프로젝트기반 빅데이터 분석 과정 5기

프로젝트 결과 보고서

2023. 03. 13

프로젝트 명	인구 공공데이터 분석을 통한 미래 예측	
개발 기간	2023. 02. 24 ~ 2023. 03. 10	
팀 장	김지수	jsjh5272@naver.com
팀 원	박상욱	probono411@naver.com
	신혜진	wls106505@naver.com
	전은수	dmstn7992@naver.com

목 차

Ⅰ 분석 결과 요약

1 분석 결과 개요

❍ 프로젝트 명: 인구 공공데이터 분석을 통한 미래 예측

❍ 개발 기간: 2023. 02. 24 ~ 2023. 03. 10

❍ 개발 목표: 0세와 총인구 관계를 이용하여 예측모델 개발

❍ 주요 분석 방법

- 상관분석
- 단순선형회귀분석

2 분석 결과 및 시사점

❍ 전국 시도, 읍면동 단위로 0세부터 101세이상 인원수 데이터를 가지고 다양한 시각화 표현을 할 수 있었고 0세(주민등록기준) 인구수(출생)가 총인구와의 관계를 살펴 보았음

일반적인 기준으로도 출생이 총인구수에 가장 큰 영향이 있다고 볼 수 있다. 실제 어느 정도인지를 상관 분석과 단순회귀분석을 통해서 결과를 확인하니 이변 없이 높은 상관관계가 있음을 확인 함

단순선형회귀분석을 통해 0세와 총인구의 통계적 유의미한 결과를 얻었고 98%의 높은 설명력을 가지고 있다는 결론을 얻음

최저 기온 변화에 따라 최고 기온도 같은 방향으로 변화가 있다는 것을 확인 할 수 있음

II 서 론

1 분석 개요

‘현대 경영학의 아버지’라고 불리는 피터 드러커는 ‘인구 통계의 변화는 미래와 관련된 것 가운데 정확한 예측을 할 수 있는 유일한 사실’이라고 했습니다. 그 만큼 인구 데이터는 다양한 인사이트를 제공합니다.

행정안전부(www.mois.go.kr)에서 다양한 조건으로 인구 데이터를 제공받을 수 있습니다. 인구 공공 데이터를 활용하여 다양한 데이터 분석을 진행하기로 함.

2 분석 개발 현황

❍ 업무 분장

과제 정의	- 요구사항 및 이슈 파악 - 수행 방안 설계	팀원 전원
데이터셋 선택	- 원시 데이터 확인	팀원 전원
데이터 전처리	- 데이터 재처리를 통한 데이터셋 정제 - 추가로 요구되는 데이터셋이 필요한 경우 데이터 선택 프로세스를 재실행	박상욱
데이터 변환	- 효율적인 데이터 분석을 위한 데이터 변경	신혜진
데이터 마이닝	- 데이터 분석 기법 및 실행 상관분석, 단순선형회귀분석을 이용한 변수들 간의 관계 확인	전은수
데이터 마이닝 결과 평가	- 결과에 대한 해석 및 평가 - 분석 목적과의 일치성 확인	팀원 전원
결과 보고	- 결과보고서 작성 - 최종 보고회	발표자 김지수

❍ 개발 일정 : 2023.02.24 ~ 2023.03.10

구 분	2월					3월									
	24	25	26	27	28	01	02	03	04	05	06	07	08	09	10
1. 분석대상 비즈니스의 이해(과제 정의)															
■ 요구사항 및 이슈 파악	•	•	•												
■ 과제 관련 개념 및 이론 정리	•	•	•												
■ 목적 정의 및 수행방안 설계			•	•											
■ 프로젝트 계획서 작성 및 제출				•	•	•									
2. 데이터셋 선택															
■ 데이터 수집 및 선택						•	•	•							
■ 목표 데이터를 구성						•	•	•							
3. 데이터 전처리															
■ 데이터셋 정제							•	•	•						
4. 데이터 변환															
■ 변수 생성 및 선택							•	•	•						
■ 데이터 변환							•	•	•						
5. 데이터 마이닝															
■ 단순선형회귀분석								•	•	•	•	•	•	•	
■ 데이터 시각화								•	•	•	•	•	•	•	
■ 예측모델								•	•	•	•	•	•	•	
6. 데이터 마이닝 결과 평가															
■ 결과에 대한 해석 및 평가														•	•
7. 결과 보고															
■ 결과보고서 작성													•	•	•
■ 최종보고회															

※ 세부 일정은 추진상황에 따라 변경 가능

3 데이터 수집 개요

❍ 활용 데이터

데이터명	출처	주요 항목 및 특징
주민등록인구통계	행정안전부	- 연령별 인구현황(0세 ~ 100세 이상 / 남, 여 구분)

III 데이터 수집 및 전처리

1 데이터 수집

❍ 데이터 수집

- 원시 데이터 수집

1) 행정안전부_주민등록 인구통계(정책자료 - 통계)

https://jumin.mois.go.kr/#

❍ 데이터 불러오기

- 연령별 인구현황(0세 ~ 100세 이상) : 2023년 01월 기준 / age.csv

```
[ ] import csv
    import pandas as pd

    # 구글 드라이브 연동 하기
    from google.colab import drive
    drive.mount('/gdrive', force_remount = True)

    df = pd.read_csv('/gdrive/My Drive/age.csv', encoding = 'cp949')
    df.head()
```

```
Mounted at /gdrive
```

	행정구역	2023년01월_계_총인구수	2023년01월_계_연령구간인구수	2023년01월_계_0세	2023년01월_계_1세	2023년01월_계_2세	2023년01월_계_3세	2023년01월_계_4세	2023년01월_계_5세	2023년01월_계_6세	...	2023년01월_계_91세	2023년01월_계_92세	2023년01월_계_93세
0	서울특별시 (1100000000)	9,424,873	9,424,873	40,241	43,826	43,908	47,646	50,383	54,606	61,607	...	7,987	6,515	5,430
1	서울특별시 종로구 (1111000000)	141,223	141,223	436	491	469	562	536	622	729	...	165	134	108
2	서울특별시 종로구 청운효자동 (1111051500)	11,600	11,600	40	48	50	52	53	64	85	...	16	8	7
3	서울특별시 종로구 사직동 (1111053000)	9,131	9,131	21	41	30	41	41	52	71	...	13	10	16
4	서울특별시 종로구 삼청동 (1111054000)	2,325	2,325	4	4	4	7	5	15	8	...	4	3	5

5 rows × 104 columns

- 연령별 인구현황(구분 : 계) : age.csv

- 연령별 인구현황(0세 ~ 100세 이상) : 2023년 01월 기준 / gender.csv

```
[20]  import csv
      import pandas as pd

      # 구글 드라이브 연동 하기
      from google.colab import drive
      drive.mount('/gdrive', force_remount = True)

      df = pd.read_csv('/gdrive/My Drive/gender.csv', encoding = 'cp949')
      df.head()
```

```
Mounted at /gdrive
```

	행정구역	2023년01월_남_총인구수	2023년01월_남_연령구간인구수	2023년01월_남_0세	2023년01월_남_1세	2023년01월_남_2세	2023년01월_남_3세	2023년01월_남_4세	2023년01월_남_5세	2023년01월_남_6세	...	2023년01월_여_91세	2023년01월_여_92세	2023년01월_여_93세
0	서울특별시 (1100000000)	4,567,739	4,567,739	20,597	22,518	22,568	24,815	26,116	27,892	31,517	...	5,814	4,858	4,031
1	서울특별시 종로구 (1111000000)	68,314	68,314	216	271	234	275	276	311	376	...	109	98	74
2	서울특별시 종로구 청운효자동 (1111051500)	5,327	5,327	16	25	28	27	23	40	46	...	11	6	5
3	서울특별시 종로구 사직동 (1111053000)	4,064	4,064	12	23	13	20	20	28	36	...	10	9	10
4	서울특별시 종로구 삼청동 (1111054000)	1,110	1,110	2	2	1	3	2	5	6	...	3	1	3

5 rows × 207 columns

- 연령별 인구현황(구분 : 남/여) : gender.csv

2 데이터 전처리

❍ 데이터 셋 확인

- 데이터의 행, 열의 개수 및 결측치 개수 등을 확인

1) 연령별 인구현황(0세 ~ 100세 이상) : 2023년 01월 기준 / age.csv

[구분 : 계]

```
[22]  import csv
      import pandas as pd

      # 구글 드라이브 연동 하기
      from google.colab import drive
      drive.mount('/gdrive', force_remount = True)

      df = pd.read_csv('/gdrive/My Drive/age.csv', encoding = 'cp949')
      df.info()

      Mounted at /gdrive
      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 3874 entries, 0 to 3873
      Columns: 104 entries, 행정구역 to 2023년01월_계_100세 이상
      dtypes: int64(1), object(103)
      memory usage: 3.1+ MB
```

2) 연령별 인구현황(0세 ~ 100세 이상) : 2023년 01월 기준 / gender.csv

[구분 : 남/여]

```
[23]  import csv
      import pandas as pd

      # 구글 드라이브 연동 하기
      from google.colab import drive
      drive.mount('/gdrive', force_remount = True)

      df = pd.read_csv('/gdrive/My Drive/gender.csv', encoding = 'cp949')
      df.info()

      Mounted at /gdrive
      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 3874 entries, 0 to 3873
      Columns: 207 entries, 행정구역 to 2023년01월_여_100세 이상
      dtypes: int64(8), object(199)
      memory usage: 6.1+ MB
```

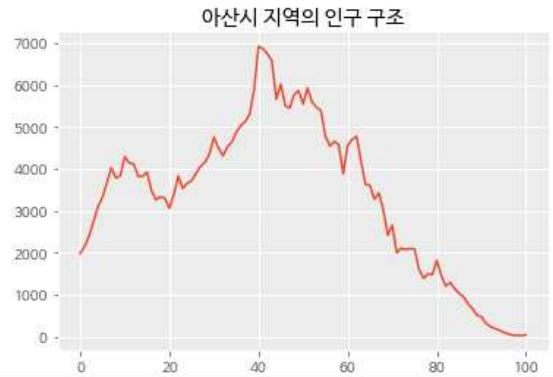
❍ 탐색적 자료 분석

- 시각화를 이용한 탐색적 자료 분석 시행

1) 특정 지역의 인구 구조

=> 행정구역명을 입력 받아 해당 지역의 0세부터 100세 이상의 인구 구조를 시각화 할 수 있음

```
[24] import csv
     import matplotlib.pyplot as plt

     # 구글 드라이브 연동 하기
     from google.colab import drive
     drive.mount('/gdrive', force_remount = True)

     # 내 드라이브 루트(최상위)에 업로드 필요
     f = open('/gdrive/My Drive/age.csv', 'r', encoding = 'cp949')
     data = csv.reader(f, delimiter = ',')

     result = []  # 빈 리스트 - 시각화 위해 자료를 리스트에 넣어줌
     name = input('인구 구조가 알고 싶은 지역의 이름(읍면동 단위)을 입력해 주세요 :')

     for row in data:
       if name in row[0]:
         for i in row[3:]:  # 0세 부터 101세 이상 자료만 추출
           result.append(int(i.replace(',', '')))  # int - 정수로 변환(시각화 위한 형변환)
         break

     plt.style.use('ggplot')  # 차트 스타일(격자 무늬)
     plt.rc('font', family = 'NanumbarunGothic')
     plt.title(name + ' 지역의 인구 구조')
     plt.plot(result)
     plt.show()

     Mounted at /gdrive
     인구 구조가 알고 싶은 지역의 이름(읍면동 단위)을 입력해 주세요 :아산시
```

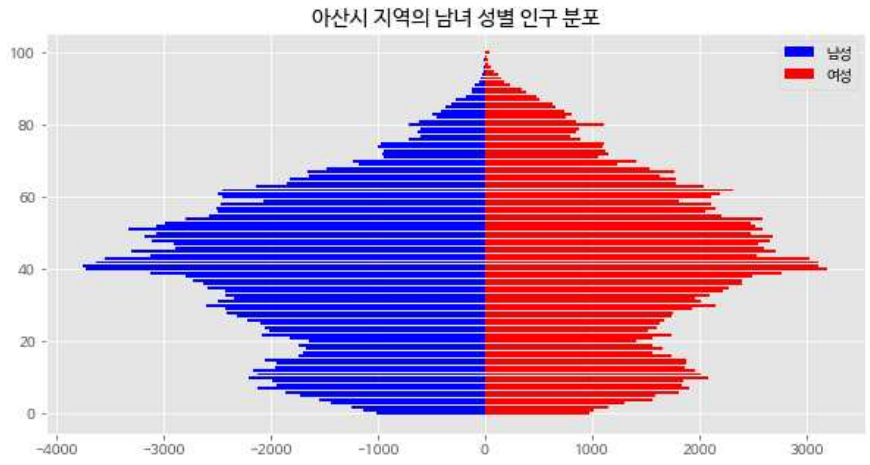


2) 남녀 성별 인구 분포 시각화

=> 항아리 형태의 시각화를 통해 남녀 인구 분포를 확인 할 수 있음

```
[4]  import csv
     import matplotlib.pyplot as plt

     # 구글 드라이브 연동 하기
     from google.colab import drive
     drive.mount('/gdrive', force_remount = True)

     # 내 드라이브 루트(최상위)에 업로드 필요
     f = open('/gdrive/My Drive/gender.csv', 'r', encoding = 'cp949')
     data = csv.reader(f, delimiter = ',')

     m = []; f = []
     name = input('찾고 싶은 지역의 이름을 알려주세요 : ')

     for row in data:
       if name in row[0]:
         for i in row[3:104]:
           m.append(-int(i.replace(',', '')))
         for i in row[106:]:
           f.append(int(i.replace(',', '')))
         break

     plt.style.use('ggplot')
     plt.figure(figsize = (10, 5))  # 그래프 사이즈 조절하기
     plt.rcParams['axes.unicode_minus'] = False
     plt.rc('font', family = 'NanumbarunGothic')
     plt.title(name + ' 지역의 남녀 성별 인구 분포')
     plt.barh(range(101), m, label = '남성', color = 'blue')
     plt.barh(range(101), f, label = '여성', color = 'red')
     plt.legend()
     plt.show()
```

```
Mounted at /gdrive
찾고 싶은 지역의 이름을 알려주세요 : 아산시
```



- 탐색적 자료 분석을 통해 0세부터 100세 이상의 인구구조를 행정구역명을 입력하여 살펴 볼 수 있었고 남녀의 차이를 알 수 있는 항아리 형태의 시각화를 통해 비교할 수 있음, 20세 미만의 인구가 많을수록 30대~40대 인구구조가 높은 것을 알 수 있고 자녀와 부모의 차이 정도로 확인함.

❍ 데이터 정제

- 결측값 처리
  1) 행을 기준으로 가장 아래쪽 행 값이 비어 있는 데이터 확인
  2) 결측치가 전체 데이터 출력시 에러 발생하여 Excel 프로그램에서 csv 파일을 불러와 해당하는 행을 삭제 처리함
- 이상값 처리
  1) 이상치로 파악되는 값이 없음을 확인

Ⅳ 모델 구축 및 검증

1 모델 구축

1) 변수 선택

- 상관분석

  상관분석을 통해 0세와 총인구의 선형적인 상관관계를 보이는지 확인함.

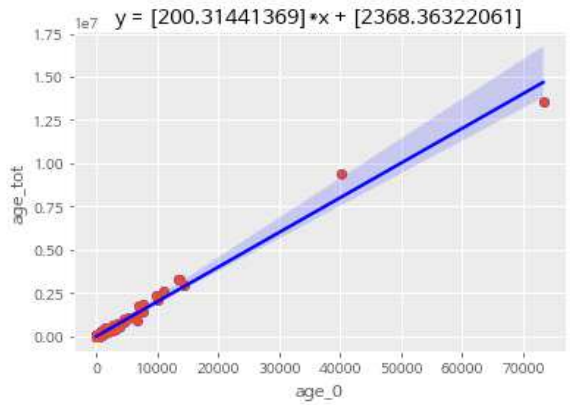
```
[12]  import csv
      import pandas as pd
      import numpy as np

      # 구글 드라이브 연동 하기
      from google.colab import drive
      drive.mount('/gdrive', force_remount = True)

      df = pd.read_csv('/gdrive/My Drive/age.csv', encoding = 'cp949', thousands = ',')
      df.rename(columns = {'2023년01월_계_총인구수' : 'age_tot'}, inplace = True)
      df.rename(columns = {'2023년01월_계_0세' : 'age_0'}, inplace = True)

      Y = df.age_tot.values
      X = df.age_0.values

      cov = (np.sum(X * Y) - len(X) * np.mean(X) * np.mean(Y)) / len(X)
      print(cov)

      print(np.cov(X, Y)[0, 1])

      corr = cov / (np.std(X) * np.std(Y))
      print(corr)

      print(np.corrcoef(X, Y)[0, 1])

      Mounted at /gdrive
      455017632.6269858
      455135117.16936314
      0.9922658208847495
      0.9922658208847498
```

- 단순선형회귀분석

  통계적 가설 검정에서 유의 확률을 보기 위해 회귀분석을 통해 p값(유의 확률)을 확인함.

```
[18]  import csv
      import pandas as pd
      from sklearn.linear_model import LinearRegression
      lr = LinearRegression()

      # 구글 드라이브 연동 하기
      from google.colab import drive
      drive.mount('/gdrive', force_remount = True)

      df = pd.read_csv('/gdrive/My Drive/age.csv', encoding = 'cp949', thousands = ',')
      df.rename(columns = {'2023년01월_계_총인구수' : 'age_tot'}, inplace = True)
      df.rename(columns = {'2023년01월_계_0세' : 'age_0'}, inplace = True)

      Y = df.age_tot.values
      print(Y)
      X = df.age_0.values
      print(X)

      X = X.reshape(-1,1)
      Y = Y.reshape(-1,1)

      lr.fit(X, Y)

      print(lr.coef_[0])
      print(lr.intercept_)

      import matplotlib.pyplot as plt
      import seaborn as sns
      plt.title('y = {}*x + {}'.format(lr.coef_[0], lr.intercept_))
      sns.regplot(x = 'age_0', y = 'age_tot', data = df, color = 'blue')
      plt.scatter(X, Y)
      plt.show()
```

```
Mounted at /gdrive
[9424873  141223   11600 ...   13748   12308    3887]
[40241   436    40 ...   101    68     8]
[200.31441369]
[2368.36322061]
```



회귀식 : Target(총인구) = 200.21441369 * 0세 + 2368.36322061

2) 모델 구축

- 회귀분석(Regression Analysis)

상관분석은 변수들이 서로 얼마나 밀접하게 직선적인 관계를 가지고 있는지를 분석하는 통계적 기법이며, 회귀분석은 한 개 또는 그 이상의 변수들(독립변수)에 대하여 다른 변수(종속변수) 사이의 관계를 수학적인 모형을 이용하여 설명하고 예측하는 분석기법 입니다.

상관분석에서는 산점도의 점들의 분포를 통해 일정한 패턴을 확인한 후, 상관계수를 구하여 두 변수 간의 선형 관계를 알 수 있습니다.

여기서 더 나아가, 이 일정한 패턴을 활용하여 무엇인가를 예측하는 분석인 회귀분석을 사용할 것임.

2 모델 검증

1) 모델 검증 방법

- 0세와 총인구는 상관분석을 통해 도출된 선형 상관관계를 보이는 변수, 단순선형회귀분석에서 통계적으로 유의한지 변수가 유의하게 영향을 미치는 지 그리고 얼마만큼의 설명력을 가지는 등의 여부를 확인함

2) 모델 평가 과정

- 상관분석

```
[13]  import scipy.stats as stats
      stats.pearsonr(X, Y)

      (0.9922658208847497, 0.0)
```

=> scipy 패키지의 stats.pearsonr()을 이용하면 상관계수와 p-value를 동시에 얻을 수 있습니다.
뒤 결과 값이 p-value인데, 귀무가설 "상관관계가 없다"에 대한 검정 결과 p-value가 0.0라는 0이 나왔으므로 귀무가설을 기각할 수 있음을 알 수 있습니다.

- 단순선형회귀분석

회귀식 : Target(총인구) = 200.21441369 * 0세 + 2368.36322061

```
[17] import statsmodels.api as sm
     results = sm.OLS(Y, sm.add_constant(X)).fit()

     results.summary()
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.985
Model:	OLS	Adj. R-squared:	0.985
Method:	Least Squares	F-statistic:	2.474e+05
Date:	Thu, 02 Mar 2023	Prob (F-statistic):	0.00
Time:	06:46:36	Log-Likelihood:	-46326.
No. Observations:	3874	AIC:	9.266e+04
Df Residuals:	3872	BIC:	9.267e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2368.3632	612.209	3.869	0.000	1168.081	3568.646
x1	200.3144	0.403	497.411	0.000	199.525	201.104

Omnibus:	5964.882	Durbin-Watson:	1.519
Prob(Omnibus):	0.000	Jarque-Bera (JB):	68406211.814
Skew:	8.345	Prob(JB):	0.00
Kurtosis:	653.775	Cond. No.	1.53e+03

A. F-statistic : 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 의미가 있는지 파악

=> F-statistic의 p-value 값은 Prob(F-statistic)으로 표현되는데, 이는 0.00으로 0.05보다 작기에 이 회귀식은 회귀분석 모델 전체에 대해 통계적으로 의미가 있다고 볼 수 있습니다.

B. P-Value : 각 변수가 종속변수에 미치는 영향이 유의한지 파악

=> 중간쯤에 보면 coef와 변수 x1의 p-value 값이 나와있습니다. 여기서 x1은 0세이고 이 변수의 p-value가 0.000으로 표기 되어 있기에 0.05보다 작으므로 Target을 설명하는데 유의하다고 판단할 수 있습니다.

C. 수정된 R제곱 : 회귀직선에 의하여 설명되는 변동이 총변동 중에서 차지하고 있는 상대적인 비율이 얼마인지 나타냄

즉, 회귀직선이 종속변수의 몇%를 설명할 수 있는지 확인

=> 제일 위 부분에 R-squared와 Adj. R-squared가 표기되어 있는데, 값이 0.985정도로 이는 98%만큼의 설명력을 가진다고 판단할 수 있습니다. 참고로, 0에 가까울 수록 예측값을 믿을 수 없고 1에 가까울 수록 믿을 수 있다고 보면 됩니다.

V 결론

1 분석 결과

- 변수 중(행정구역명, 총인구수, 0세 인구수 ~ 100세 이상 인구수 등) 0세와 총인구의 관계를 살펴보았으며 상관 분석을 통해서 높은 상관관계가 있음을 확인 하였다. 회귀 분석을 통해 회귀식이 도출 되었고 총인구 1명 올라가는데 0세 인구수 값이 상당한 영향을 주고 있는 것도 확인이 되었다.

2 시사점 및 개선점

1) 시사점

- 일반적으로 출생과 관련된 0세 인구수 변화에 따라 총인구의 변화가 높다는 것은 어느 정도 예측을 할 수 있다. 수치로 계산을 했을 때 크기 정도가 궁금했고 실제 결과를 보았을 때 큰 영향을 주고 있다는 것은 확인할 수 있었음

2) 개선점

- 행정안전부 제공 자료를 조건에 따라 다양한 변수 값을 이용할 수 있다. 다양한 변수의 관계를 가지고 단순선형회귀분석에서 확장된 다중선형회귀분석 이용해 여러 변수 간의 관계 분석이 필요함

VI 산출물 목록

1 산출물 목록

목록 코드	목록 명	작성일자	작성자	비고
CNR-001	age.csv(csv파일)	23/03/10	박상욱	https://drive.google.com/file/d/1TyJIZt4YOolMl2EiUw4Lsqd_-XDqovmN/view?usp=sharing
CNR-002	gender.csv(csv파일)	23/03/10	신혜진 전은수	https://drive.google.com/file/d/1fGdUt_jzJUd8S7hrZDbvbjTjYO-srmtV/view?usp=sharing
CNR-003	단순선형회귀(ipynb파일)	23/03/10	김지수	https://colab.research.google.com/drive/1JHezQWsv11Mm1Jg77N6BJIJ_UOPPhT21?usp=sharing