

라이브 코칭 4회차

Data Science 2024

잠시 후 오후 8시, 코칭스터디(Data Science 2024) LIVE가 진행됩니다

오늘의 순서

#4회차 라이브
K-Beauty 온라인 판매 분석

코칭스터디 공지사항

라이브 코칭

QnA

- ✓ 4주차 학습계획/미션
- ✓ 수료안내

✓ 라이브 코칭

- ✓ 라이브 코치에게 물어봐
- ✓ 실시간 QnA

공지사항

4주차 학습계획/미션

부스트코스 - [스터디강좌]에서 확인하기

필독 | 오리엔테이션

FAQ | 자주묻는 질문

학습 :: 미션 :: 라이브 안내

4주차 | K-beauty 온라인 판매분석

- 4주차 | 학습 계획 & 학습 범위 
- 4주차 | 미션 
- 4주차 | 라이브 코칭 :: 실시간 시청
- 4주차 | 라이브 코칭 :: VOD 다시보기

♥ ♥ 4주차 학습 계획

1) 파이썬으로 시작하는 데이터 사이언스 강좌 수강하기(아래의)

- (1) K-beauty 온라인 판매분석 강좌 수강하기
- (2) QUIZ 4 풀기
 QUIZ 4 : <https://www.boostcourse.org/ds112/quiz/60790>

2) 퀴즈 인증 제출하기(08월 25일 일요일 23:59까지!)

- 위의 QUIZ 4 풀이 후 화면 캡쳐해서 슬랙에 업로드하기
- 슬랙(코칭스터디 <Data Science>) → 본인 팀 채널에 업로드 (00코치_01~10팀)

3) 라이브 코치님께 질문 남기기(08월 22일 목요일 23:59까지)

- 라이브 코치님께 궁금하신 사항이 있으신 분들은 자유롭게 남겨주세요!
- 슬랙(코칭스터디 <Data Science>) → 03-코치에게-들어봐 채널에 남겨주세요!

  4주차 미션 내용을 알려드립니다  

부스터 여러분들, 4주차 강의는 잘 들으셨나요?!

학습한 내용을 토대로 풀이하여야 할 4주차 미션 내용을 아래와 같이 공개합니다!

미션 내용을 팀원들과 함께 풀이해주세요!(적극적인 토론이 필요합니다!)

* 매주 일요일 23:59까지 리드부스터가 제출해주세요! 모두들 화이팅입니다!

◆ 미션에 도전하기 전에 먼저!!

이번 미션에 활용되는 데이터를 다운로드 받기 위해, 주피터 노트북에서 다음 셀을 먼저 실행해주세요.

나의 컴퓨터 환경에 데이터를 저장하지 않아도, 웹 사이트에서 바로 데이터를 받아올 수 있습니다.

- 원활한 피드백을 위해 미션을 제출할 때에도 아래 코드를 꼭 포함해서 제출해주세요!
- [참고] 한글폰트 설정 : <https://github.com/ychol-kr/koreanize-matplotlib>

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# 한글폰트 사용을 위해 설치
# 아래 모듈을 설치하고 불러오면 별도의 한글폰트 설정이 필요 없습니다.
# !pip install koreanize-matplotlib
```

```
import koreanize_matplotlib

df = pd.read_csv('...')
```

리드부스터 활동일지

8월 25일 일요일
23시 59분까지

💡 슬랙에서 과업을 인증하면 해당 주차에 체크 ✅ 해주세요.

⚠ 작성하는 공간이 아닙니다

퀴즈	1주차	2주차	3주차	4주차	수료 기준: 75% 이상 0%
홍길동_리더	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
미션	1주차	2주차	3주차	4주차	수료 기준: 75% 이상 0%
홍길동_리더	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
라이브	1주차	2주차	3주차	4주차	수료 기준: 75% 이상 0%
홍길동_리더	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

우수 참가자 리워드

#열정적인 여러분을 위한 깜짝 이벤트



8월 30일 금요일

#합동왕(님)
우수마련상(M)
최다 신청상(상)

발표

#스통원(대인) 3명

자유게시판, QnA 등
동료부스터 질문에 적극적으로 답변
좋은 학습 자료를 아낌없이 공유
슬랙에서 열심히 소통



열심히 달려갈 여러분을 위해
수료 리워드까지
준비했으니, 끝까지 함께
달려보아요 🍀

수료 리워드 받기

배송지 미리 입력하고 수료 후 리워드 받아요 ❤️✉️

#09/01(일) 자정까지!



<https://forms.gle/U4zXc72nmcdXaGeF8>

코칭스터디 수료 관련 안내

리드부스터 여러분!

마지막 4주차 활동일지

08월 25일 (일) 23:59 전까지

작성 완료해주세요!

4주차 활동일지 기준으로 수료 여부를 확인합니다.
제출 전에 팀원과 최종 활동일지를 공유해주세요.

수료 안내 콘텐츠 이용 기간

한 달 동안만 볼 수 있는 콘텐츠

미션/라이브/주차별 우수미션

10월 30일까지 공개!

평생 볼 수 있는 콘텐츠

강의/퀴즈

코칭스터디 강좌 내 저작권이 있는 미션, 퀴즈, 라이브 VOD, OT자료, 주차별 우수미션은 공식적인 코칭스터디 운영 기간 종료 후 **10월 30일까지** 공개되어 있습니다!
이후에는 확인 불가합니다.

※수료증 발급을 위하여 강좌 자체의 접근은 가능합니다

※프로그램 종료 후 공개 기간 동안 과업을 복습할 수 있어요.

※어려워 도전하지 못한 과업, 레벨 업을 위해 다시 도전해보는 것을 적극 추천드려요.

수료 안내 콘텐츠 이용 기간

#04_주차별 우수미션

**우수미션 저작권 O
10월 30일까지 공개!**

우수미션은 저작권이 있기 때문에 공식적인 코칭스터디 운영 기간이 종료 후 10월 30일까지 공개되어 있습니다!
(10월 30일 이후 우수미션 관련 채널 비공개 예정)

※ 이후에도 워크스페이스 접근은 가능하며, 자료 공유 & 소통 등 다양한 활동 가능합니다.
(슬랙 내 게시글은 작성일로 부터 90일 이후 게시글이 순차적으로 사라집니다.)



코칭스터디 운영진 아웅 오후 12:09

[전체공지] @channel

[1주차 우수미션 공개]

안내사항

- 공유드린 자료는 동료 부스터들이 열심히 작성한 내용입니다.
- 무단 배포와 수정을 금지합니다.
- 저작권 문제가 있으니 주의 부탁드립니다.
- 우수미션은 학습 목적으로만 활용해주세요.
- 코칭스터디 종료 한 달 후까지 확인 가능하니, 미리 확인 후 학습해주세요!
- 우수미션 링크 : <https://docs.google.com/spreadsheets/d/1ZUcvt9pOj3JJ91lZC4lrUoL-kGLY5wJoY2Jhnp5oud0/edit?usp=sharing>

코칭스터디 수료증

발급 안내

4주간 여정 고생하셨습니다.

#수료증 꼭 발급 받아가세요:)



#부스터

코칭스터디 수료증



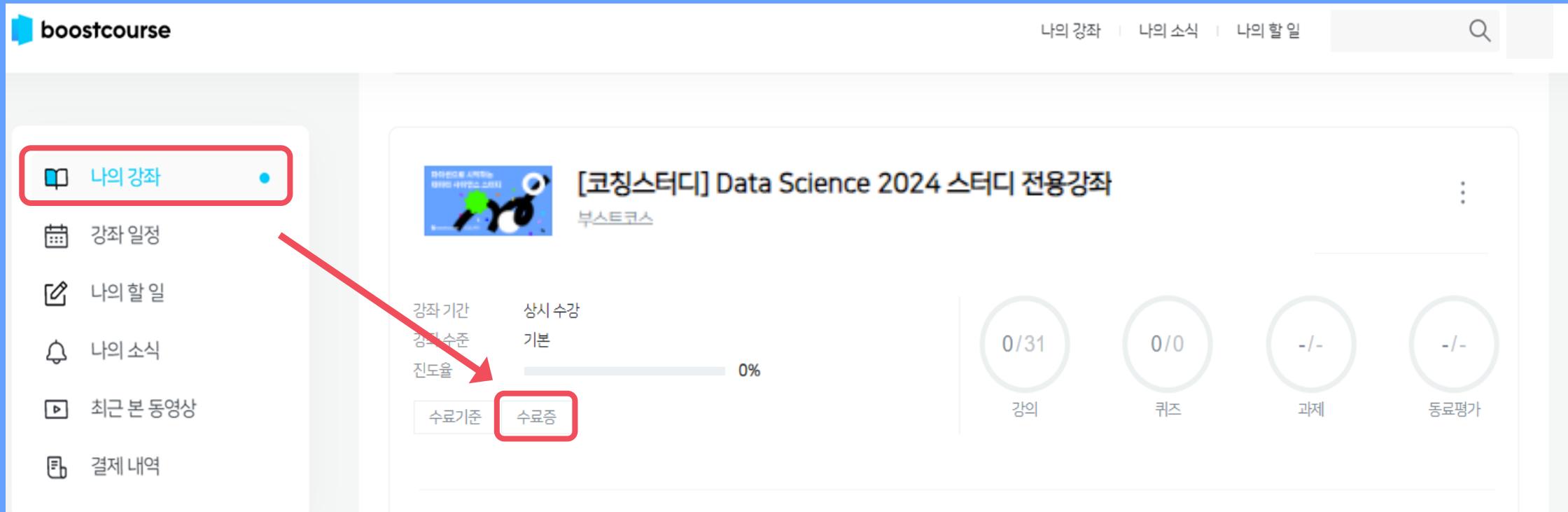
#리드부스터

코칭스터디 수료증

리드부스터 수료증

수료증 발급 스터디강좌

#8월 30일 금요일 19시 이후



The screenshot shows the boostcourse platform interface. On the left, a sidebar menu is visible with the following items:

- 나의 강좌** (My Courses) - highlighted with a red box and a red arrow pointing from the top-left.
- 강좌 일정
- 나의 할 일
- 나의 소식
- 최근 본 동영상
- 결제 내역

The main content area displays a course titled "[코칭스터디] Data Science 2024 스터디 전용강좌" (hosted by 부스트코스). The course details include:

- 강좌 기간: 상시 수강
- 강좌 수준: 기본
- 진도율: 0%

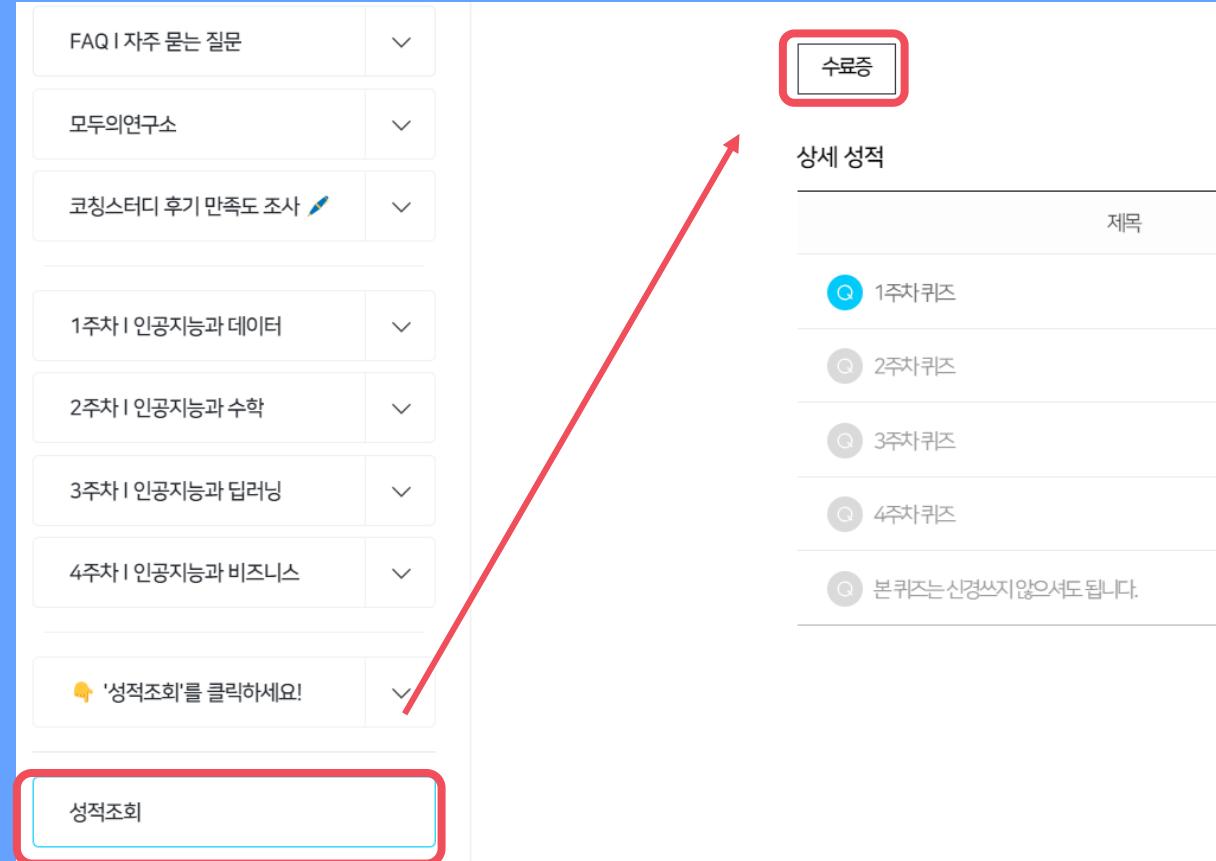
Below these details are two buttons: "수료기준" and "수료증". The "수료증" button is highlighted with a red box and a red arrow pointing from the bottom-left.

On the right side, there are four circular progress indicators:

- 강의: 0/31
- 퀴즈: 0/0
- 과제: -/-
- 동료평가: -/-

수료증 발급 스터디강좌

#8월 30일 금요일 19시 이후



FAQ | 자주 묻는 질문

모두의연구소

코칭스터디 후기 만족도 조사

1주차 | 인공지능과 데이터

2주차 | 인공지능과 수학

3주차 | 인공지능과 딥러닝

4주차 | 인공지능과 비즈니스

👉 '성적조회'를 클릭하세요!

성적조회

수료증

상세 성적

제목

- Q 1주차 퀴즈
- Q 2주차 퀴즈
- Q 3주차 퀴즈
- Q 4주차 퀴즈

본 퀴즈는 신경쓰지 않으셔도 됩니다.

리드부스터 수료증 발급 리드부스터강좌

#8월 30일 금요일 19시 이후

나의 강좌

+ 강좌 참여하기

참여중인 강좌 종료된 강좌

수강순 | 가나다순

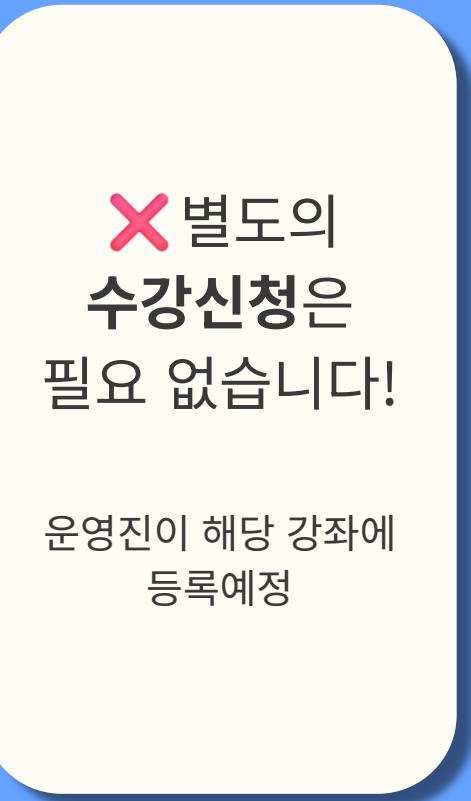
[코칭스터디] Data Science 2024 - 리드부스터용

부스트코스 | 부스트코스

강좌 기간: 상시 수강
강좌 수준: 기본
진도율: 0%

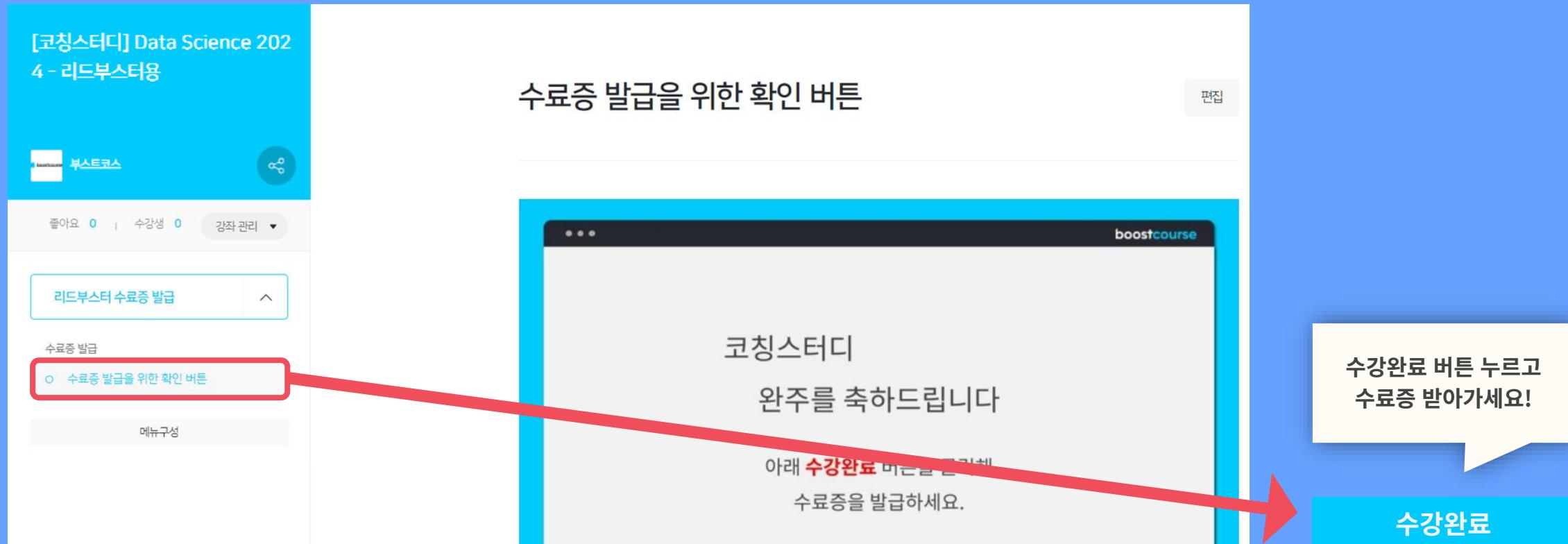
강의: 0/31
퀴즈: 0/0
과제: -/-
동료평가: -/-

수료기준 수료증



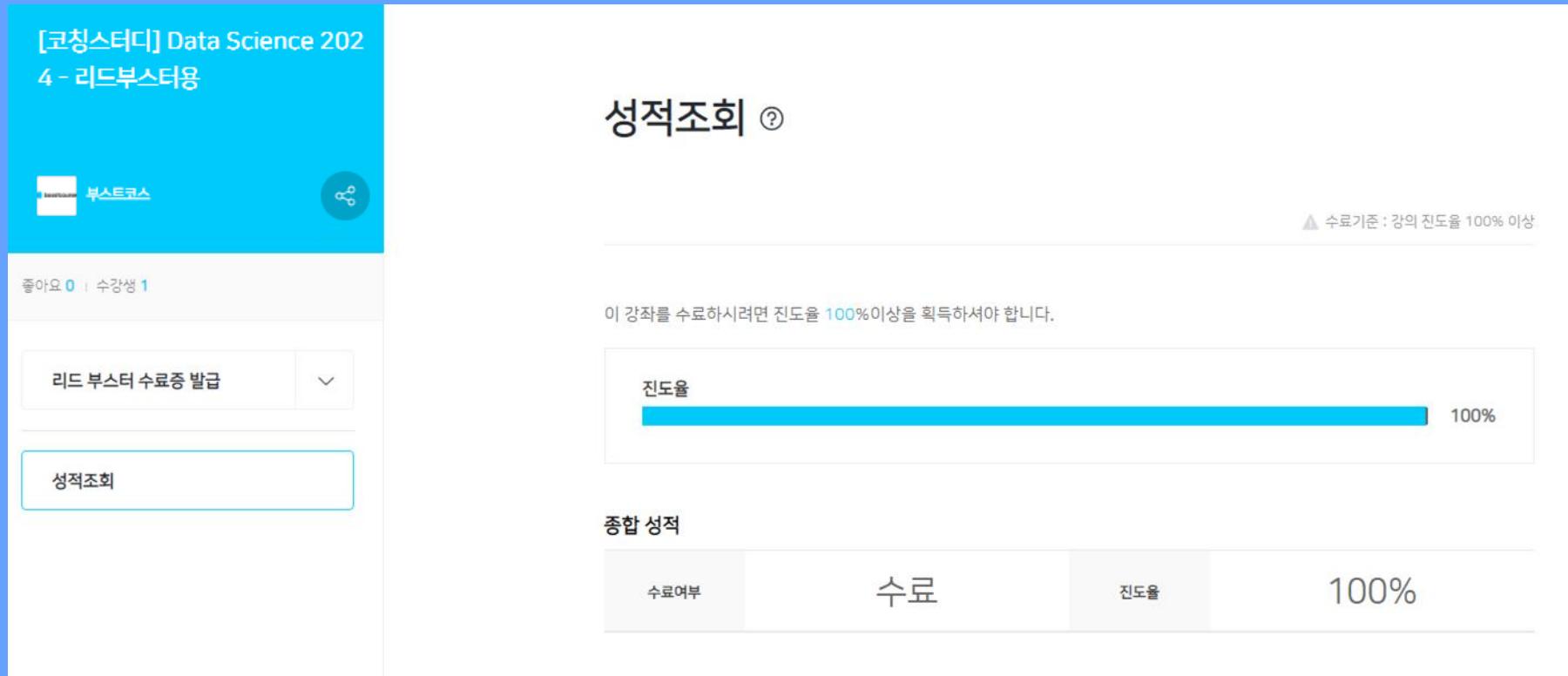
리드부스터 수료증 발급 리드부스터강좌

#8월 30일 금요일 19시 이후



리드부스터 수료증 발급 리드부스터강좌

#8월 30일 금요일 19시 이후



[코칭스터디] Data Science 202
4 - 리드부스터용

부스트코스

좋아요 0 | 수강생 1

리드 부스터 수료증 발급

성적조회

성적조회 ②

수료기준 : 강의 진도를 100% 이상

이 강좌를 수료하시려면 진도를 100% 이상을 획득하셔야 합니다.

진도율

100%

종합 성적

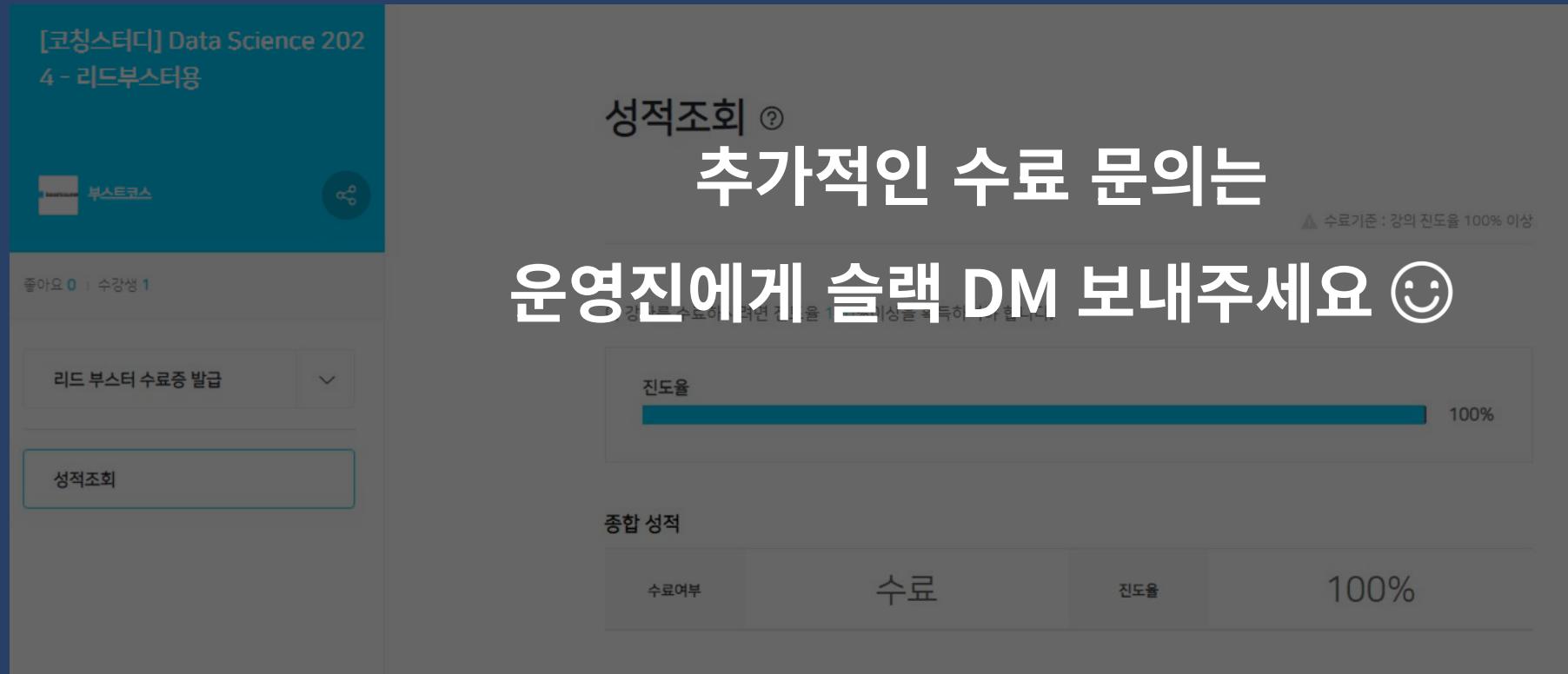
수료여부	수료	진도율	100%
------	----	-----	------

수료증 버튼 누르고
수료증 받아가세요!

수료증

리드부스터 수료증 발급 리드부스터강좌

#8월 30일 금요일 19시 이후



만족도 조사 시행 안내

*8/23(금)부터~9/2(월)까지

이번 주 금요일부터!

만족도 조사 스터디강좌

#8/23(금)부터~9/2(월)까지

코칭스터디 공지사항

⭐필독|코칭스터디 오리엔테이션

FAQ | 자주 묻는 질문

코칭스터디 후기 만족도 조사 

모두의연구소

코칭스터디 후기 만족도 조사 

기본 만족도 조사

[코칭스터디] 2024년 8월
디 전용강좌

└ 기본 만족도 조사

답자 정보 조사 

└ 커뮤니케이션 만족도 조사

세부 만족도 조사

[코칭스터디] [Let's AI 2024]에 적극 참여해 주셔서 진심으로 감사드립니다! ❤️

└ 미션 만족도 조사

동안 포기하지 않고 끝까지 원료해주신 부스터분들을 위하여
코칭스터디 리워드로 모두의연구소 풀잎스쿨 쿠폰을 발행해드립니다!

└ 코딩코치 만족도 조사

부스터님들에게 보다 나은 교육 환경을 제공하기 위해
코칭스터디 [Let's AI 2024]를 진행하신 여러분들의 경험을 듣고자 합니다.

└ 라이브코치 & 라이브 강의 만족도 조사

직접으로 의견 남겨주시면 감사하겠습니다! 😊

└ 플랫폼 만족도 조사

제출 횟수 : 0 / 1

저

여러분 9/6(금) 저녁에 바쁘세요?
모두연 강남캠퍼스에서 아웅다웅이랑 놀아요 😊

3주차 미션 정리

미션 간단 풀이

3주차 미션 솔루션 공유

3주차 미션 출제 의도

1번 연령대별 허리둘레에 대한 기술통계를 구하기

2번 "음주여부", "흡연상태", "연령대코드(5세단위)", "성별코드" 상관계수를 구와 시각화 해보기

3번 흡연하는 사람과 음주하는 사람들의 수의 차이 알기

4번 체중이 120Kg 이상인 사람의 "총콜레스테롤", "감마지티피" 값을 음주여부에 따라 산점도로 시각화 해보기

5번 연령대별로 시력은 얼마나 차이날지 알아보기

3주차 미션 - 1번

기술 통계



Q1.

연령대별 허리둘레에 대한 기술통계를 구하려고 합니다.

다음 제공되는 딕셔너리를 통해 연령대코드(5세단위)를 "연령대"로 만들고 아래와 같은 기술통계값을 구해주세요!

연령대	count	mean	std	min	25%	50%	75%	max
20~24세	23244	75.1522	12.2518	47.5	67.5	73.4	81	999
25~29세	64898	77.7048	16.7357	48	69	76.5	84.2	999
30~34세	77517	81.0893	22.9881	49	72	80.1	88	999
35~39세	84621	82.094	14.5221	9.2	75	82	89	999
40~44세	130912	80.4883	10.8031	42.1	73	80	87	999
45~49세	118357	80.8224	9.52162	40	74	81	87	137
50~54세	129833	81.0628	9.09544	6.5	75	81	87	142
55~59세	112175	81.7999	8.7304	32	76	82	87.5	139
60~64세	106491	82.7228	8.59618	0	77	83	88	137
65~69세	53624	83.5885	8.44354	50	78	83.5	89	129
70~74세	51586	84.0634	8.53964	51	78	84	90	129.8
75~79세	25972	84.2001	8.77231	50	78	84	90	122
80~84세	16205	83.7514	9.04109	38	78	84	90	120
85세+	4125	81.7367	17.326	34	75	81.5	88	999

3주차 미션 - 1번

기술 통계

```
age_code = {1: '0~4세',
 2: '5~9세',
 3: '10~14세',
 4: '15~19세',
 5: '20~24세',
 6: '25~29세',
 7: '30~34세',
 8: '35~39세',
 9: '40~44세',
 10: '45~49세',
 11: '50~54세',
 12: '55~59세',
 13: '60~64세',
 14: '65~69세',
 15: '70~74세',
 16: '75~79세',
 17: '80~84세',
 18: '85세+'}

df["연령대"] = df["연령대코드(5세단위)"].replace(age_code)
df.groupby("연령대")["허리둘레"].describe()
```

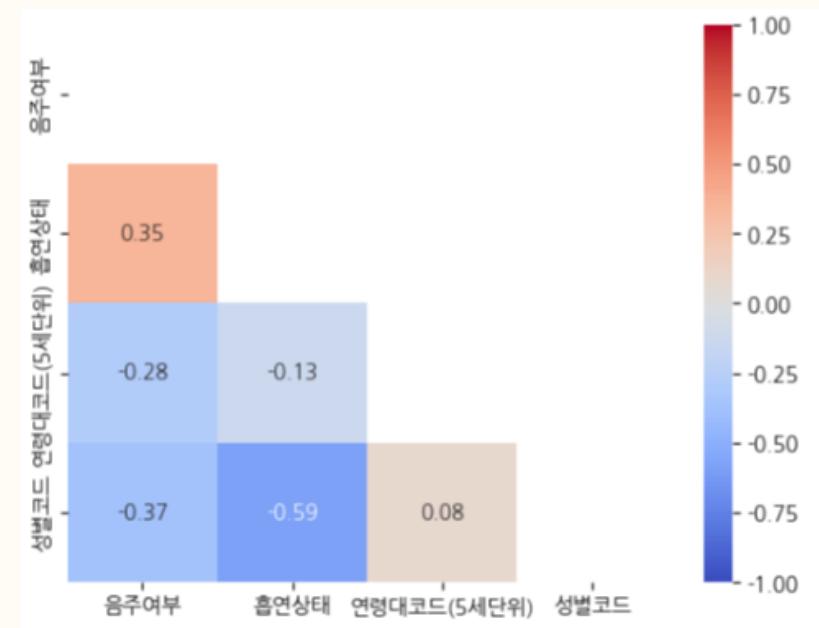
	count	mean	std	min	25%	50%	75%	max
연령대								
20~24세	23244.0	75.152220	12.251781	47.5	67.5	73.4	81.0	999.0
25~29세	64898.0	77.704783	16.735734	48.0	69.0	76.5	84.2	999.0
30~34세	77517.0	81.089268	22.988111	49.0	72.0	80.1	88.0	999.0
35~39세	84621.0	82.094012	14.522095	9.2	75.0	82.0	89.0	999.0
40~44세	130912.0	80.488308	10.803098	42.1	73.0	80.0	87.0	999.0
45~49세	118357.0	80.822449	9.521622	40.0	74.0	81.0	87.0	137.0
50~54세	129833.0	81.062754	9.095438	6.5	75.0	81.0	87.0	142.0
55~59세	112175.0	81.799905	8.730398	32.0	76.0	82.0	87.5	139.0
60~64세	106491.0	82.722769	8.596176	0.0	77.0	83.0	88.0	137.0
65~69세	53624.0	83.588500	8.443542	50.0	78.0	83.5	89.0	129.0
70~74세	51586.0	84.063372	8.539639	51.0	78.0	84.0	90.0	129.8
75~79세	25972.0	84.200127	8.772306	50.0	78.0	84.0	90.0	122.0
80~84세	16205.0	83.751435	9.041091	38.0	78.0	84.0	90.0	120.0
85세+	4125.0	81.736703	17.325969	34.0	75.0	81.5	88.0	999.0

3주차 미션 - 2번

상관계수와 시각화



Q2.
"음주여부", "흡연상태", "연령대코드(5세단위)", "성별코드"에 대한
상관계수를 구하고 시각화 해주세요.



3주차 미션 - 2번

상관계수와 시각화

```
corr = df[["음주여부", "흡연상태", "연령대코드(5세단위)", "성별코드"]].corr()  
corr
```

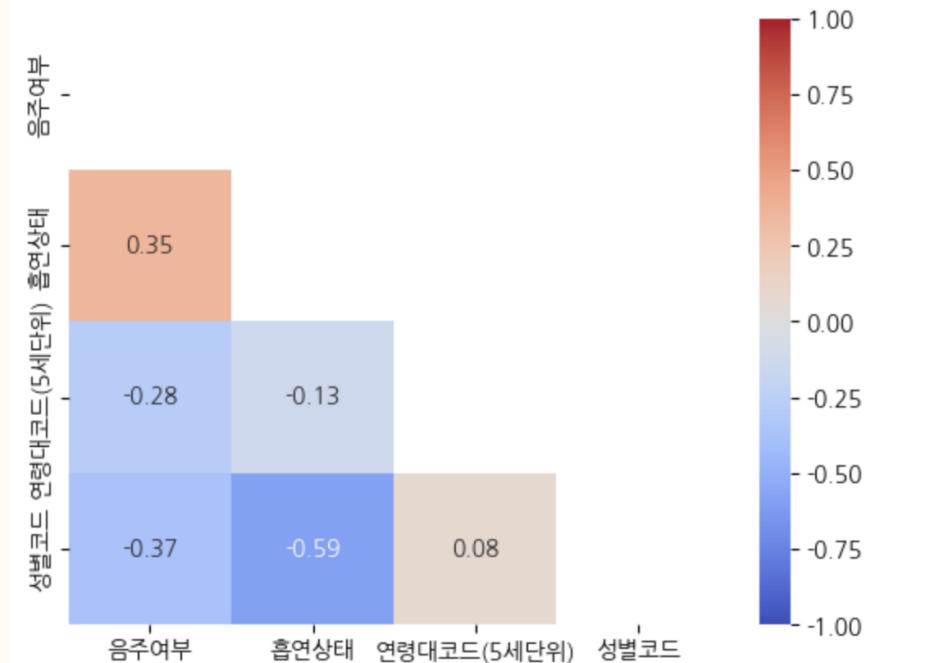
	음주여부	흡연상태	연령대코드(5세단위)	성별코드
음주여부	1.000000	0.352014	-0.283296	-0.368630
흡연상태	0.352014	1.000000	-0.125714	-0.588491
연령대코드(5세단위)	-0.283296	-0.125714	1.000000	0.080093
성별코드	-0.368630	-0.588491	0.080093	1.000000

```
mask = np.triu(np.ones_like(corr))  
mask
```

```
array([[1., 1., 1., 1.],  
       [0., 1., 1., 1.],  
       [0., 0., 1., 1.],  
       [0., 0., 0., 1.]])
```

```
sns.heatmap(corr, annot=True, cmap="coolwarm", vmax=1, vmin=-1, mask=mask)
```

<Axes: >



3주차 미션 - 3번

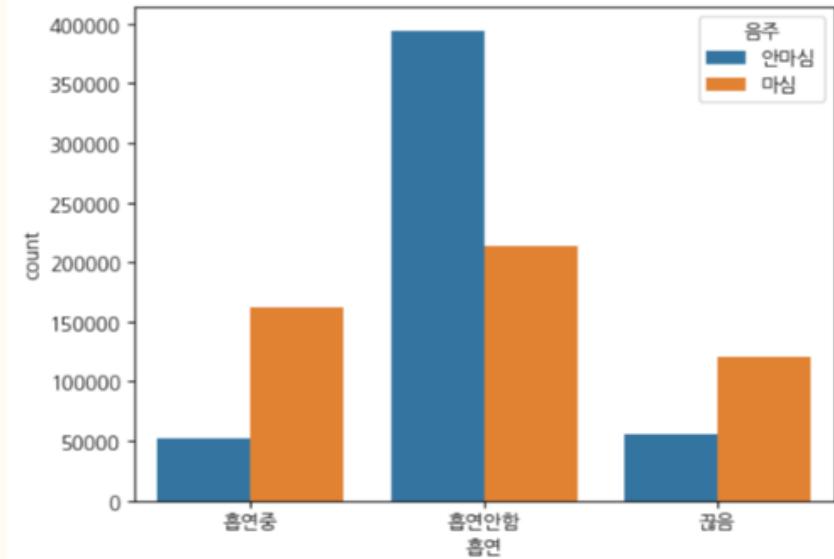
데이터 비교



Q3.
흡연하는 사람과 음주하는 사람들의 수는 얼마나 차이가 있을까요?

결과 예시

음주	끊음	흡연안함	흡연중
마심	120779	213743	162166
안마심	55334	394503	52845



3주차 미션 - 3번

데이터 비교

```
# 흡연 1(피우지 않는다), 2(이전에 피웠으나 끊었다), 3(현재도 피우고 있다)
# 음주 0(마시지 않는다), 1(마신다)
```

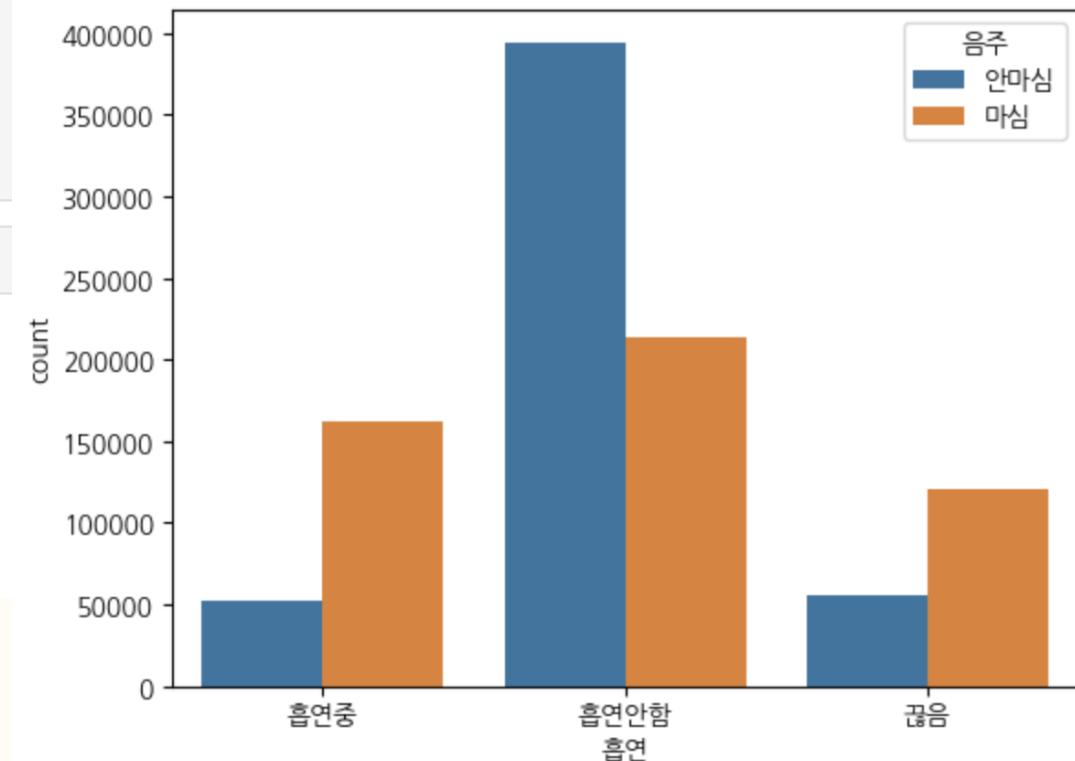
```
smoke = {1 : "흡연안함", 2: "끊음", 3: "흡연중"}
drink = {0: "안마심", 1: "마심"}
df[["흡연"]] = df[["흡연상태"]].replace(smoke)
df[["음주"]] = df[["음주여부"]].replace(drink)
```

```
pd.crosstab(df[["음주"]], df[["흡연"]])
```

	흡연	끊음	흡연안함	흡연중
음주				
마심	120779	213743	162166	
안마심	55334	394503	52845	

```
sns.countplot(data=df, x="흡연", hue="음주")
```

```
<AxesSubplot:xlabel='흡연', ylabel='count'>
```

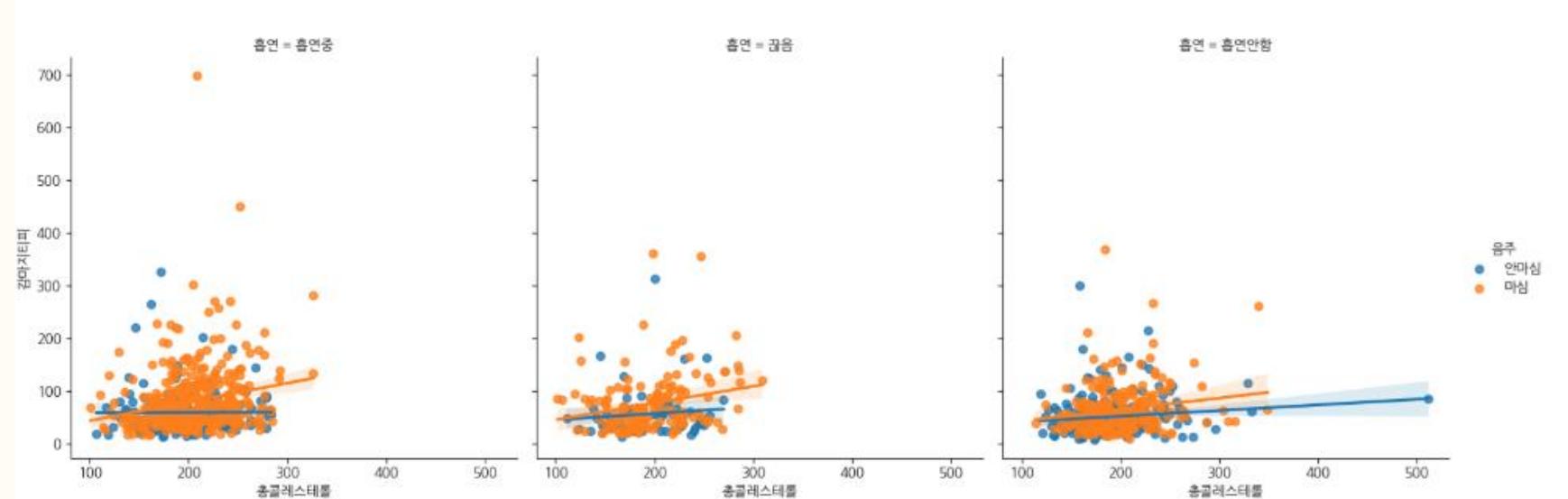


3주차 미션 - 4번

조건 있는 산점도 그리기

📌 Q4.

체중이 120Kg 이상인 데이터를 찾아 "총콜레스테롤", "감마지티피" 값을 음주여부에 따라 산점도로 시각화해주세요!



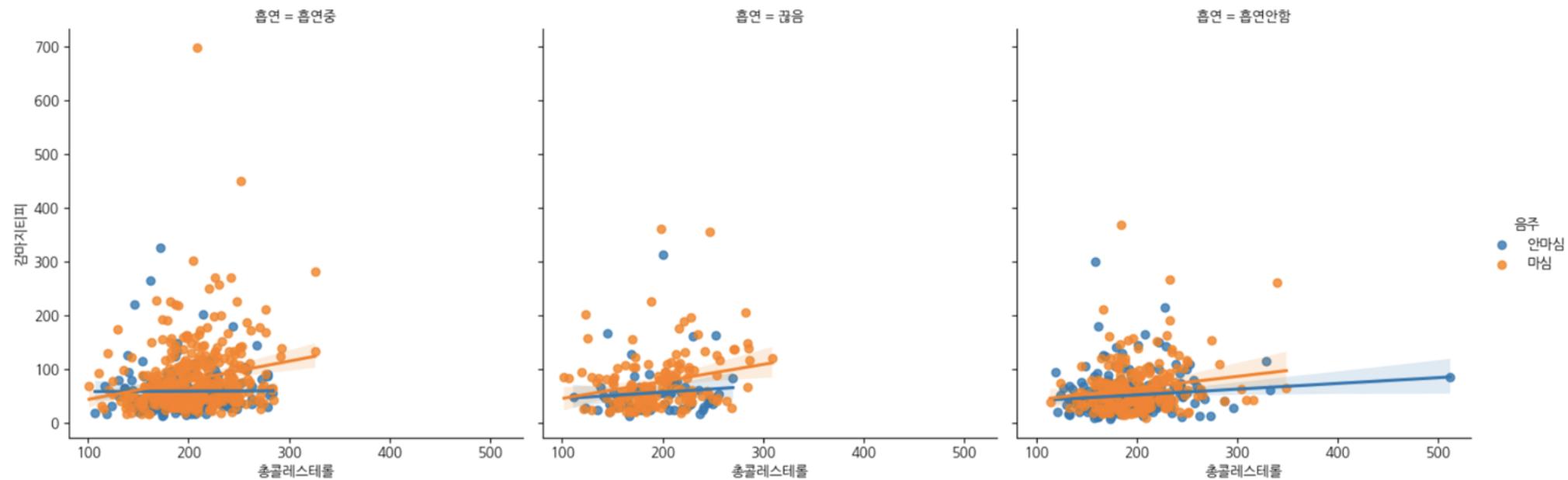
3주차 미션 - 4번

조건 있는 산점도 그리기

```
df_w150 = df[df["체중(5Kg 단위)"] >= 120]
```

```
sns.lmplot(data=df_w150, x="총콜레스테롤", y="감마지티피", hue="음주", col="흡연")
```

```
<seaborn.axisgrid.FacetGrid at 0x7ff0f2994d10>
```

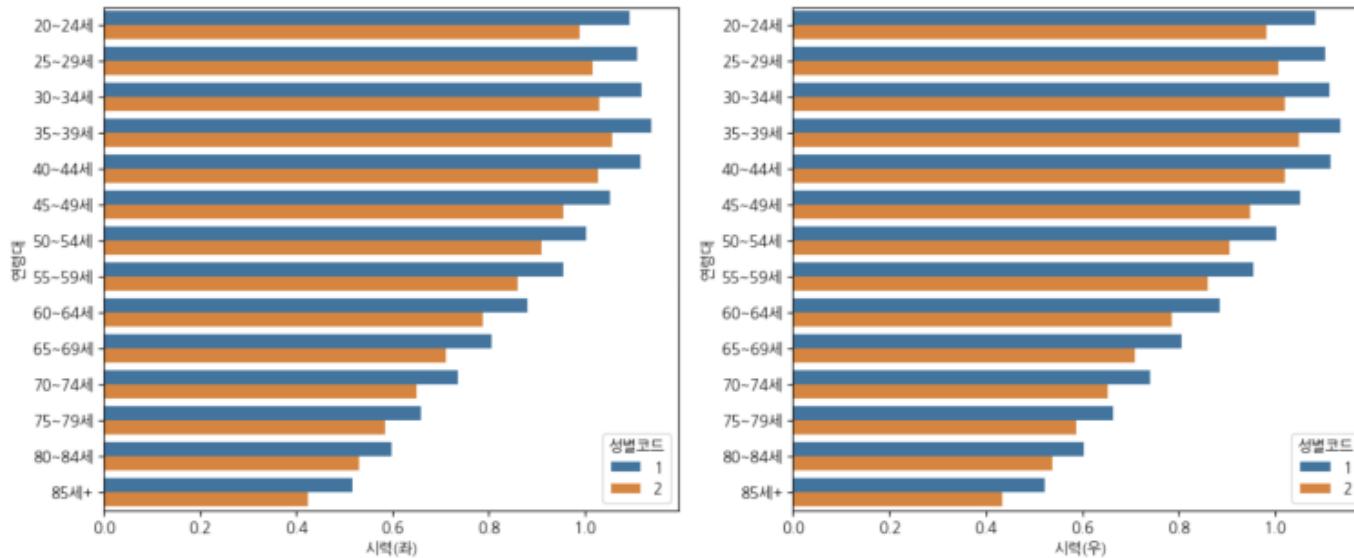


3주차 미션 - 5번

차이 시각화

📌 Q5.

연령대별로 시력은 얼마나 차이가 날까요? 연령대, 성별 좌우 평균 시력을 시각화 해주세요!



3주차 미션 - 5번

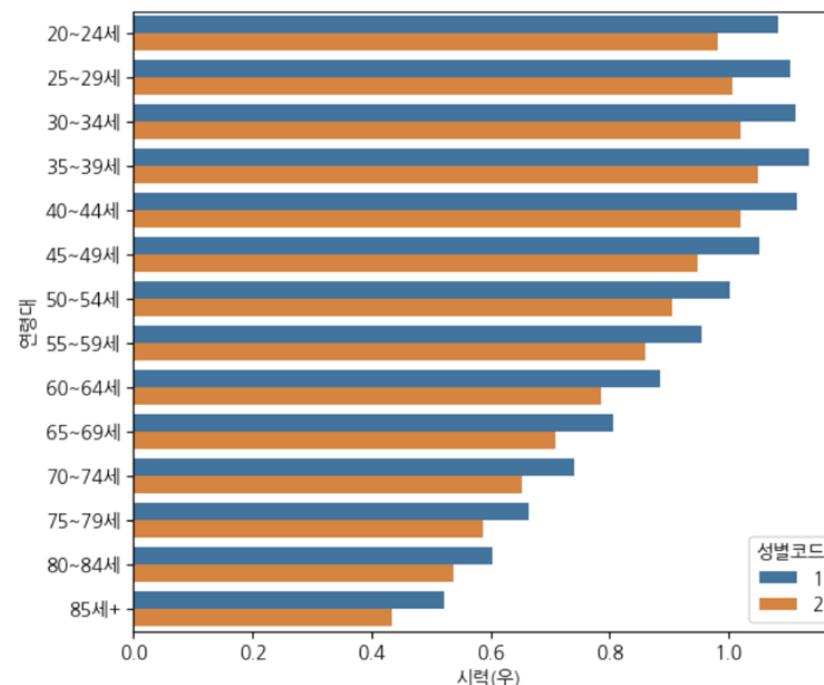
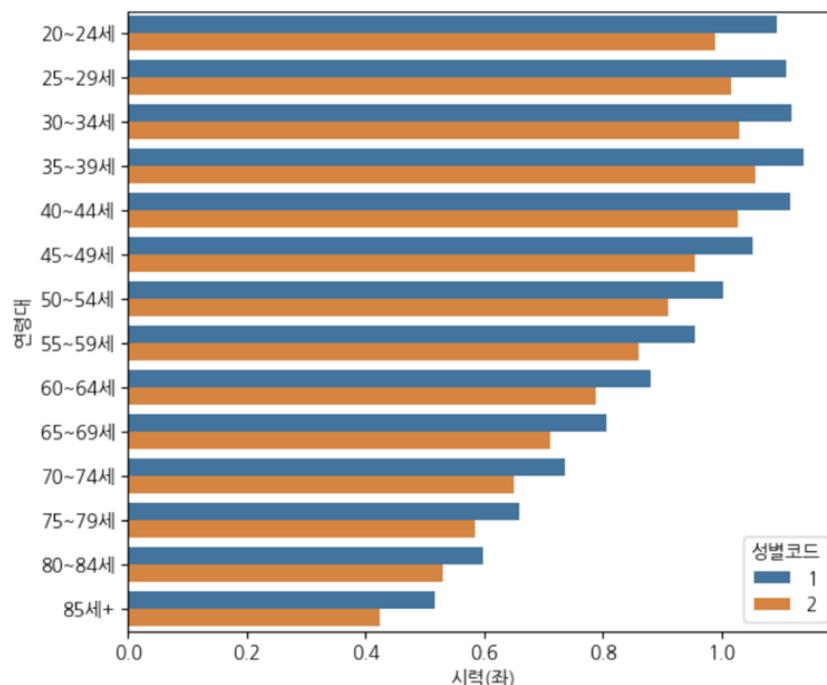
차이 시각화

```
df_eye = df[(df["시력(좌)"] != 9.9) & (df["시력(우)"] != 9.9)]  
df_eye.shape
```

(994187, 39)

```
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(15, 6))  
sns.barplot(data=df_eye.sort_values("연령대"), x="시력(좌)", y="연령대", errorbar=None, hue="성별코드", ax=axes[0])  
sns.barplot(data=df_eye.sort_values("연령대"), x="시력(우)", y="연령대", errorbar=None, hue="성별코드", ax=axes[1])
```

<AxesSubplot:xlabel='시력(우)', ylabel='연령대'>



4주차 학습 내용

어떤 내용을 배우게 될까요?

4주차 학습 내용 정리

도움말 보기 Shift + TAB

```
In [1]: from pandas import Series, DataFrame  
       import pandas as pd  
  
In [ ]: example_obj = Series()  
  
Init signature: Series(data=None, index=None, dtype=None, name=None, copy=False, fastpath=False)  
Docstring:  
One-dimensional ndarray with axis labels (including time series).  
  
Shift + TAB
```

?

?로 도움말을 볼 수도 있습니다.
??로 소스코드를 볼 수도 있습니다.
예시) pd.DataFrame?

4주차 학습 내용 정리

Pandas Cheat Sheet

Data Wrangling
with pandas Cheat Sheet
<http://pandas.pydata.org>

Creating DataFrames

```
df = pd.DataFrame({
    "a": [4, 5, 6],
    "b": [7, 8, 9],
    "c": [10, 11, 12],
}, index=[1, 2, 3])
Specify values for each column.
```

```
df = pd.DataFrame([
    [4, 7, 10],
    [5, 8, 11],
    [6, 9, 12]],
    index=[1, 2, 3],
    columns=['a', 'b', 'c'])
Specify values for each row.
```

```
df = pd.DataFrame(
    {"a": [(4, 5, 6),
           (7, 8, 9),
           (10, 11, 12)],
     "index": [1, 2, 3],
     "columns":['a', 'b', 'c']})
Create DataFrame with a MultiIndex
```

Method Chaining

Most pandas methods return a DataFrame so that another pandas method can be applied to the result. This improves readability of code.

```
df = (pd.melt(df)
      .rename(columns={
          'variable': 'var',
          'value': 'val'})
      .query('val > 200')
)
```

Tidy Data – A foundation for wrangling in pandas

In a tidy data set: & Each variable is saved in its own column. Each observation is saved in its own row. $M * A$

Reshaping Data – Change layout, sorting, reindexing, renaming

Subset Observations - rows

Subset Variables - columns

Subsets - rows and columns

Using query

Logic in Python (and pandas)

<	Less than	I=	Not equal to
>	Greater than	df.column.isin(values)	Group membership
==	Equals	pd.isnull(obj)	Is NaN
<=	Less than or equals	pd.notnull(obj)	Is not NaN
>=	Greater than or equals	~df.any(), df.all()	Logical and, or, not, xor, any, all

regex (Regular Expressions) Examples

'\.'	Matches strings containing a period ''
'Length\$'	Matches strings ending with word 'Length'
'^Sepal'	Matches strings beginning with the word 'Sepal'
'^([A-Z]\$ [^,]*[,_][^,]*\$)'	Matches strings beginning with 'x' and ending with 1,2,3,4,5
'^(?i:Species)\$.*'	Matches strings except the string 'Species'

Tidy Data – A foundation for wrangling in pandas

Tidy data complements pandas's **vectorized operations**. pandas will automatically preserve observations as you manipulate variables. No other format works as intuitively with pandas.

Summarize Data

Handling Missing Data

Make New Columns

Group Data

Windows

Plotting

Combine Data Sets

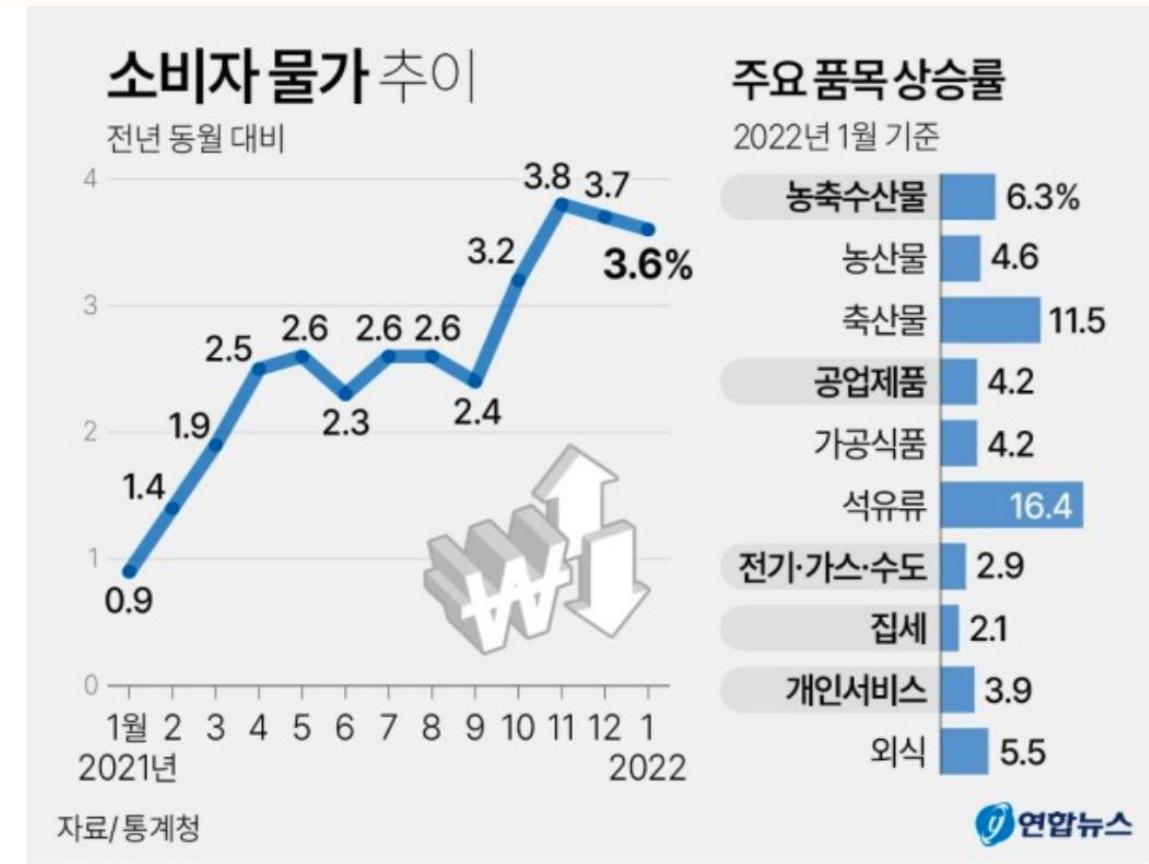
Standard Joins

Filtrering Joins

Set-like Operations

4주차 학습 내용 정리

통계청 데이터 활용 예시



김영은 기자 20220204

4주차 학습 내용 정리

Tidy Data

Hadley Wickham

From Wikipedia, the free encyclopedia

Hadley Alexander Wickham (born 14 October 1979) is a statistician from New Zealand and Chief Scientist at RStudio Inc.^{[2][4][5][6]} and an adjunct Professor of statistics at the University of Auckland,^[7] Stanford University,^[8] and Rice University.^[9] He is best known for his development of open-source software for the R statistical programming language for data visualisation, including `ggplot2`,^[1] and other tidyverse packages, which support a `tidy data` approach to data science.^{[10][11][12]}

Contents [hide]



Hadley Wickham은 "**Tidy Data**" 를
각 변수가 열이고 각 관측치가 행이 되도록
배열된 데이터로 정의했습니다.

4주차 학습 내용 정리

Tidy Data



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II.

<http://www.jstatsoft.org/>

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

**Hadley Wickham은
“Tidy Data(깔끔한 데이터)”를
통계 소프트웨어 저널에서 소개했습니다.**

4주차 학습 내용 정리

Tidy Data

religion	income	freq
Agnostic	<\$10k	27
Agnostic	\$10-20k	34
Agnostic	\$20-30k	60
Agnostic	\$30-40k	81
Agnostic	\$40-50k	76
Agnostic	\$50-75k	137
Agnostic	\$75-100k	122
Agnostic	\$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The `column` has been renamed to `income`, and `value` to `freq`.

**Tidy Data(깔끔한 데이터)로 만들기 위해서
각 변수가 열이고 각 관측치가 행이 되도록
배열하면 다음의 모습이 됩니다.**

* 열Column에 있던 값들이 행Row으로 녹아내렸어요.

4주차 학습 내용 정리

Pandas melt

Pandas는 Tidy Data(깔끔한 데이터)를
만들기 위해 melt라는 기능을 제공합니다.

Melt

df3

	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150

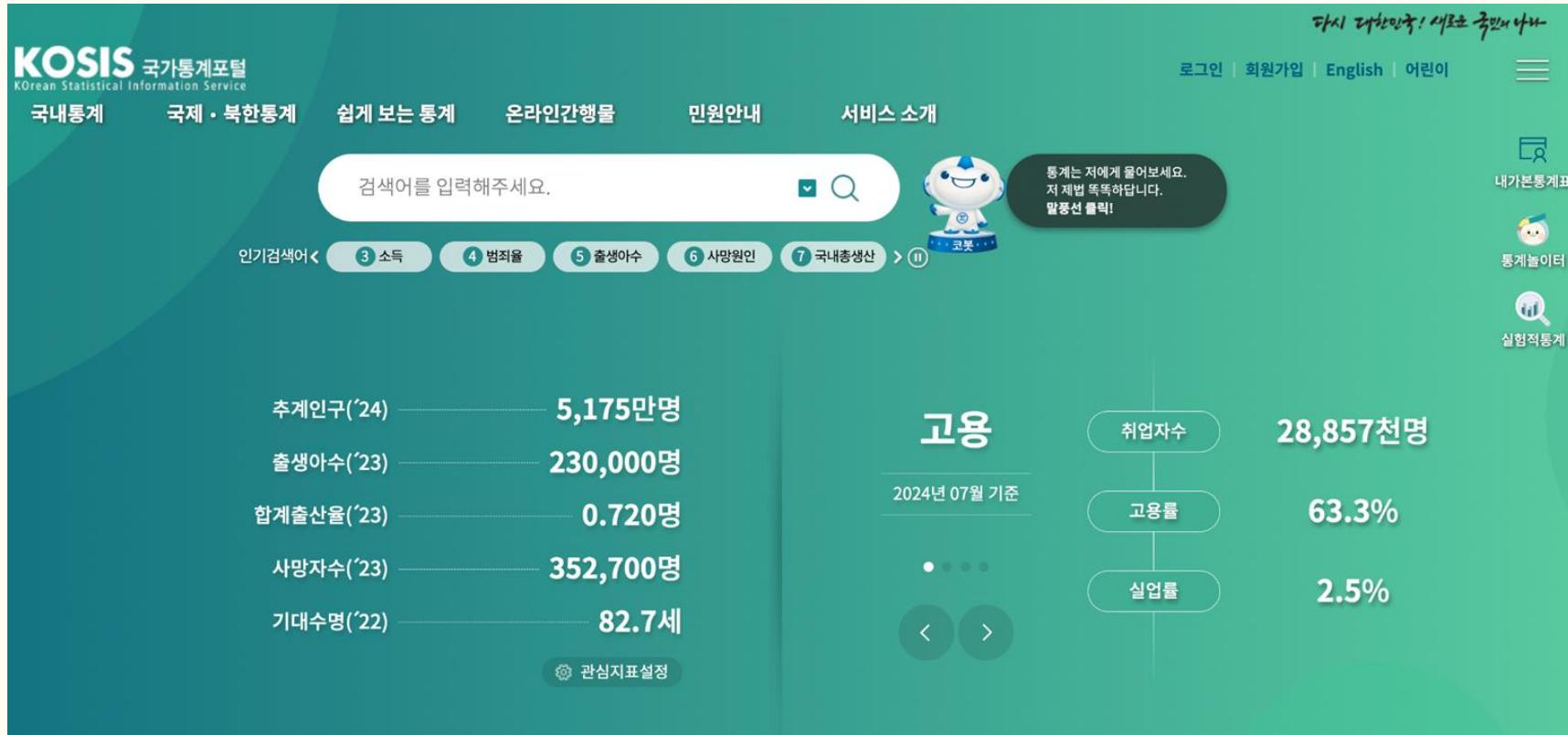


df3.melt(id_vars=['first', 'last'])

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150

4주차 학습 내용 정리

KOSIS 국가통계포털



4주차 학습 내용 정리

KOSIS

The screenshot shows the KOSIS homepage with a search query entered: "규모별 사회보험가입률, 상여금·퇴직(연)금 적용(가입)률, 노조가입률". The results table for 2020 includes the following data:

고용형태	규모	2020				
		고용보험 (%)	건강보험 (%)	국민연금 (%)	산재보험 (%)	상여금 (%)
전체근로자(특수형태제외)	전체	90.3	91.1	91.3	97.8	51.6
	5인미만	73.7	78.0	78.8	91.5	31.1
	5~29인	94.5	92.6	92.6	99.7	48.7
	30~299인	97.2	96.8	96.6	99.8	59.3
	300인미만	89.5	89.8	90.1	97.5	46.8
	300인이상	89.5	95.5	99.7	98.4	99.8
정규근로자	전체	94.4	98.5	98.3	97.9	61.5
	5인미만	82.4	93.9	94.5	91.1	43.8
	5~29인	97.6	99.5	99.2	99.7	56.4
	30~299인	99.0	99.8	99.6	100.0	66.0

4주차 학습 내용 정리

KOSIS 행렬전환은 Pandas의 melt 와 유사합니다.

The screenshot shows the KOSIS statistics portal with two main windows. The left window displays a table titled '국가(대륙)별/상품군별 온라인쇼핑 해외직접판매액' (International Online Shopping Direct Purchase Amount by Country and Product Category). The right window shows the 'Melt' transformation interface. A blue arrow points from the 'Melt' interface back to the original table, illustrating how the transformation process works.

국가(대륙)별	상품군별	판매유형별	2022.2/4	2022.3/4 p)
합계	합계	계	505,954	399,881
		면세점	300,602	211,727
	컴퓨터 및 주변기기	면세점 이외	205,352	188,154
		계	2,867	1,573
	가전·전자·통신기기	면세점	0	0
		면세점 이외	2,867	1,573
	소프트웨어	계	4,298	3,281
		면세점	2,195	1,316
	서적	면세점 이외	2,103	1,965
		계	63	24
	면세점	면세점	0	0
		면세점 이외	63	24
	서적	면세점	9,196	10,727
		계	0	0

4주차 학습 내용 정리

분기, 연도 파생변수 만들기

```
df[ "분기" ] = df[ "기간" ].apply(lambda x : x.split()[1].split("/")[0])
df[ "분기" ] = df[ "분기" ].astype(int)
df[ "연도" ] = df[ "연도" ].astype(int)
df.head()
```

	국가(대륙)별	상품군별	판매유형별	기간	million	연도	분기
0	합계	합계	계	2014 1/4	148272	2014	1
1	합계	합계	면세점	2014 1/4	-	2014	1
2	합계	합계	면세점 이외	2014 1/4	-	2014	1
3	합계	컴퓨터 및 주변기기	계	2014 1/4	4915	2014	1
4	합계	컴퓨터 및 주변기기	면세점	2014 1/4	-	2014	1

4주차 학습 내용 정리

금액을 수치데이터로 나타내기

```
df[ "million" ] = df[ "million" ].replace( "-", np.nan).astype( float )  
df
```

	국가(대륙)별	상품군별	판매유형별	기간	million	연도	분기
0	합계	합계	계	2014 1/4	148272.0	2014	1
1	합계	합계	면세점	2014 1/4	NaN	2014	1
2	합계	합계	면세점 이외	2014 1/4	NaN	2014	1
3	합계	컴퓨터 및 주변기기	계	2014 1/4	4915.0	2014	1
4	합계	컴퓨터 및 주변기기	면세점	2014 1/4	NaN	2014	1

4주차 학습 내용 정리

파생변수 만들기

```
df['쿼터'] = df['기간'].map(lambda period : period.split()[1])  
df
```

	국가(대륙)별	상품군별	판매유형별	기간	million	연도	분기	쿼터
49	미국	컴퓨터 및 주변기기	면세점	2014 1/4	NaN	2014	1	1/4
50	미국	컴퓨터 및 주변기기	면세점 이외	2014 1/4	NaN	2014	1	1/4
52	미국	가전·전자·통신기기	면세점	2014 1/4	NaN	2014	1	1/4
53	미국	가전·전자·통신기기	면세점 이외	2014 1/4	NaN	2014	1	1/4
55	미국	소프트웨어	면세점	2014 1/4	NaN	2014	1	1/4

4주차 학습 내용 정리

map, apply

	Series	DataFrame	사용예
map	O	X	<code>df[“컬럼명”].map(함수 or dict)</code>
apply	O	O	<code>df.apply(함수)</code>

4주차 학습 내용 정리

Series Accessor 시리즈 접근자

Data Type	Accessor
Datetime, Timedelta, Period	dt
String	str
Categorical	cat
Sparse	sparse

4주차 학습 내용 정리

dt Accessor

<code>Series.dt.date</code>	Returns numpy array of python <code>datetime.date</code> objects.
<code>Series.dt.time</code>	Returns numpy array of <code>datetime.time</code> objects.
<code>Series.dt.timetz</code>	Returns numpy array of <code>datetime.time</code> objects with timezone information.
<code>Series.dt.year</code>	The year of the datetime.
<code>Series.dt.month</code>	The month as January=1, December=12.
<code>Series.dt.day</code>	The day of the datetime.
<code>Series.dt.hour</code>	The hours of the datetime.
<code>Series.dt.minute</code>	The minutes of the datetime.
<code>Series.dt.second</code>	The seconds of the datetime.
<code>Series.dt.nanosecond</code>	The nanoseconds of the datetime.
<code>Series.dt.week</code>	(DEPRECATED) The week ordinal of the year.

4주차 학습 내용 정리

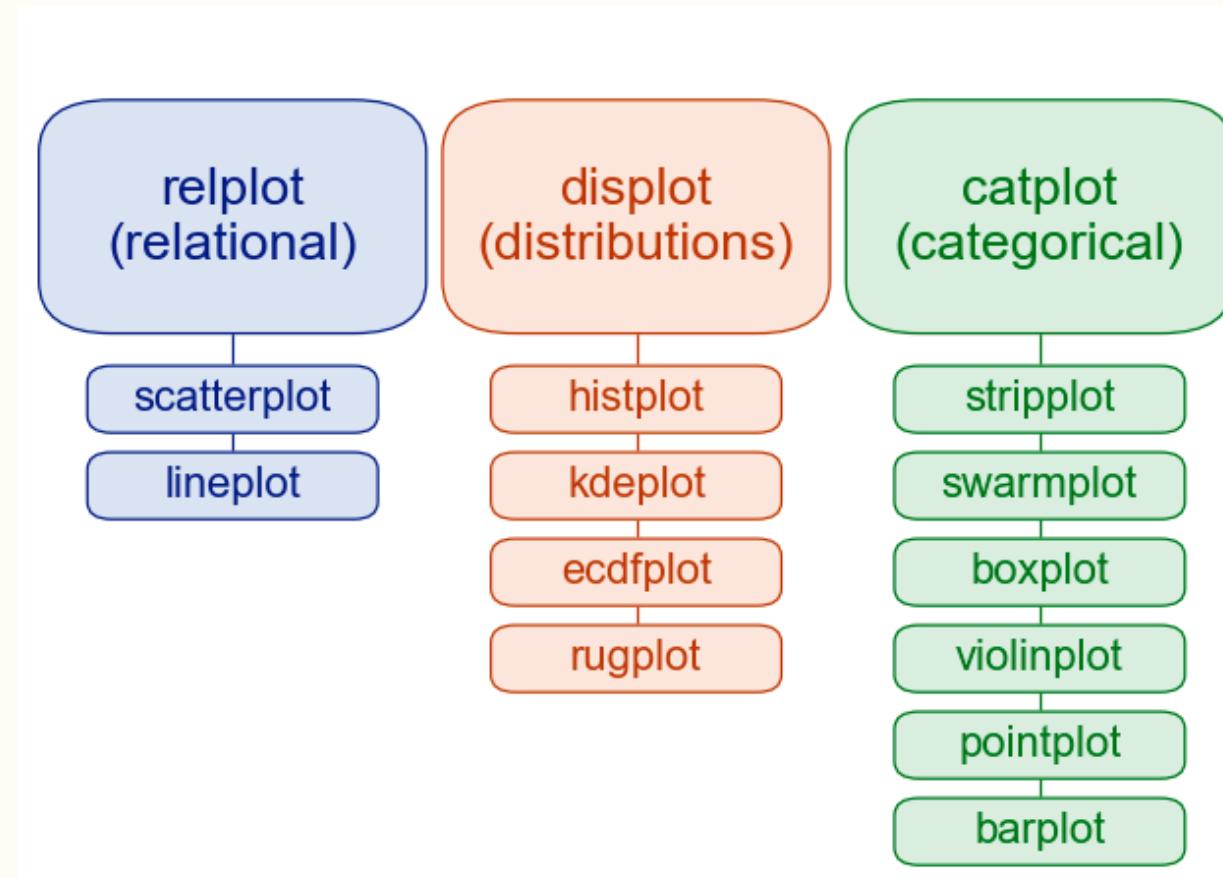
메서드	기능
<code>Series.str.capitalize()</code>	첫 글자를 대문자로 만든다.
<code>Series.str.cat([others, sep, na_rep, join])</code>	주어진 구분자로 문자열을 연결한다.
<code>Series.str.contains(pat[, case, flags, na, ...])</code>	문자열에 특정 표현이나 정규표현식에 만족하는 값이 있는지 확인한다.
<code>Series.str.count(pat[, flags])</code>	특정 패턴이 등장하는 빈도를 계산한다.
<code>Series.str.endswith(pat[, na])</code>	특정 패턴의 값으로 끝나는지를 확인한다.
<code>Series.str.find(sub[, start, end])</code>	패턴이 등장하는 가장 첫 인덱스 번호를 반환한다.
<code>Series.str.get(i)</code>	특정 인덱스에 해당되는 값을 반환한다.
<code>Series.str.join(sep)</code>	특정 구분자로 문자열을 연결한다.
<code>Series.str.len()</code>	문자열의 길이를 구한다.
<code>Series.str.lower()</code>	소문자로 변환한다.
<code>Series.str.lstrip([to_strip])</code>	왼쪽의 공백문자를 제거한다.
<code>Series.str.match(pat[, case, flags, na])</code>	정규표현식의 일치여부를 확인한다.

4주차 학습 내용 정리

메서드	기능
<code>Series.str.pad(width[, side, fillchar])</code>	입력된 길이가 되도록 앞 문자를 공백 문자로 채운다.
<code>Series.str.repeat(repeats)</code>	값을 지정한 횟수만큼 반복해서 생성한다.
<code>Series.str.replace(pat, repl[, n, case, ...])</code>	특정 패턴을 대체한다.
<code>Series.str.split([pat, n, expand, regex])</code>	구분자로 문자를 나눈다.
<code>Series.str.startswith(pat[, na])</code>	특정 패턴 혹은 문자로 시작하는지를 확인한다.
<code>Series.str.strip([to_strip])</code>	앞뒤 공백을 제거한다.
<code>Series.str.swapcase()</code>	대문자라면 소문자로 소문자라면 대문자로 대소문자를 반대로 변환한다.
<code>Series.str.title()</code>	단어의 첫글자를 대문자로 만든다. <code>capitalize()</code> 는 문장의 첫글자만 대문자로 만들지만 <code>title()</code> 은 단어의 첫글자를 대문자로 만든다.
<code>Series.str.upper()</code>	영문자를 모두 대문자로 변경한다.
<code>Series.str.wrap(width, **kwargs)</code>	지정된 너비로 줄 바꿈한다.
<code>Series.str.zfill(width)</code>	'0'문자를 앞에 추가하여 지정한 길이의 문자가 되도록 채운다. '9'라는 문자에 width 값을 3으로 채우면 '009'가 된다.
<code>Series.str.pad(width[, side, fillchar])</code>	입력된 길이가 되도록 앞 문자를 공백 문자로 채운다.

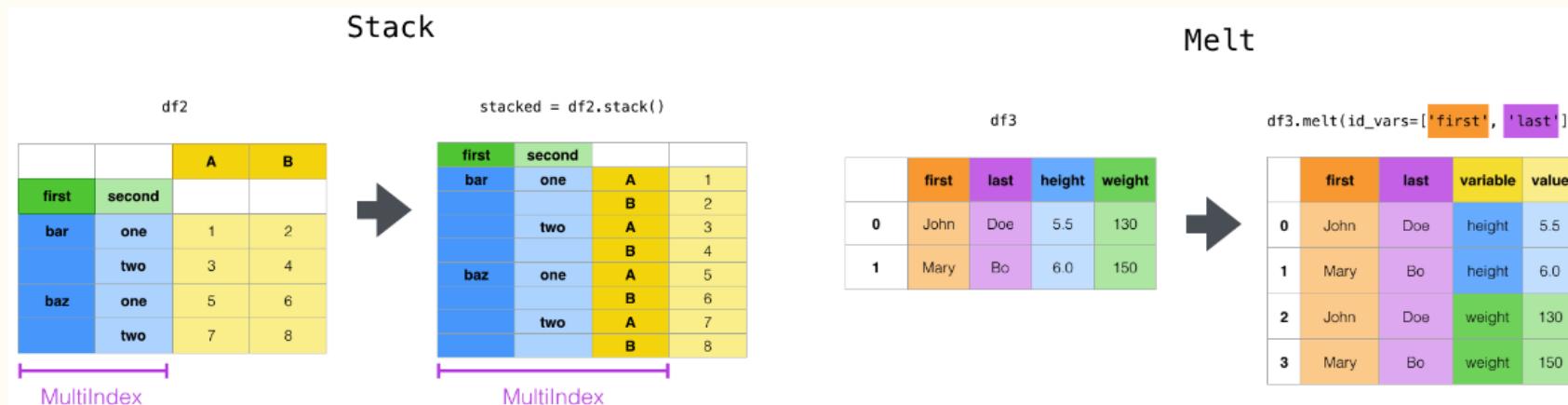
4주차 학습 내용 정리

Seaborn을 통한 Subplots



알아두면 쓸데있는 신비한 프로그래밍

	stack()	melt()
DataFrame	O	O
Series	X	X
기능	컬럼 => 인덱스, 값으로 녹임	컬럼=>변수, 값으로 녹임
반환결과	Series	DataFrame



알아두면 쓸데있는 신비한 프로그래밍

The screenshot shows a web browser displaying the website clauswilke.com/dataviz/. The page has a sidebar on the left containing a search bar and a navigation menu with links like 'Welcome', 'Preface', '1 Introduction', and many numbered sections from 2 to 15. The main content area features a large title 'Fundamentals of Data Visualization' and the author's name 'Claus O. Wilke'. Below this is a section titled 'Welcome' with a descriptive paragraph about the website. To the right of the text is an image of the book 'Fundamentals of Data Visualization' by Claus O. Wilke, published by O'Reilly.

Fundamentals of Data Visualization

Claus O. Wilke

Welcome

This is the website for the book "Fundamentals of Data Visualization," published by O'Reilly Media, Inc. The website contains the complete author manuscript before final copy-editing and other quality control. If you would like to order an official hardcopy or ebook, you can do so at various resellers, including [Amazon](#), [Barnes and Noble](#), [Google Play](#), or [Powells](#).

The book is meant as a guide to making visualizations that accurately reflect the data, tell a story, and look professional. It has grown out of my experience of working with students and postdocs in my laboratory on thousands of data visualizations. Over the years, I have noticed that the same issues arise over and over. I have attempted to collect my accumulated knowledge from these interactions in the form of this book.

O'REILLY®

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

Claus O. Wilke

알아두면 쓸데있는 신비한 프로그래밍

☰ SBS프리미엄 뉴스 지식뉴스 팟캐스트 커뮤니티 스퀴즈 SBS NEWS > 로그인

MABU NEWS

마부뉴스

하나의 이슈를 데이터로 깊이 있게 살펴보는 뉴스레터, 환경도 놓칠 수 없죠

구독하기 +

내진 설계, 일본 87% vs. 대한민국 16%... 이런 상황에 대지진 닥치면?
2024-08-16 07:00

"동일본 대지진의 10배 피해" 대지진, 일본에 발생할 수 있다?
2024-08-15 07:00

여성으로 태어나 자랐더라도 남성호르몬 많으면 여성 아니라고?
2024-08-09 07:00

XY 염색체를 가지고 있다면 여성 경기에 출전하면 안 될까?
2024-08-08 07:00

준비 안 된 '친환경' 올림픽, 그린워싱 아닌가요?
2024-08-02 07:00

센강에서 파리올림픽 수영 경기가 열려도 괜찮은 걸까?
2024-08-01 07:00

'출퇴근 미어터진다' 성수역, 근데 더 심한 지하철 역이 있다고?
2024-07-26 07:00

"지하철 터진다!" 출퇴근 시간 1만 명 넘게 몰리는 그 역, 이대로 괜찮을까?
2024-07-25 07:00

미션 출제 의도

4주차

4주차 미션의 출제 의도와 문제 소개

4주차 미션 출제 의도

1번 [파생변수와 기술통계] 분기에 대한 파생변수 구하고 기술통계 구하기

2번 [pivot_table] 국가(대륙)별 연도별 판매액의 합계를 분석

3번 [groupby] 2020년의 온라인 해외판매 상품군 지역별 합계

4번 [시각화] 주요 판매 국가와 상품군 연도별 온라인 직접 판매액 합계 시각화

5번 [자유주제] KOSIS 데이터 분석하기

4주차 미션 - 1번

파생변수와 기술통계

- ❖ Q1. 시점 컬럼에서 연도와 분기에 대한 파생변수를 생성하고 기술통계를 구해주세요.

4주차 미션 - 2번

pivot_table로 연도별 판매액의 합계 구하기

- ❖ Q2. pivot_table을 사용하여 국가(대륙)별 연도별 판매액의 합계를 분석해 주세요.

4주차 미션 - 3번

groupby 상품군 지역별 합계

☞ Q3. groupby를 사용하여 2020년의 온라인 해외판매 상품군을 지역별 합계를 구해 분석해 주세요.

4주차 미션 - 4번

연도별 온라인 직접 판매액 합계를 시각화

- ❖ Q4. 주요 판매 국가와 상품군에 대해 2021년까지의 연도별 온라인 직접 판매액 합계를 시각화 해주세요.

4주차 미션 - 5번

KOSIS 자유주제

❖ Q5. KOSIS에 있는 데이터를 하나 정해 자유롭게 분석해 주세요!

코치에게 물어봐

+실시간 QnA

코치에게 물어봐

슬랙 #코치에게-물어봐 채널에 남겨주신 질문에 대해 답변해드려요

?
주은미_리더

현재 BIO 도메인의 30대 초반 Senior Researcher 입니다.

아무래도 현업에서 실험을 통한 실측 데이터를 많이 수집하다 보니, 데이터를 활용/예측하여 사업적으로도 유용한(돈을 벌 수 있는) 인사이트를 도출하고 싶습니다.

이번 년도에 특히 생각이 확고해져 현재 빅데이터 분석기사를 준비하고 있고, 현업에서 관련 경험을 쌓아 이를 토대로 빅데이터 석사 진학을 준비하고자 합니다.

- 30대 중순에 학위 수여가 가능할텐데, 제 나이에도 이직을 통한 직무 전환이 가능할지 걱정입니다. 주변에 유사한 사례를 경험해 보셨을까요?
- 또한, 데이터 사이언스 관련 업무를 수행하기 위해서는 컴퓨터 전공자를 더 많이 우대하는 것 같습니다. 다른 도메인에서 데이터 사이언스를 학습하여 해당 분야에 실무를 하는 사례가 있을까요?

코치에게 물어봐

슬랙 #코치에게-물어봐 채널에 남겨주신 질문에 대해 답변해드려요

?

사과맛배

- 업무 자동화에 대해 관심이 있는데요, 지금까지 배운 내용들을 자동화를 통해 효율적으로 활용할 방법이 있는지 궁금합니다. (ex. data 업데이트시 분석 output도 자동 업데이트 등)
- 현업에서 어떻게 pandas를 이용하여 업무 효율화를 하고 계신지 관련 사례가 있다면 공유받고 싶습니다!

코치에게 물어봐

슬랙 #코치에게-물어봐 채널에 남겨주신 질문에 대해 답변해드려요

؟ 향기

안녕하세요 심리학과 4학년 1학기 재학중인 대학생입니다.

현재 취업 방향을 전산/인사관리 분야로 희망 중이라 몇 가지 질문 사항이 있습니다.

- 대학원 진학은 아직 계획중에 있지 않아서 학부 졸업 후 바로 취업을 계획 중에 있습니다.
이런 상황에서 어떤 내용을 공부하면 실무 시에 유의하게 활용 가능할까요?
- 전산/인사관리 직종 종사를 희망할 시에 관련 자격증은 어떤걸 취득하는 것이 좋을까요?
현재 사회조사분석사 2급 실기 / ADsP 를 위주로 공부중이며, 올해 자격증 취득 목표로 하고 있습니다. 이 외에 필요한 것이나 취득해 놓으면 이후에 수월하다 여겨질 자격증이 있을까요?
- 별개로 AI나 빅데이터 관련 자격증을 공부할 때는 어떤 요소를 위주로 공부하면 될까요?

끝까지 포기하지 않고
완주한 부스터 여러분!
진심으로 감사드립니다:)
🍀 다음에 또 만나요 🍀