About Discrimination in Hiring

20180490 이재헌

Introduction: KLIPS

- Data source: Korean Labor and Income Panel Study(KLIPS), 7th wave

KLIPS is:

- A longitudinal study of a representative sample of Korean households and individuals living in urban areas.
- Conducted annually to track the characteristics of households as well as the economic activities, labor movement, income, expenditures, education, job training, and social activities of individuals.

Introduction: Details on the data

- Data Description: Participants' experiences of hiring discrimination were measured.

- How the data looks like

index	disc_hire	gender	age	edu_cat	mariage	emp_fin	income_quartile	birth_region	health	disability	residence	disc_wage	disc_jobedu	disc_promotion	disc_resign
0	1.0	0	3	1	2	0	0	0	1	0	1	1	0	0	0
1	0.0	1	3	1	2	1	1	0	2	0	1	0	0	0	0
2	0.0	1	1	2	0	1	1	0	2	0	1	0	0	0	0
3	0.0	1	1	2	1	0	3	0	1	0	1	0	0	0	0
4	1.0	0	3	0	1	0	1	0	2	0	1	0	0	0	0

Introduction: Details on the data

	Variable name	Description	Possible answers
1	disc_hire	Response to the question, "Have you ever experienced discrimination in getting hired?"	0:'No', 1:'Yes', NA:'Not Applicable'
2	Gender	Gender	0:male, 1:female
3	Age	Age	0:16–24, 1:25–34, 2:35–44, 3:45–54, 4:55–64, 5:65+ years old
4	Edu_cat	Education level	0:middle school graduate or less, 1:high school graduate, 2:college graduate or more
5	Marriage	Marital status	0:never married, 1:currently married, 2:previously married
6	Emp_fin	Employment status	0:permanent, 1:non-permanent
7	Income_quartile	Total household income divided by the square root of the number of household members	0:Q1, 1:Q2, 3:Q3, 4:Q4 (4 categories based on the quartiles)
8	Birth_region	Birth region	1:Jeolla-do, 0:other regions
9	Self-rated health	Response to the question, "How would you rate your health?"	0:'very good', 1:'good', 2:'poor', 3:'very poor'
10	Disability	Response to the question "Do you have any impairment or disability?"	0:'No', 1:'Yes'
11	Residence	Residential areas	1:Seoul, 2:Pusan, 3:Daegu, 4:Daejeon, 5:Incheon, 6:Gwangju, 7:Ulsan, 8:Kyunggi, 9:Kangwon, 10:Choongbuk, 11:Choongnam, 12:Jeonbuk, 13:Jeonnam, 14:Kyungbuk, 15:Kyungnam
12	disc_wage	Experience of discrimination in receiving income	0:'No', 1:'Yes', 2:'Not Applicable'
13	disc_jobedu	Experience of discrimination in training	0:'No', 1:'Yes', 2:'Not Applicable'
14	disc_promotion	Experience of discrimination in getting promoted	0:'No', 1:'Yes', 2:'Not Applicable'
15	disc_resign	Experience of discrimination in being fired	0:'No', 1:'Yes', 2:'Not Applicable'
16	disc_edu	Experience of discrimination in obtaining higher education	0:'No', 1:'Yes', 2:'Not Applicable'
17	disc_home	Experience of discrimination at home	0:'No', 1:'Yes', 2:'Not Applicable'
18	disc_social	Experience of discrimination at general social activities	0:'No', 1:'Yes', 2:'Not Applicable'

Background (Motivation)

- The KLIPS dataset is intended to help design and implement more reasonable and accurate employment policies.

- By controlling for unobservable individual effects, it permits accurate understanding of the dynamic changes in the decision-making process and behavioral pattern of individuals and households in response to changes in the environment such as time and government policies.

Background (Motivation)

Examples of Use

- **Understanding an individual's lifetime career path** from the first job out of school to turnover and/or job separation and retirement
- **Understanding women's labor activities** from the time of labor market entry to exit, along with the factors that have an impact in the process
- **Understanding an individual's process of building human capital** through education, training and company training, and its impact on entry into the labor market
- **Understanding the impact of the government's employment policies** on building human capital
  Investigating on dynamic changes in household economy and generational changes

- Therefore, we can make <u>better employment policies</u> and <u>understand better what factors are related to discrimination</u> in hiring by analyzing the data

Objective

- Figuring out

- 1. What variables are related to hiring discrimination and how they are related. (Q. 1,4,5)

- 2. How the data can be clustered (Q. 2,3)

Question 1

- What are the important variables that are associated with the experience of hiring discrimination?
- How are those variables related to the experience of hiring discrimination?

Q1: Data Preprocessing

- 원래는 disc_hire 가 N/A 인 사람들의 자료를 제외하고 3479명의 자료를 그대로 사용하려고 했습니다.

- 하지만 disc_wage, disc_jobedu, disc_promotion, disc_resign, disc_edu, disc_home and disc_social 의 총 7가지 variable에 Not Applicable 로 응답한 사람의 값이 2로 저장되어 있었습니다.

- No = 0, Yes = 1 이기 때문에 N/A = 2를 쓰는 것은 이치에 맞지 않다고 생각해서 데이터를 제외하려고 하였습니다. 하지만 제외할 데이터의 수가 너무 많아지고(약 1000개) disc_hire의 평균이 크게 바뀌었습니다. 그래서 대신 N/A를 0과 1 사이의 값으로 변환하기로 결정하였습니다.

Q1: Data Preprocessing – Mean Est.

- N/A 처리 방법:
- 각 variable에 대해서 N/A(=2) 를 제외한 0과 1의 값들로 평균을 낸 후, 이 평균을 N/A 대신 넣어주었습니다.
- 이렇게 하면 scaling 한 이후에 이 값들은 모두 0이 되기 때문에 regression에 영향을 주지 않을 수 있다고 생각했습니다.

- Ex) (1, 0, 0, 2) → (1, 0, 0, 1/3)

Q1: Data Preprocessing – Mean Est.

- Data after preprocessing

	gender	age	edu_cat	mariage	emp_fin	income_quartile	birth_region	health	disability	residence	disc_wage	disc_jobedu	disc_promotion	disc_resign
0	0	3	1	2	0	0	0	1	0	1	1.0	0.000000	0.00000	0.000000
1	1	3	1	2	1	1	0	2	0	1	0.0	0.000000	0.00000	0.000000
2	1	1	2	0	1	1	0	2	0	1	0.0	0.000000	0.00000	0.000000
3	1	1	2	1	0	3	0	1	0	1	0.0	0.000000	0.00000	0.000000
4	0	3	0	1	0	1	0	2	0	1	0.0	0.000000	0.00000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3571	0	1	2	1	0	3	0	2	0	5	0.0	0.000000	0.00000	0.000000
3572	0	1	1	1	1	0	0	1	0	3	0.0	0.021856	0.00000	0.020421
3573	0	1	1	1	0	0	1	0	0	1	0.0	0.000000	0.00000	0.000000
3574	0	1	2	1	0	0	0	1	0	8	0.0	0.000000	0.06526	0.020421
3575	1	1	2	1	0	3	0	1	0	1	0.0	0.000000	0.00000	0.000000

Q1: Finding Important Variables – corr.

- 방법 1: disc_hire와의 correlation이 높은 variable 찾기
- Correlation의 절댓값으로 정렬하였습니다.

```
Correaltion of disc_hire with repect to each variable
disc_wage  :  0.6295803472606901
disc_social  :  0.3177017098600349
disc_promotion  :  0.2330898418994605
disc_resign  :  0.22434538252118089
disc_jobedu  :  0.21561331302082498
edu_cat  :  -0.18122465963652767
income_quartile  :  -0.1794744229947014 5
disc_home  :  0.16181032007132162
emp_fin  :  0.15331820400340926
disc_edu  :  0.09986301712643753
health  :  0.09002393573477176
age  :  0.08808585005683363
disability  :  0.07985001614343912
birth_region  :  -0.03616662107234033
gender  :  0.028481598991618836
residence  :  0.023151828427013427
mariage  :  0.015606196448611167
```

Ranking	Variable
1	disc_wage
2	disc_social
3	disc_promotion
4	disc_resign
5	disc_jobedu
6	edu_cat
7	income_quartile

Q1: Finding Important Variables – corr.

Ranking	Variable
1	disc_wage
2	disc_social
3	disc_promotion
4	disc_resign
5	disc_jobedu
6	edu_cat
7	income_quartile

```
Correaltion of disc_hire with repect to each variable
disc_wage  :  0.6295803472606901
disc_social  :  0.3177017098600349
disc_promotion  :  0.2330898418994605
disc_resign  :  0.22434538252118089
disc_jobedu  :  0.21561331302082498
edu_cat  :  -0.18122465963652767
income_quartile  :  -0.17947442299470145
```

- Disc___ variable은, 입사를 한 이후에 경험하는 것들입니다. 채용과정에서 차별을 받았으면 입사 이후에도 차별받은 경우가 많다는 것을 알 수 있습니다.
- 채용 과정에서 영향을 주는 것은 최종학력(edu_cat)과 소득(income_quartile)이라고 해석할 수 있습니다.
- 학력이 낮을수록, 소득이 적을수록 입사할 때 차별을 받았다는 것을 알 수 있습니다.

Q1: Finding Important Variables – Logis.

- 방법 2: disc_hire와의 logistic regression 후 p value가 작은 variable 찾기
- 각 variable의 scale이 다르기 때문에 먼저 scaling 하였습니다.
- 각 variable별로 정규화를 하였습니다. $Z = \frac{X-\mu}{\sigma}$

- Scaled Data

3571	-0.809863	-0.826216	1.151205	0.364598	-0.555769	1.185739	-0.487481	0.886686	-0.16203	-0.266745	-0.426822	-1.569191e-01	-2.808207e-01	-0.152124	-0.107663	-0.146478
3572	-0.809863	-0.826216	-0.159345	0.364598	1.799310	-1.655951	-0.487481	-0.587875	-0.16203	-0.706037	-0.426822	2.490930e-17	-2.808207e-01	0.000000	-0.107663	-0.146478
3573	-0.809863	-0.826216	-0.159345	0.364598	-0.555769	-1.655951	2.051361	-2.062435	-0.16203	-1.145329	-0.426822	-1.569191e-01	-2.808207e-01	-0.152124	-0.107663	-0.146478
3574	-0.809863	-0.826216	1.151205	0.364598	-0.555769	-1.655951	-0.487481	-0.587875	-0.16203	0.392193	-0.426822	-1.569191e-01	-5.971782e-17	0.000000	-0.107663	-0.146478
3575	1.234777	-0.826216	1.151205	0.364598	-0.555769	1.185739	-0.487481	-0.587875	-0.16203	-1.145329	-0.426822	-1.569191e-01	-2.808207e-01	-0.152124	-0.107663	-0.146478

Q1: Finding Important Variables – Logis.

- Logistic Regression 을 한 후, p value가 작은 순으로 정렬했습니다.

index	coef	std err	z	P>|z|
disc_wage	1.3801	0.083	16.651	0.0
disc_resign	2.934	0.348	8.419	0.0
disc_home	0.2272	0.05	4.567	0.0
disc_social	0.3747	0.051	7.374	0.0
income_quartile	-0.1575	0.046	-3.42	0.001
emp_fin	0.1284	0.046	2.807	0.005
disc_jobedu	0.2425	0.086	2.814	0.005
mariage	-0.1382	0.054	-2.559	0.01
residence	0.092	0.043	2.145	0.032
age	0.1216	0.058	2.085	0.037
disc_edu	0.0683	0.05	1.362	0.173
health	0.0582	0.045	1.288	0.198
gender	-0.0518	0.045	-1.15	0.25
disc_promotion	-0.0694	0.061	-1.14	0.254
disability	0.0366	0.044	0.822	0.411
birth_region	-0.0332	0.043	-0.772	0.44
edu_cat	0.0032	0.053	0.061	0.952

- H0: coeff = 0.
- P value가 작다는 것이 무조건 important variable임을 보장하지 않지만 important variable이면 p value가 작다.

Q1: Finding Important Variables – Logis.

- Logistic Regression 을 한 후, p value가 작은 순으로 정렬했습니다.

```
results_scaled.pvalues.sort_values()
```

```
disc_wage           2.976096e-62
disc_resign         3.793729e-17
disc_social         1.655163e-13
disc_home           4.950384e-06
income_quartile     6.259342e-04
disc_jobedu         4.892201e-03
emp_fin             5.004861e-03
mariage             1.048899e-02
residence           3.191839e-02
age                 3.710601e-02
disc_edu            1.731857e-01
health              1.975961e-01
gender              2.503133e-01
disc_promotion      2.544269e-01
disability          4.109147e-01
birth_region        4.400138e-01
edu_cat             9.515474e-01
dtype: float64
```

Ranking	Variable
1	disc_wage
2	disc_resign
3	disc_social
4	disc_home
5	Income_quartile
6	disc_jobedu
7	emp_fin

Q1: Finding Important Variables – Logis.

Ranking	Variable
1	disc_wage
2	disc_resign
3	disc_social
4	disc_home
5	Income_quartile
6	disc_jobedu
7	emp_fin

index	coef
disc_wage	1.3801
disc_resign	2.934
disc_home	0.2272
disc_social	0.3747
income_quartile	-0.1575
emp_fin	0.1284
disc_jobedu	0.2425
mariage	-0.1382

- Disc_ __ variable은, 입사를 한 이후에 경험하는 것들입니다. 채용과정에서 차별을 받았으면 입사 이후에도 차별받은 경우가 많다는 것을 알 수 있습니다.
- 채용 과정에서 영향을 주는 것은 소득(income_quartile)과 정규직 여부(emp_fin)이라고 해석할 수 있습니다.
- 학력이 낮을수록, 비정규직일수록 입사할 때 차별을 받았다는 것을 알 수 있습니다.
- edu_cat의 p value는 크게 나왔습니다.

Q1: Finding Imp. Var. : Corr vs Logis.

[Correlation]

Ranking	Variable
1	disc_wage
2	disc_resign
3	disc_social
4	disc_home
5	Income_quartile
6	disc_jobedu
7	emp_fin

[Logistic Regression]

Ranking	Variable
1	disc_wage
2	disc_social
3	disc_promotion
4	disc_resign
5	disc_jobedu
6	edu_cat
7	income_quartile

- 두 방법에서 모두 disc_wage가 가장 중요한 variable이라는 것을 알 수 있습니다.
- Logistic regression의 결과가 최종 답안입니다.

Q2. PCA

- What are the important principle components (PC) that explain a large portion of variation in the given 12 explanatory?
- How would you interpret those PCs?

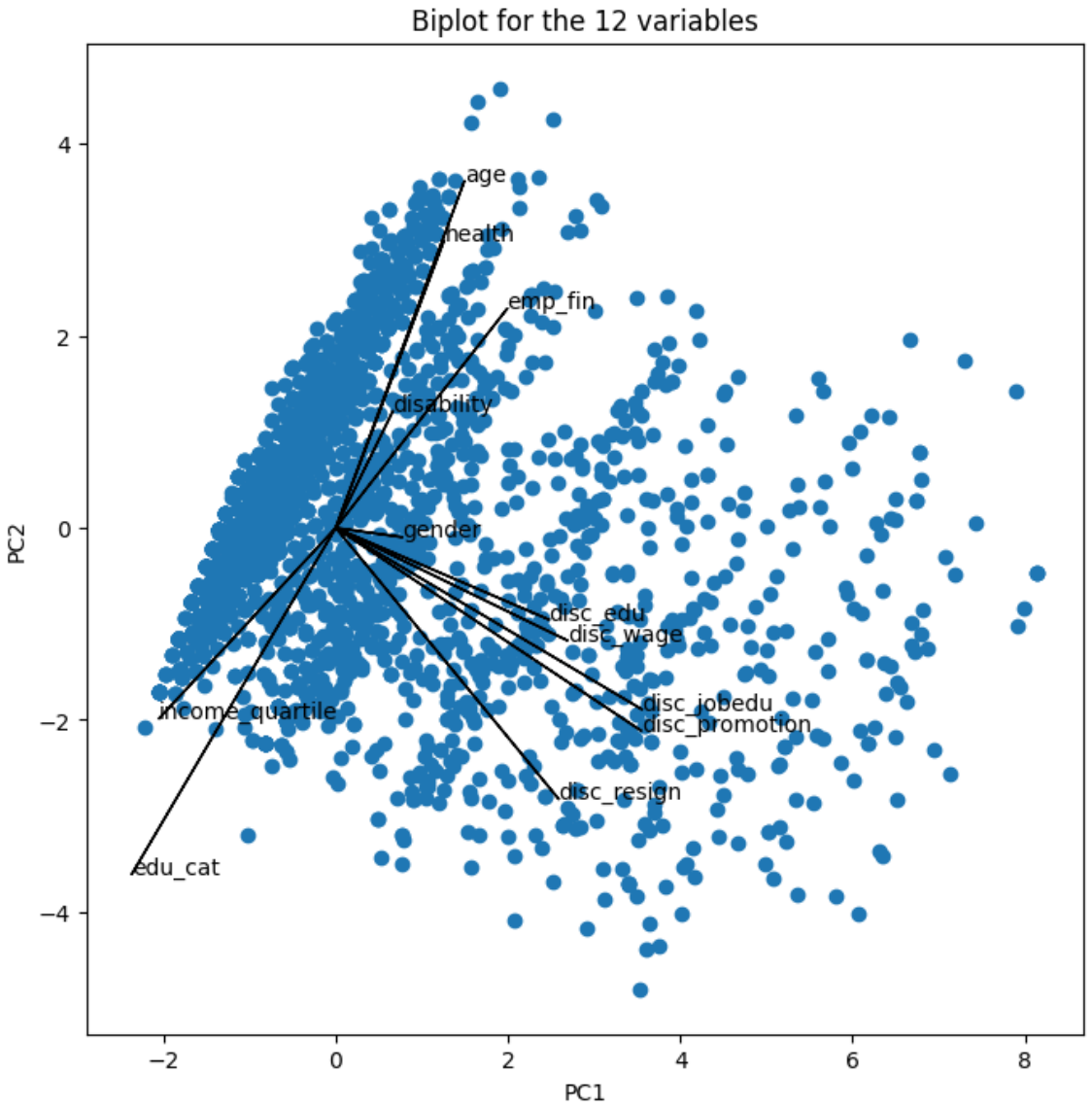
Q2. PCA: Preprocessing

- Q1에서와 같이 N/A data를 각 변수의 N/A(=2) 값을 제외한 평균으로 대체하였습니다.

- 그런 다음, PCA를 진행하기 위해 Scaling을 해주었습니다. (with Standard Scaler, $Z = \frac{X-\mu}{\sigma}$)

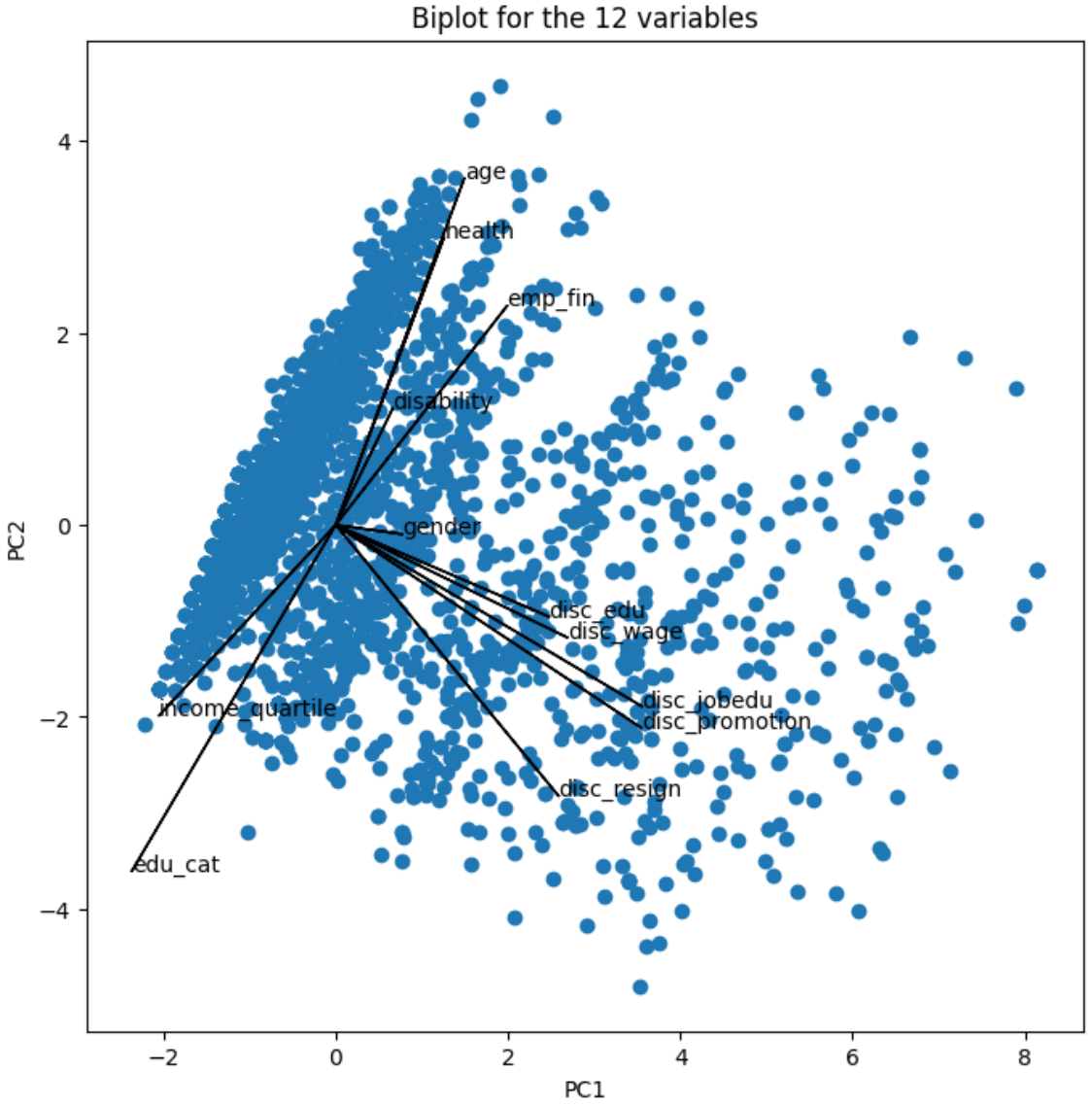
Q2. PCA: Biplot

- Interpretation on PC1
- disc_edu, disc_wage 등 직장 내에서 차별 받았던 경험이 PC1 에 가장 많이 contribute 했음을 볼 수 있습니다.
- 또한 disc_wage, disc_edu 등 차별받았던 경험 벡터들의 방향이 유사하므로 이들의 correlation이 높음을 알 수 있습니다. 어떤 차별을 받았으면 다른 차별도 받았을 확률이 높다는 뜻입니다.



Q2. PCA: Biplot

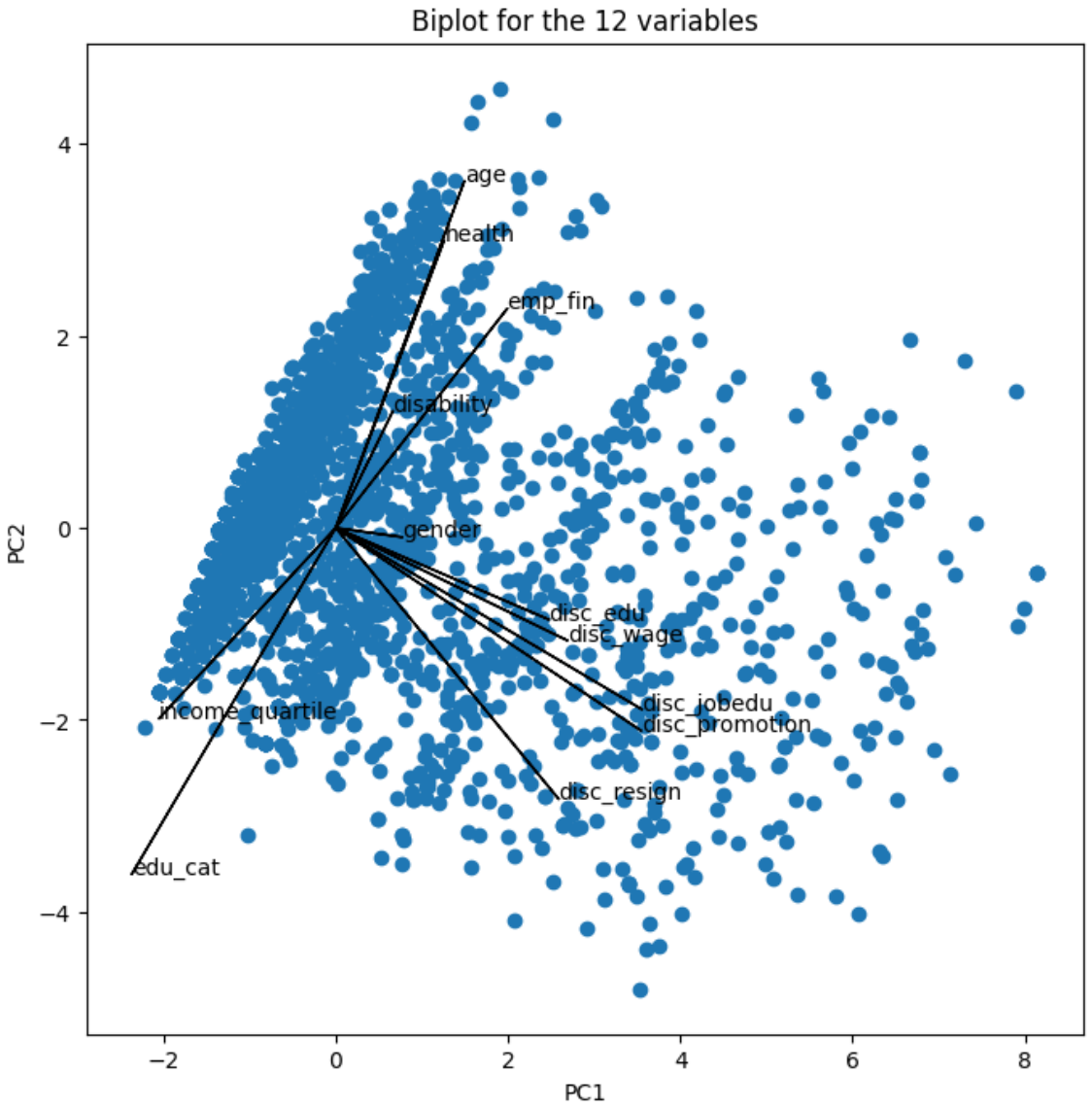
- Interpretation on PC2
- Age, health, edu_cat 이 PC2 에 가장 많이 contribute 했음을 볼 수 있습니다.
- Age와 health, emp_fin의 correlation이 높습니다. 나이가 많을수록 건강하지 않고, 비정규직일 확률이 높습니다.
- 또한 income_quartile 과 edu_cat이 correlation이 높습니다. 학력이 낮을수록 수입이 낮을 확률이 높습니다.



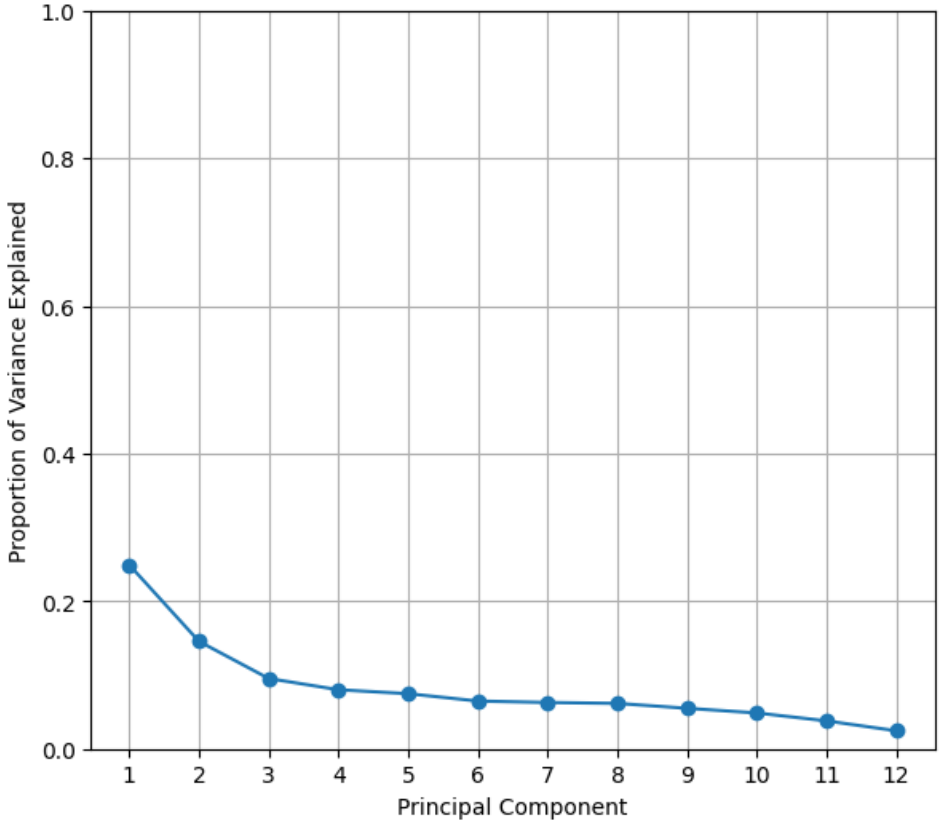
Q2. PCA: Biplot

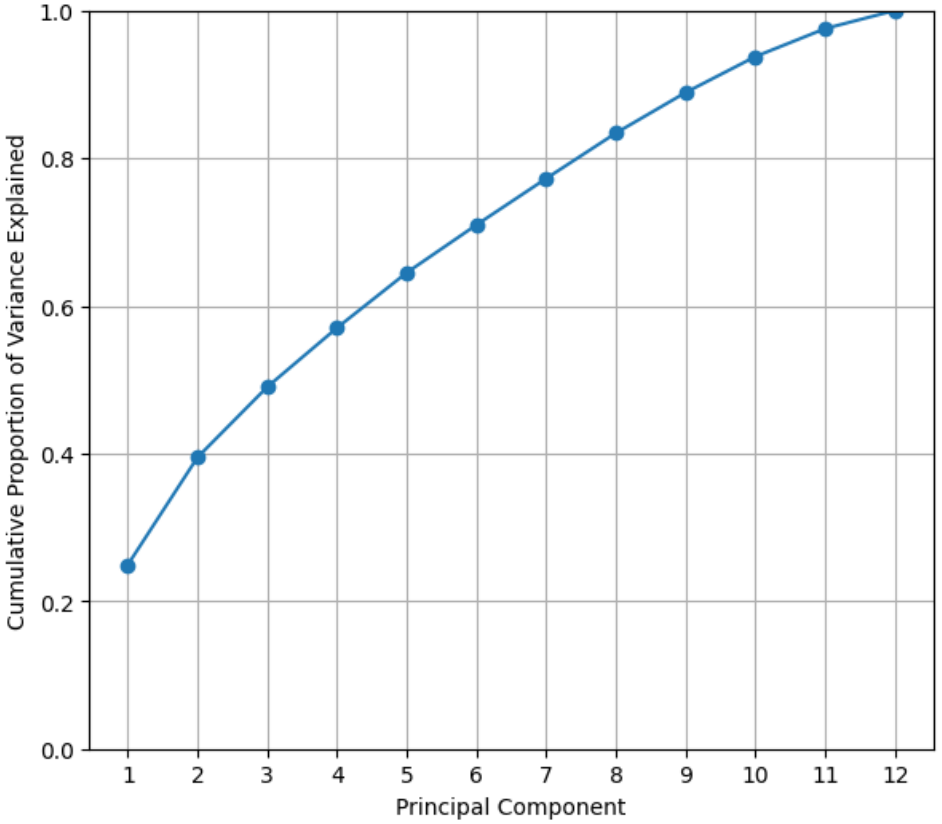
- Interpretation on PC2

- 차별 받았던 경험 disc_wage, disc_edu 등과 age, health 등의 차별과 관계 없는 변수들과의 correlation이 작습니다.



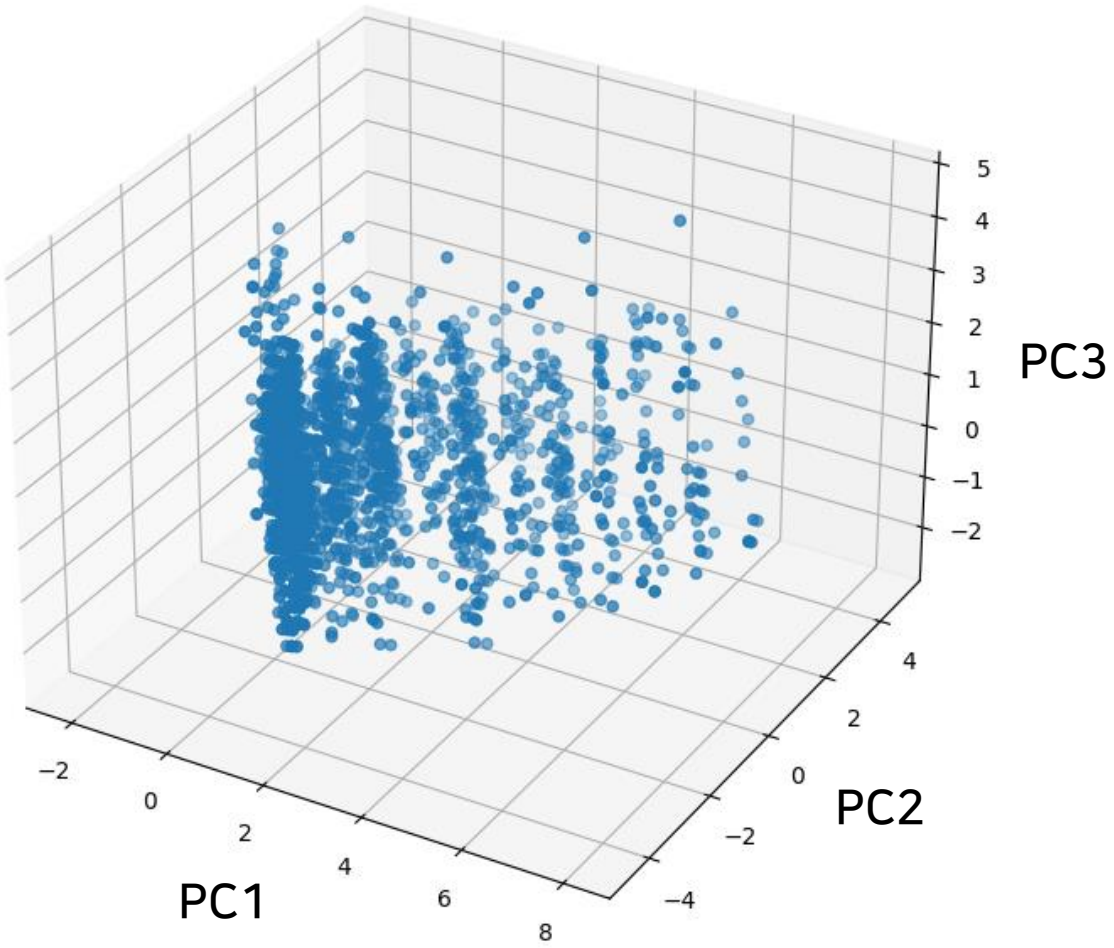
Q2. PCA: Explained Variance

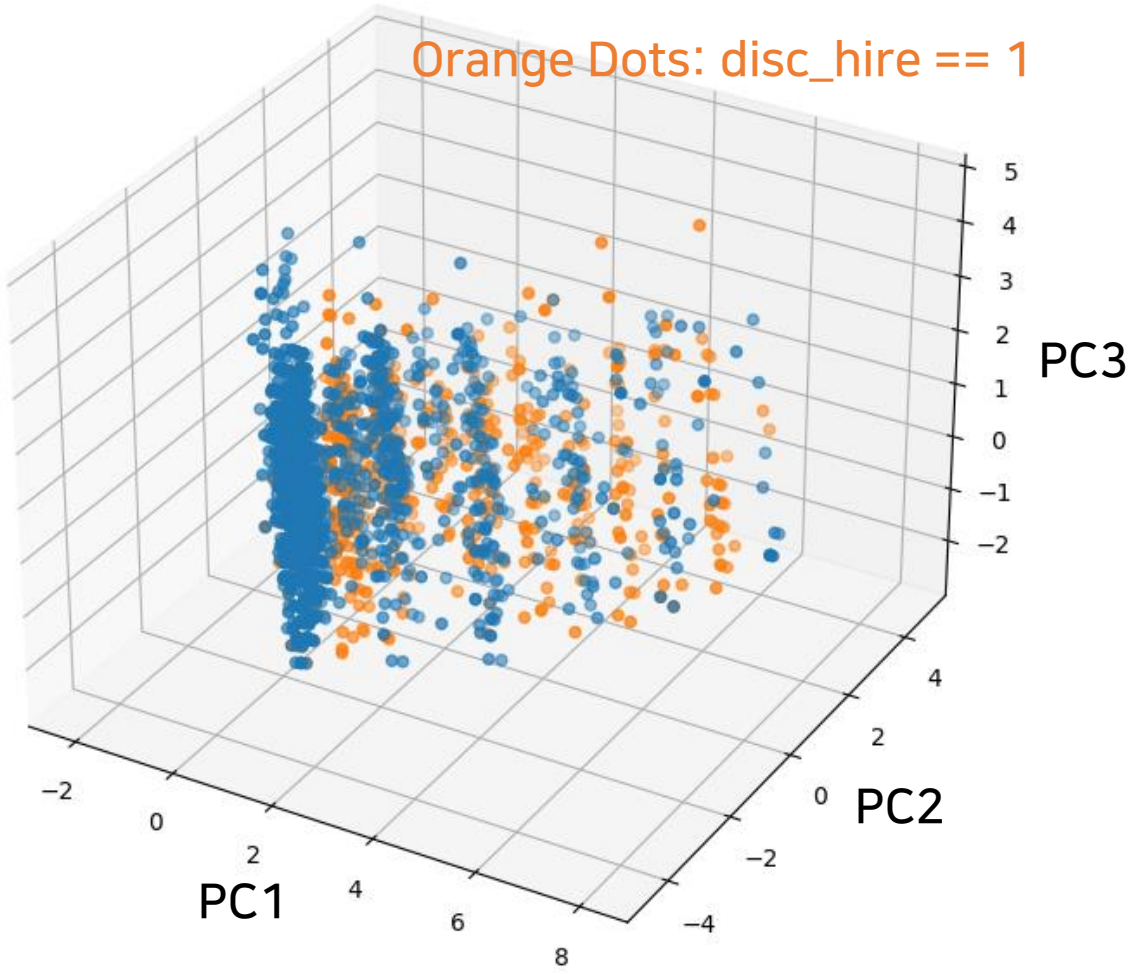




- Criteria: Sum of variance ratio > 50% ⇒ n = 4 로 Q.3 진행

Q2. PCA: Visualization on 3D space



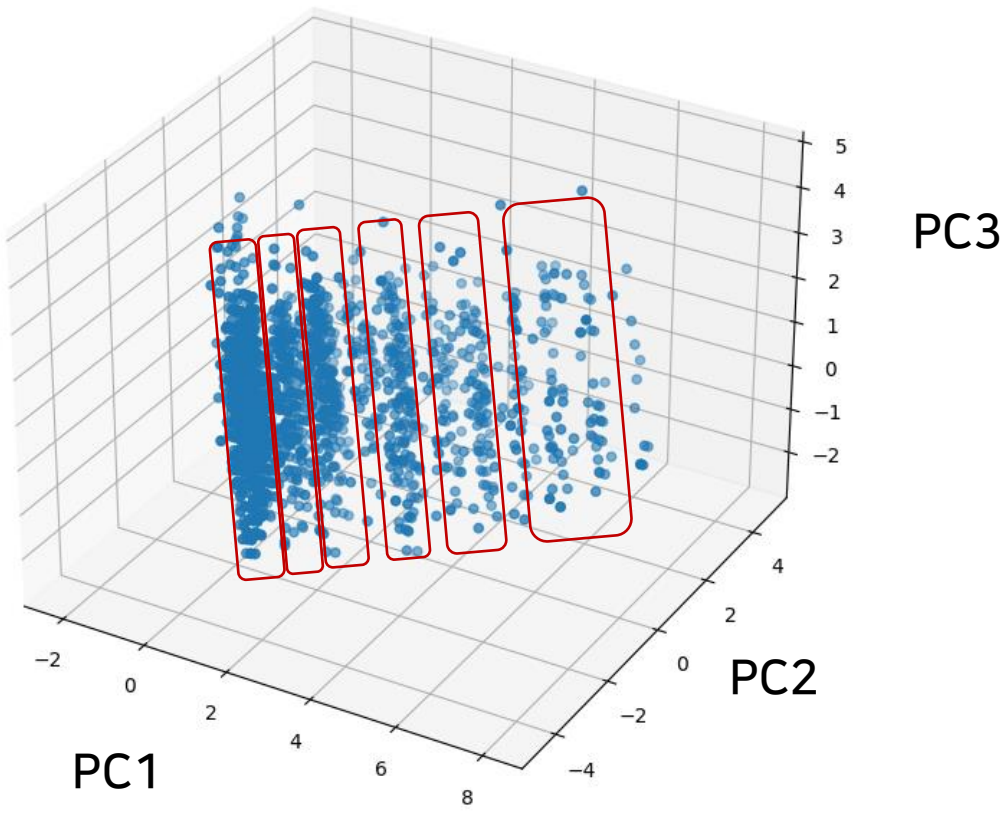


Q3. Clustering with PCA

- Identify subgroups (clusters) based on the 12 variables you use for PCA to answer question 2.
- Also, identify subgroups based on the important PCs you find to answer question 2.
- Compare the clustering results.

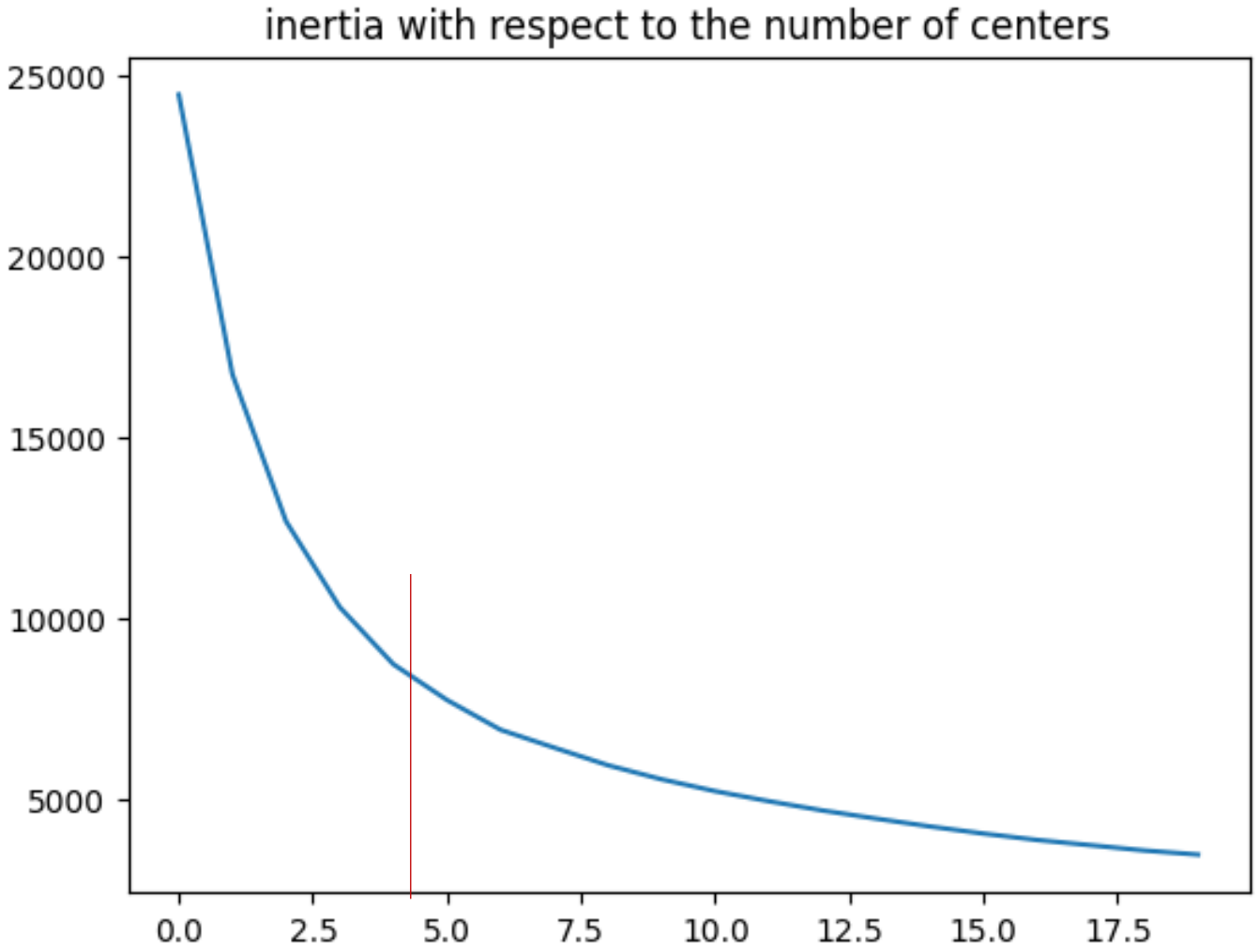
Q3. Clustering with PCA

- From Q.2, I chose n = # of PC = 4



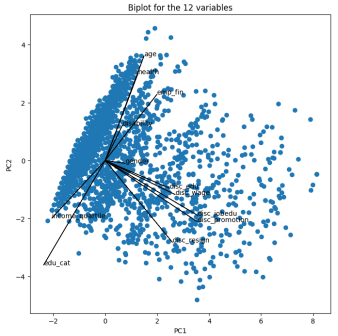
- 눈으로 봤을 때 cluster가 보이긴 하지만, 이렇게 나눌만한 기준을 세울 수 없었습니다.
- 그래서 K means clustering 방법으로 접근하였습니다.

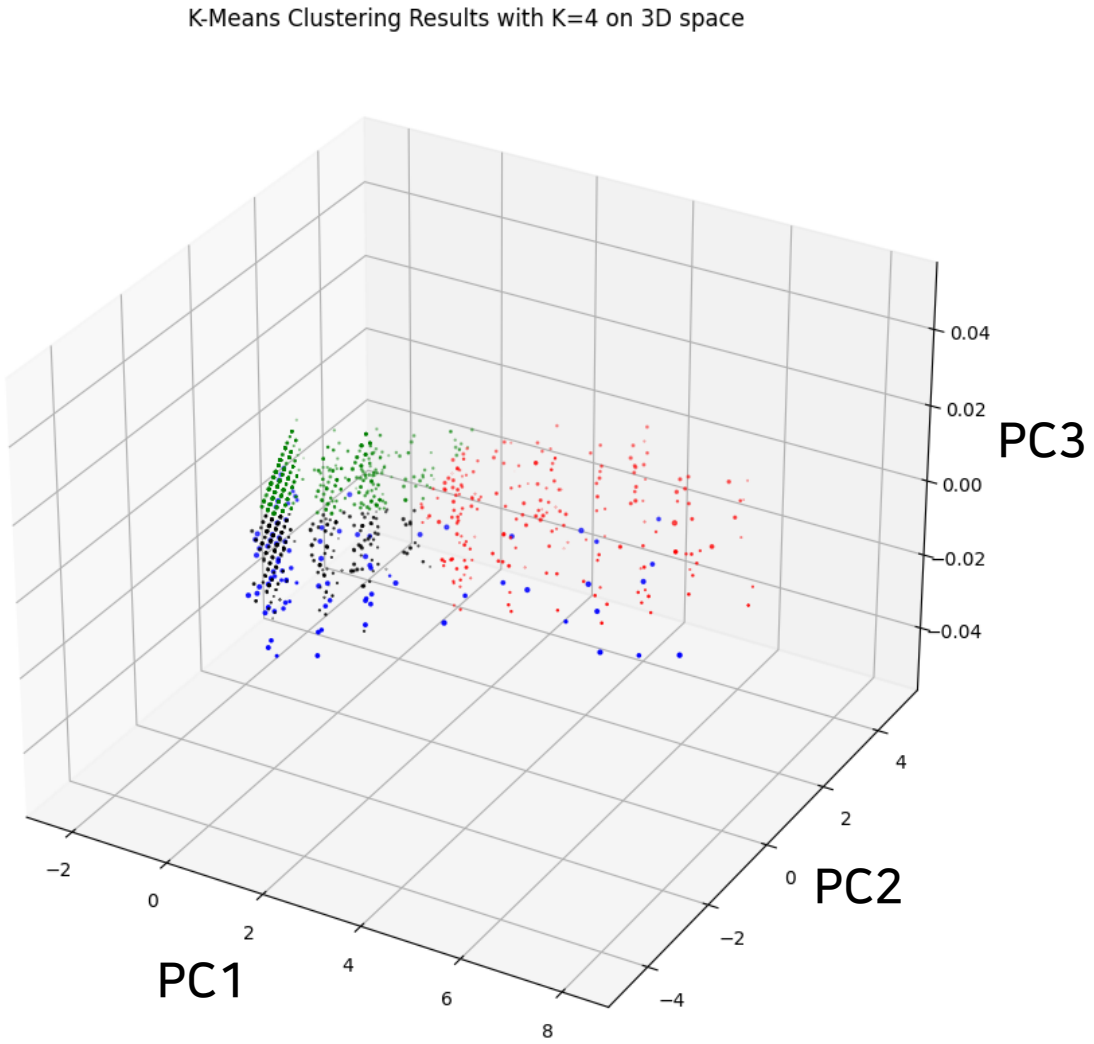
Q3. Clustering with PCA



- Inertia = $\sum d^2(x, c_k)$ 의 변화 비교하여 K 값 설정
- Inertia의 변화를 기준으로 K = 4로 설정 (코드 참조)

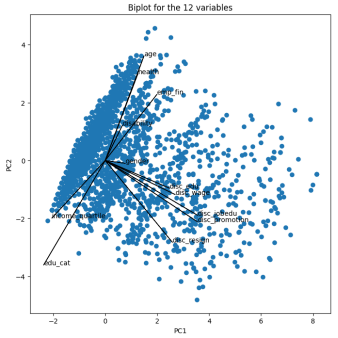
Q3. Clustering with PCA: K means

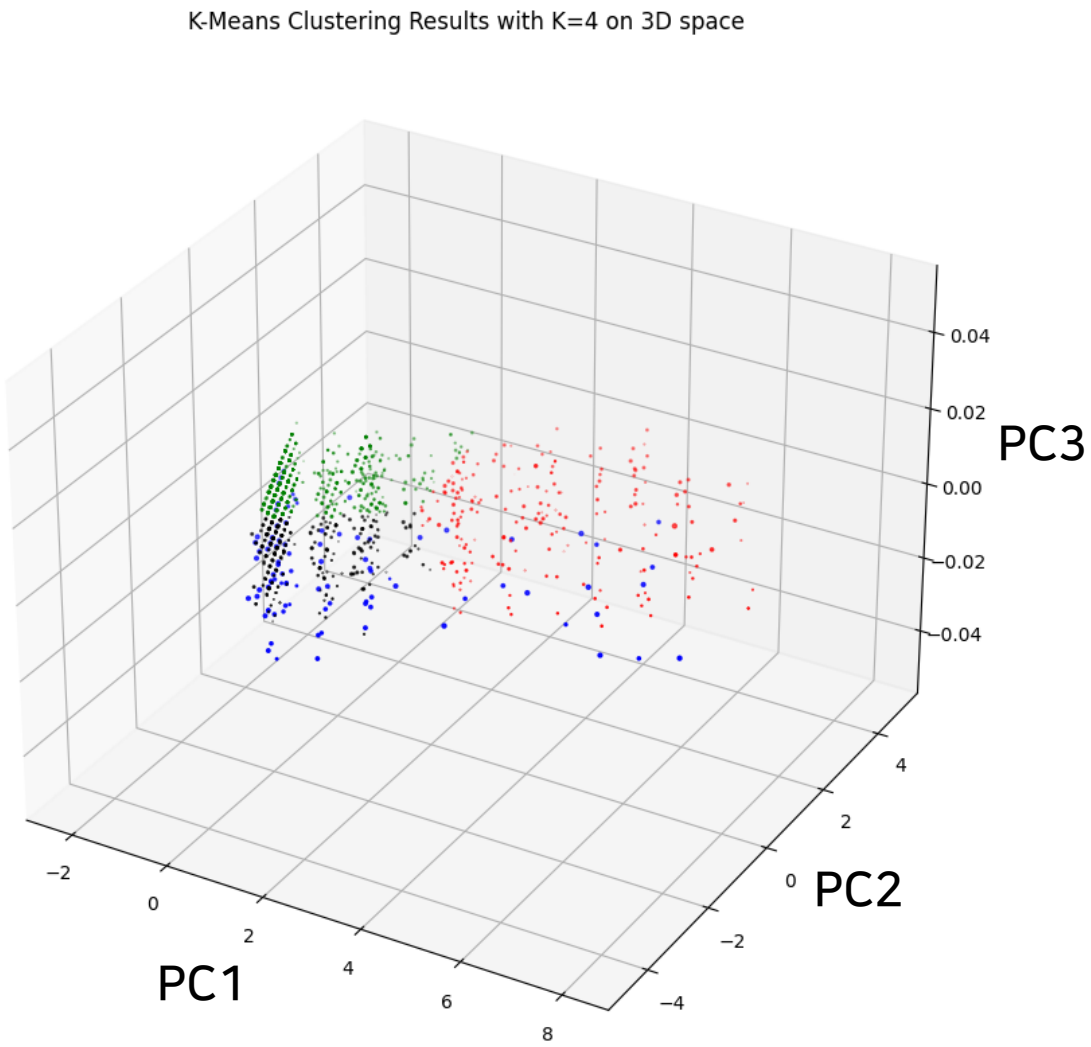




- K means 알고리즘을 사용해서 clustering 한 결과, 왼쪽과 같이 4개의 cluster를 형성하였습니다. (빨, 초, 파, 검)

Q3. Clustering with PCA: K means

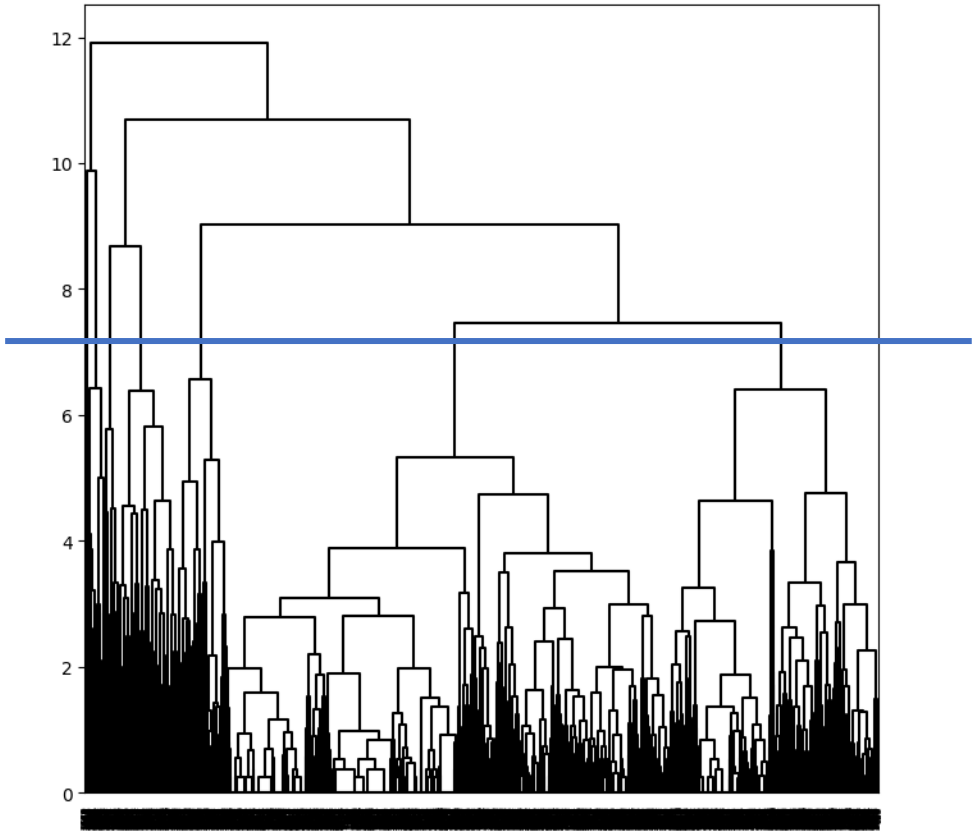




- 초록색: 차별 경험이 적고, 나이가 많고 건강하지 않으며, 비정규직인 사람들
- 검은색: 차별 경험이 적고, 소득이 많으며, 학력이 높은 사람들
- 파란색: 젊고, 소득이 많고, 학력이 높은 사람들
- 빨간색: 차별 경험이 있는 사람들

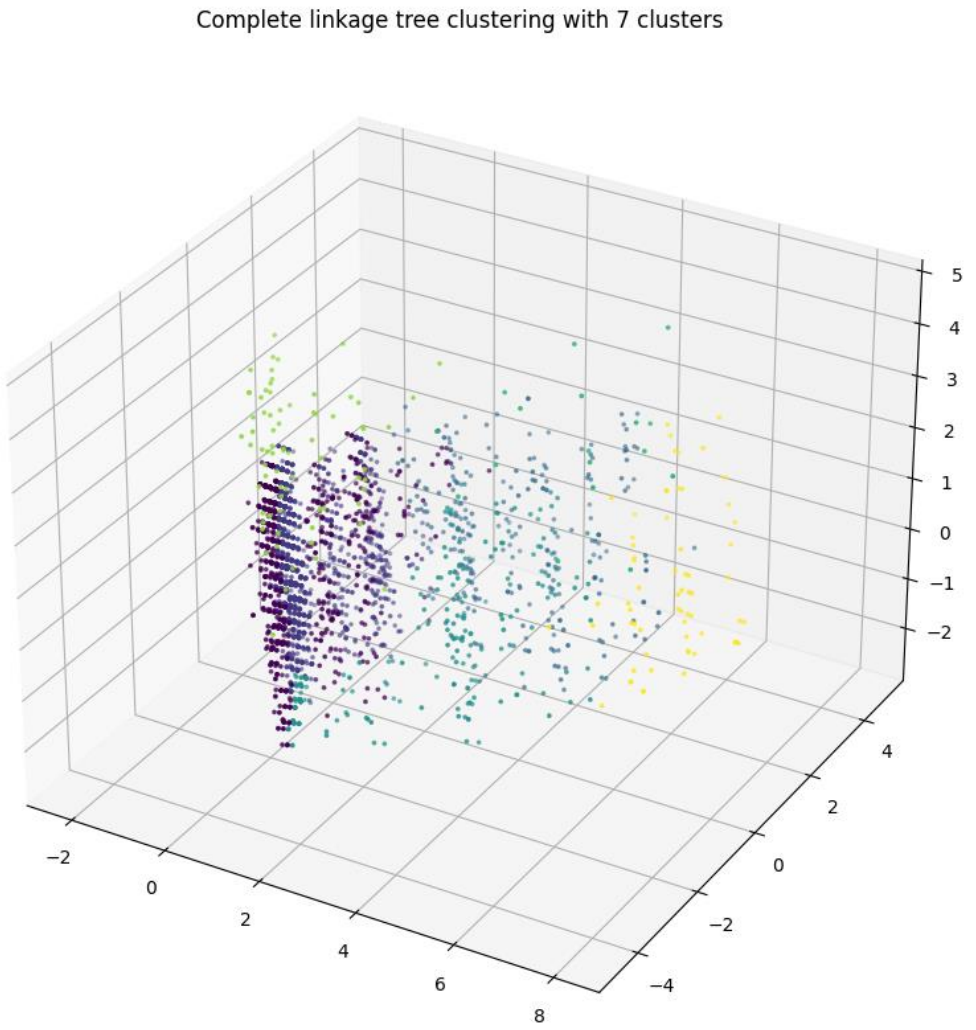
Q3. Clustering with PCA: Hierarchical

- Clustering with complete linkage hierarchical method



- Height = 7에서 tree cut
- Average, single linkage 로도 시도 해보았지만, 성능이 좋지 않았습니다.

Q3. Clustering with PCA: Hierarchical



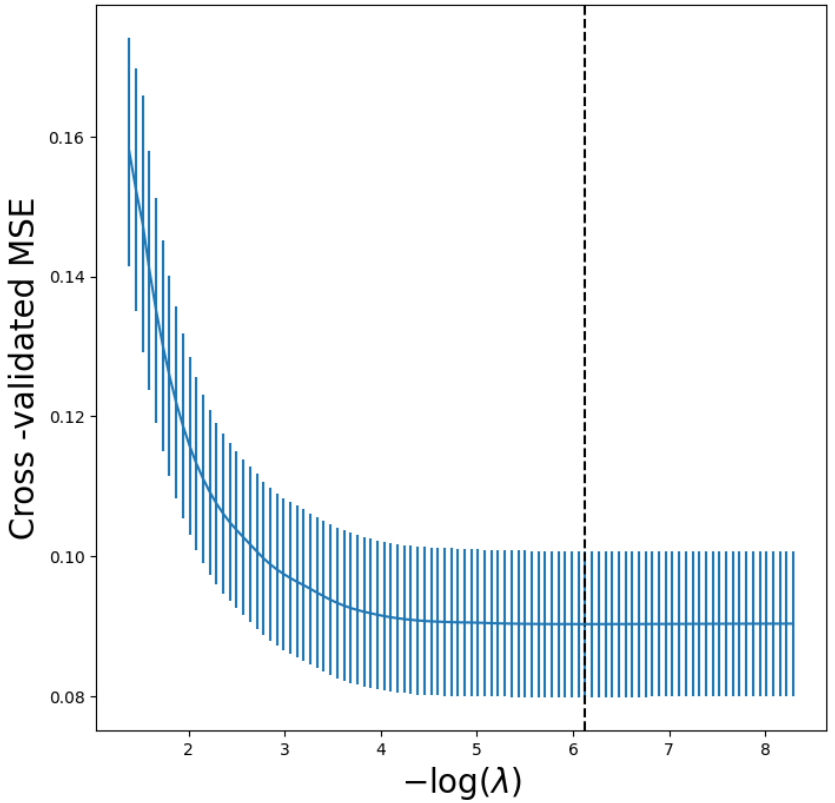
- Cluster가 총 7개 나왔습니다.
- 처음에 눈으로 봤을 때 확인했던 cluster와 비슷하게 clustering 이 되었습니다.
- 그림에서 왼쪽에 있는 cluster일수록 차별 경험이 없는 사람들, 오른쪽에 있는 cluster일수록 차별 경험이 많은 사람들입니다.

Q4. Discrimination Under reporting

- Is there a difference in under reporting of hiring discrimination between males and females?

Q4. Lasso Cross Validation

- Q.1 에서 사용하였던 disc_hire 가 N/A가 아닌 자료들을 training set으로 사용하여 LASSO parameter λ를 tuning 하였습니다.



- 실행 결과, λ = 0.00218 를 얻었습니다.

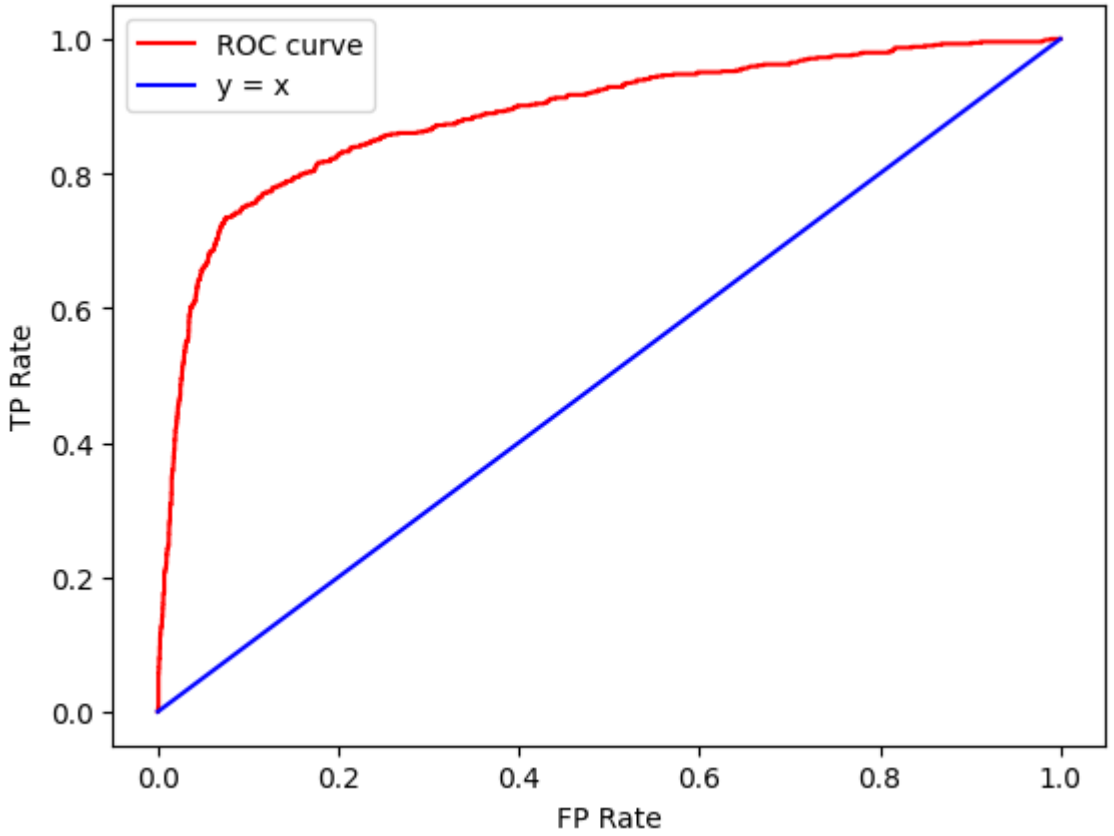
```
tuned_lasso.coef_
```

```
array([-0.003403,  0.009698, -0.001013, -0.014724,  0.02313 , -0.025819,
       -0.002628,  0.007461,  0.003798,  0.013161,  0.213954,  0.013982,
        0.      ,  0.026307,  0.00479 ,  0.029658,  0.049524])
```

- 일부 LASSO parameter가 0이 된 것을 확인할 수 있습니다.

Q4. Lasso Cross Validation

- ROC and AUC of LASSO



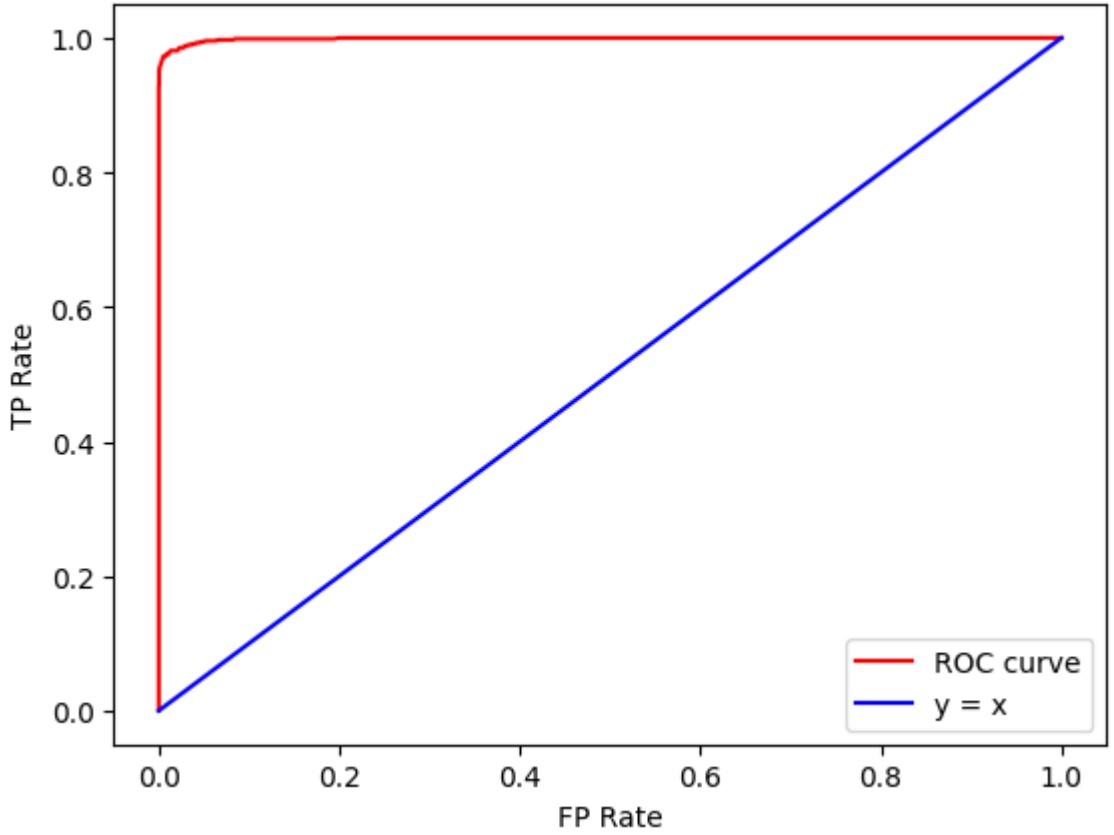
- AUC of LASSO
  = 0.88674

Q4. Random Forest Cross Validation

- Random forest 의 parameter 수의 후보: all, sqrt, log2
- “Sqrt”의 경우가 Cross validation set에서 best performance를 보였습니다.
- 따라서, model = RF(max_features=‘sqrt’)

Q4. Random Forest Cross Validation

- ROC and AUC of Random Forest



- AUC of Random Forest = 0.99887

Q4. Random Forest Prediction

- Random forest model로 예측한 결과는 아래와 같습니다.

- male_report_ratio = 0.3125
- female_report_ratio = 0.8182

- 남성의 경우, under reporting이 있는 것을 확인했습니다.
- (남성의 경우, 차별을 받는다고 말을 안 하는 경향이 있습니다.)

Q5.

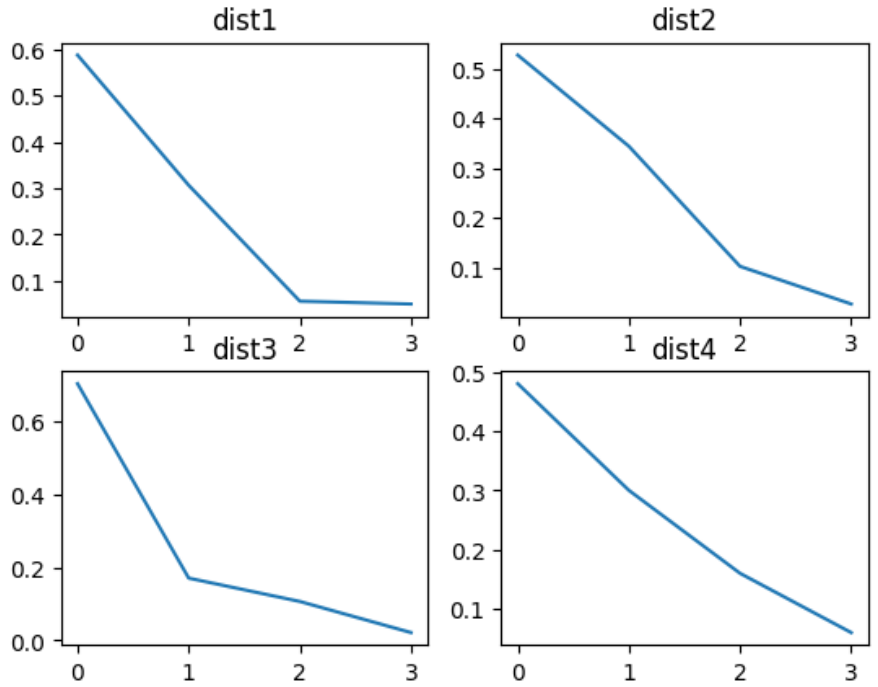
- Is there an association between the experience of hiring discrimination and health?

Q5. Relationship btw disc_hire & health

- Distribution of health with respect to the given groups

```
print(dist1, dist2, dist3, dist4)
```

```
[0.587898 0.30648  0.055854 0.049767] [0.527697 0.344023 0.102041 0.026239] [0.702128 0.170213 0.106383 0.021277] [0.48 0.3  0.16 0.06]
```



Q5. Relationship btw disc_hire & health

- Overall - method: Chi-square test
- Result: Reject the null hypothesis
- Pairwise multiple testing – method: Tukey HSD
- Results: as follows

```
Overall Chi-Square Test: Chi2 = 44.1060153352378, p-value = 1.349338405457055e-06
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj   lower  upper  reject
-------------------------------------------------
    G1     G2   0.1534    0.0  0.0791 0.2278   True
    G1     G3  -0.0067 0.9999 -0.2632 0.2498  False
    G1     G4   0.1916 0.1961 -0.0572 0.4404  False
    G2     G3  -0.1602 0.3985 -0.4231 0.1028  False
    G2     G4   0.0381 0.9808 -0.2173 0.2936  False
    G3     G4   0.1983 0.4752  -0.156 0.5526  False
-------------------------------------------------
```

Q5. Relationship btw disc_hire & health

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
G1	G2	0.1534	0.0	0.0791	0.2278	True
G1	G3	-0.0067	0.9999	-0.2632	0.2498	False
G1	G4	0.1916	0.1961	-0.0572	0.4404	False
G2	G3	-0.1602	0.3985	-0.4231	0.1028	False
G2	G4	0.0381	0.9808	-0.2173	0.2936	False
G3	G4	0.1983	0.4752	-0.156	0.5526	False

- H_0: G2(No)=G3(N/A but Yes)
- H_0: G3 (N/A but Yes) =G4 (N/A but No)
- 두 귀무가설 모두 reject 되어야 하지만 accept 되었다. 따라서 disc_hire와 health의 관계는 없다.

Conclusion

- Q1: 학력이 낮을수록, 비정규직일수록 입사할 때 차별을 받았다는 것을 알 수 있습니다. 따라서 이를 시정할 수 있는 제도를 만들어야 할 것입니다.

- Q4: 남성이 여성보다 차별 받는다고 말하는 경우가 적습니다. 남성이 말을 하지 않는다고 해서 차별을 받고 있지 않다고 생각하면 안 될 것입니다.

- Q5: 차별 받는 것과 건강한 정도는 유의미한 상관관계가 없습니다.

Thank you so much for listening!