데이터 사이언티스트가 뭐야?

2020 M1_개인주제프로젝트 조이

```
    'Name':['Woojung', 'Wooyoung'],
    'Occupation':['Happy Person', 'Nurse'],
    'Born':['2002-03-22', '2004-11-03'],
    'Died':['3000-03-22', '3000-11-03'],
    'Age':[50, 48]}
 )
print(mysisnme)

mysisnme = pd.DataFrame(
    data={'Occupation' : ['Happy Person', 'Nurse'],
          'Born': ['2002-03-22','2004-11-03'],
          'Died': ['3000-03-22', '3000-11-03'],
          'Age':  [50, 48]},
index=['Woojung','Wooyoung'],
```

DATA SCIENTAIST

목차

CHAPTER_1

CHAPTER 1

1) 프로젝트 진행 동기

2) 데이터 사이언스란?

3) 데이터 사이언티스트는 뭐하는 사람이야?

프로젝트를 진행하게 된 동기

‘내가 하고 싶은 일은 무엇일까?’ 내가 중, 고등학교를 다닐 때 항상 하던 생각이다. 하지만, 누가 나에게 하고 싶은 일이 무엇인지 물어보면 나는 대답을 하지 못했다. 하고 싶은 일과 직업에 대한 생각을 놓지 못했고, 항상 그것을 찾고 싶어 했다. 거캠에 들어오고 나서도 그 고민은 여전했다. 이런 고민들을 가지고 작년 3모듈을 지냈다. 그리고 다음 모듈인 4모듈에 D랩에 가게 되었다. D랩에서 R이라는 데이터 분석 툴을 이용해서 수업을 한 적이 있는데. 그때 R에 대해 관심이 생겼다. 많은 데이터 속에서 필요한 조건을 코드로 입력해서 인사이트를 뽑는 게 흥미롭고 재밌었다. R로 인해 데이터사이언스에 관심이 생겨 그것에 대해 알아보면서 , 데이터 사이언티스트라는 직업을 되었다. 그에 대해 알아보다 보니, 이 직업에 대한 흥미가 생기기 시작했고, 직업으로 삼고 싶다는 생각이 들었다. 데이터사이언티스트라는 직업에 대한 관심이 그저 충동적인 관심이었는지, 아니면 진짜 내가 이걸 직업으로 삼고 싶은 건지 알아보고 싶어서 이번 프로젝트를 진행하게 되었다.

데이터 사이언스란?

데이터 사이언스는 큰 데이터 안에서 쉽게 알 수 없으면서 유용한 패턴을 뽑아내기 위한 일련의 규칙, 문제의 정의, 알고리즘과 처리과정 등을 아우르는 개념이다.

데이터 사이언티스트는 뭐하는 사람이야?

데이터 사이언티스트는 수많은 데이터 속에서 가치 있는 데이터를 추출해 분석해 결과를 현업에 적용하고, 미래를 예측하기도 한다. 수많은 데이터 속에서 인사이트를 추출하는 것은 어려운 일이다. 파이썬이나 R 툴을 이용해서 데이터를 분석한다. 분석 뿐만 아니라 분석한 데이터를 회사나 조직 전반에 걸쳐 실행 가능한 전략적인 인사이트를 제공한다. 하지만 분석 뿐 아니라 또 다른 중요한 기술이 있는데 그것은 바로 커뮤니케이션 기술이다.

커뮤니케이션 기술이 왜 필요할까? 분석한 데이터를 현업에 적용하기 위해서는 커뮤니케이션 기술을 가져야한다. 의미 있는 인사이트라고 해도 최고 경영자를 이해, 설득시키지 못하면 무용지물이 되기 때문이다. 데이터 사이언티스트는 현황분석 보다는 주로 산업별 전문 지식을 갖고 예측 최적화 분석을 한다. 단계별로 분석 기술을 비교하자면 현황분석, 원인분석, 예측분석, 예측 최적화 분석이 있다. 현황분석은 과거 데이터를 바탕으로 한 일반적인 기초 통계를 통해 전반적인 상황을 파악하고 이해하고 확인하는 작업을 말한다.

다음으로 원인 분석은 과거에 왜 그런 결과가 나왔는지 원인을 분석하고 확인하는 작업을 말한다. 세 번째로 예측 분석은 앞으로 발생할 가능성이 있는 사안들을 추측하는 것을 가리킨다. 마지막으로 예측 최적화 분석이다. 예측 최적화 분석은 분석에서 찾아낸 인사이트를 설득을 통해 현업에 적용하고 기업이나 기관에 도움이 될 데이터 만한 성과를 내는 것이다. 한 마디로 원하는 결과를 어떻게 일어나게 할 것인가? 의 질문을 던지는 분석 과정이다.

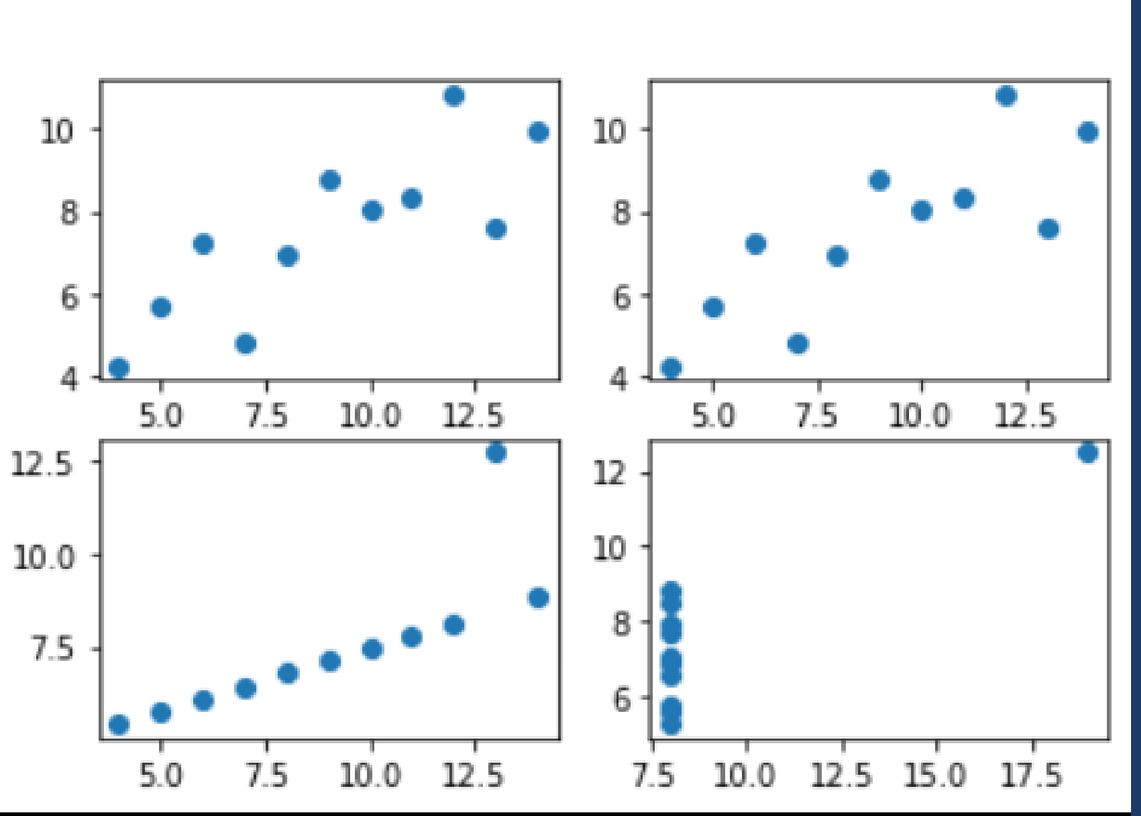
단계별 분석 기술 비교 예를 들면 어느 백화점 판매 데이터에서 지점별 한 달 매출을 합산하여 과거 데이터와 비교 분석을 했을 때 어떤 지점의 매출이 왜 적게 나왔는지 파악하는 것은 '현황분석'이고, 과거 매출 데이터와 다양한 경제 변수, 내부 비즈니스 환경을 고려하여 내년 매출을 추측하는 것은 '예측 분석'이다. 그리고 내년 매출과 함께 내년 반품률을 예측하고, 제한된 자원 안에서 최적의 구매 예산을 도출해 내기 위하는 것은 '예측 최적화 분석'이다.

CHAPTER_2

CHAPTER 2
1) 데이터 사이언티스트가 되기 위해서는?
2) 나의 적성과 직업군

프로그램언어 및 기술

먼저 데이터 사이언티스트가 되기 위해서는 여러 능력과 역량이 필요하다. 첫 번째로는 데이터를 분석하기 위한 능력을 기를 수 있는 프로그램언어 및 기술을 배우는 것이 중요하다. 프로그램 언어의 대표적인 예로는 파이썬과 R프로그래밍, JAVA 등이 있다. 이러한 프로그램언어를 익혀 컴퓨터에게 원하는 값을 얻기 위해 명령을 하게 되는데 명령을 하는 과정을 코딩이라고 한다. 수많은 데이터 속 내가 원하는 인사이트를 얻기 위해서는 컴퓨터에게 정확한 명령을 내려 코딩을 할 수 있기 때문에 프로그램언어 및 기술이 필요하다.



기초 통계학

데이터사이언스 상위단계에 가면 (ex. 머신러닝) 코딩을 할 때 통계적 지식이나 이론이 필요해진다. 이때 머신러닝은 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자 하는 기술 및 기법이다.

커뮤니케이션 능력

세 번째는 앞서 말했듯이 커뮤니케이션 능력이 필요하다. 코딩을 통해 인사이트를 추출했을 때 추출한 인사이트를 최고 경영자에게 설득시키지 못하면 무용지물이 된다. 커뮤니케이션 기술의 중요성은 2014년 미국 라바스톰 애널리틱스사에서 데이터 사이언티스트를 대상으로 실시한 설문 조사에서도 확인된다. “분석 과정에서 가장 큰 도전이 무엇인가” 라는 질문에 가장 많이 나온 응답은 “데이터에서 얻은 인사이트에 대한 신뢰를 얻는 것” 이었다.
이를 통해 알 수 있듯이 데이터 과학자에게 가장 큰 도전은 분석도 보다 분석에서 찾아낸 인사이트를 설득을 통해 현업에 적용하고 기업이나 기관에 도움이 될 만한 성과를 내는 것이다.

나의 직업과 적성

나는 먼저 나의 직업적성을 알기 위해 커리어넷이라는 사이트에서 직업적성 검사를 진행했다. 적성검사는 지금 현재 내가 잘하고 있거나 앞으로 발전할 가능성이 높은 능력을 알기 위해 진행하는 것이다. 검사를 통해서는 자신의 적성 영역과 그 영역에 잘 맞는 직업에 대해서 알 수 있다.

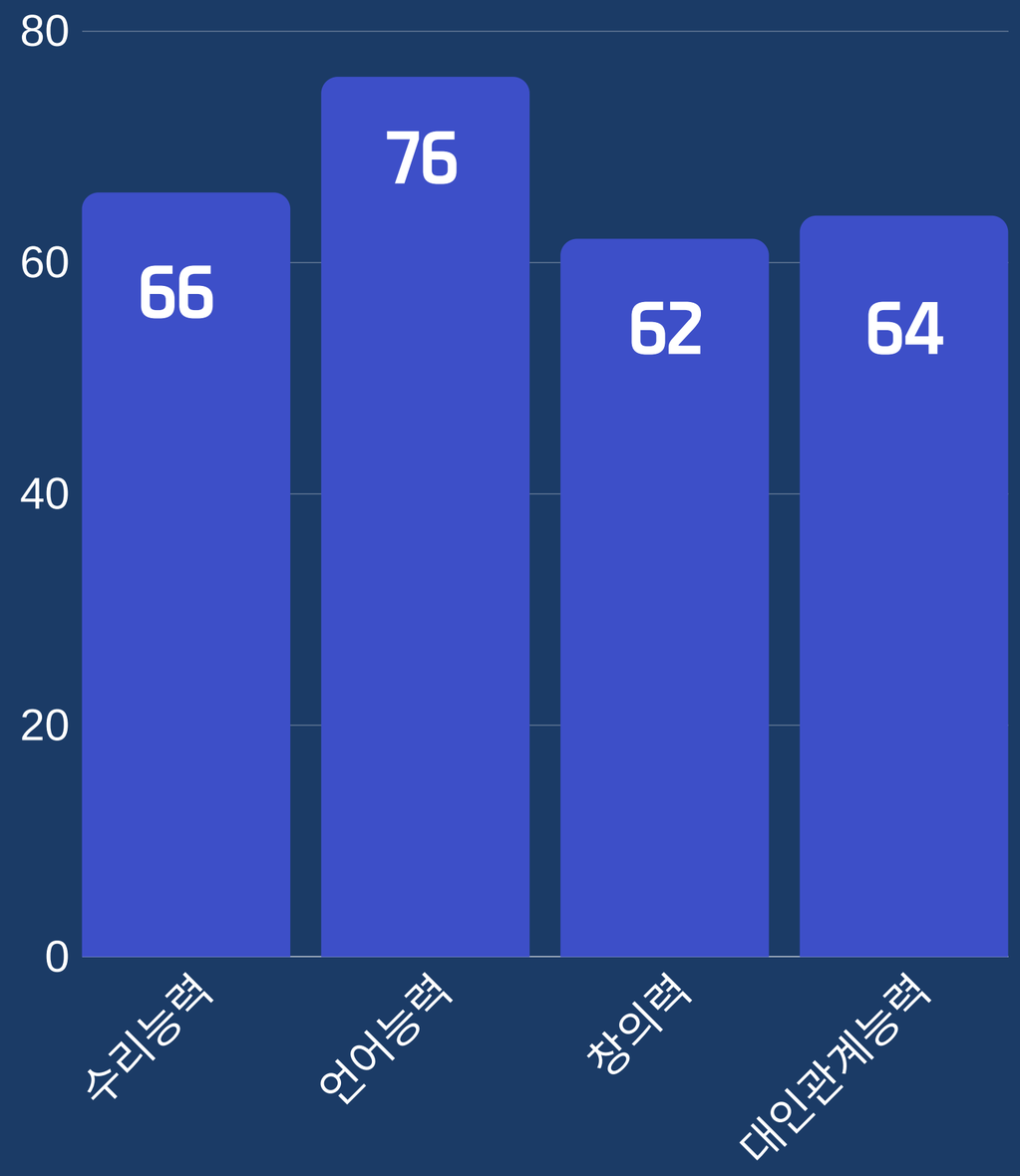
<직업적성검사 주요결과>

수리 능력

수학적 개념을 이해하고 계산하고 해결하는 능력을 가지고 있습니다. 논리적인 사고로 분석하고 응용하는 것을 잘합니다.

언어 능력

말이나 글로 생각과 감정을 표현하는 능력이 뛰어납니다. 다른 사람의 말과 글을 잘 이해할 수 있습니다.



커리어넷 직업적성검사 결과

(단위: %)

나의 직업적성결과 중 일부분이다. 그래프를 보면 수리능력, 언어능력, 창의력 등의 능력이 높게 나왔다. 앞서 말한 데이터 사이언티스트가 되기 위한 능력 중 수리와 언어 능력 등이 겹치게 나왔다. 이러한 능력을 더 기를 수 있는 방법이 어떤 것이 있는지 알고, 능력을 발전시켜야 겠다.

CHAPTER_3

CHAPTER 3
1) 데이터 사이언티스트가 되기 위해 내가 하고 있는 일
2) 앞으로 해야할 일

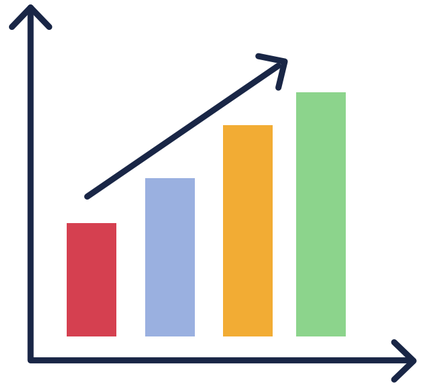
데이터 사이언티스트가 되기 위해 내가 하고 있는 일

프로그램언어
공부하기

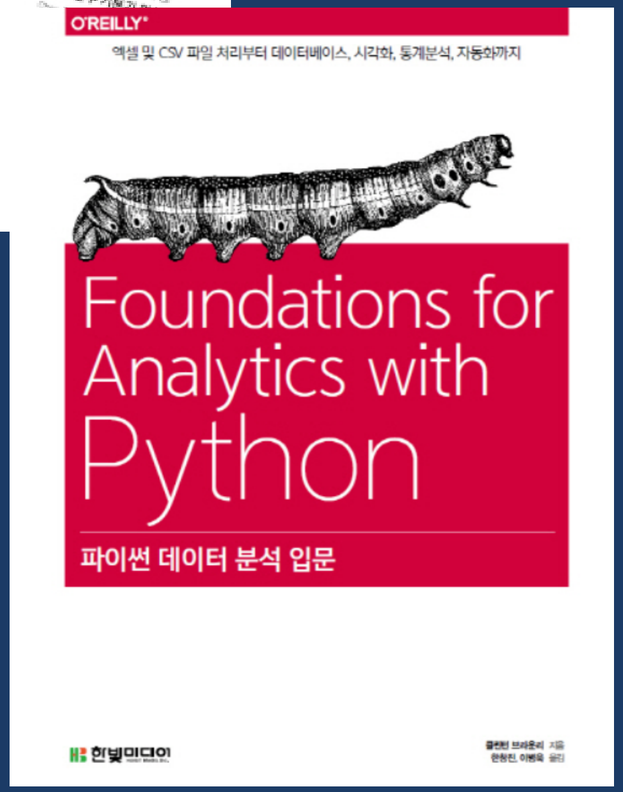
python™

커뮤니케이션
능력 기르기

통계학
공부하기



프로그램 언어 공부하기



앞서 말한 데이터 사이언티스트가 갖춰야할 역량 중에 내가 지금 당장 할 수 있는게 무엇인가 생각해 보았다. 그러다가 디랩에서 R 프로그램언어를 공부한 것이 생각났다. 나는 R 프로그램 언어이외에 다른 프로그램 언어가 배우고 싶어서 파이썬 책을 찾아보았다. 먼저보이는 '데이터 분석을 위한 판다스 입문' 이라는 책을 끝내고, 현재는 '파이썬 데이터 분석 입문' 이라는 책을 공부하고 있다. '데이터 분석을 위한 판다스 입문' 이라는 책을 공부하면서 여러 함수와 데이터를 다루는 방법을 배웠다.

```python
mysisnme = pd. DataFrame({
    'Name':['Woojung', 'Wooyoung'],
    'Occupation':['Happy Person', 'Nurse'],
    'Born':['2002-03-22', '2004-11-03'],
    'Died':['3000-03-22', '3000-11-03'],
    'Age':[50, 48]}
 )
print(mysisnme)

mysisnme = pd.DataFrame(
    data={'Occupation' : ['Happy Person', 'Nurse'],
          'Born': ['2002-03-22','2004-11-03'],
          'Died': ['3000-03-22', '3000-11-03'],
          'Age':  [50, 48]},
index=['Woojung','Wooyoung'],
columns=['Occupation','Born', 'Age','Died']
```

```python
kk = df.iloc[:, 0:6:2]
print(kk.head())
```

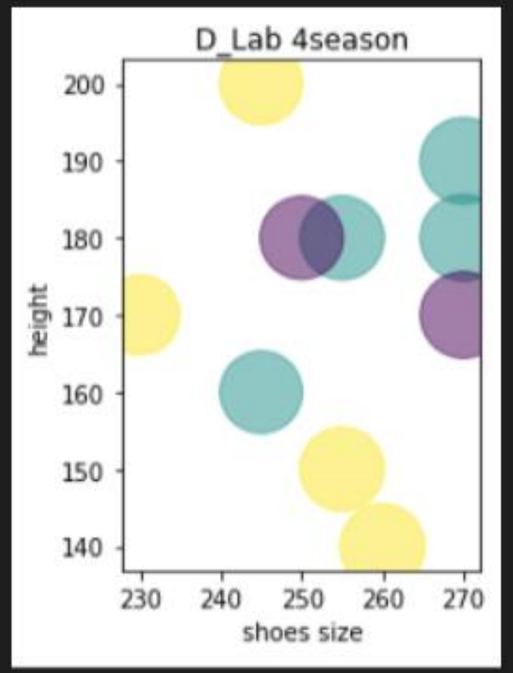
```
country  year       pop
0  Afghanistan  1952   8425333
1  Afghanistan  1957   9240934
2  Afghanistan  1962  10267083
3  Afghanistan  1967  11537966
4  Afghanistan  1972  13079460
```

```python
print(df.iloc[[0, 99, 999], [0, 3, 5]])
```

```python
print(df.loc[[0, 99, 999], ['country', 'lifeExp', 'gdpPercap']])
```

```
country  lifeExp    gdpPercap
0    Afghanistan   28.801   779.445314
99    Bangladesh   43.453   721.186086
999      Mongolia   51.253  1226.041130
         country  lifeExp    gdpPercap
0    Afghanistan   28.801   779.445314
99    Bangladesh   43.453   721.186086
999      Mongolia   51.253  1226.041130
```

```
Text(0, 0.5, 'height')
```



그동안 공부했던 파이썬의 일부분이다. 나는 데이터 프레임 만들기, 데이터 시각화하기, 집게 메서드, 데이터 필터링, 누락값확인 등등의 공부를 했고, 데이터 분석 공부를 하면서 흥미롭고 재밌는 기분을 느꼈다.

앞으로 내가 해야할 일

앞서 말한 책 두 권 중 '파이썬 데이터 분석 입문' 이라는 책을 다 끝내서 파이썬 프로그램 언어를 다룰 수 있는 실력을 높여야겠다. 또한 파이썬 프로그램으로 사최수프를 하는 데에 도움을 주고 싶다. 또한 프로그램언어 이외에 커뮤니케이션 능력, 통계학을 공부할 것이다. 이번 프로젝트에서는 파이썬을 연습만 했다면 다음 프로젝트에서는 파이썬을 이용한 결과물을 낼 것이다. 데이터를 찾아서 그것을 가공하고 내가 원하는 인사이트를 추출해서 프로젝트에 도움을 주고싶다.

참고 문헌

-도서

1) 처음 배우는 데이터 과학 통계, 수학, 머신러닝, 프로그래밍까지 데이터 과학자를 꿈꾸는 히치하이커를 위한 최고의 안내서- 필드 케이디

2) 데이터 과학 무엇을 하는가? 현직 데이터 과학자가 알려주는 실무 적용 방법!- 김옥기
(6, 7페이지 내용 참고)

-논문

1) 데이터 사이언티스트의 역량과 빅데이터 분석성과의 PLS
경로모형분석 : Kaggle 플랫폼을 중심으로

2) 발전을 위한 빅데이터 활용과 데이터 사이언티스트 양성 (9,10, 11 페이지 내용 참고)