열린국회정보공개포털 DB를 활용한

국회의원 법안발의에 영향을 미치는 변수 분석

- ‘출신지역’ 변수를 중심으로-



|  |  |
| --- | --- |
| 과 목 명 | 한국어인공지능과 데이터 |
| 지도교수 | 최동혁 |
| 학번/학과/이름 | 2017311443 강로이 / 2019314960 김민제 |
| 제 출 일 | 2022.06.10 |

**<목 차>**

1. 서론과 연구문제

2. 선행연구

3. 수집 데이터

4. 분석 방법

4.1. 데이터 수집 및 전처리

4.2. 분석 방법 및 결과

5. 시각화 및 2차 분석

5.1. Boxplot 시각화

5.2. 시각화와 전체 코드

5.3 수정 분석 및 결론

6.조원 역할 분담

7. 수업에 대한 피드백

1. 연구 주제 및 서론

한국사회에 사는 사람이라면 ‘학연 지연 혈연’이 중요하다는 이야기를 많이 듣고 자랐을 터이고, 이러한 연고주의가 인사발령이나 조직운영의 자율권을 많이 제한하고 있다는 것은 공공연한 사실이다. 또한 연고주의는 사회생활에서의 불이익을 준다는 인식도 있지만, 사실 더 나아가서 인류학적 관점을 보았을 때, 개개인의 가치관과 정체성 형성에 빼놓을 수 없는 요인이다.

본 연구는 ‘학연’,‘지연’,‘혈연’ 중 ‘지연’에 초점을 맞추어 , “열린국회정보공개포털 DB를 활용한 국회의원 법안발의에 영향을 미치는 변수 분석- ‘출신지역’ 변수를 중심으로” 라는 주제로 연구를 진행하였다. 이 주제를 선정한 이유는, 대한민국은 대한민국은 학연 지연 혈연이 중요하다는 말이 있듯이 앞선 연구들은 학연,지연,혈연이 당선에 영향을 미친다는 것들이 존재했다. 하지만 이것을 데이터를 기반으로 하여 ‘법안 발의에 학연,지연,혈연이 영향을 미치는가’에 대해 근거가 있을까라는 의문을 가지게 되었다.

연구의 본격적인 진행을 위해 ‘법안 공동 발의에 지역 요인이 미치는 영향이 존재한다.’라는 가설을 설정하였다.

1. 선행연구

이전 선행연구를 탐색해본 결과, 본 연구 주제와 유사한 논문이 하나 있었는데, 연세대학교 보건대학원 보건정책관리학과 남혜진님의 ‘ 국회의원 입법활동 결정요인에 관한 연구 -제17대 국회를 중심으로’라는 논문이 있었다. 이 논문에서는 연구대상자 즉 국회의원의 성별, 연령, 학력 등의 인구사회학적 특성에 대한 다중회귀분석을 진행한 후, 국회의원 업무 특성에 대한 로지스틱 회귀 분석을 하였으며 그 결과 연령과 당선횟수가 입법 활동에 영향을 미치는 주요한 요인이라는 것을 알 수 있었다.

여기서 더 나아가 본 연구는 입법활동만 보는 것이 아니고 이전에 없었던, ‘법안 발의와 지연의 관계성’에 대해 새롭게 탐구하는 의의가 있다. 본 연구는 “N대에 걸쳐 시간이 흐름에 따라 지연과의 관계가 강화되었는지” 탐구하고자 한다.

본 연구는 우리나라 지역구의원 253명, 비례대표의원 47명을 대상으로 진행하였으며, 지역 변수에 대한 조사이기 때문에 비례대표나 출신지 정보가 없는 경우 지역구의원으로 당선되었을 때의 지역 정보로 대체했다.

1. 수집데이터

연구에 필요한 데이터로 필요한 것은 역대 국회의원의 지역정보에 대한 수집을 진행하였다. 1조는 ‘열린국회정보’포털에서 역대국회의원 5670명의 csv데이터를 수집하였고, 누락된 정보는 크롤링하여서 채웠다. 또한 법률안 발의자명단 918556116개의 법안 csv 데이터를 통해 법안당 대표발의자 1명, 공동발의자 n명의 정보를 얻을 수 있었다.

수집데이터를 정제하고 전처리하는 과정에서 결측값은 0으로 처리하였다.

역대국회의원 정보를 봤을때 출생지가 누락된 데이터가 1000건 정도 존재하였는데, 출생지 자체가 미파악이거나, 비례대표로 선발된 경우가 그러하였다. 하여 이 데이터들은, 출생지 대신에 첫 당선된 선거구를 입력하여 처리하였다.

1. 분석 방법

데이터 분석은 jupyter notebook에서 진행하였으며 '역대국회의원 정보' csv와 '법안발의데이터' csv를 등록해 활용하였고, 각 두 csv를 등록할 때 두 csv 파일 모두 분석에 불필요한 셀(국회의원 생년월일 등)은 제거하였다.

이후 데이터 분석에 있어서 지역별의 일치성을 나타내는 기준으로 무엇이 가장 적합할지를 고려했을 때 가장 일반적인 행정구역 8도를 기준으로 정하였다.

Csv에서 볼 수 있는 지역들은 '황해', '전남', '대전' 등으로 다양해서 분류에 어려움이 있었다. 그래서 모든 지역을 'def.replace' 코드를 통해 8도로 통일하였다. 예를 들어 '대전'은 '충청', '전남'은 '전라' 로 하여 광역시까지도 8도에 포함시켰다. 단, '서울'은 인구수가 많아 서울 그대로 진행하였다. (‘그림1’ 참고)



▲그림1

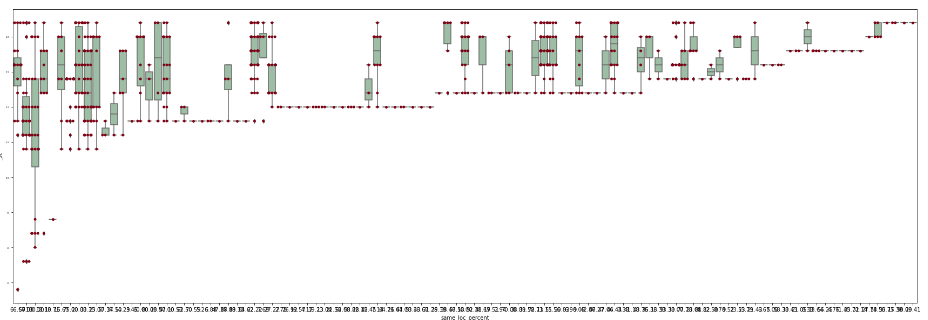
주제에 맞는 진행을 위해 법안당 '대표발의자의 지역'과 '공동발의자들의 지역' 일치성 확인을 위해 인적사항 데이터프레임에 '출생지 간편'이라는 셀을 추가해서 각 국회의원마다 어느 지역 출신인지 한눈에 볼 수 있도록 매칭하였다.

하여 예를 들어서 한 국회의원 법안에 대표발의자가 1명이고, 공동발의자가 5명이라면 대표발의자와 공동발의자1의 지역 비교, 대표발의자와 공동발의자2의 지역 비교, 대표발의자와 공동발의자3의 지역 비교 등을 통해 지역(8도)이 일치하는지 퍼센트를 비교하여 50% 이상이면 일치율이 높은 법안, 50% 미만이면 낮은 법안으로 기준을 정하였다.

1. 시각화

다음단계인 시각화에서는 시간에 걸쳐, 대수별로 지역이 일치하는 법안이 전체에서 차지하는 비중이 어떻게 되는지 한 눈에 볼 수 있도록 구상하였다. 본 ANOVA 분석의 귀무가설은 ‘’

Boxplot으로 시각화하기 위해 데이터셋을 ‘n대 국회’와 ‘법안 당 공동발의자의 지역정보 일치수준’ 의 두 칼럼으로 이루어진 DataFrame으로 재구성하였고, 그 그림1과 같이 시각화가 이루어졌다. ‘x축’은 ‘국회의원 대수’, ‘y축’은 ‘이를 통해 알 수 있는 것은, ’21대, 최근에 이르면서 지연의 영향력이 점차 커진다'라는 결론을 낼 수 있었다. 그러나, 시기에 따라 표본이 너무 다양하고 ‘지연 영향력을 크게 받은 안건’ 뿐 아니라 ‘받지 않은 안건’의 수 역시 늘어났기 때문에 일원분산분석이 요구 된다.(‘그림2’ 참고)



▲그림2

하여 1차적으로 결과를 분석하였을 때 알 수 있는 것은,

1)“지연이 ‘법안 발의에 영향을 미치는 정도가 있었다.

2) 하지만 해당 독립변수가 종속변수에 지배적으로 영향력을 발휘하는가를 따져 보았을 때 아닐 수 있다는 것

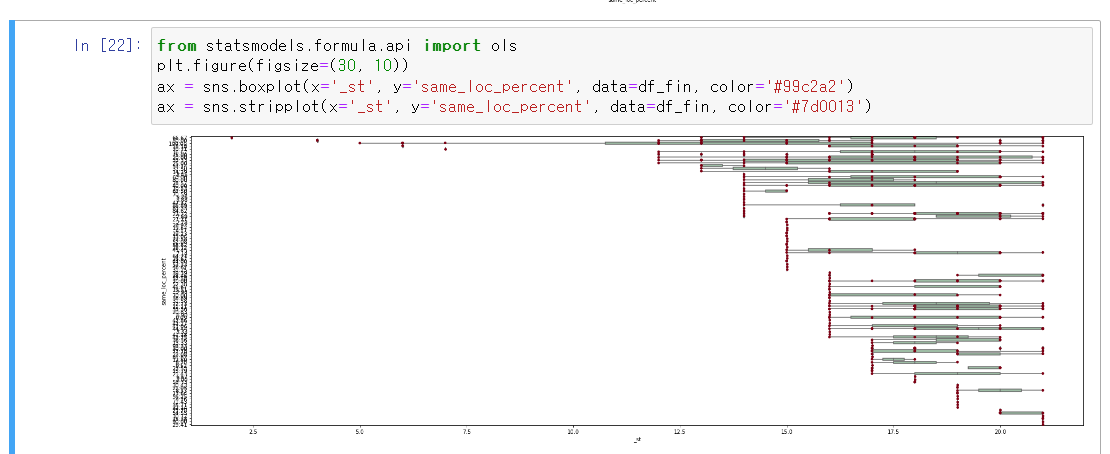
3)정당 정보가 지역 개념을 반영하고 있을 경우, 데이터의 사각지대가 발생할 수 있다는 것

그 결과 초기에 수립한 가설이 거짓임을 알 수 있었다.

그럼에도 불구하고 지연이 영향력을 가지는 변수임은 사실이고, 국회의원 법안 발의 시 지역에 따라 편파적인 판단을 하지 않을 장치가 필요해 보인다.

1차 분석 후에도 개선해야할 부분들이 많이 발견됐다. 이 보완점들을 수정하여서 2차 분석을 재진행 하였다.

첫째, 시각화에 있어서는, X축과 Y축의 위치를 바꾸어서 시간대별로 비중의 변화를 볼 수 있도록 하였다. (그림3 참고)



▲그림3

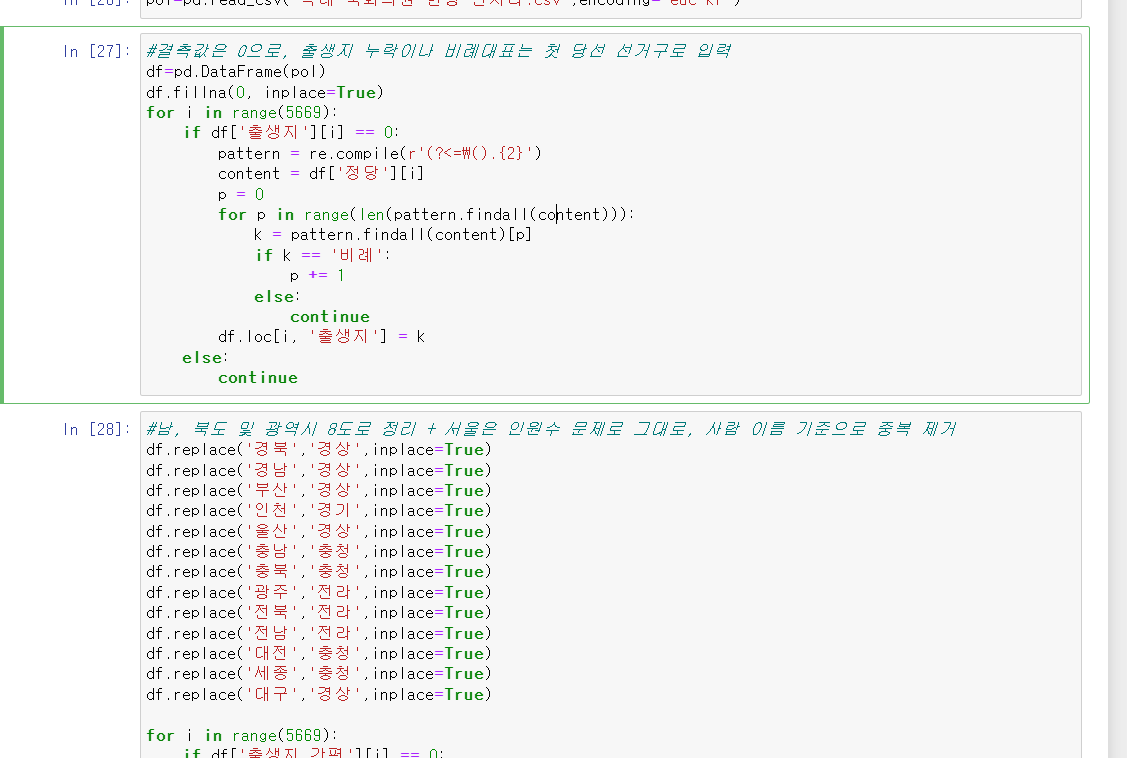
둘째, 독립변수와 종속변수에 대한 명확한 정의를 내렸다. 귀무가설은 ‘시간에 따른 법안 공동 발의에 지역 요인이 미치는 영향이 관계가 없다.’ 이고, 대립가설은 ‘시간에 따라 법안 공동 발의에 지역 요인이 미치는 영향이 증가한다.’이기 때문에 독립변수는 ‘법안당 가장 많은 지역의 비중’이고, 종속변수는 ‘해당 동의 비율을 얻은 법안이 발의된 국회의원 대 수’ 로 설정을 하였다.

셋째는 8도로 나누는 것에 대한 설득력이었는데, 실제로 지역감정이 정당 선호도에 영향을 많이 미친다는 연구 결과와 여론조사는 수없이 많다. 예를 들어 한 조사[[1]](#footnote-1)에 의하면, 강원도 내의 영서,영남 지역의 지역감정은 존재하지 않으나, 다른 도 지역 사이에는 지역감정이 존재한다는 것을 알 수 있었다. 또한 금번 대선 때에도 지역감정으로 인해 ’지역 몰표현상‘[[2]](#footnote-2)이 나타나는데, 지역감정은 대한민국의 정치발전을 가장 저해하는 요인 중 하나이다. 하여 이 8도의 분류를 통해 법안당 각 지역의 비중을 조사한다면 지역감정이 영향을 주는 비율도 살펴볼 수 있을 것이다.

본 연구를 통해 최종 결론을 내자면 “지연이 ‘법안 발의에 영향을 미치는 정도가 있었다는 것을 알 수 있었다. 귀무가설은 ‘시간에 따른 법안 공동 발의에 지역 요인이 미치는 영향이 관계가 없다.’ 이고, 대립가설은 ‘시간에 따라 법안 공동 발의에 지역 요인이 미치는 영향이 증가한다.’이기 때문에 독립변수는 ‘법안당 가장 많은 지역의 비중’이고, 종속변수는 ‘해당 동의 비율을 얻은 법안이 발의된 국회의원 대 수’ 로 설정을 하였을 때, 시간의 흐름에 따라 지역의 비중이 법안 발의에 주는 영향이 증가하고 있음을, 하여 지연과 법안발의는 유의미한 상관관계에 있다는 것을 시각화를 통해 알 수 있었다. 하지만 지연의 기준을 8도로 설정했기에, 지역감정이 영향을 주는 것은 맞지만 ‘같은 도 출신’으로만 해당 독립변수가 종속변수에 지배적으로 영향력을 발휘하는가를 따져 보았을 때 아닐 가능성을 배제할 수는 없다. 같은 정당인지 아닌지 지역에 포함될 수 있기 때문이다. 지연이 영향력을 가지는 변수임은 사실이고, 국회의원 법안 발의 시 지역에 따라 편파적인 판단을 하지 않을 장치를 만들기 위해서는 법안발의 시 공동발의자들 사이에서 같은 지역 출신이 과반수를 넘지 않도록 분배하는 제도가 필요해보인다.

5.시각화와 전체 코드

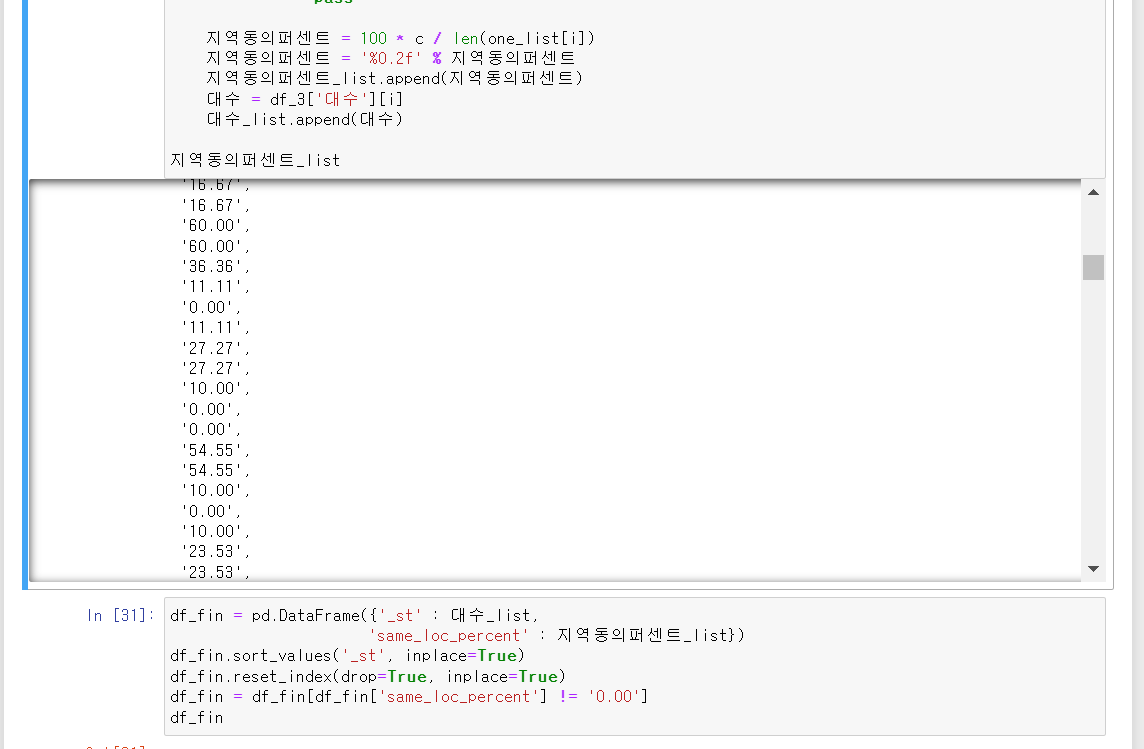




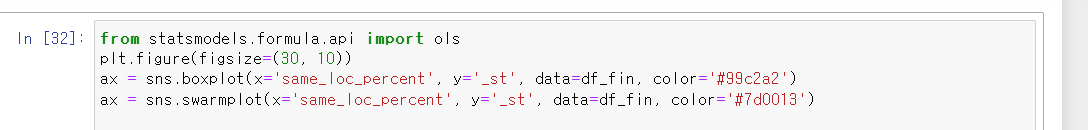


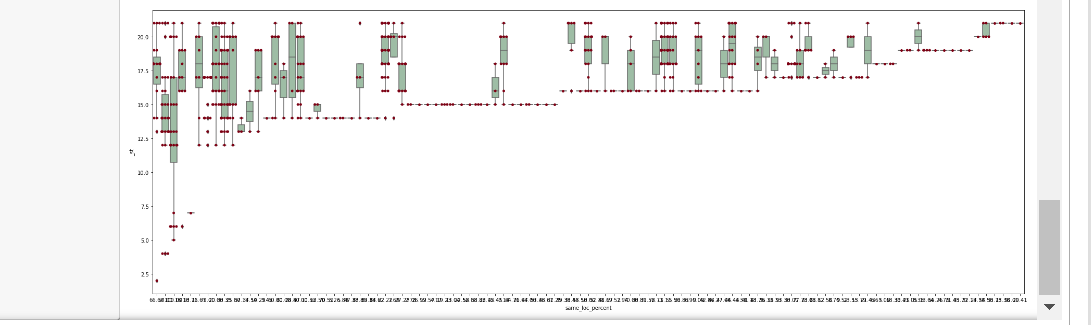


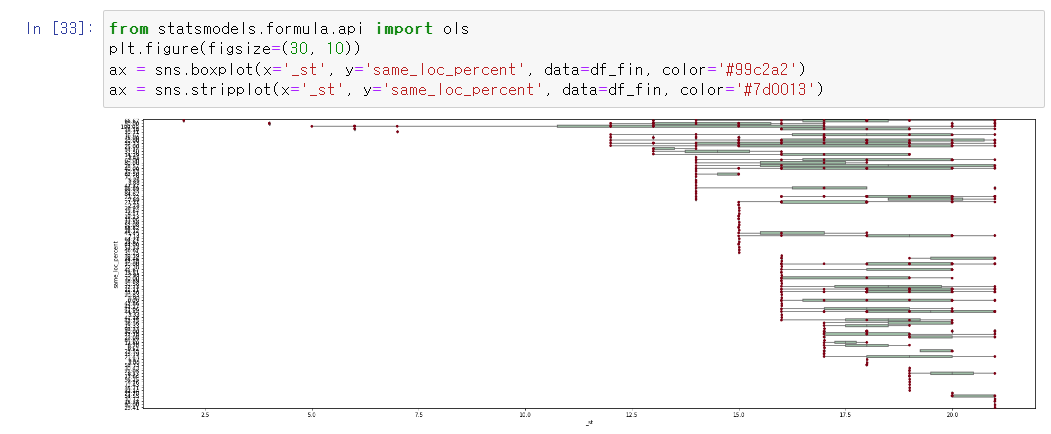


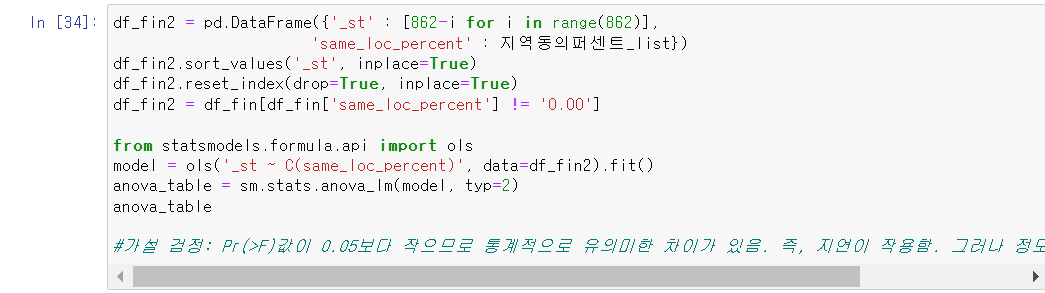


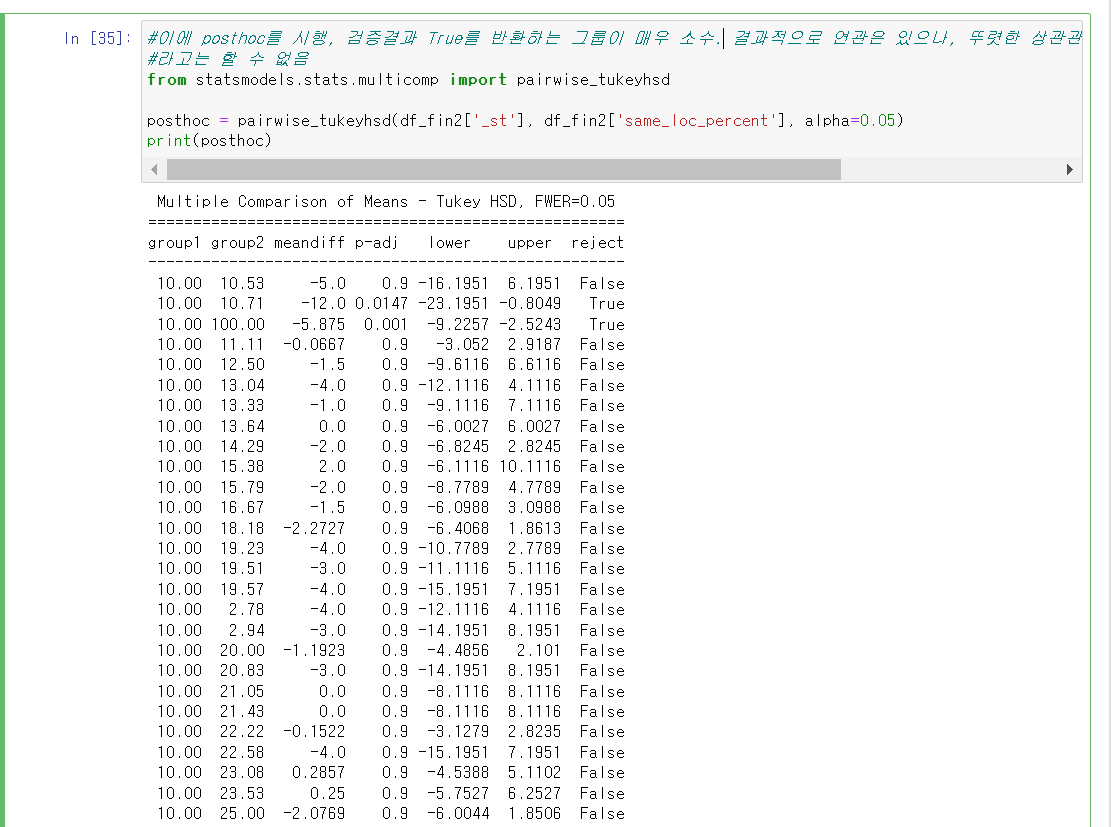












\*전체 내용은 분량관계상 생략함

6. 역할분담

역할분담은 다음과 같이 분배되었다.

• 자료조사: 강로이, 김민제

• 데이터 수집: 강로이, 김민제

• 데이터 정제: 김민제

• EDA: 김민제

• 분석: 김민제

• 시각화: 김민제

• 발표자료 제작: 강로이, 김민제

• 발표: 강로이, 김민제

• 보고서작성: 강로이

7.수업에 대한 피드백

• 수업에서 좋았던 점

- 학생들이 스스로 주제를 선정하여 데이터 프로젝트를 직접 이끌어갈 수 있도록 주도적인 프로젝트 진행이 가능한 분위기를 조성해 주셨다.

- 섬세한 피드백이 큰 도움이 됐다.

- 비교적 소수 인원으로 진행되는 강의였으나, 낙오자 없이 모든 수강생이 실습을 함께 진행할 수 있도록 강의 속도를 조절하며 커리큘럼을 운영하신 것이 매우 좋았다.

• 수업에서 개선되어야 할 점

- 수업시간 내에 조별로 함께 실습할 수 있는 시간이 있으면 나오는 에러사항들을 즉각적으로 교수님께 피드백 받을 수 있을 것이다.

- 해당 강의 이후 연강 등의 문제로 인해 오피스 아워를 활용하지 못하는 학생의 경우 아쉬움이 있었다. 실습 위주의 강의 특성상 대면으로 질의를 할 수 있는 시간이 보장되는 것이 중요하다고 생각한다.

- 매주 있었던 실습의 난이도는 유지하되, 실습 과제의 빈도를 매주에서 격주로 변경하는 것이 낫다고 생각된다.

1. http://www.kado.net/news/articleView.html?idxno=118825 [↑](#footnote-ref-1)
2. 전남일보 https://www.jeonmin.co.kr/news/articleView.html?idxno=351148 [↑](#footnote-ref-2)