1. 문제 1유형 1번

문제 요약 및 목표

우리는 나라별 커피 원두 소비량 데이터를 이용해 다음과 같은 목표를 달성하려고 합니다:

1. 대륙별로 `Arabica` 커피 원두 소비량의 평균을 계산하고, 그 중에서 가장 높은 대륙을 찾습니다.
2. **(1)**에서 찾은 대륙에서 `Arabica` 커피 원두 소비량이 가장 많은 **5**개 나라를 찾습니다.
3. 그 다섯 나라의 소비량을 높은 순서대로 나열합니다.

단계별 풀이 방법

1. 대륙별 평균 **`Arabica`** 소비량 계산

- 의미: `Arabica`는 커피의 한 종류인데, 이 커피의 소비량이 대륙별로 평균적으로 얼마나 되는지를 계산해 보자는 겁니다.
- 방법:
  - 데이터를 `continent`(대륙) 기준으로 묶습니다.
  - 각 대륙별로 `Arabica` 소비량의 평균을 계산합니다.
- 예시: 만약 아시아와 유럽이 있다면, 아시아의 모든 나라들의 `Arabica` 소비량을 더해서 나라 수로 나눈 값을 아시아의 평균으로 계산합니다.

2. 평균 소비량이 가장 높은 대륙 찾기

- 의미: `Arabica`를 가장 많이 마시는 대륙이라고 볼 수 있습니다.
- 방법:
  - (1)에서 계산한 대륙별 평균 값들 중에서 가장 큰 값을 찾아 이름을 확인합니다.

3. 해당 대륙에서 **`Arabica`** 소비량이 높은 다섯 개 나라 찾기

- 의미: `Arabica` 소비량이 높은 대륙을 찾았으니, 이제 그 대륙 내에서 `Arabica`를 가장 많이 소비하는 나라 **5**개를 찾는 과정입니다.
- 방법:
  - (2)에서 찾은 대륙에 해당하는 나라들만 필터링합니다.
  - 그 나라들 중에서 `Arabica` 소비량이 가장 높은 순서대로 다섯 개 나라를 선택합니다.

4. 다섯 나라를 소비량 순서대로 나열

- 의미: 이제 `Arabica` 소비량이 가장 많은 다섯 개 나라를 찾아냈으니, 그 나라들을 소비량이 많은 순서대로 나열하여 보여줍니다.
- 방법:
  - 선택된 다섯 개 나라를 `Arabica` 소비량 기준으로 내림차순**(**큰 값부터 작은 값으로**)** 정렬합니다.

요약

- 데이터를 그룹으로 묶고, 평균을 계산하는 방법을 배우면 대륙별 소비량 비교가

가능합니다.
- 특정 기준으로 데이터를 필터링하는 방법을 알면 원하는 대상을 찾을 수 있습니다.
- 상위 몇 개의 값을 찾는 방법은 데이터를 분석할 때 자주 사용됩니다.

2. 문제 **1**유형 **2**번

문제 요약 및 목표

주어진 데이터는 각 나라별로 세단 판매량, **SUV** 판매량, 총 판매량을 포함하고 있습니다. 여기서 해결해야 할 문제는:

1. **SUV** 판매 비율이 세 번째로 높은 나라의 세단 판매량을 구합니다.
2. 세단 판매량이 네 번째로 높은 나라의 SUV 판매량을 구합니다.
3. 이 두 값을 더한 합을 구합니다.

단계별 풀이 방법

1. SUV 판매 비율 계산

- 의미: SUV 판매 비율은 각 나라에서 전체 판매량 중 SUV가 차지하는 비율을 의미합니다.
- 계산 방법:
  - `SUV 판매 비율` = `suv_sales` / `total_sales`
  - 각 나라에 대해 계산한 SUV 판매 비율을 새로운 열로 추가합니다.

2. SUV 판매 비율이 세 번째로 높은 나라 찾기

- 의미: 모든 나라의 SUV 판매 비율 중 세 번째로 높은 값을 가진 나라를 찾는 것입니다.
- 방법:
  - (1)에서 계산한 SUV 판매 비율을 기준으로 내림차순으로 정렬합니다.
  - 세 번째 위치에 있는 나라의 세단 판매량(`sedan_sales`)을 선택합니다.

3. 세단 판매량이 네 번째로 높은 나라 찾기

- 의미: 모든 나라의 `sedan_sales`를 기준으로 네 번째로 높은 값을 가진 나라를 찾는 것입니다.
- 방법:
  - `sedan_sales`를 기준으로 내림차순으로 정렬합니다.
  - 네 번째 위치에 있는 나라의 SUV 판매량(`suv_sales`)을 선택합니다.

4. 두 값을 합산

- 의미: (2)와 (3)에서 구한 값을 더하여 최종 결과를 도출합니다.
- 방법:
  - (2)에서 구한 세단 판매량과 (3)에서 구한 SUV 판매량을 더합니다.

요약

- **SUV** 판매 비율을 계산한 뒤, 이를 기준으로 나라를 정렬하여 특정 위치에 있는 값을 찾는 방법을 익힙니다.
- 데이터를 정렬하고 순위를 찾는 방법은 다양한 데이터 분석 문제에서 활용됩니다.
- 데이터를 이해하고 작은 문제들로 나누어 해결하는 과정이 중요합니다.

3. 문제 **1**유형 **3**번

문제 요약 및 목표

1. Co와 Nmch 값을 최대**-**최소 정규화하여 각 변수를 동일한 스케일로 맞춥니다.
2. 정규화된 Co와 Nmch의 표준 편차를 구한 후, 이 두 표준 편차의 차이를 계산합니다.
3. 두 변수 중 변동성이 큰 변수가 무엇인지 판단합니다.

단계별 풀이 방법

1. 최대**-**최소 정규화 **(Min-Max Normalization)**

- 의미: 정규화는 각 변수의 값을 0과 1 사이의 범위로 변환하여 서로 다른 스케일을 맞추는 과정입니다. 이를 통해 값의 크기에 관계없이 비교가 가능합니다.
- 계산 방법:
    - 정규화된 값 = (x−min(x))/(max(x)−min(x))(x - \min(x)) / (\max(x) - \min(x))(x−min(x))/(max(x)−min(x))
    - 각 변수의 최솟값과 최댓값을 구한 후, 이 공식을 적용하여 모든 데이터를 0~1 사이로 변환합니다.

2. 표준 편차 계산

- 의미: 표준 편차는 데이터가 평균값을 중심으로 얼마나 퍼져 있는지를 나타내는 값입니다. 즉, 데이터의 변동성이 어느 정도인지 측정할 수 있습니다.
- 계산 방법:
    - 정규화된 Co와 Nmch 각각에 대해 표준 편차를 계산합니다.
    - 두 표준 편차의 차이를 구하여 변동성 차이를 확인합니다.

3. 변동성이 큰 변수 찾기

- 의미: 변동성이 큰 변수란, 값들이 평균을 중심으로 더 많이 퍼져 있는 변수를 의미합니다.
- 방법:
    - 표준 편차가 더 큰 변수가 변동성이 더 큽니다.
    - 표준 편차가 큰 변수를 선택합니다.

요약

1. 정규화: 데이터를 0~1 사이로 변환하여 비교할 수 있게 합니다.
2. 표준 편차: 데이터의 변동성을 측정하여 어떤 변수가 더 넓게 퍼져 있는지를 확인합니다.
3. 변동성 판단: 표준 편차가 큰 변수가 변동성이 더 큽니다.

4. 문제 **2**유형

문제 요약 및 목표

- 목표: 주어진 데이터를 이용하여 각 지하철역의 이용객 수(`num_people`)를 예측하는 머신러닝 모델을 구축하는 것입니다.
- 주어진 데이터: 각 날짜별 지하철역의 특성(날씨, 가시거리, 강수량, 온도 등)과 지하철역 이름, 이용객 수가 포함된 데이터입니다.
- 채점 기준: 모델의 성능은 예측값과 실제값의 차이를 나타내는 RMSE(Root Mean Squared Error, 평균 제곱근 오차)를 기준으로 평가됩니다.

단계별 풀이 방법

1. 데이터 준비 및 이해

- 의미: 데이터를 분석하기 전에 변수들을 이해해야 합니다.
- 주요 변수:
  - `num_people`: 예측해야 하는 목표 변수(타겟)입니다.
  - `date`, `day_of_week`, `month`, `station_name`: 날짜와 지하철역에 관련된 변수들입니다.
  - `visibility`, `precipitation`, `temperature`: 날씨와 관련된 변수들입니다.
- 방법:
  - 데이터를 불러와 각 변수의 의미를 확인하고, 데이터의 결측치 및 이상치를 탐색합니다.

2. 데이터 전처리

- 의미: 모델이 데이터를 잘 학습할 수 있도록 변수를 적절히 변환하고, 결측치를 처리해야 합니다.
- 방법:
  - 날짜 정보를 사용하여 `month`와 `day_of_week` 같은 변수를 파생 변수로 사용합니다.
  - `station_name` 같은 범주형 변수를 숫자형으로 변환하기 위해 **One-Hot Encoding**을 적용할 수 있습니다.
  - `visibility`, `precipitation`, `temperature`의 결측치를 처리하거나 적절히 변환합니다.

3. 모델 선택 및 학습

- 의미: 데이터를 바탕으로 이용객 수를 예측할 수 있는 모델을 선택하여 학습합니다.
- 방법:
  - 주로 사용되는 모델: 랜덤 포레스트**(Random Forest)**, **Gradient Boosting**, 선형 회귀**(Linear Regression)** 등.
  - 데이터를 훈련 데이터와 테스트 데이터로 나누어, 훈련 데이터로 모델을 학습하고, 테스트 데이터로 성능을 평가합니다.
  - **RMSE**를 기준으로 모델의 성능을 평가합니다.

4. 예측 및 결과 저장

- 의미: 학습된 모델을 이용하여 테스트 데이터의 `num_people` 값을 예측하고, 이를 저장하여 제출합니다.
- 방법:

    - 모델을 이용해 예측값을 계산하고, 이를 CSV 파일로 저장합니다.
    - CSV 파일에는 테스트 데이터의 날짜, 역 이름, 예측된 `num_people`을 포함해야 합니다.

5. 성능 평가 및 개선

- 의미: RMSE 값을 줄이기 위해 모델의 성능을 개선할 수 있는 방법을 탐색합니다.
- 방법:
    - 하이퍼파라미터 튜닝: 모델의 파라미터를 조정하여 성능을 향상시킵니다.
    - 데이터 전처리 개선: 날씨 데이터와 요일 등을 더 세분화하여 추가적인 정보를 모델에 제공합니다.
    - 더 복잡한 모델 사용: 예를 들어, **XGBoost**나 **LightGBM**과 같은 모델을 사용해 볼 수 있습니다.

요약

- 데이터 전처리: 데이터를 이해하고 모델에 맞게 변환합니다.
- 모델 선택 및 학습: 적절한 모델을 선택해 데이터를 학습하고, 예측 결과를 생성합니다.
- 성능 평가: RMSE를 이용해 모델의 정확도를 평가합니다.
- 결과 제출: 예측 결과를 CSV 파일로 저장하여 제출합니다.

5. 문제 **3**유형 **1**번

문제 요약 및 목표

1. 로지스틱 회귀 수행:
    - 고객 이탈 여부(`churn`)를 예측하기 위해 로지스틱 회귀 모델을 사용합니다.
    - 모든 변수들을 포함하여 모델을 학습한 후, 유의하지 않은 변수를 제거합니다.
2. 유의한 변수만을 사용한 로지스틱 회귀:
    - 1단계에서 찾은 유의한 변수만으로 로지스틱 회귀를 다시 수행하고, 이 모델의 회귀 계수의 평균을 계산합니다.
3. 변수 변화에 따른 오즈비 계산:
    - 로지스틱 회귀 결과를 바탕으로 특정 변수(`calls`)가 증가할 때 **오즈비(Odds Ratio)**가 어떻게 변하는지 계산합니다.

단계별 풀이 방법

1. 로지스틱 회귀 모델 구축 **(**상수항 포함**)**

- 의미: 로지스틱 회귀는 종속변수가 이진형(예: 0/1, 유지/이탈)인 경우, 그 확률을 예측하는 모델입니다. 이 문제에서는 고객의 `churn` 여부를 예측합니다.
- 방법:
    - 모든 변수(`age`, `gender`, `calls`, 등)를 사용하여 로지스틱 회귀 모델을 만듭니다.
    - 회귀 결과에서 **p-value**를 확인하여 각 변수의 유의성을 평가합니다.
    - **p-value**가 0.05보다 크면, 해당 변수는 통계적으로 유의하지 않다고 판단할 수 있습니다.

2. 유의한 변수만을 사용하여 로지스틱 회귀 다시 수행

- 의미: 1단계에서 통계적으로 유의하지 않은 변수를 제거한 후, 유의한 변수만을 이용해 모델을 다시 학습합니다.
- 방법:
    - 유의한 변수만을 추출하여 새로운 로지스틱 회귀 모델을 학습합니다.
    - 새 모델의 회귀 계수들을 계산하고, 이들의 평균을 구합니다.

3. `calls` 변수의 변화에 따른 오즈비 계산

- 의미: 오즈비는 특정 변수의 값이 변할 때, 결과가 긍정적으로 변화할 확률의 변화 정도를 나타냅니다. 여기서는 `calls` 변수가 5 증가할 때 오즈비의 변화를 계산합니다.
- 방법:
    - 로지스틱 회귀 모델의 `calls`에 대한 회귀 계수를 사용하여 오즈비를 계산합니다.
    - `calls`가 1 증가할 때의 오즈비는 exp(βcalls)\exp(\beta_{\text{calls}})exp(βcalls)입니다.
    - `calls`가 5 증가할 때의 오즈비는 exp(5×βcalls)\exp(5 \times \beta_{\text{calls}})exp(5×βcalls)로 계산할 수 있습니다.

요약

- 로지스틱 회귀: 고객 이탈 여부를 예측할 수 있는 모델을 만듭니다.

- 유의성 평가: p-value를 통해 통계적으로 중요한 변수를 판별합니다.
- 오즈비 계산: 특정 변수의 변화가 결과에 미치는 영향을 파악합니다.

6. 문제 3유형 2번

문제 요약 및 목표

- 목표: 주어진 데이터를 사용해 PIQ를 예측하는 모델을 구축하고, 이를 통해 테스트 데이터를 이용한 결정계수(R^2)를 계산하거나 새로운 데이터를 사용해 PIQ를 예측합니다.
- 주어진 데이터: `PIQ`(목표 변수), `brain_size`(뇌 크기), `height`(키), `weight`(몸무게), `education_level`(교육 수준), `reading_habits`(독서 습관), `age`(나이) 등을 포함합니다.

단계별 풀이 방법

1. 다중선형회귀 모델 구축 및 유의한 변수 찾기

- 의미: 다중선형회귀 분석은 여러 독립변수(`brain_size`, `height`, `weight`, 등)를 사용해 종속변수(`PIQ`)를 예측하는 방법입니다.
- 방법:
  - 모든 독립변수를 포함한 다중선형회귀 모델을 만듭니다.
  - 모델의 회귀 계수를 통해 각 변수의 영향을 파악하고, **p-value**를 이용해 유의한 변수를 확인합니다.
  - **p-value**가 0.05보다 작을 경우, 해당 변수는 종속변수에 통계적으로 유의미한 영향을 미친다고 판단할 수 있습니다.

2. 테스트 데이터로 결정계수**(R^2)** 계산

- 의미: R^2 값은 모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표입니다. 1에 가까울수록 모델이 데이터를 잘 설명합니다.
- 방법:
  - 데이터를 훈련 데이터와 테스트 데이터로 나눕니다.
  - 훈련 데이터로 모델을 학습하고, 테스트 데이터로 예측한 값과 실제 값을 비교하여 R^2 값을 계산합니다.

3. 새로운 데이터로 **PIQ** 예측

- 의미: 새로운 입력 데이터를 사용해 PIQ를 예측하는 과정입니다.
- 방법:
  - 새로운 데이터를 모델에 입력하여 PIQ 값을 예측합니다.
  - 이 예측된 값은 모델이 추정한 새로운 데이터의 PIQ입니다.

요약

- 다중선형회귀 분석: 여러 변수를 사용해 목표 변수(PIQ)를 예측합니다.
- 유의한 변수 찾기: p-value를 통해 통계적으로 중요한 변수를 확인합니다.
- 결정계수**(R^2)**: 모델의 예측 성능을 평가하는 지표입니다.
- 새로운 데이터 예측: 모델을 이용해 새로운 데이터의 PIQ를 예측할 수 있습니다.