Final Exam:

R PROGRAMMING LANGUAGE

Thời gian làm bài : từ khi nhận đề đến 19h45, THứ 7 ngày 26/08/2023

- HV được sử dụng tài liệu.
- HV sẽ bị trừ điểm nếu bài làm giống nhau.

Câu 1: ames_housing

Cho dữ liệu ames_housing.csv. Hãy thực hiện các yêu cầu được liệt kê bên dưới:

1. Đọc dữ liệu, hiển thị thông tin chung của dữ liệu : head(), tail(), str(), summary()

In [ ]:
```
1
```

2. Cho biết số dòng, số cột của dữ liệu

In [ ]:
```
1
```

3. Cho biết có bao nhiêu loại Garage.Type, đó là những loại nào, mỗi loại đếm được bao nhiêu mẫu.

In [ ]:
```
1
```

4. Xóa các cột Garage.Qual, Garage.Cond, Pool.QC, Fence, Misc.Feature trong dữ liệu

In [ ]:
```
1
```

5. Tìm max, min của Gr.Liv.Area theo Electrical và Heating (sử dụng group_by() và summarize()).

In [ ]:
```
1
```

6. Vẽ biểu đồ thể hiện mối liên hệ của SalePrice và X1st.Flr.SF. Nhận xét biểu đồ.

In [ ]:
```
1
```

7. Vẽ pie chart thể hiện % giữa Y(1) và N(0) của cột Central.Air

In [ ]:
```
1
```

8. Cho biết năm xây của các căn nhà cũ nhất và mới nhất (theo Year.Built). Liệt kê các căn nhà cũ nhất, mới nhất với 3 thông tin Id, Year.Built, SalePrice

In [ ]:
```
1
```

9. Thống kê số lượng các căn nhà được xây theo từng năm. In head và tail. Cho biết năm nào có nhiều nhà được xây nhất?

In [ ]:
```
1
```

10. Trực quan hóa kết quả của câu thống kê trên với 10 năm gần đây nhất bằng barplot; với tất cả các năm bằng line.

In [ ]:
```
1
```

11. Vẽ boxplot của cột SalePrice.

In [ ]:
```
1
```

Câu 2: canxi

Cho dữ liệu canxi.xlsx. Hãy thực hiện các yêu cầu được liệt kê bên dưới

1. Đọc dữ liệu. Xem thông tin dữ liệu với head(), tail(), str(), summary().

In [ ]:
```
1
```

2. Vẽ biểu đồ phân phối tần suất của knowledge_score. Nhận xét.

In [ ]:
```
1
```

3. Thực hiện các thống kê cơ bản cho knowledge_score và calcium_intake (mean, median, max, min)

In [ ]:
```
1
```

4. Vẽ boxplot cho knowledge_score và cho calcium_intake. knowledge_score có outlier hay không? calcium_intake có outlier hay không? Nhận xét

In [ ]:
```
1
```

5. Vẽ biểu đồ thể hiện mối quan hệ giữa knowledge_score và calcium_intake. Nhận xét.

In [ ]:
```
1
```

Câu 3: Cho dữ liệu fruit_data_with_colors.txt

Hãy thực hiện các yêu cầu sau:

1. Đọc dữ liệu. Xem thông tin dữ liệu với head(), tail(), str(), summary().

In [ ]:
```
1
```

2. Hãy cho biết kiểu dữ liệu của cột fruit_name và fruit_subtype. Nếu kiểu dữ liệu không phải là factor thì hãy chuyển thành factor

In [ ]:
```
1
```

3. Có bao nhiêu loại fruit_name? Đó là những loại nào? Có bao nhiêu loại fruit_subtype? Đó là những loại nào?

In [ ]:
```
1
```

4. Hãy lọc ra tất cả các dòng dữ liệu có fruit_name là 'apple' chứa vào dataframe df_apple. Hãy cho biết có bao nhiêu dòng dữ liệu thỏa điều kiện này?

In [ ]:
```
1
```

5. Hãy lưu dataframe df_apple vào tập tin apple.csv

In [ ]:
```
1
```

6. Hãy lọc ra tất cả các dòng dữ liệu có fruit_subtype là 'golden_delicious' chứa vào dataframe df_golden_delicious. Hãy cho biết có bao nhiêu dòng dữ liệu thỏa điều kiện này?

In [ ]:
```
1
```

7. Hãy lưu dataframe df_golden_delicious vào tập tin golden_delicious.xlsx

In [ ]:
```
1
```

8. Hãy lưu dataframe df_golden_delicious vào tập tin golden_delicious.xml

In [ ]:
```
1
```

9. Nhóm theo fruit_subtype, hãy thống kê số lượng mẫu theo từng subtype; max và min từng cột mass, width, height. (Gợi ý: dùng group_by() và summarize())

In [ ]:
```
1
```

10. Hãy lưu kết quả thống kê trên vào file subtype_summarize.json

In [ ]:
```
1
```

11. Đọc file subtype_summarize.json vừa lưu. In nội dung.

In [ ]:
```
1
```