The Data

For this assessment you should use the "Survival from Malignant Melanoma" data set which is available on canvas. The data consists of measurements made on patients with malignant melanoma. Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977. The surgery consisted of complete removal of the tumour together with about 2.5cm of the surrounding skin. Among the measurements taken were the thickness of the tumour and whether it was ulcerated or not. These are thought to be important prognostic variables in that patients with a thick and/or ulcerated tumour have an increased chance of death from melanoma. Patients were followed until the end of 1977. The data frame contains the following columns.

- `time` - Survival time in days since the operation.
- `status` - The patients status at the end of the study. 1 indicates that they had died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma.
- `sex` - The patients sex; 1=male, 0=female.
- `age` - Age in years at the time of the operation.
- `year` - Year of operation.
- `thickness` - Tumour thickness in mm.
- `ulcer` - Indicator of ulceration; 1=present, 0=absent.

Outline

Your assessment should include, at least, the following:

(i) Appropriate **summary statistics** for each of the variables in the data set and a **commentary** on the values of these statistics.

[10 Marks]

(ii) Appropriate **graphical summaries** of each of the variables in the data set and a **commentary** on any emerging aspects or trends.

[10 Marks]

(iii) A **regression analysis and appropriate correlation computations** for the relationship between the following variables

```
time      ~ thickness
time      ~ age
thickness ~ age
```

[20 Marks]

(iv) A **commentary** on any observed relationships between the variables in part (iii).

[10 Marks]

(v) Appropriate **two sample significance tests** for the variables in part (iii) grouped by **gender**.

[20 Marks]

(vi) For the three variables in part (iii) (grouped by **gender** as appropriate), **QQ-plots** and a **commentary** about the underlying distribution of the variables.

[10 Marks]

(vii) A **discussion** of the insights generated from the data as well as a **recommendations** on any aspects that should be investigated in more detail.

[20 Marks]

Submission

You should upload your complete assessment to the portal on canvas before the due date shown on your own canvas. I do not reproduce the due date here because for some of you it will be different due to extensions, extenuating circumstances etc.