Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. Python 설치 방법

1) 크롬 브라우저에서 anaconda.com로 접속한다.
2) Download 화면에서 Windows / macOS / Linux 중 PC운영체제에 맞는 버전을 다운 받는다.
   ※ OS가 32-Bit인지, 64-Bit인지도 확인해야 한다.

2. Jupyter Notebook 사용법

1) 폴더 생성: New → Folder 클릭한다.
2) 폴더 이름 변경: 변경할 이름의 폴더 체크 → Rename 클릭하여 이름 변경 → Rename 클릭한다.
3) 실행 환경: New → Python 3 선택한다.
4) 결과 출력: Ctrl + Enter
5) 결과 출력 + 빈 셀 생성: Shift + Enter

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 변수

- 숫자로 시작할 수 없다.
- _(언더 스코어)로 시작할 수는 있으나, 그 외 특수문자는 사용할 수 없다.
- 시스템 예약어나 연산자도 사용할 수 없다.
- 대소문자를 구분하여 동일한 이름에 대소문자를 다르게 사용한 경우 다른 변수로 생각한다.
- 변수로 사용할 이름을 입력하고 등호(=)를 입력한 다음 변수에 넣어 두고 싶은 내용을 입력하면 변수를 정의할 수 있다.

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 함수

- 함수는 짧은 명령어로 어떤 특정한 작업을 하도록 한다.
- 파이썬을 설치하면 파이썬 기본 코드가 내장한 함수를 바로 사용할 수 있으며, 이외에도 패키지를 설치하여 사용 가능한 함수를 확장할 수 있다.

3. 데이터 타입

List, Tuple, Set, Dict(dictionary)

4. 연산자

사칙연산자, 비교연산자, 멤버십연산자

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 데이터 파일 업로드

- read_csv는 pandas 패키지에서 csv 파일을 읽을 때 사용할 수 있는 함수 가운데 하나이다.

2. 데이터 업로드 중 발생하는 문제 해결 방법

- 디폴트 인코딩 기준은 UTF-8이다.
- 데이터를 읽을 때 인코딩 기준이나 separator를 바꿀 수 있다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

- **pandas 패키지**
  - 엑셀 같은 테이블 구조를 가진 DataFrame 타입은 편리함과 다양한 기능으로 데이터 분석에 널리 사용된다.
  - 데이터프레임의 특정 컬럼만 추출할 때는 대괄호를 사용하거나 점 연산자를 사용한다. (bike_data['Distance'] 또는 bike_data.Distance)
  - 특정 컬럼의 값을 조건으로 추출할 때 대괄호 안에 조건을 제시하는 방법을 사용할 수 있다. (bike_data[bike_data.Momentum == 'WWN'])

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 데이터 전처리

1) 개념

- 데이터 분석 작업 전에 데이터를 분석하기 좋은 형태로 만드는 과정이다.

2) 종류

① 데이터 정제(Data Cleansing): 없는 데이터는 채우고, 잡음은 제거하며,
모순된 데이터는 정합성이 맞는 데이터로 교정하는 작업

② 데이터 통합(Data Integration): 여러 개의 데이터 Source를 통합하는 작업

③ 데이터 축소(Data Reduction): 샘플링 등을 통해 데이터 볼륨을 줄이거나 분석 대상 속성을 줄이는 작업

④ 데이터 변환(Data Transformation): 데이터 정규화(Normalization) or 집단화(aggregation)하는 작업

3) 특성

- Case-by-case로 데이터마다 전처리 방법이 다양하다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 데이터 정제

1) 결측값 : 데이터가 비어있는 경우를 말한다.

2) 이상값 : 특정 범위를 벗어나는 극단값을 말한다.

→ 분석의 품질을 높이기 위해서는 적절한 처리 후, 데이터 분석을 해야 한다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

• 데이터 결합

1) 필요성

- 데이터가 여러 개의 테이블에 나뉘어 저장되어 있는 경우가 많으므로, 실제 사용해야 하는 시점에 데이터를 연결하여 사용해야 한다.

2) 방법

- 두 테이블 간에 공통으로 존재하는 컬럼을 Key로 선정한다.
- Join Type을 기준으로 데이터를 결합한다.

3) Join Type 종류

① inner: 교집합
② outer: 합집합
③ Left: 왼쪽 테이블에 있는 데이터 기준으로 오른쪽 데이터를 가져오는 방식
④ right: 오른쪽 테이블에 있는 데이터 기준으로 왼쪽 데이터를 가져오는 방식

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 데이터 시각화의 필요성

1) 분석 결과를 효과적으로 전달하는 방법 중에 하나이다.

2) 구체적인 분석에 앞서 데이터 전체에 대한 이해를 할 수 있다.

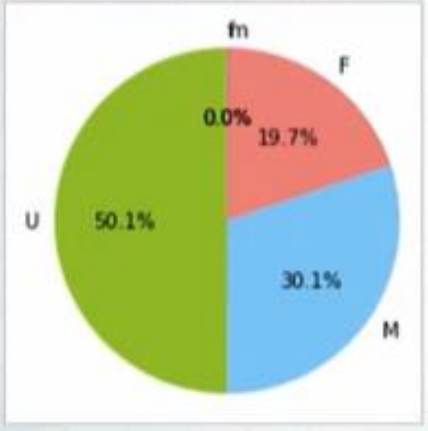
3) 수치 데이터로 확인할 수 없는 데이터의 특성을 확인할 수 있다.

Summary

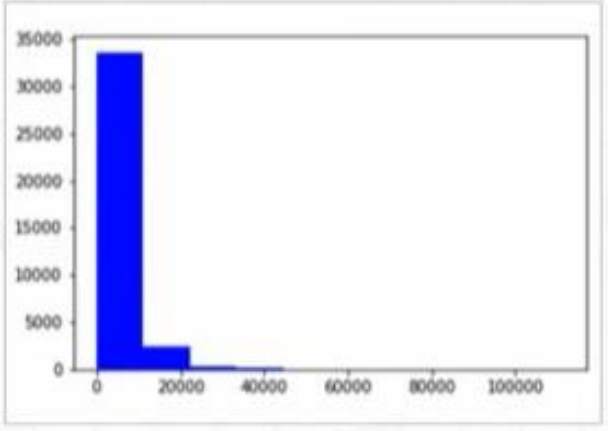
아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

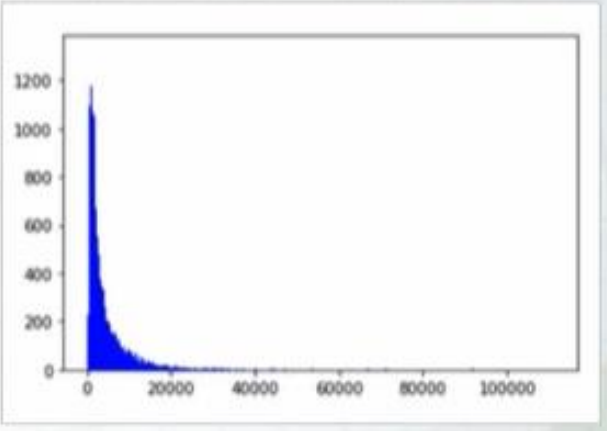
2. 그래프 유형

1) 파이 차트



2) 히스토그램



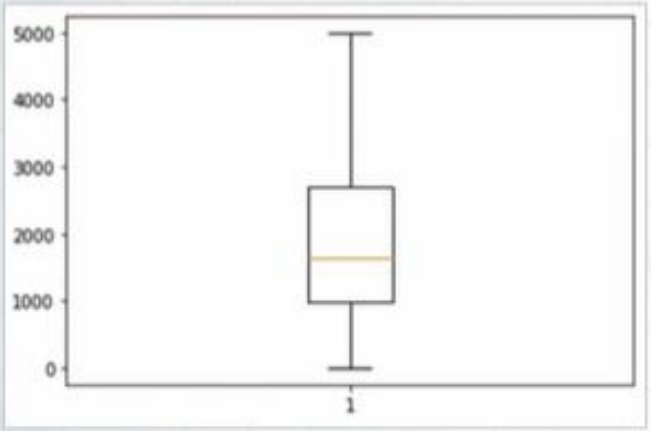


Summary

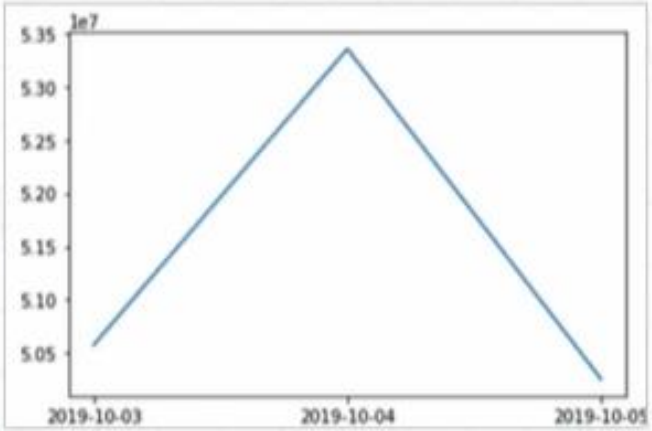
아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 그래프 유형

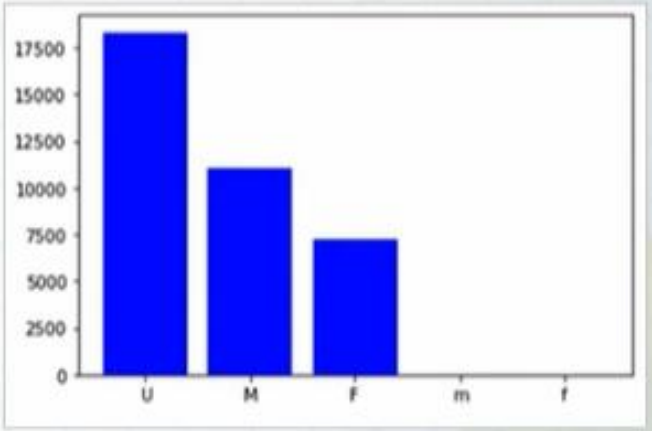
3) 상자 수염 그림



4) 선그래프



5) 막대그래프



Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 탐색적 데이터 분석에 빈번히 활용되는 함수

1) value_counts( )

특정 컬럼 내용을 구성하는 값과 각 값의 빈도를 보여준다.

2) pivot_table

데이터를 원하는 기준으로 요약 정리하는 방법이다.

3) melt

pivot에 의해 요약 테이블로 바꾸었던 모양을 다시 요약 전 테이블처럼 만드는 작업이다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 정규분포

통계의 이론적 중심에 있는 분포이며, 종모양 좌우대칭의 형태로 자연 현상에서 종종 발견된다.

2. 중심극한정리

어떤 모집단에서 표본을 취하고 평균을 구할 때, 횟수가 충분히 크면 표본 평균은 정규분포를 따른다.

3. Histogram

정규분포, 중심극한정리를 확인할 때 활용 할 수 있다.

4. Q-Q plot

이론적인 값과 실제 데이터 값을 비교하여 확인할 수 있다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 가설검정

- 귀무가설의 설정: '없다', '같다'와 같이 기존의 주장
- 대립가설의 설정: '작다', '크다', '같지 않다' 등 귀무가설에 대한 반대의 가설을 표현

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. t-검정(t-test)

두 집단 간의 평균에 대한 검정를 통계적 유의성으로 검정하는 방법

1) 일표본 t-검정(One sample t-test): 표본의 평균이 모집단 평균과 같은지 검정
(Ex. 모집단의 평균 vs 3만 시간(LED조명의 평균수명))

2) 이표본 t-검정(Two sample t-test): 두 표본의 평균이 같은지 검정
(Ex.마포구 이동거리 vs 영등포구 이동거리)

3) 대응표본 t-검정(Paired t-test): 대응하는 두 표본의 평균 차이가 특정 값과 같은지 검정
(Ex. A의 영어성적(Before) vs A의 시험성적(After))

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. Two Sample t-test와 ANOVA

1) 유사성

- 평균의 비교

2) 종류

- Two Sample t-test : 2개 데이터 그룹의 평균을 상호 비교
- ANOVA(분산분석) : 3개 이상의 데이터 그룹의 분산을 비교

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 등분산 분석, 분산분석, 사후분석

1) 등분산 분석

- 여러 그룹의 분산이 같은가를 확인
- 귀무가설 : "모든 그룹의 분산은 같다."

2) 분산분석

- 귀무가설과 대립가설 확인
- 귀무가설 : "모든 그룹의 평균은 같다."
- 대립가설 : "어떤 그룹의 평균은 같지 않다."

3) 사후분석

- 어떤 그룹의 평균이 다른지 확인

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

3. Bartlett's Test, One Way ANOVA, Tukey's HSD Test

1) Bartlett's Test(등분산 분석)

모든 그룹의 분산은 같다는 가정하에 분산분석을 진행할 수 있다.

2) One Way ANOVA(분산분석)

귀무가설 기각의 경우, 모든 그룹의 평균은 같지 않으니, 사후분석을 통해 어떤 그룹이 평균이 같은지 확인할 수 있다.

3) Tukey's HDS Test(사후분석)

- p-adj 값과 reject의 True / False 확인
- 유의수준 0.05 가정 하에 True는 평균이 같지 않은 그룹, False는 평균이 같은 그룹

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 카이제곱분포(Chi-square Distribution)

1) 정규분포의 분산에 대한 확률분포이다.

2. 카이제곱 독립성검정

1) 두 개의 범주형 데이터 사이의 관련성을 확인하는 검정

2) 방법: scipy.stats 패키지의 chi2_contingency 활용

- 귀무가설: "Age_Group은 Membership_type에 독립적이다"
  = "Age_Group은 Membership_type와 연관성이 없다."
- p-value 확인: 유의수준보다 낮은 확률의 값이라면 귀무가설 기각
  "멤버십 가입 형태가 연령대에 독립적이지 않다."

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 산점도(Scatter plot)

1) 데이터 간의 상관관계를 시각적으로 빠르게 확인할 수 있는 방법이다.

2) 해석 시, 유의사항

- 흩어짐 정도가 상관계수의 고저를 나타내며 흩어질수록 상관계수가 낮고, 모여있을수록 상관계수는 높게 나타난다.
- 모양이 우상향이면 양의 관계이고, 우하향이면 음의 관계이다.

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 상관분석

1) 두 수치형 데이터 간의 직선적 관계를 알려준다.

2) 직선적 관계와 양의 관계인지 음의 관계인지를 알려주지만
원인-결과 관계를 얘기하는 것은 아니다.

3) 상관계수가 데이터 안에 존재하는 진짜 상관성과 일치하지 않는 경우가 종종 있다.

4) 귀무가설은 "상관관계가 없다."이다.

3. 상관분석 수행

1) scipy.stats가 제공하는 pearsonr 함수

2) pandas의 corr 함수

Summary

아래 내용은 학술적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 회귀분석

1) 목적

① 데이터의 관계에 대해서 설명한다.

② 데이터 예측에 활용한다.

2) 개념

- 독립변수 x와 종속변수 y의 관계를 설명하는 선형 식을 찾는 것
- y(종속변수) = 기울기 * x(독립변수) + 절편

→ 수치형 데이터 x, y로부터 관계를 분석하여 적절한 기울기와 절편을 찾는 것이 회귀분석의 결과

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 회귀분석 결과 해석

1) R-squared: R제곱, 혹은 결정계수

회귀식이 실제 관찰된 값을 얼마나 설명하는지(=설명력)를 의미한다.

2) Prob(F-statistic): F 검정통계량 추정치의 p-value

- 회귀식이 통계적으로 유의한지 알려준다.
- 회귀분석의 귀무가설: 회귀식이 존재하지 않는다. 또는 회귀식의 기울기가 0이다.
- F에 대한 p-value가 유의수준 0.05보다 작은 값을 갖게 되면, 귀무가설을 기각한다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

2. 회귀분석 결과 해석

3) P > |t| : t값의 p-value

- 독립변수에 대한 기울기 값이다.
- 독립변수가 종속변수의 변화에 영향을 주는지 통계적 유의성을 확인할 수 있는 지표이다.
- 기울기값은 coef라고 되어 있는 계수값으로 알 수 있다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

1. 다중회귀분석

복수의 독립변수를 사용하는 회귀분석이다.

2. 머신러닝

1) 많은 데이터를 사용해서 종속변수를 설명할 수 있는 특징, 패턴, 수식 등을 찾아내도록 하는 것이다.

2) 프로세스

- 1단계 : 회귀분석 모델을 만든다.
- 2단계 : 회귀분석 모델을 활용하여 예측값을 계산한다.
- 3단계 : 예측 결과의 정확도를 계산한다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

• 로지스틱 회귀분석

1) 의미

- 선형관계를 로그 및 역함수 변환을 통하여 분류로 변환한다.

2) 프로세스

- 1단계 : Train 함수를 활용하여 로지스틱 회귀분석 모델을 만든다.
  → p-value값을 확인하여 통계적 유의성을 확인한다.
- 2단계 : Predict 함수를 활용하여 로지스틱 회귀분석 모델의 예측값을 계산한다.
- 3단계 : Evaluate 함수를 활용하여 예측 결과의 정확도를 계산한다.
  → accuracy, precision, recall, f1-score의 지표로 정확도를 확인할 수 있다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

• 의사결정나무

1) 의사결정나무의 출발점을 'Root노드'라 하며 분류 기준이 되는 Feature의 기준 값에 따라 좌우로 분기한다.
2) Stopping Rule에 따라 Split을 멈추었을 때 마지막 단계에 있는 노드를 'Leaf노드'라고 한다.
3) 각 노드의 첫 줄에는 split의 기준이 되는 split feature와 split value가 제시되고 해당 노드의 class 값이 표시된다.
4) 예측 시에는 Root노드부터 시작해서 새로운 데이터를 한 행씩 Split Feature의 기준값과 비교하고 좌우로 분기해 가면서 내려가다 보면 도달하는 Leaf노드의 class가 해당 데이터의 예측 Class가 된다.

Summary

아래 내용은 학문적 개념이 아닌 본 과정의 이해를 돕기 위해 강의내용을 요약한 것입니다.

- **K-Means 클러스터링**
  - 데이터에 종속변수가 없는 비지도학습으로 유사한 것끼리 군집을 형성하는 기법이다.
  - 유사성 측정의 기준이 되는 Feature를 정하는 것과 몇 개의 그룹으로 나눌 것인지 K값을 정해주는 것이 중요하다.
  - 클러스터 결과가 항상 만족스러운 것은 아닐 수 있으며 K값을 변경하며 분석 목적에 맞는 결과를 도출한다.