

TRƯỜNG ĐẠI HỌC VĂN LANG
 ĐƠN VỊ: KHOA KỸ THUẬT CƠ – ĐIỆN VÀ MÁY TÍNH

ĐỀ THI/ĐỀ BÀI, RUBRIC VÀ THANG ĐIỂM
THI KẾT THÚC HỌC PHẦN
Học kỳ 3, năm học 2023-2024

I. Thông tin chung

Tên học phần:	Phân tích dữ liệu trực quan		
Mã học phần:	71DSDV40013	Số tín chỉ:	3
Mã nhóm lớp học phần:	233_71DSDV40013_01		
Hình thức thi: Tiểu luận nhóm (không TT)	Thời gian làm bài:	14	ngày
<input type="checkbox"/> Cá nhân	<input checked="" type="checkbox"/> Nhóm		
<i>Quy cách đặt tên file</i>	<i>Mã SV_Ho và ten SV_Nhom</i>		

1. Format đề thi

- Font: Times New Roman
- Size: 13
- Quy ước đặt tên file đề thi/đề bài:
- + **Mã học phần**_Tên học phần_Mã nhóm học phần_TIEUL_De 1

2. Giao nhận đề thi

Sau khi kiểm duyệt đề thi, đáp án/rubric. **Trưởng Khoa/Bộ môn** gửi đề thi, đáp án/rubric về Trung tâm Khảo thí qua email: khaothivanlang@gmail.com bao gồm file word và file pdf (*nén lại và đặt mật khẩu file nén*) và nhắn tin + họ tên người gửi qua số điện thoại **0918.01.03.09** (Phan Nhật Linh).

II. Các yêu cầu của đề thi nhằm đáp ứng CLO

(Phần này phải phối hợp với thông tin từ đề cương chi tiết của học phần)

Ký hiệu CLO	Nội dung CLO	Hình thức đánh giá	Trọng số CLO trong thành phần đánh giá (%)	Câu hỏi thi số	Điểm số tối đa	Lấy dữ liệu đo lường mức đạt PLO/PI
(1)	(2)	(3)	(4)	(5)	(6)	(7)
CLO1	Nắm vững các kiến thức nền tảng xác suất thống kê	Tiểu luận	15%	1,2,3,4	2,2,2,4	PI1.1
CLO2	Vận dụng các kiến thức chuyên sâu về xác suất và ngôn ngữ R nhằm trực quan và phân tích dữ liệu	Tiểu luận	40%	4	4	PI2.3
CLO3	Sử dụng thành thạo ngôn ngữ R trong trực quan hoá dữ liệu. Vận dụng kỹ năng lập luận, phân tích nhằm tìm ra ý nghĩa hành động từ dữ liệu	Tiểu luận	15%	1,2,3,4	2,2,2,4	PI5.3
CLO4	Xây dựng kế hoạch động nhóm và quản lý tiến độ công việc hiệu quả	Tiểu luận	15%	1,2,3,4	2,2,2,4	PI6.1
CLO5	Có khả năng đọc tài liệu và tự nghiên cứu về các phương pháp trực quan hóa dữ liệu	Tiểu luận	15%	1,2,3,4	2,2,2,4	PI9.3

Chú thích các cột:

(1) Chỉ liệt kê các CLO được đánh giá bởi đề thi kết thúc học phần (tương ứng như đã mô tả trong đề cương chi tiết học phần). Lưu ý không đưa vào bảng này các CLO không dùng bài thi kết thúc học phần để đánh giá (có một số CLO được bố trí đánh giá bằng bài kiểm tra giữa kỳ, đánh giá qua dự án, đồ án trong quá trình học hay các hình thức đánh giá quá trình khác chứ không bố trí đánh giá bằng bài thi kết thúc học phần). Trường hợp một số CLO vừa được bố trí đánh giá quá trình hay giữa kỳ vừa được bố trí đánh giá kết thúc học phần thì vẫn đưa vào cột (1)

(2) Nêu nội dung của CLO tương ứng.

(3) Hình thức kiểm tra đánh giá có thể là: trắc nghiệm, tự luận, dự án, đồ án, vấn đáp, thực hành trên máy tính, thực hành phòng thí nghiệm, báo cáo, thuyết trình, ..., phù hợp với nội dung của CLO và mô tả trong đề cương chi tiết học phần.

(4) Trọng số mức độ quan trọng của từng CLO trong đề thi kết thúc học phần do giảng viên ra đề thi quy định (mang tính tương đối) trên cơ sở mức độ quan trọng của từng CLO. Đây là cơ sở để phân phối tỷ lệ % số điểm tối đa cho các câu hỏi thi dùng để đánh giá các CLO tương ứng, bảo đảm CLO quan trọng hơn thì được đánh giá với điểm số tối đa lớn hơn. Cột (4) dùng để hỗ trợ cho cột (6).

(5) Liệt kê các câu hỏi thi số (câu hỏi số ... hoặc từ câu hỏi số... đến câu hỏi số...) dùng để kiểm tra người học đạt các CLO tương ứng.

(6) Ghi điểm số tối đa cho mỗi câu hỏi hoặc phần thi.

(7) Trong trường hợp đây là học phần cốt lõi - sử dụng kết quả đánh giá CLO của hàng tương ứng trong bảng để đo lường đánh giá mức độ người học đạt được PLO/PI - cần liệt kê ký hiệu PLO/PI có liên quan vào hàng tương ứng. Trong đề cương chi tiết học phần cũng cần mô tả rõ CLO tương ứng của học phần này sẽ được sử dụng làm dữ liệu để đo lường đánh giá các PLO/PI. Trường hợp học phần không có CLO nào phục vụ việc đo lường đánh giá mức đạt PLO/PI thì để trống cột này.

III. Nội dung đề bài

1. Đề bài

Bài tập 1: Thực hiện các câu hỏi bên dưới về nội dung Thay đổi khí hậu.

(2 điểm)

Sinh viên sử dụng các thư viện sau:

`library(tidyverse)`

`library(dslabs)`

`data(temp_carbon)`

`data(historic_co2)`

1. Lượng khí thải carbon năm vừa rồi lớn hơn năm đầu tiên bao nhiêu lần?

Gợi ý:

- Tìm năm đầu tiên (năm đầu tiên có chứa dữ liệu về khí thải carbon) và năm vừa rồi (năm cuối cùng có chứa dữ liệu về khí thải carbon).
 - Sử dụng hàm filter để lọc các năm (biến `year`) mà không có dữ liệu khí thải carbon (biến `carbon_emissions`).
 - In ra các năm (hàm `pull()`) và tìm năm đầu tiên và năm vừa rồi (hàm `min()`, `max()`)
 - Tìm lượng khí thải cacbon của những năm đã tìm ở trên.
 - Tính tỉ lệ lượng khí thải cacbon của năm vừa rồi và năm đầu tiên.
2. Hãy so sánh nhiệt độ đã tăng lên như thế nào giữa những năm gần đây và những năm xa nhất trong tập dữ liệu?

Gợi ý:

- Tìm năm gần đây và năm xa nhất như câu 1. Lưu ý rằng lọc các năm theo biến nhiệt độ thay đổi (biến `temp_anomaly`)
 - In ra nhiệt độ thay đổi của năm gần đây và năm xa nhất.
3. Vẽ time series plot của lượng phát thải carbon từ tập dữ liệu `temp_carbon`. Trục tung là lượng khí thải carbon (tấn) của mỗi năm.
4. Vẽ line plot về nồng độ CO2 qua các năm (Tô màu bằng cách sử dụng biến `source`).

Bài tập 2: Sinh viên trả lời các câu hỏi bên dưới về nội dung Properties of Stars.

(2 điểm)

Sinh viên sử dụng các thư viện sau:

`library(tidyverse)`

`library(dslabs)`

`data(stars)`

`options(digits = 3)`

1. Tính Mean magnitude?
2. Tính độ lệch chuẩn của magnitude?
3. Vẽ density plot của magnitude.
4. Từ câu 3 sinh viên hãy cho biết có bao nhiêu peaks trong dữ liệu?
5. Vẽ phân bố (distribution) của star temperature.

6. Cho biết Statement nào biểu diễn tốt nhất đặc tính của phân bố temperature (temperature distribution)?
7. Vẽ scatter plot của dữ liệu với temperature trên trục x và magnitude trên trục y .

Bài tập 3: Sinh viên trả lời các câu hỏi bên dưới về nội dung Titanis Survival. (2 điểm)

Sinh viên sử dụng các thư viện như sau:

```
options(digits = 3) # report 3 significant digits
library(tidyverse)
library(titanic)
```

```
titanic <- titanic_train %>%
  select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>%
  mutate(Survived = factor(Survived),
         Pclass = factor(Pclass),
         Sex = factor(Sex))
```

1. Vẽ density plot của độ tuổi (biến *age*) và được tô màu điểm bởi tình trạng survival (biến *Survived*).
2. Dựa vào câu 1, hãy cho biết giữa nam và nữ, nhóm nào có tỉ lệ sống cao hơn?
3. Lọc (filter) dữ liệu để loại bỏ người không trả tiền vé.
4. Vẽ boxplot của giá vé (fare) dựa trên thông tin về tình trạng sống sót (survival status).
5. Hãy biến đổi (transform) biến giá vé (fares) về thang đo log2.
6. Cho biết biến Pclass dùng phân loại tầng lớp hành khách. Hãy vẽ 3 biểu đồ barplot cho các lớp hành khách tương ứng.
 - Vẽ barplot cơ bản về lớp hành khách và tô màu nhằm phân biệt các tình trạng sống sót khác nhau được biểu diễn bởi biến *filled by survival*.
 - Vẽ lại barplot như trên nhưng sử dụng đối số `position = position_fill()` để mô tả mối liên hệ giữa mỗi nhóm thay vì dùng hàm đếm (`count`).
 - Vẽ barplot của người sống sót được tô màu bởi lớp hành khách sử dụng `position = position_fill()`.

Bài tập 4:
(4 điểm)

Dùng thuật toán phù hợp để crawl dữ liệu kết quả kỳ thi tốt nghiệp THPT năm 2018. Sử dụng package *ggplot2* để trực quan hoá dữ liệu và phân tích dữ liệu của kỳ thi.

Lưu ý:

1. Sinh viên nên phân tích theo hướng dữ liệu thu thập được đã truyền tải những nội dung gì quan trọng/hấp dẫn đến người đọc).
2. Mỗi nhóm sinh viên (2-5 bạn/nhóm) chỉ chọn 1 môn thi để phân tích dữ liệu.

2. Hướng dẫn thể thức trình bày đề bài

1. Trang bìa: Học phần, Tên đề tài, Họ tên sinh viên, Giảng viên
2. Mục lục
3. Header: Tên môn học
4. Footer: Tên sinh viên, Mã sinh viên, đánh số trang/Tổng số trang

5. Độ dài: 10 – 20 trang**6. Canh lề:**

Lề trên: Cách mép trên từ 20 – 25mm (2cm – 2.5cm).

Lề dưới: Cách mép dưới từ 20 – 25mm (2cm – 2.5cm).

Lề trái: Cách mép trái từ 30 – 35 mm (3cm – 3.5cm).

Lề phải: Cách mép phải từ 15 – 20 mm (1.5cm – 2cm).

3. Rubric và thang điểm

Tiêu chí	Trọng số (%)	Tốt 100%	Khá 75%	Trung bình 50%	Yếu 25%	Kém 0%
Hình thức trình bày	10%	Căn chỉnh hợp lý	1 – 3 đoạn căn chỉnh không hợp lý	4 – 5 đoạn căn chỉnh không hợp lý	6 – 8 đoạn căn chỉnh không hợp lý	Hơn 8 đoạn căn chỉnh không hợp lý
Nội dung lý thuyết	30%	Trình bày đầy đủ các nội dung lý thuyết được sử dụng	Thiếu 1 nội dung	Thiếu 2 nội dung	Thiếu 3 nội dung	Thiếu hơn 3 nội dung
Giới thiệu về bộ dữ liệu	10%	Giới thiệu rõ: nguồn và thời gian thu thập, tên và thang đo của các biến	Không giới thiệu 1 nội dung	Không giới thiệu 2 nội dung	Không giới thiệu 3 nội dung	Không giới thiệu cả 4 nội dung
Kết quả phân tích	30%	Chính xác	1 – 3 kết quả sai	4 – 5 kết quả sai	6 – 8 kết quả sai	Hơn 8 kết quả sai
Nhận xét kết quả	20%	Khớp với kết quả phân tích	Có 1 nhận xét không khớp	Có 2 nhận xét không khớp	Có 3 nhận xét không khớp	Có hơn 3 nhận xét không khớp

TP. Hồ Chí Minh, ngày tháng năm 2024

Người duyệt đề

Giảng viên ra đề

TS. Nguyễn Quốc Dũng

TS. Trương Quốc Trí