2021147510 김주은

주제: 워라밸을 결정하는 주된 요인은 무엇인가?

: 워라밸을 결정할 수 있는 요인들과 워라밸 사이의 상관관계 분석과 다중 회귀 분석을 중심으로

데이터: 잡플래닛을 직접 크롤링하여 얻은 데이터

(기존에는 노동 통계 통계 포털에 나와있는 산업_규모별 근로시간 데이터와 사업체 노동력 조사 데이터를 사용하려고 했으나, 워라밸과 관련된 평가 항목이 임금과 이직률 이외에는 없기도 하고, 두 개의 데이터의 행을 기준으로 join 을 해서 사용하는 경우 데이터가 20 개밖에 되지 않기 때문에 잡플래닛의 데이터를 직접 크롤링해서 사용하기로 하였음.)

데이터 형식: 총 700 개의 기업(행)에 대하여 8 개의 열의 데이터를 조사하였음.

	Company Na	Company ID	복지 및 급여	업무와 삶의 균	사내문화	승진 기회 및 가	경영진	연봉	입사자 수	퇴사자 수
0	(주)에스큐제약	348308	5.0	5.0	5.0	5.0	5.0	2,715	0	1
1	세이지리서치(주	351150	4.6	4.8	4.7	4.5	4.7	5,692	1	0
2	(주)팀엘리시움	329950	4.7	4.8	4.9	4.7	4.7	3,556	0	0
3	월급쟁이부자들(	394321	4.7	3.8	4.7	4.6	4.8	4,008	9	7

열 항목:

5 개의 점수의 경우 잡플래닛 이용자들이 직접 설문에 참여하여 기록한 별점의 평균값임.

-복지및급여 점수/업무와 삶의 균형(워라밸) 점수/사내문화 점수/승진 기회 및 가능성 점수/경영진 점수/연봉/입사자/퇴사자

행 항목: 잡플래닛의 상위 인기 700 개 기업

1.입력 데이터 처리(가공 및 생성)

크롤링 코드 : crolling.py

Pandas /Selenium/ BeautifulSoup 을 사용하여 crolling 을 진행하였음.

데이터를 직접 생성한 이유:

고용 노동 통계 포털에 존재하는 현재 데이터의 경우에는 임금 및 근로시간 데이터와 이직률 데이터가 다른 데이터로 존재했는데, 두 개의 행이 일치하지 않아 Inner join 을 하여 데이터를 일치시키는 경우 산업 분야 1 개당 1 개의 데이터밖에 존재하지 않아 데이터가 일반성을 가지지 못한다고 판단하였음.

또한 워라밸의 요소로 판단할 수 있는 데이터로 근무시간이 전부였기 때문에 워라밸의 임계를 찾고자 하는 주제에는 부족한 데이터로 느껴졌음.

데이터 생성 과정:

1) 잡플래닛 데이터의 경우에는 특정 정보를 가져오기 위하여 로그인이 필요하므로, 셀레니움을 통하여 로그인을 한 이후에 크롤링을 진행하였음.

```
driver = webdriver.Chrome()
usr = "kjeiun@yonsei.ac.kr"
pwd = "010525kk!!"

driver.get("https://www.jobplanet.co.kr/users/sign_in?_nav=gb")
time.sleep(5)

login_id = driver.find_element(By.CSS_SELECTOR, "input#user_email")
login_id.send_keys(usr)
login_pwd = driver.find_element(By.CSS_SELECTOR, "input#user_password")
login_pwd.send_keys(pwd)

login_id.send_keys(Keys.RETURN)
```

2) 총 100 페이지에 달하는 기업 순위 리스트에서 각 기업의 고유 company id 와 company name 을 가져와서 list 에 dictionary 형태로 저장하였음.

```
for page_num in range(1, 101):
    # 웹 페이지 열기
    url = f'https://www.jobplanet.co.kr/companies?sort_by=review_survey_total_avg_cache&page={page_num}'
    driver.get(url)
    driver.implicitly_wait(5)     # 웹 페이지가 로딩될 때까지 기다리기 (예: 5초 기다림)
    html = driver.page_source     # HTML 가져오기
    soup = BeautifulSoup(html, 'html.parser')    # Beautiful Soup을 사용하여 HTML 파싱
    a_element = soup.select(
        '#listCompanies > div > div.section_group > section > div > div > dl.content_col2_3.cominfo > dt > a')

    print(f'Page {page_num}, len: {len(a_element)}')
    for i in range(len(a_element)):
        company_name = a_element[i].get_text(strip=True)
        company_id_href = a_element[i]['href']
        company_id = company_id_href.split('/')[2]
        company_list.append(
            {'Company Name': company_name, 'Company ID': company_id})
    df = pd.DataFrame(company_list)
    df.to_excel('Company1000.xlsx', index=False)
```

3) 해당 dictionary 에 저장된 company id 를 이용하여 '/reviews/' url 에 접근하여 5 개의 평가 점수를 획득하고

4) 연봉 입사자 퇴사자 정보는 다른 url 에 존재하기 때문에, 'salaries' url 에 접근하여 데이터를 가져왔음.

데이터 가공

1) 데이터를 가져오는 과정에서 특정 정보가 없거나 exception error 가 발생하는 경우 데이터를 가져오지 못하였으므로, 정보가 없는 데이터를 담은 행의 경우에는 아예 데이터 파일에서 삭제하였음.

2) 퇴사자 수 자체는 회사 규모에 영향을 받으므로, (퇴사자 수)/(퇴사자 수 + 입사자수) 를 하나의 지표로 사용하기 위해 새로운 행을 추가하였음.

2.문제 해결 방법(알고리즘) 개요

1)가설 또는 질문:

1.복지및급여 점수, 사내문화 점수, 승진기회 및 가능성 점수, 경영진 점수는 업무와 삶의 균형 점수에 대하여 양의 상관관계를 가지며, 가장 큰 양의 상관관계를 가지는 요소는 복지및급여와 관련된 점수일 것이다.

2. (퇴사자수)/(입사자수+퇴사자수) 의 경우에는 워라밸점수와 음의 상관관계를 가진다.

2)알고리즘: 다중 선형 회귀 분석 / 상관계수,산점도 시각화 / 다중공산성 확인

1) 다중 선형 회귀 분석

1. target_data 는 "업무와 삶의 균형(워라밸)"으로 설정하고 , 나머지 변인들은 x_data 데이터 프레임에 불러온다.

```
# 데이터 불러오기
company_data = pd.read_excel("700_after_preprocess1.xlsx")
cause_data = company_data.drop(['복지 및 급여'], axis=1)

# 다중 선형회귀분석

# target에 대하여 원인을 따질 데이터
x_data = company_data[["복지 및 급여", "사내문화",
                       "승진 기회 및 가능성", "경영진", "퇴사율"]]  # 변수 여러개
target = company_data[["업무와 삶의 균형"]]
```

2. 선형 회귀분석을 위한 상수항을 계산하고, 상수항을 추가하여 x_data1 에 담는다.

3. statsmodel 라이브러리를 사용하여 ols 검정을 진행한다.

```
# for b0, 상수항 추가
x_data1 = sm.add_constant(x_data, has_constant="add")

# OLS 검정
multi_model = sm.OLS(target, x_data1)
fitted_multi_model = multi_model.fit()
results = fitted_multi_model
```

4. 진행된 값을 바탕으로 summary 데이터를 엑셀파일로 저장한다.

```
# 회귀분석 결과를 데이터프레임으로 저장
result_df = pd.DataFrame(fitted_multi_model.summary(
).tables[1].data[1:], columns=fitted_multi_model.summary().tables[1].data[0])
result_df.set_index('coef', inplace=True)

# 결과를 엑셀 파일로 저장
result_path = 'regression_results_cleaned.xlsx'
result_df.to_excel(result_path)
```

2) 상관계수, 산점도 시각화

1. seaborn 라이브러리를 사용하여 1)에서 ols 검정을 통하여 계산된 x_data1 을 이용하여 상관계수 시각화를 히트맵으로 진행한다.

```
# 상관행렬 시각화
plt.rcParams['font.family'] = 'AppleGothic'  # 폰트를 설정해줘야 한글이 깨지지 않음
cmap = sns.light_palette("darkgray", as_cmap=True)
sns.heatmap(x_data1.corr(), annot=True, cmap=cmap)
plt.show()
plt.savefig('corr_heatmap.png')
```

2.마찬가지로 seaborn 라이브러리를 이용하여 변수들의 산점도를 시각화 한다.

```
# 변수끼리 산점도를 시각화
sns.pairplot(x_data1)
plt.show()
plt.savefic('scatter.png')
```

3. VIF 를 이용한 다중공산성 체크

```
# 3. VIF를 이용한 다중공산성 체크
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(
    x_data1.values, i) for i in range(x_data1.shape[1])]
vif["features"] = x_data1.columns

print(vif)
```

```
# 3. VIF를 이용한 다중공산성 체크 , x_data1에서 다중공산성이 높은 데이터는 제거하고 다시 실행
vif = pd.DataFrame()

# 다중공산성이 높은 경영진은 제거
x_data2 = x_data1.drop("경영진", axis=1)
vif["VIF Factor"] = [variance_inflation_factor(
    x_data1.values, i) for i in range(x_data2.shape[1])]
vif["features"] = x_data2.columns
```

1,2 에서 이용한 x_data1 을 이용하여 VIF 벡터를 추출하고 다중공산성을 체크할 수 있다. VIF 값이 1 에 가까울 수록 다중공산성의 문제가 낮다고 판단할 수 있다.

다중 공산성이 있는 경우에는 회귀 계수의 추정이 불안정하므로, 다중공산성 체크를 이용하여 강한 상관관계가 있는 feature 은 제거하고 다시 회귀분석을 하는 것이 도움이 된다.

3.프로그램의 전체적인 구조

[1]프로그램구조:

1-1) 데이터:

- Company1000.xlsx : 잡플래닛에서 기업 1000개의 이름과, 기업 고유 id를 담은 데이터
- jobPlanetData700.xlsx : 잡플래닛에서 크롤링한 원본 데이터
- after_preprocess.xlsx : data_preprocessing.py를 통해 원본 데이터를 전처리하여 얻은 데이터 ->. algorithm.py의 분석 대상파일

1-2) .py 파일

- crolling.py : 잡플래닛을 크롤링하기 위한 파이썬 코드
- data_preprocessing.py : 크롤링 이후 데이터를 전처리하기 위한 코드
- algorithm.py : 다중회귀분석, 데이터 시각화, 다중공산성 검사를 진행하기 위한 알고리즘 파일

1-3) 결과 파일

- regression_results_clean1.xlsx:, 2: 다중회귀분석모델의 summary 정보가 담긴 엑셀파일

(p-value, correlation coefficient value 등)

-산점도1.png, 산점도2.png : 산점도 시각화 이미지

-상관관계_히트맵1.png,, 상관관계_히트맵2.png, : 상관계수행렬 히트맵 시각화 이미지

[2]프로그램 특징:

1.잡플래닛의 데이터 총 700개를 8개의 측면에서 직접 크롤링 했다는 점이 가장 큰 특징이다.

2. 다중 회귀분석을 진행하고, correlation coefficient 행렬과 산점도를 통한 시각화를 이용하여 target 변수와 나머지 변수들 사이의 관계를 포착하였다.

3. 단순이 초기 다중회귀분석을 진행하는 것에 그치치 않고, p-value값을 관찰하고, VIF 값을 관찰하여 다중공산성이 있는 데이터를 제거하고 다중회귀분석을 다시 진행하였다는 점에서 해당 데이터와 관찰의 유효성을 검증하였다.

[3]필요한 환경 설정

사용된 라이브러리:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns  # heatmap 만들기 위한 라이브러리
import matplotlib.font_manager as fm
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

```
import pandas as pd
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
import time
from selenium.webdriver.common.by import By
```

```
usr = "kjeiun@yonsei.ac.kr"
pwd = "010525kk!!"
```

: 잡플래닛 아이디와 비밀번호

4.데이터분석 결과

1. 최초시행 target_column = “업무와 삶의 균형(워라밸)”

[다중회귀분석 Summary 표]

	coef	std err	t	P>|t|	[0.025	0.975]
const	**2.5454**	0.234	10.870	0.000	2.085	3.005
복지 및 급여	**0.2082**	0.045	4.612	0.000	0.119	0.297
사내문화	**0.5362**	0.069	7.786	0.000	0.401	0.671
승진 기회 및 가능성	**-0.3231**	0.062	-5.226	0.000	-0.445	-0.202
경영진	**-0.1287**	0.072	-1.779	0.076	-0.271	0.013
퇴사율	**0.0549**	0.079	0.692	0.490	-0.101	0.211

- 해당 표에 대한 관찰을 통해 target value 에 대하여 상관계수가 가장 큰 변수는 '사내문화' 임을 알 수 있었다. 복지 및 급여라고 예상한 가설과는 달리, 실제로 워라밸에 가장 큰 영향을 미치는 요인은 사내문화였다. 특히 다른 변수들의 상관계수값이 0.33 미만인 것을 고려하였을 때, 꽤나 큰 상관계수 값을 가지는 것을 확인할 수 있다.
- 또한 1 번가설에서 퇴사율을 제외한 나머지 변수들은 양의 상관관계를 가질 것이라 판단했는데, 이는 예측과 동일했다.

- 그러나 퇴사율의 경우에 음의 상관관계를 예측했으나 실제로는 0.05 의 매우 작은 양의 상관관계가 관측된것을 볼 수 있다.

-p-value 값이 0 에 가까운것으로 보아 꽤 유효한 다중회귀분석일 것임을 예측할 수 있다. 다만 퇴사율의 경우에 0.5 에 해당하기 때문에 퇴사율과 워라밸 사이에는 큰 상관관계가 없다고 생각할 수 있다. 실제 상관계수 값도 매우 작게 나온것을 관찰할 수 있다.

[VIF value]
:더 좋은 회귀분석을 위하여 다중공산성 체크를 진행하고 특정 feature 을 제거한 이후에 다시 진행하였다.

	VIF Factor	features
0	154.079468	const
1	1.060240	복지 및 급여
2	1.953435	사내문화
3	2.018108	승진 기회 및 가능성
4	2.895297	경영진
5	1.014703	퇴사율

경영진의 경우 다중공산성이 높게 나오는 요소이기 때문에 이것은 제거하고 다시 다중선형회귀를 진행하였다.

2. 다중공산성이 높은 데이터(경영진)는 제거하고 다시 분석 실행

	coef	std err	t	P>|t|	[0.025	0.975]
const	**2.5945**	0.233	11.135	0.000	2.137	3.052
복지 및 급여	**0.2042**	0.045	4.520	0.000	0.115	0.293
사내문화	**0.4675**	0.057	8.182	0.000	0.355	0.580
승진 기회 및 가능성	**-0.3849**	0.051	-7.507	0.000	-0.486	-0.284

퇴사율	**0.0628**	0.079	0.791	0.429	-0.093	0.219

-경영진 별점 데이터는 제거하고 다시 데이터를 분석해본 결과 다중회귀 분석의 값은 다음과 같았다. 대체로 target_value(워라밸)과 비슷한 상관계수값을 가지는 것을 확인해 볼 수 있다.

[비교]

1. [VIF Value]

경영진 데이터를 제거하고 나니 확실하게 다중공산성문제가 해결된 것을 확인할 수 있다.

[제거 전]

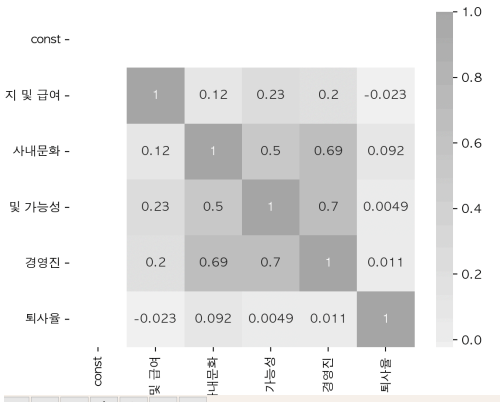
	VIF Factor	features
0	154.079468	const
1	1.060240	복지 및 급여
2	1.953435	사내문화
3	2.018108	승진 기회 및 가능성
4	2.895297	경영진
5	1.014703	퇴사율

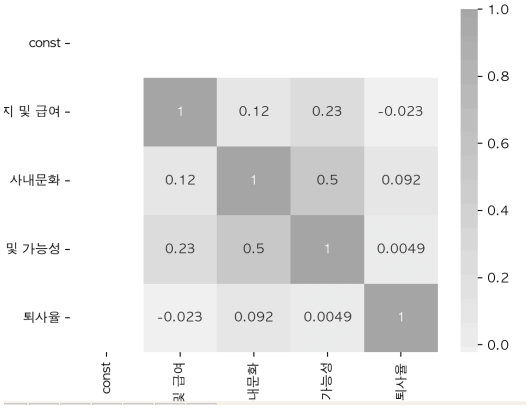
[제거 후]

	VIF Factor	features
0	151.933328	const
1	1.057619	복지 및 급여
2	1.339110	사내문화
3	1.382024	승진 기회 및 가능성
4	1.011492	퇴사율

2. 상관계수 히트맵

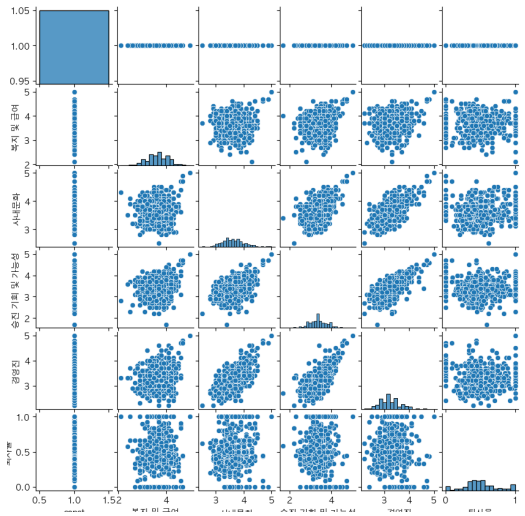
히트맵 또한 변수들 사이의 상관관계 값이 작아져서 다중공산성이 많이 제거된 것을 확인할 수 있었다. (앞이 경영진데이터 제거전, 뒤가 제거 후)

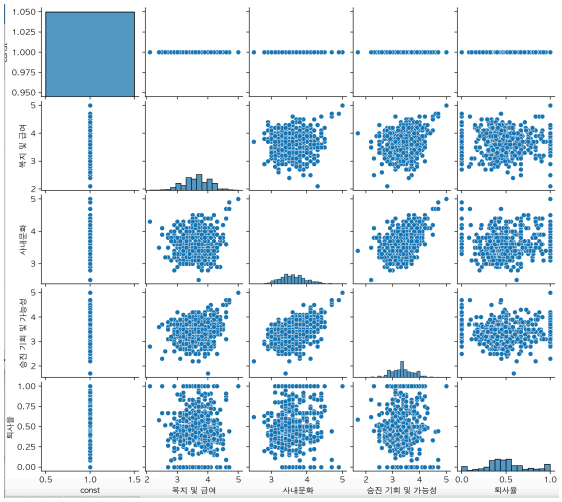




3.산점도

최대한 둥근 모양인 것이 상관관계가 적다는 것 -> 다중공산성이 낮다는 것
확실히 앞의 산점도 그래프에 비하여 뒤에있는 산점도 그래프에서 변수들 사이에상관관계가 덜 두드러지는 것을 확인할 수 있었다.





결론

1. 퇴사율을 제외한 나머지 데이터들의 경우에는 워라밸과 양의 상관관계를 가진다는 가설이 맞았다.

2. 워라밸에 가장 큰 영향을 미치는 요인이 복지 및 급여라고 예상하였는데, 실제로는 사내 문화가 가장 큰 상관관계를 가지고 있었다. 이는 워라밸이라는 개념이 근로시간이나 급여 복지 이외에도 사내문화에까지 확장되는 개념인 것과도 관련이 있는 것 같다. 사실 데이터 분석 프로젝트를 진행하기 전에 잡코리아에서 진행한 설문조사에서 직장분위기가 가장큰 요소로 꼽혔는데, 실제 데이터 분석에서도 그러한 결과가 나와 사내 분위기가 직장인들에게 미치는 심리적인 중요성이 커졌다고 생각한다. (잡코리아 관련 기사 링크 : https://www.jobkorea.co.kr/goodjob/tip/view?News_No=21586&schCtgr=120002&schTxt=%EC%95%84%EC%A7%81%EB%8F%84%EC%A7%81%EC%9E%A5%EC%9D%B8&Page=1&Tip_Top=1)

3. 마지막으로 퇴사율과 워라밸 사이에는 음의 상관관계가 있을 것이라고 가정을 했었는데, 실제로 퇴사율과 워라밸 사이에는 큰 상관관계가 존재하지 않는 것으로 보였다. 이는 예측컨데 워라밸의 정도가 퇴사에까지는 큰 영향을 미치지 못한다고 판단할 수 있었다.