2022 데이터 분석기초 기말고사

데이터 설명

- 서울시에서 제공한 모기지수 자료 및 기상청 자료의 일부입니다.
- 모기지수는 높을 수록 모기가 발생할 확률이 높아지는 값입니다.
- 모기지수와 날씨 정보와의 관계를 확인해보려 합니다.
- 데이터는 총 2,241개의 행과 9개의 열로 구성되어 있습니다.
- 모기지수는 날씨변수들을 사용하여 계산된 값입니다.
- 모든 값은 발생일 기준으로 측정, 계산됐습니다.

모기지수 발생일	모기지수	평균기온	최저기온	최고기온	일강수량	평균 증기압	평균 전운량	상대습도
2022-08-02	51.5	26.8	25.5	28.5	33.5	31.6	9.9	90
2022-08-01	59.3	28.6	25.4	32.4	30	31.8	8.9	82.4
2022-07-31	58.8	27.2	25.4	29	47.2	30.7	9.8	85.8
2022-07-30	56.8	30.9	27.3	36.1	1	30.2	4.5	68.4
2022-07-29	55.9	30.7	27.3	34.3	0	29.8	5	68.1
2022-07-28	55.1	29.7	25.1	34.4	0	26.3	3.5	63.8

1 데이터 전처리

- 회귀 문제를 위한 시작 단계 문제
  - 동작 1: 데이터에서 NaN값이 하나라도 포함된 행을 drop하세요.
  - 동작 2: 모기지수가 100 이하일 경우 0, 100 초과일 경우 1을 가지는 '모기라벨' 이라는 열을 만드세요.
  - 동작 3: '모기지수 발생일' 열은 사용하지 않으므로 drop하세요.
  - 동작 4: 각 컬럼에 대한 평균 값을 출력하세요

1

데이터 전처리 출력 예시

```
> 모기지수          26.933970
평균기온          12.854975
최저기온           8.803618
최고기온          17.559799
일강수량           3.807337
평균 증기압        11.859698
평균 전운량         7.392060
상대습도          63.452663
모기라벨           0.000000
dtype: float64
```

※ 열 순서를 지켜주세요.

값은 정답과 무관합니다

2 이상 데이터 탐지하기

※ 1번의 전처리가 적용된 데이터가 주어집니다.

- 각 열의 데이터중 어떤 열에 이상치 데이터가 탐지되는지 알아보고자 합니다.

  아래 동작을 차례대로 정확히 수행하세요.

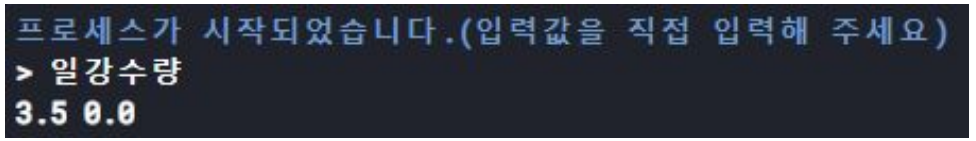
  1. 사용자로부터 하나의 열 이름을 입력받으세요.
  2. 해당 열에서 아래 식의 결과값을 초과하는 값을 가진 데이터는 이상치 데이터로,

     이상치 데이터를 포함한 행을 모두 지우세요.

     a. (제3사분위수) + 1.5 * (제3사분위수 - 제1사분위수)
  3. 이상치 데이터를 지우기 전과 후의 해당 열 평균값을

     아래와 같이 변수를 출력하여 **소수점 한자리까지** 각각 출력하세요.

     a. “{:.1f}".format(변수)



입출력 예시

3 파생 변수 만들기

※ 1번의 전처리가 적용된 데이터가 주어집니다.

- 미 국립기상국에서 고안한 '열 지수'를 계산해 모기지수와의 관계를 확인해보려 합니다.
- **식에서 T는 화씨온도(섭씨 * 1.8 + 32), R은 상대습도를 나타냅니다.**

(열지수 식)

$$\begin{aligned}\mathrm{HI} = &-42 + 2.05\mathrm{T} + 10.14\mathrm{R} - 0.23\mathrm{T} * \mathrm{R} - 0.0684T^2 - 0.0548R^2 \\ &+ 0.00123T^2 * R + 0.000853T * R^2 - 0.00000199T^2 * R^2\end{aligned}$$

3 파생 변수 만들기

- 동작 1: R이 13% 미만이고 80 < T < 112 인 경우 HI에서 $(\frac{(13-\mathrm{R})}{4}) * \mathrm{SQRT}(\frac{(17-\mathrm{ABS}(\mathrm{T}-95))}{17})$를 빼세요
- 동작 2: R이 85% 초과이고 80 < T < 87인 경우 HI에 $((\frac{(\mathrm{R}-85)}{10}) * (\frac{(87-T)}{5}))$ 를 더하세요.
- 동작 3: T가 80 미만일 경우 HI를 $0.5 * (\mathrm{T} + 61 + ((\mathrm{T}-68) * 1.2) + (\mathrm{R} * 0.094))$로 변경하세요.
- 동작 4: 열지수가 평균 이상인 행들에 대해서 열지수와 모기지수의 상관관계를 출력하세요.

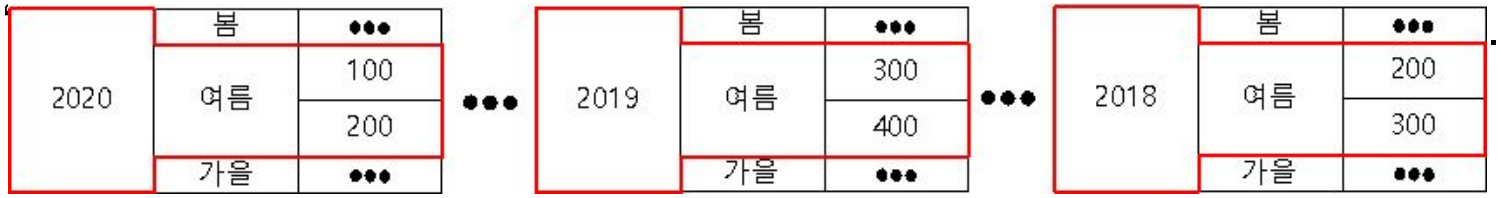
4 계절 별 모기지수 변화

※ 1번의 전처리가 적용되었으며, 단 '모기지수 발생일'은 본 문제에서만 사용되도록 남아 있습니다.

- 계절에 따른 모기지수의 변화를 살펴보고자 합니다. 아래의 동작을 수행해 주세요.
  - 봄은 {3, 4, 5}월, 여름은 {6, 7, 8}월, 가을은 {9, 10, 11}월, 겨울은 {12, 1, 2}월입니다.
  - 살펴보고자 하는 계절을 입력받으세요. (봄, 여름, 가을, 겨울 중 하나).
  - 동작 1: 입력받은 계절에 대한 **평균 모기지수**를 출력하세요.
  - 동작 2: **입력받은 계절**에 대한 **연도 별 평균 모기지수**를 계산하세요.
    계산한 연도 별 평균 모기지수에 대한 내림차순으로 연도를 출력하세요.

[입출력 예시: 예시의 값은 실제 문제와 무관합니다]

  - 입력이 '

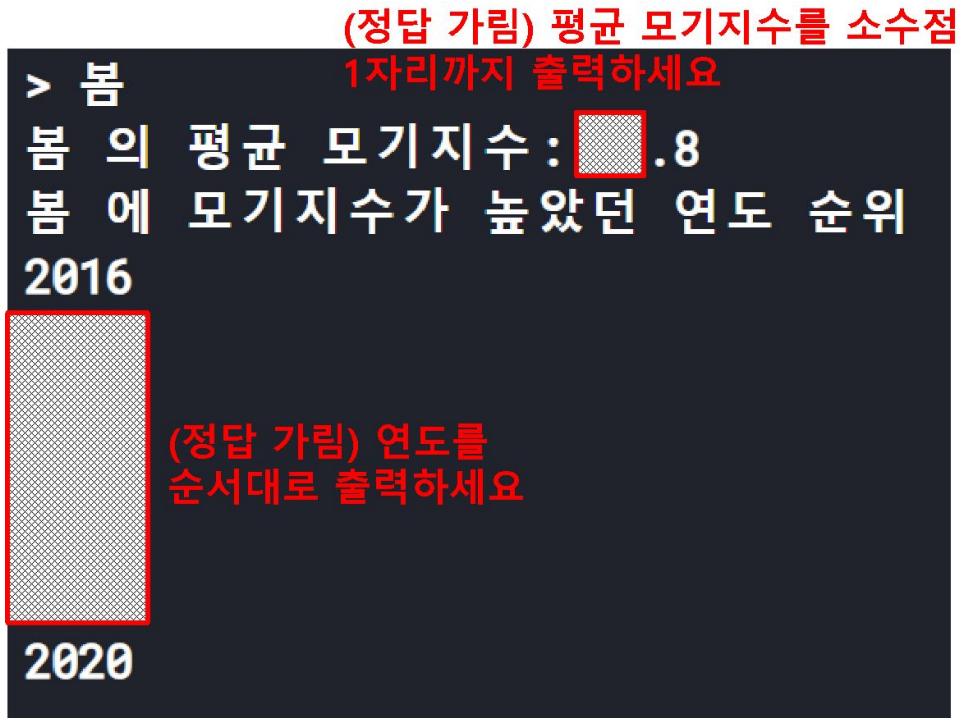


.

  - [출력 1: 250.0], [출력 2: 2019 2018 2020]

4 계절 별 모기지수 변화 출력 예시

- 예제 코드에 포함된 아래 양식을 활용하여 **소수점 한자리까지** 출력하세요.
  - “{:.1f}".format(변수)



5 회귀 모델 학습

- **문제 1번 전처리된 데이터에서 '모기지수' 열이 drop된 데이터가 주어집니다.**
- 종속변수는 '모기라벨'이며, 주어진 변수들로 모기라벨을 분류하는 로지스틱 회귀 모델을 만드는 것이 목표입니다.
- 지금까지 배운 여러 가지 기법들을 사용하여 모델의 성능을 높이는 것에 도전해봅시다.
- 데이터 시각화를 포함한 여러 데이터 분석 기법을 통해 데이터 분석을 진행해보고 전처리, 파생변수 생성, 독립변수 선택 등으로 모델의 score를 높여 봅시다.
- colab에서 시각화 및 분석을 자유롭게 할 수 있도록 Train 데이터는 아이캠퍼스에 공유됩니다.

5 회귀 모델 학습 – 주의사항

- Test 데이터에 대한 수정은 독립 변수 선택 및 파생변수 생성 이외에는 허용되지 않습니다.
- Test 데이터에 대한 전처리, Test 데이터에 대한 길이(데이터 수) 조정 등은 허용되지 않습니다.
- Test 데이터에 대한 허용 범위 이외의 수정이 확인될 경우 불이익을 받을 수 있습니다.
- Train score가 0.80 이하일 경우 0점으로 채점됩니다.
- 채점함수가 복잡하니 변수명을 꼭 지켜주시기 바랍니다.

Model

채점함수 :

```
calculation(len train_X keys()), lr.score train_X, train_Y), lr.scor (test_X, test_Y))
```

Train data　　Train data, Train label　　Test data, Test label

5 회귀 모델 학습 – 채점 기준

※ Train score가 0.80을 넘지 않으면 0점입니다.

채점 기준	배점 (%)
0.75 ≤ Test score	100
0.73 ≤ Test score < 0.75	80
0.71 ≤ Test score < 0.73	60
0.67 ≤ Test score < 0.71	40
0.60 ≤ Test score < 0.67	20
Test score < 0.60	0