제주 지역의 경제활동인구와 연간 소득의 관계 분석

2019108280 컴퓨터공학전공 한인철

목차

개요 및 필요성

- 소득 예측의 중요성
  - 소득 수준에 관계없이 물가는 일정해야 하지만,
    소득 수준이 낮을 수록 물가도 낮게 형성됨

    하지만, 사회 현상의 변수로 인하여 항상 일정한 결과를 도출하지 못함

소득 수준을 예측할 수 있으면, 비용 산정 계획에 영향을 줄 수 있으며,
효과적인 매출 및 수익에 이점을 얻을 수 있음

구매력평가환율이론 및 경제 예측의 필요성 - Link

관련 연구

- **Prediction on the Economic Activity Level of the Elderly in South Korea**
  - 고령화 시대를 문제점으로 삼아, 고령층의 경제활동 수준을 머신 러닝 기법을 활용하여 예측함

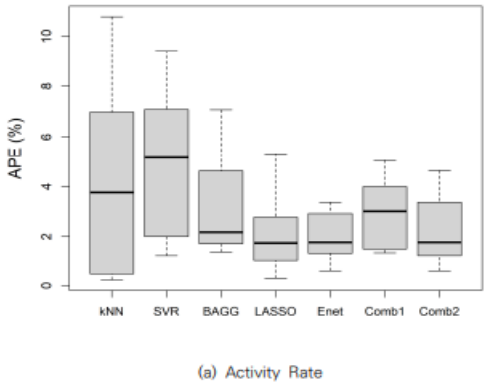
Table 2. MAPE comparison

(Unit: %)

Var.	kNN	SVR	BAGG	LASSO	Enet	Comb1	Comb2
AR	4.115	4.781	3.148	2.449	2.261	2.981	2.488
	(3.613)	(2.895)	(1.951)	(2.076)	(1.481)	(1.360)	(2.003)
ER	1.282	0.979	1.248	0.971	0.968	0.878	0.971
	(1.351)	(1.353)	(0.733)	(0.883)	(0.929)	(0.942)	(0.883)
EPR	4.500	5.651	3.943	3.135	2.833	3.538	3.045
	(3.856)	(2.740)	(3.297)	(2.491)	(2.148)	(2.122)	(2.464)

* AR(Activity Rate), ER(Employment Rate), EPR(Employment-Population Ratio), standard deviations in parentheses, 5-fold cross validation is used

Comparision ML



(a) Activity Rate

Box plot

데이터 시각화 – 데이터 전처리

```
import numpy as np
import plotly_express as px
import plotly.figure_factory as ff
import matplotlib.pyplot as plt
import plotly.graph_objects as go
from plotly.subplots import make_subplots
from sklearn.preprocessing import LabelEncoder
from statsmodels.compat import lzip
import statsmodels.api as sm
import pandas as pd
from sklearn.preprocessing import scale
from pandas import DataFrame
from sklearn.preprocessing import StandardScaler

from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
```

필요 라이브러리 모듈 호출

```
df = pd.read_csv('Korea Income and Welfare.csv')

df=df.drop(df[df.income <0].index)

# 제주 지역 분할
df.loc[df['region'] == 7, 'region'] = 'Jeju'

# 성별 분할
df.loc[df['gender'] == 1, 'gender'] = 'male'
df.loc[df['gender'] == 2, 'gender'] = 'female'

# 종교 분할
df.loc[df['religion'] == 1, 'religion'] = 'religious'
df.loc[df['religion'] == 2, 'religion'] = 'non-religious'

df = df[df["region"] == "Jeju"]
```

DataFrame 생성 및 전처리

데이터 시각화 – 데이터 전처리

```
# 소득 분할

df1=df.copy()
df1_males = df[df['gender']=='male']
df1_females = df[df['gender']=='female']
df1_males=df1_males.groupby(['year']).mean()
df1_males['year']=df1_males.index
df1_females=df1_females.groupby(['year']).mean()

df1_females['year']=df1_females.index
df2=pd.concat([df1_females, df1_males])
df1_males['income females']=df1_females['income']
df1_males['income males']=df1_males['income']
df1_males.year=df1_males.year.astype(int)
df=df[df['year']==2018]
df.reset_index(inplace=True)
df1_males['ratio females']=(df1_males['income females']/df1_males['income males']).round(2)
```

소득 및 성별 관계 데이터 분할

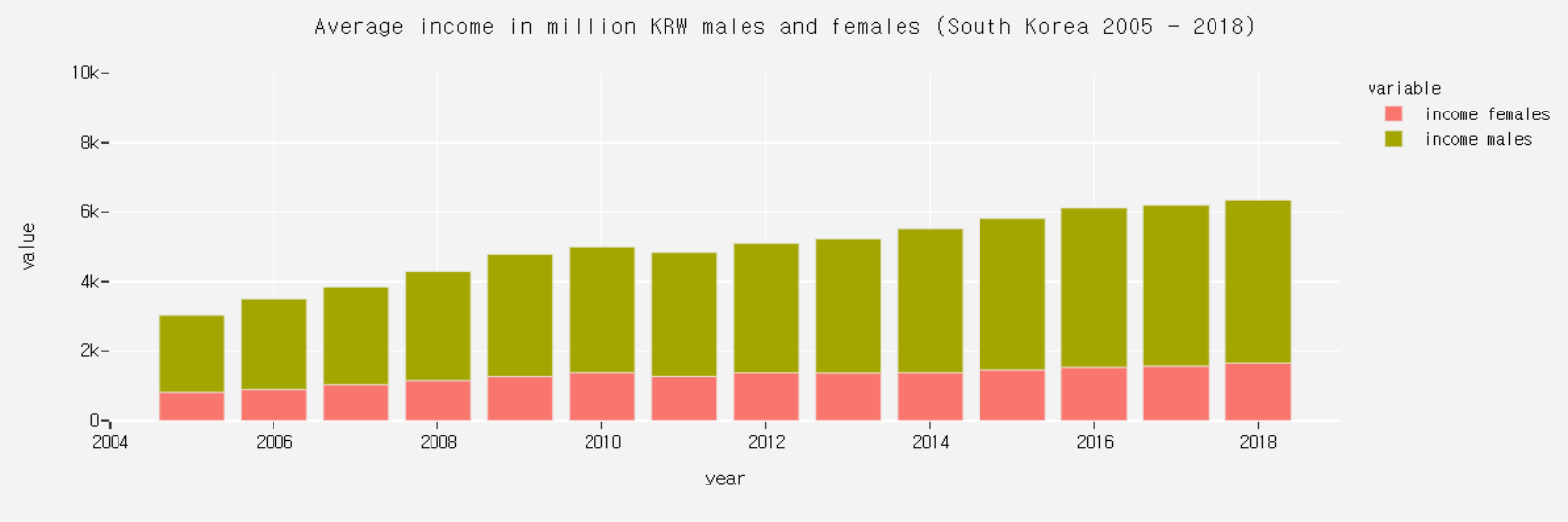
Log Income을 사용한 이유

수익 간의 계산을 용이하게
만들기 위해 Log 함수를 사용함

데이터 시각화

```
fig1 = px.bar(df1_males,x = 'year', y = ['income females','income males'],range_y=[0,10000],range_x=[2004,2019]
            ,height=400,template='ggplot2')

fig1.update_layout(paper_bgcolor='rgb(243, 243, 243)',plot_bgcolor='rgb(243, 243, 243)'
,title_text='Average income in million KRW males and females (South Korea 2005 - 2018)',
    font=dict(
        family="Century, monospace",
        color="black"
    ))

fig1.show()
```

연도가 지나감에 따라, 평균 소득도 증가하는 모습을 보이고 있음



연도별 성별에 따른 평균 소득

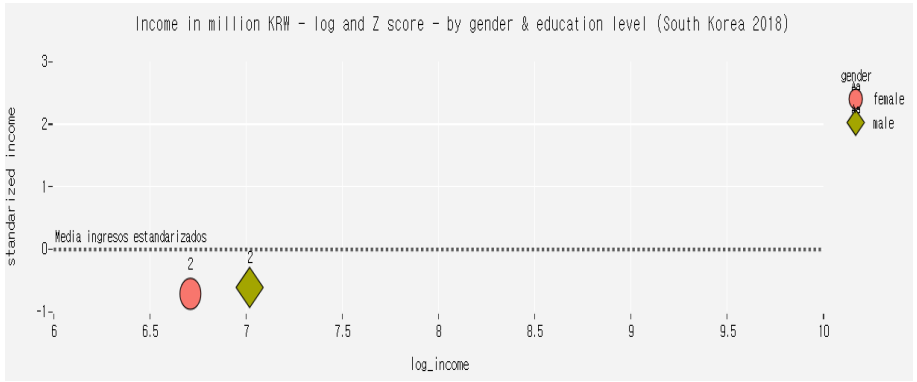
데이터 시각화

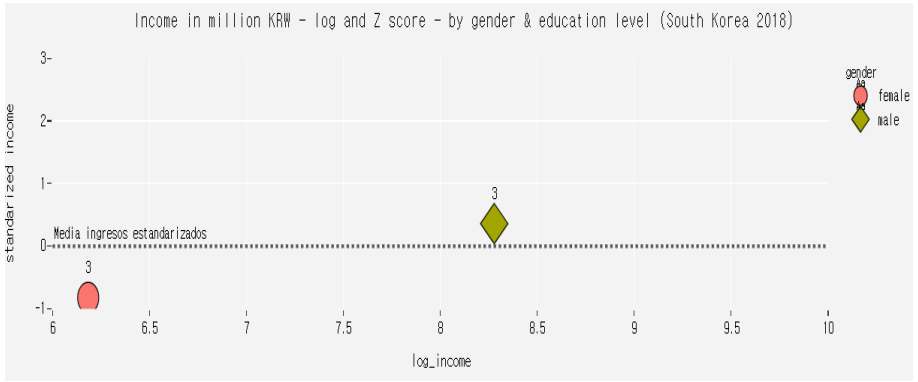
```
df.sort_values(by='education_level',inplace=True)
fig1 = px.scatter(df,x = 'log_income', y = 'standarized income',color='gender',animation_frame='education_level',symbol='gender',
                  animation_group='education_level',text='education_level', range_y = [-1,3], range_x= [6,10],hover_name ='gender',
                  height=400,template='ggplot2')

fig1.update_traces(dict(marker_line_width=1,marker_line_color="black",mode='markers'),
                   textposition='top center',marker_size=25)

fig1.update_layout(
    margin=dict(l=20, r=20, t=50, b=20),
    paper_bgcolor='rgb(243, 243, 243)',plot_bgcolor='rgb(243, 243, 243)'
    ,title_text='Income in million KRW - log and Z score - by gender & education level (South Korea 2018)',
    font=dict(
        family="Century, monospace",
        color="black"))


fig1.add_shape( # add a horizontal "target" line
    type="line", line_color="black", line_width=3, opacity=0.65, line_dash="dot",
    x0=0, x1=1, xref="paper", y0=0, y1=0, yref="y"
)

fig1.add_annotation(x=6.4, y=0.2,
            text="Media ingresos estandarizados",
            showarrow=False,
            yshift=0)

fig1.layout.updatemenus[0].buttons[0].args[1]["frame"]["duration"] = 2000

fig1.show()
```

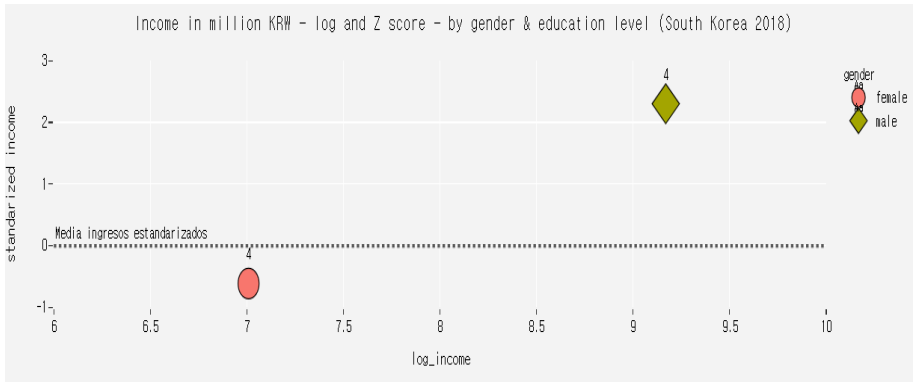
Education Level (졸업 기준)

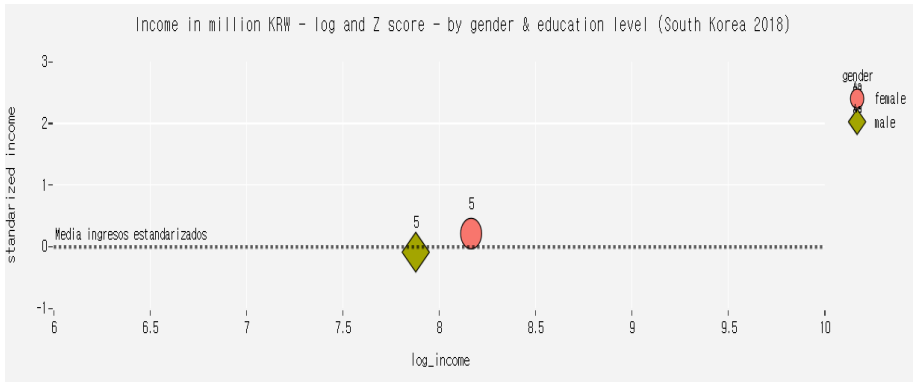
- 1 : 무 학력 (8세 이하)
- 2 : 무 학력 (8세 초과)
- 3 : 초등학교
- 4 : 중학교
- 5 : 고등학교
- 6 : 전문 대학 (2년제)
- 7 : 종합 대학 (4년제)
- 8 : MA (석사 학위)
- 9 : DA (박사 학위)

데이터 시각화





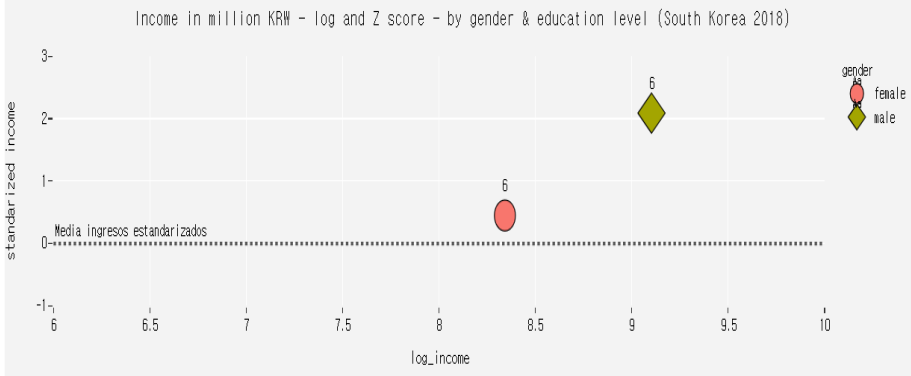


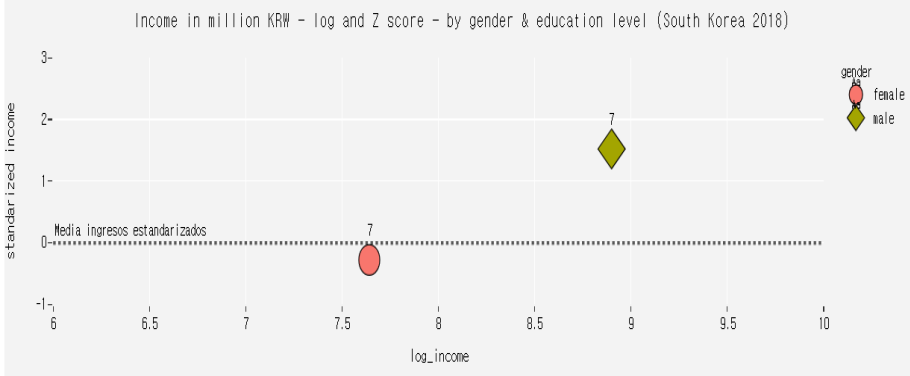


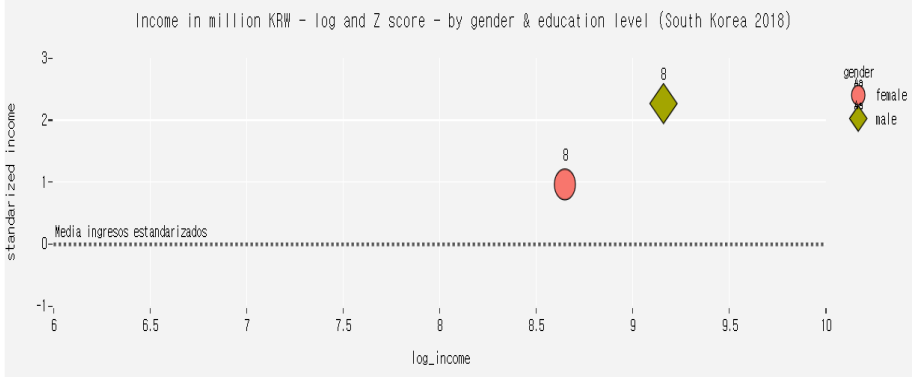
학업 수준 별 성별에 따른 소득

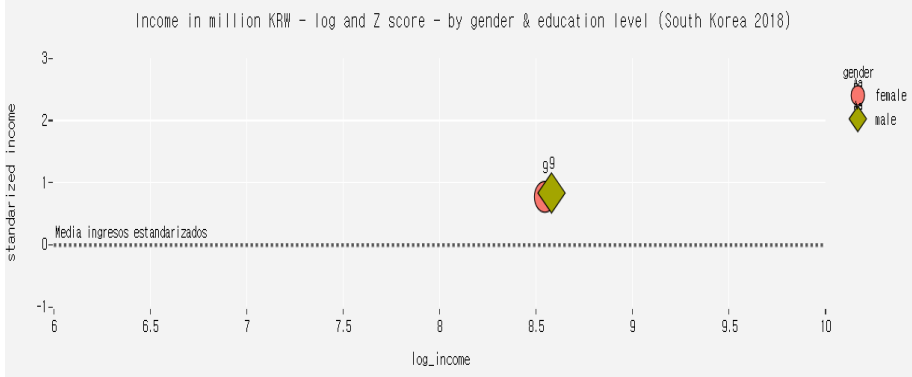
데이터 시각화

대체적으로 학력이 오를 수록 소득 수준이 높은 모습을 보임









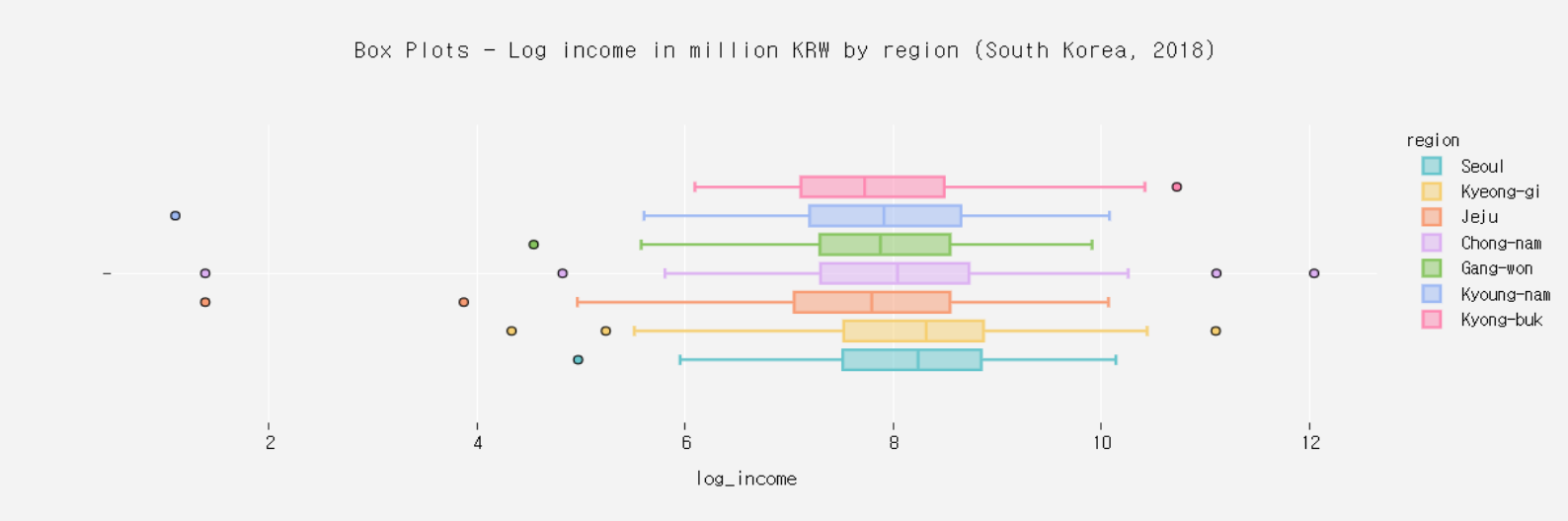
학업 수준 별 성별에 따른 소득

데이터 시각화

```
px.box(df,x="log_income",
            title="Box Plots - Log income in million KRW by region (South Korea, 2018)",template='ggplot2',color='region',
 height=400,
            color_discrete_sequence=px.colors.qualitative.Pastel).update_traces(dict(marker_line_width=1,
   marker_line_color="black")).update_traces(dict(marker_line_width=1,
            marker_line_color="black")).update_layout(
    paper_bgcolor='rgb(243, 243, 243)',plot_bgcolor='rgb(243, 243, 243)',
    font=dict(
        family="Century, monospace",
        color="black"))
```

지역 별 소득 수준은 전체적으로 비슷한 수준으로 통계됨

제주 지역의 소득 분포가 가장 넓게 분포되어 있음 : **빈부 차이가 큼**



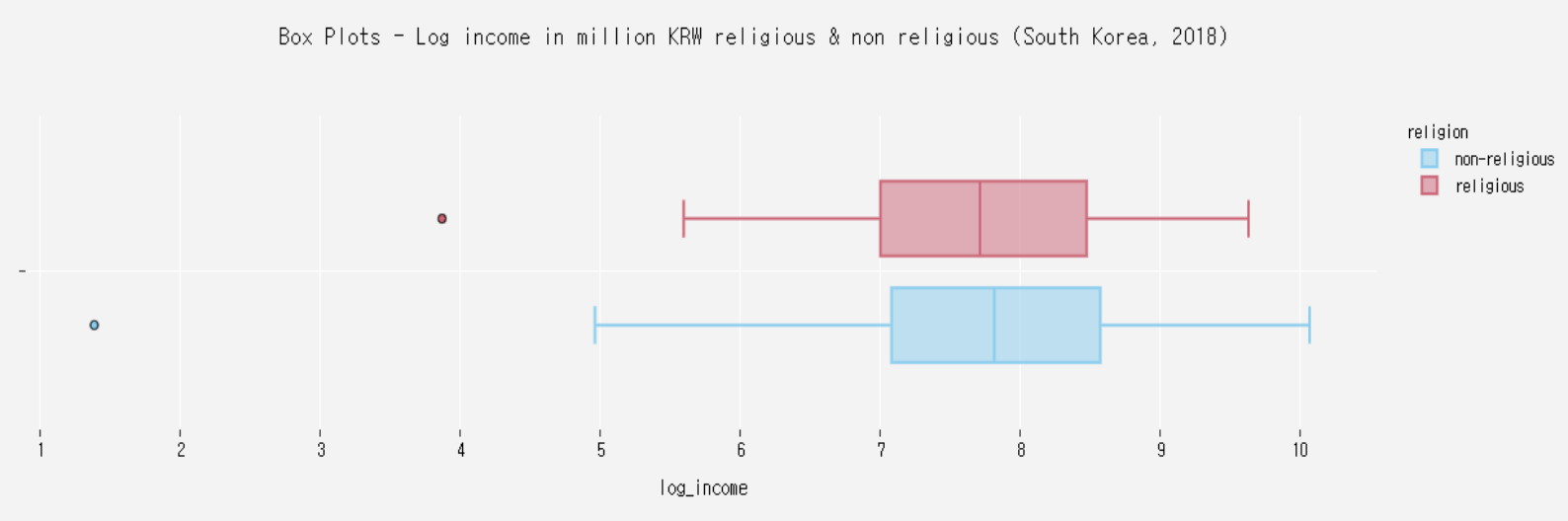
지역 별 소득 수준

데이터 시각화

```
px.box(df,x="log_income",
            title="Box Plots - Log income in million KRW religious & non religious (South Korea, 2018)",template='ggplot2',color='religion',
       height=400,
            color_discrete_sequence=px.colors.qualitative.Safe).update_traces(dict(marker_line_width=1,
   marker_line_color="black")).update_traces(dict(marker_line_width=1,
            marker_line_color="black")).update_layout(
    paper_bgcolor='rgb(243, 243, 243)',plot_bgcolor='rgb(243, 243, 243)',
    font=dict(
        family="Century, monospace",
        color="black"))
```

종교 여부에 따라
소득 평균은 비슷함

종교를 갖지 않는 분포의
빈부 차이가 큼



종교 여부 별 소득 수준

추가 작업 – 머신 러닝

```
X = df.drop(['income'], axis=1)
y = df['income']
```

```
from sklearn.model_selection import train_test_split

X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size=0.3, shuffle= True)
```

데이터 분할 및 처리

```
kn_cfr = KNeighborsClassifier()
kn_cfr.fit(X_train, y_train)

pred = kn_cfr.predict(X_valid)
score_kn = accuracy_score(y_valid, pred)
print(f'KNeighborsClassifier Accuracy : {score_kn*100:.2f}')
```

KNN ML

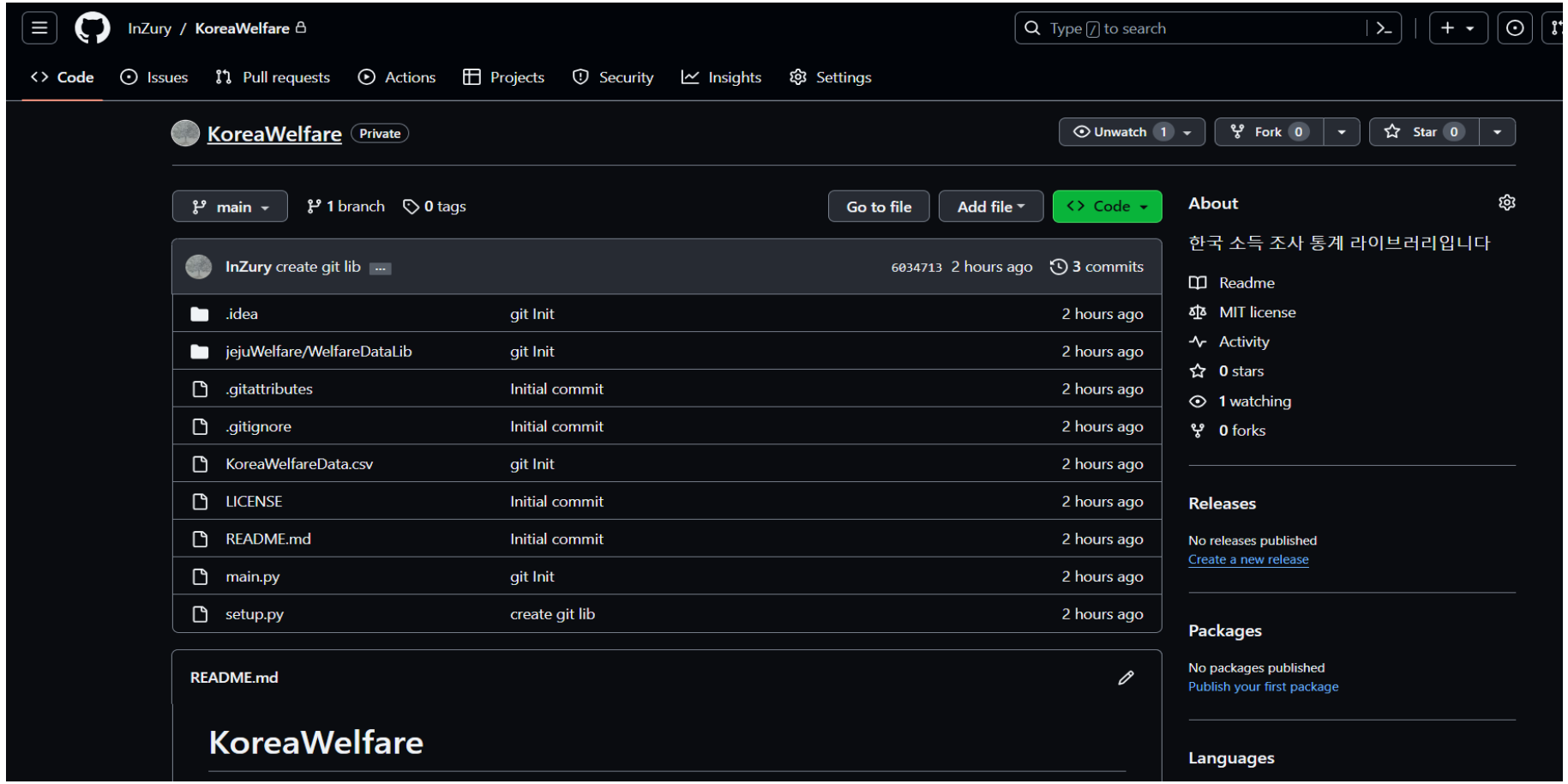
```
KNeighborsClassifier Accuracy : 81.96
```

```
rf_cfr = RandomForestClassifier(random_state=0)
rf_cfr.fit(X_train, y_train)

pred = rf_cfr.predict(X_valid)
score_rf = accuracy_score(y_valid, pred)
print(f'RandomForestClassifier Accuracy : {score_rf*100:.2f}')
```

Random Forest ML

```
RandomForestClassifier Accuracy : 84.75
```

추가 작업 – 라이브러리화



Github에 라이브러리화

주소를 이용하여 pip install 가능

결론 및 소감

- 제주 지역의 소득 편차는 타 지역에 비해 큰 편으로 확인됨
- 연도가 지나갈 수록 소득이 늘어남, 따라서 소득에 맞는
  물가 선정이 필요함

- 데이터 통계화와 수치를 직접 라이브러리로 만드는 작업을 진행하였는데
  제가 만든 코드를 다른 사람도 이용할 수 있는 배포하는 것이 신기하였다.
- 아직 코드에 오류가 있어서 버전을 수정하고 있으며, 오류 수정이 끝나면
  다시 priavate에서 public으로 수정할 예정이다.