

REPORT

데이터 사이언스



과 목 명 : 데이터 사이언스

담당교수 : 오세종

학 과 : 산업보안학과

학 번 : 32233421

이 름 : 이은지

제 출 일 : 2024.12.15

요약

분석 주제	업종별 산재 위험 현황은 지역별 부동산 실거래가에 영향을 미치는가?
데이터셋	1. 근로복지공단_업종별 산재신청 승인현황 2. 인천광역시 미추홀구_동별 사업체 현황 3. 인천광역시 미추홀구 아파트(매매) 실거래가
분석도구	Seaborn라이브러리 막대그래프, 히트맵, 버블차트, 산점도, 상관계수
분석내용 요약	1. 제조업, 운수·창고 및 통신업이 제일 산재위험도가 높다 2. 지역별 가장 많은 사업체는 건설업, 운수·창고 및 통신업, 제조업이다 3. 산재 위험 점수와 평균거래금액 간 약한 음의 상관 관계존재
결론 요약	산재 위험이 높은 산업군이 많은 동에서는 평균 거래 금액이 낮은 경향이 약하게 나타난다. 고로, 업종별 산재 위험 현황은 지역별 부동산 실거래가에 약하게 영향을 미친다고 할 수 있다.
이 분석의 장점	다양한 데이터 셋을 통합해 산업군별 사업체 현황, 산재 승인 건수, 부동산 거래등의 변수를 모두 포함한 분석을 진행해 여러 관점에서 문제를 분석했고, 상관계수로 산재위험과 평균거래금액간의 차이가 숫자적으로 유의미한지 검증하여 분석결과의 신뢰성을 높였다.

분석 주제 및 데이터 셋 소개

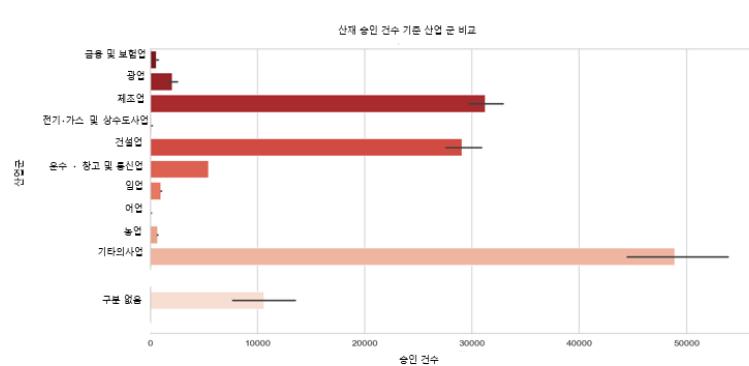
업종별 산재 위험 현황은 지역별 부동산 실거래가에 영향을 미치는가? 가 분석 주제이다.

데이터셋은, 첫번째, 근로복지공단_업종별 산재 신청 승인 현황이다. 업종별로 한 해에 얼마나 산재 신청이 들어왔고, 얼마나 승인됐는지 포함되어 있다. 두번째, 인천광역시 미추홀구_동별 사업체 현황으로 인천 광역시 미추홀구의 각 동별 어떤 사업체 들이 얼마나 있는지 현황이 포함되어 있다. 세번째, 인천광역시 미추홀구 아파트(매매) 실거래가이다. 인천광역시 미추홀구의 아파트 (매매) 실거래가가 어떻게 되는지 주소, 번지, 단지명, 층, 가격등이 포함되어 있다.

탐색적 데이터 분석 내용

1. 산재 위험도가 높은 산업군(산재 신청 승인 건수) 파악 - 기타 제외하고 제조업, 건설업,

운수·창고 및 통신, 광업 순으로 산재 신청 승인 건수가 높은 결과가 나왔다.



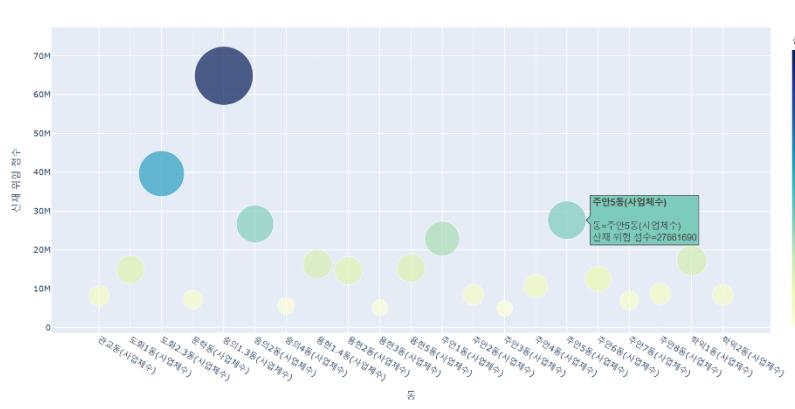
산재 신청 승인 건수가 많다는 것은 그만큼 산재가 많이 일어난다는 뜻이므로 이것들은 산재 위험도가 높은 산업군이라고 볼 수 있다.

지역별 사업체 현황								
	농업	공업	제조업	전기·가스·화생수·도·상업	건설업	운수·창고 및 운송업	금융 및 보험업	기타의 사업
승인1.3동(사업체수)	0	0	203	0	1861	810	5	67
승인2동(사업체수)	0	0	58	0	736	317	10	65
승인3동(사업체수)	0	0	33	1	71	109	5	42
총점1.4동(사업체수)	0	0	64	0	139	847	7	96
제1~4동(사업체수)	0	0	26	0	358	224	3	58
총현1.2동(사업체수)	0	0	41	0	70	20	7	42
총현2.3동(사업체수)	0	0	69	0	151	320	17	151
총현3.4동(사업체수)	0	0	98	1	196	596	13	103
총현4.5동(사업체수)	0	0	41	1	105	46	17	90

2. 지역별 사업체 현황

대개의 지역에서 건설업, 운수창고 및 통신업, 제조업이 가장 많은 사업체 수를 차지하고 있는 것을 알 수 있다. 이것으로 보아, 이 산업군들이 지역 경제 활동의 중요한 축을 담당하고 있다는 것을 알 수 있으며 많은 지역에서 물류, 제조업등이 활발하게 운영되고 있음을 알 수 있다.

동별 산재 위험 점수 (버블 차트)



3. 각 동별 산재 위험 점수

버블차트

- 각 동의 산업체별 위험

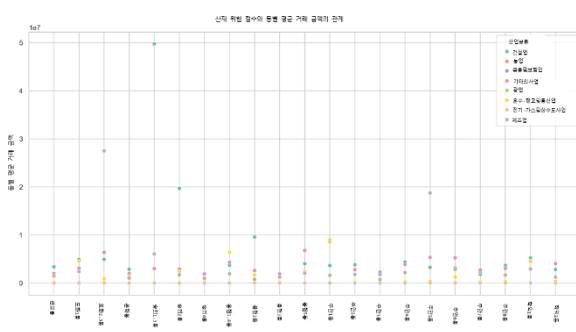
점수를((산재 위험

점수 = 산업체수 * 산재 승인 수)

를 모두 합해 각 동의 총

산업체별 위험점수를 버블 차트로
표시)

승인 1.3동이 버블 크기가 가장 크고, 그 뒤를 이어 도화 2.3동, 주안 5동 등이 크다. 이것은 이 동들이 산재 위험 점수가 크다는 것으로, 산재 위험 점수가 크다는 것은 이 동의 사업체 활동이 활발하다는 것을 의미 할 수 있다. 반대로, 버블 크기가 작은 동인 관교동과 같은 동은 산재 위험점수가 낮다. 이것은 상대적으로 산업체활동이 적다는 뜻이기도 하지만 동시에, 안정적인 산업체군이 주를 이루는 지역이라고도 볼 수 있다.



4.동과 동별 평균 거래 금액

도화 2.3동과 승인 2동이 가장 평균 거래 금액이 높은 것을 알 수 있다. 이것은 이곳에 평균 임금이 높은 사람들이 많이 산다는 것을 의미 할 수 있다.

```
>>> correlation = industry_count_per_dong[['산재 위험 점수']].corr(industry_count_per_dong['평균거래가격'])
>>> print(f"산재 위험 점수와 평균 거래 금액 간의 상관 계수: {correlation}")
산재 위험 점수와 평균 거래 금액 간의 상관 계수: -0.13599218417203693
```

5.산재 위험 점수와 평균 거래 금액 간의 상관 계수

산재 위험 점수와 평균 거래 금액 간의 상관 계수는 -0.13599218417203693으로 둘 사이에 약한 음의 상관 관계가 있다는 것을 의미 한다. 즉, 산재 위험 점수가 높을수록 평균 거래 금액이 낮은 경향이 있다는 뜻이다. 그러나 약한 상관 관계이므로, 다른 요인들(예: 지역의 상업적 발전, 교통망, 주거지 안정성 등)이 부동산 거래 금액에 더 큰 영향을 미칠 수 있다.

결론

이 분석을 통해 산재 위험 점수가 높은 산업군은 제조업과 운수·창고 및 통신업으로 나타났고, 대부분의 지역에서 건설업, 운수·창고 및 통신업, 제조업이 가장 많은 사업체 수를 차지하고 있음을 확인할 수 있었다. 또한, 산재 위험 점수와 평균 거래 금액 간에는 약한 음의 상관 관계가 나타나, 산재 위험이 높은 지역일수록 평균 거래 금액이 낮아지는 경향이 있으나, 그 영향은 크지 않음을 확인할 수 있었다. 이는 산재 위험이 부동산 가치에 미치는 영향이 제한적임을 나타낸다. 그렇기에, 지역별 부동산 가치에 영향을 미치는 다양한 외부 요인(예: 지역 개발, 인프라, 교통망 등)을 추가적으로 고려한 심층 분석이 필요하다 이를 통해 산업군 분포와 지역 경제 간의 관계를 보다 구체적으로 규명할 수 있을 것으로 기대된다..

소스파일

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import plotly.express as px

plt.rcParams['font.family'] = 'Malgun Gothic' #한글폰트변경
plt.rcParams['axes.unicode_minus'] = False #(-)부호 깨짐 방지

#1. 산재 위험도가 높은 산업군(산재 신청 승인 건수) 파악
#- seaborn라이브러리 막대그래프
highrisk=pd.read_csv('근로복지공단.csv')
highrisk.head()
sns.set_theme(style="whitegrid", rc={"figure.figsize":(5,5)})
sns.barplot(data=highrisk, x='승인', y='산재 업종별',
            hue='산재 업종별', palette='Reds_r')
plt.title('산재 승인 건수 기준 산업군 비교')
plt.xlabel('승인 건수')
plt.ylabel('산업군')
plt.show()
#지역별 산업체체
incheon = pd.read_csv('인천광역시 미추홀구_동별 사업체 현황.csv') # 지역별 사업체 현황 데이터
incheon.head()
# 2. 데이터 전처리
incheon['산업분류'] = incheon['산업분류'].str.strip()

# 인천 데이터에서 필요한 열 선택 (지역, 산업군, 사업체수)
incheon_selected = incheon[['산업분류', '송의1.3동(사업체수)', '송의2동(사업체수)', '송의4동(사업체수)', '용현1.4동(사업체수)',
                            '용현2동(사업체수)', '용현3동(사업체수)', '용현5동(사업체수)', '학의1동(사업체수)', '학의2동(사업체수)']]
```

```

highrisk_sorted = highrisk.sort_values(by='승인', ascending=False)
high_risk_industries = highrisk_sorted.head(10)
merged_data = pd.merge(incheon_selected, high_risk_industries[['산재 입종별']], left_on='산업분류', right_on='산재 입종별', how='inner')

# 지역별 산업군의 사업체 수를 히트맵에 적합한 형태로 변환
heatmap_data = merged_data.drop(columns=['산업분류', '산재 입종별']).transpose()

plt.figure(figsize=(12, 8))
sns.heatmap(heatmap_data, annot=True, cmap="YlGnBu", fmt="d", cbar_kws={'label': 'abcd'})
plt.title("산재 위험도가 높은 산업군의 지역별 사업체 현황")
plt.xlabel("산업군")
plt.ylabel("지역")
plt.show()

```

```

realestate=pd.read_csv('인천광역시 미추홀구_아파트(매매)_실거래가.csv')
realestate_selected=realestate[['동', '가격격']]
```

high_risk_industries4 = highrisk_sorted.head(4) # 산업재해 높은 산업체상위4개

```

high_risk_industries4.head()

df_melted = incheon.melt(id_vars=['산업분류'],
                           value_vars=[col for col in incheon.columns if '동' in col],
                           var_name='동',
                           value_name='사업체수')
# 동별 산업군별 사업체수 합산
top_industries_per_dong = df_melted.groupby(['동', '산업분류'])['사업체수'].sum().reset_index()

```

```

realestate['가격'] = realestate['가격'].str.replace(",","").astype(float)
average_price_per_dong = realestate.groupby('동')['cost'].mean().reset_index()
print(average_price_per_dong)

# 산업군별 사업체 수
# 동별 산업군별 사업체 수 계산
industry_count_per_dong = top_industries_per_dong.groupby(['동', '산업분류'])['사업체수'].sum().reset_index()
print(industry_count_per_dong.head())

# 'highrisk'에서 산업군별 승인 건수만 추출 (산업분류, 승인)
industry_approval = highrisk[['산재 입종별', '승인']]

industry_count_per_dong = pd.merge(industry_count_per_dong, industry_approval, left_on='산업분류', right_on='산재 입종별', how='left')

print(industry_count_per_dong.head())
industry_count_per_dong['산재 위험 점수']=industry_count_per_dong['사업체수']*industry_count_per_dong['승인']
industry_count_per_dong

pivot_data = industry_count_per_dong.groupby('동')[['산재 위험 점수']].sum().reset_index()

```

```

plt.figure(figsize=(12, 8))

# 산점도: 동 이름과 산재 위험 점수, 평균 거래 금액
sns.scatterplot(data=industry_count_per_dong, x='동', y='산재 위험 점수', hue='산업분류', palette='Set2')

# 제목과 레이블 추가
plt.title('산재 위험 점수와 동별 평균 거래 금액의 관계')
plt.xlabel('동')
plt.ylabel('산재 위험 점수')

# x축 라벨 회전 (동 이름이 길어질 수 있으므로)
plt.xticks(rotation=90)

# 그래프 표시
plt.tight_layout()
plt.show()

correlation = industry_count_per_dong['산재 위험 점수'].corr(industry_count_per_dong['평균거래가격'])
print(f"산재 위험 점수와 평균 거래 금액 간의 상관 계수: {correlation}")

```

```

# 버블 차트 시작화: 동별 산재 위험 점수
fig = px.scatter(pivot_data,
                  x='동',
                  y='산재 위험 점수',
                  size='산재 위험 점수', # 버블 크기: 산재 위험 점수
                  color='산재 위험 점수', # 색상: 산재 위험 점수
                  color_continuous_scale='YlGnBu', # 색상 팔레트
                  hover_name='동', # 툴팁에 동 이름 추가
                  title='동별 산재 위험 점수 (버블 차트)',
                  size_max=60) # 버블 크기의 최대 값 설정

# 그래프 표시
fig.show()

```

```

average_price_per_dong
industry_count_per_dong['동']

#'동' 열에서 '(사업체수)'를 제거하고, 동일한 형태로 변환
industry_count_per_dong['동'] = industry_count_per_dong['동'].str.replace(r'\(.*)', '', regex=True).str.strip()
erged_data = pd.merge(industry_count_per_dong, average_price_per_dong, on='동', how='left')

# 결과 확인
print(erged_data.head())
# 결과 확인
#print(industry_count_per_dong['동'].unique())

industry_count_per_dong

```