Week 4 Report

저번 과제에서 불러온 dataset(LCA, PW, PERM)들의 유의미한 분석을 위해 각 dataset의 numerical data type을 추출하여 각 column간의 Pearson correlation coefficient를 계산, 이를 기반으로 column간의 유의미한 상관관계를 확인하였다. 이번에는 string data type의 data를 분석하여 데이터 분석의 insight를 주는 각 dataset의 핵심적인 column을 선벌해 보려 한다.

**LCA** **EDA**

LCA의 경우 LABOR CONDITION APPLICATION의 약자로 특정 유형의 비이민 취업 비자를 신청할 외국인 근로자를 대신하여 미국 고용주가 제출하는 신청서이다. 3개의 dataset 중 96개의 가장 많은 columns를 가지고 있다

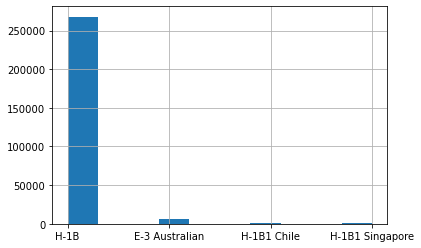


<LCA 96 columns>

이 중 유용한 정보를 직감적으로 선별하여 EDA를 수행해 보았다. 수월한 EDA를 위해 결측치(NULL)에 대해서는 대체값을 추가하는 것이 아닌 drop시키는 정책으로 진행하였다. 또한 data, email, phone number, address와 같은 개인 정보는 데이터 간의 연관성이 희박할 가능성이 높기에 분석에 제외하였다.

1. **VISAL\_CLASS**

우선 해당 년도의 LCA를 신청한 VISA에 대해 historam을 파악해 보았다.

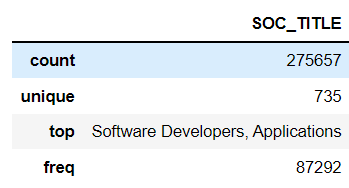
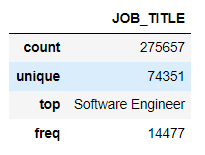


<LCA ‘VISA\_CLASS’ histogram>

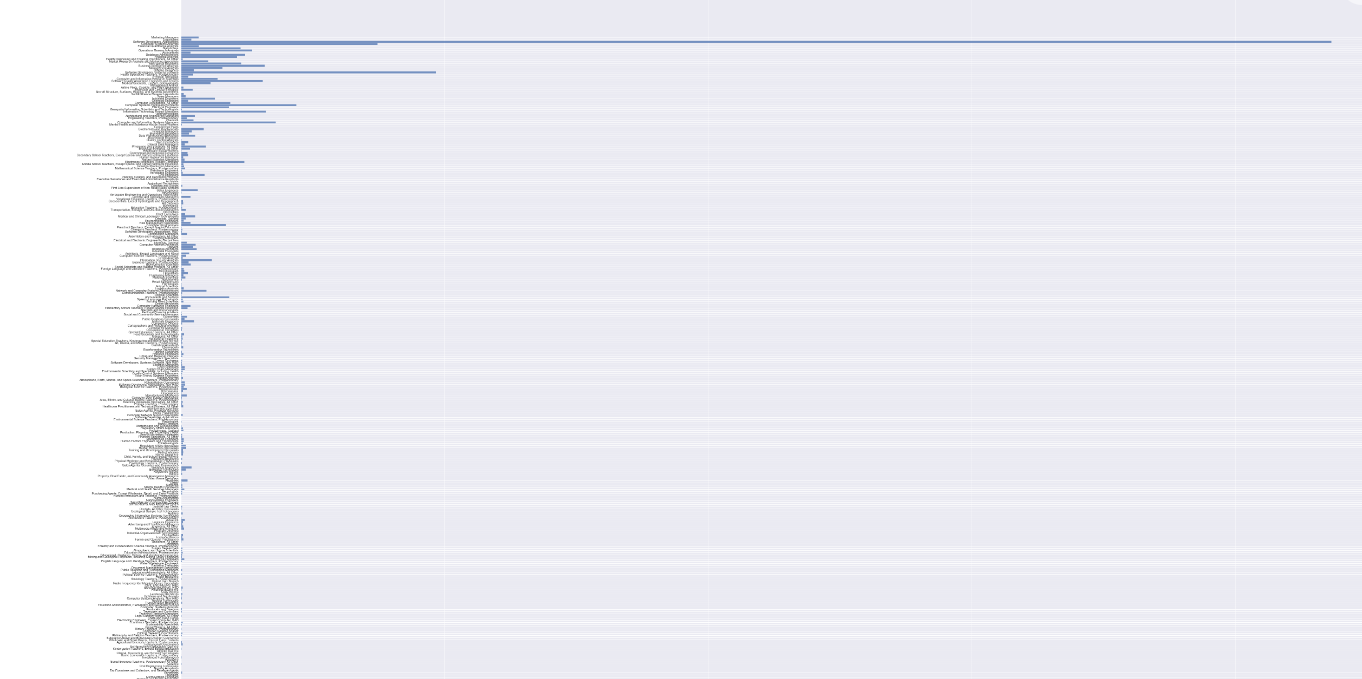
H-1B가 압도적으로 많은 것을 알 수 있다. 이 dataset의 분석은 H-1B VISA에 대한 분으로 일반화 시킬 수 있다.

1. **JOB\_TITLE**

Unique가 74351로 매우 많은 종류의 직업군이 존재하며 이를 computer resource의 한계로plot하기가 매우 어려웠다. 가장 많은 LCA Job Title을 Software Engineer로 전체의 약 20%를 차지한다. 그렇기에 JOB\_TITLE 보다는 unique가 735로 적은 SOC\_TITLE을 이용하여 분석을 해보았다.



<LCA ‘JOB\_TITLE’ & ‘SOC\_TITLE’ description>



<LCA ‘SOC\_TITLE’ histogram 1>

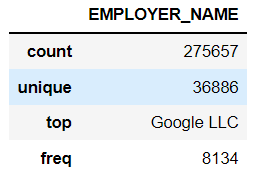


<LCA ‘SOC\_TITLE’ histogram 2>

무척 많은 종류의 SOC가 존재 했지만 histogram에서 볼 수 있듯이 Software Developers, Application이 압도적으로 높았으며 그 뒤로 Software Developers, System software가 뒤를 이었다. 그 외에도 Computer System Engineering, Computer and information System Manager등 다수의 컴퓨터 계열의 직종이 상위권을 차지하고 있는 것을 알 수 있다. 추후 데이터를 면밀히 분석해 볼 필요가 있지만 LCA의 절반 이상이 computer 계열 직종에 종사하는 사람들임을 추측 할 수 있다.

1. **Employer**
   1. EMPLOYER\_NAME

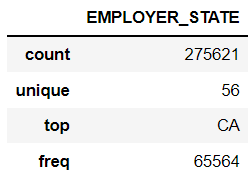
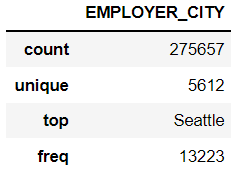
36886의 unique로 많은 종류로 구분된다. 최빈값은 Google LLC로 약 22%의 비율을 차지하고 있다. 종류가 너무 많은 관계로 차후에 데이터를 분할하여 분석을 진행해 보겠다.



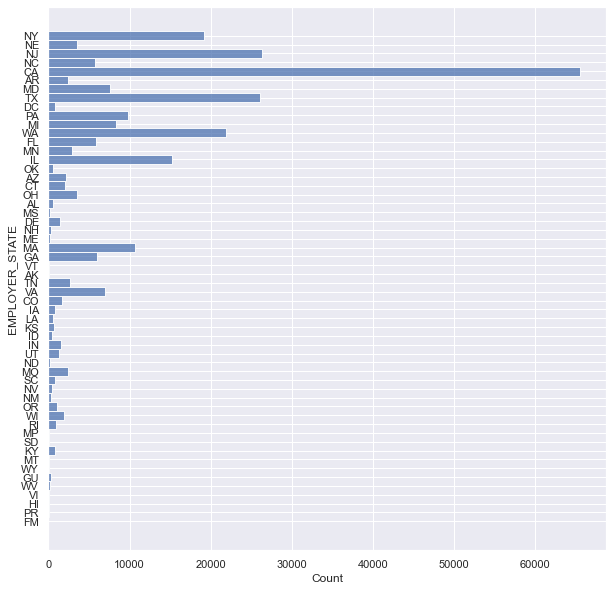
<LCA ‘EMPLOYER\_NAME’ description>

* 1. EMPLOYER\_CITY

5612의 unique로 적지 않은 분류이다. 그렇기에 이보다 좀더 상위 개념인EMPLOYER\_STATE를 분석해 보았다.



<LCA ‘EMPLOYER\_CITY’ & ‘EMPLOYER\_STATE’ description>

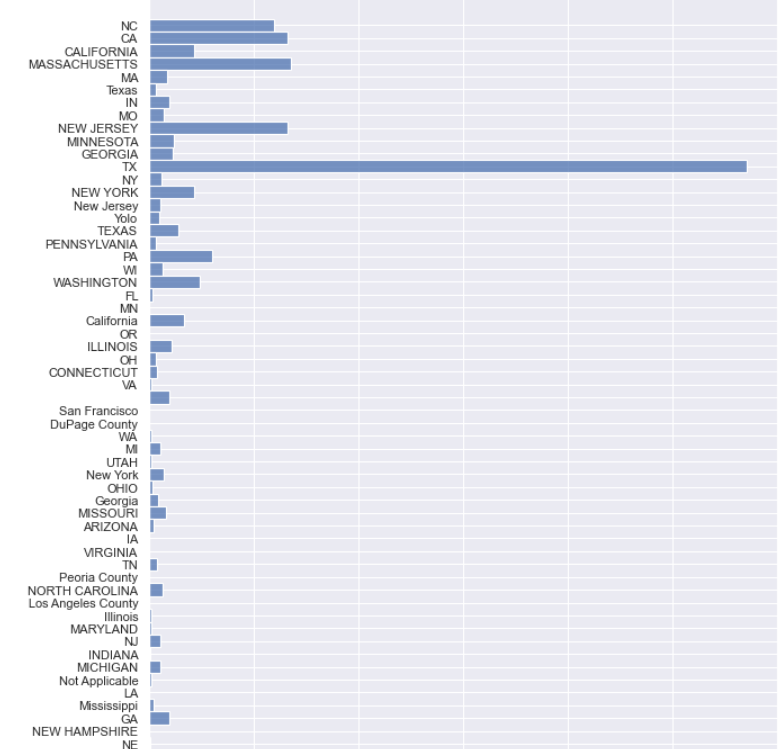


<LCA ‘EMPLOYER\_STATE’ histogram>

가장 많은 EMPLOYER\_STATE 실리콘 밸리가 위치해 있는 캘리포니아 주 였으며 이를 통해 LCA의 대부분이 computer 관련 직종자이고 실리컨 밸리가 위치한 캘리포니아 주로 유입된다는 것을 확인할 수 있다.

* 1. EMPLOYER\_PROVINCE

위 column의 경우 정확인 의미를 파악하기 어려웠다. 일반적으로 State와 province는 동치의 개념으로 이해하고 있지만 LCA에서는 다른 추가적의 정의로 사용되고 있는 것 같았다.



<LCA ‘EMPLOYER\_PROVINCE’ histogram>

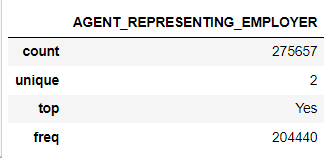
Texas 주가 압도적으로 많았다. 하지만 데이터 자체의 의미는 이해할 수 없었다. 추가적인 정보 검색이 필요해 보인다.

1. **Agent**

여기서 agent의 정의가 employer를 의미하는 건지 employee를 의미하는 건지 혼란스러웠다. 맥락상으로는 Employee에 가깝다고 고려되지만 아래 지표인 AGENT\_REPRESENTING\_EMPLOYER에서 agent가 employer를 의미하는 것 같아 혼돈이 있다.

* 1. AGENT\_REPRESENTING\_EMPLOYER

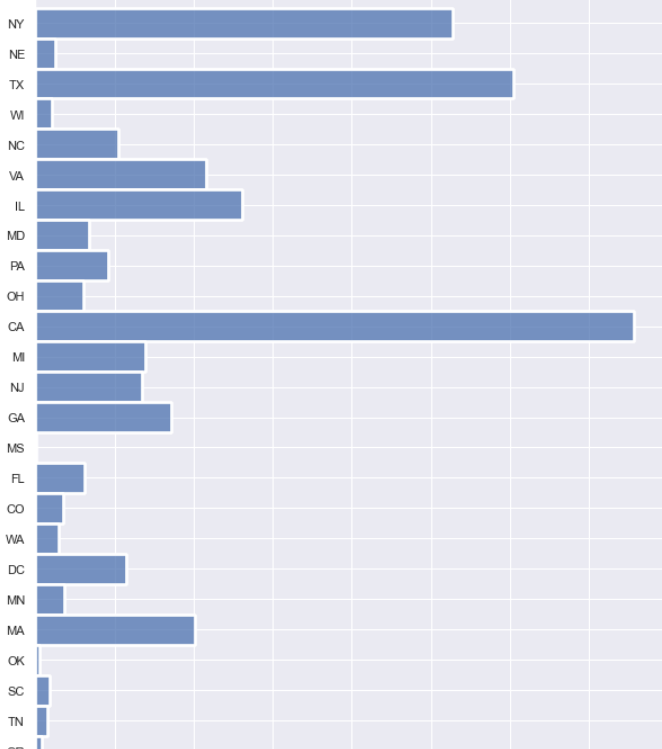
Agent가 employer를 대표하는지 묻는 지표이다.



<LCA ‘AGENT\_REPRESENTING\_EMPLOYER’ description>

* 1. AGENT\_ATTORNEY\_STATE

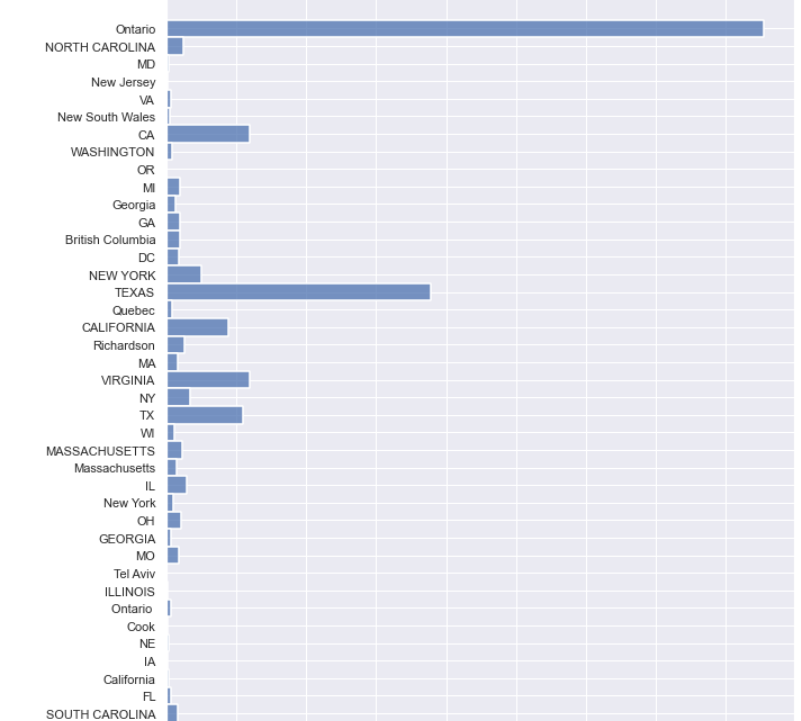
Employ와 마찬가지로 캘리포니아 주가 가장 많았다.



<LCA ‘AGENT\_ATTORNEY\_STATE histogram>

* 1. AGENT\_ATTORNEY\_PROVINCE

특이하게도 캐나다의 Ontario가 가장 많았다. Province에 대한 정의는 이후 찾아볼 예정

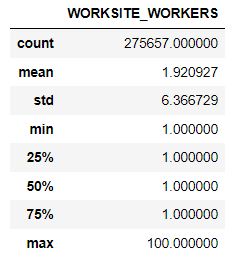


<LCA ‘AGENT\_ATTORNEY\_PROVINCE histogram>

1. **Worksite**

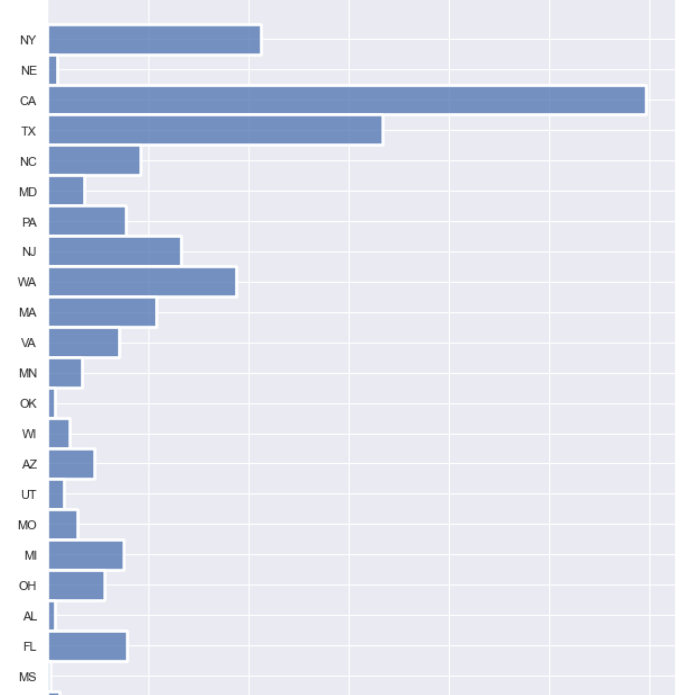
작업 현장에 대한 정보를 나타낸다

* 1. WORKSITE\_WORKERS



<LCA ‘WORKSITE\_WORKERS’ description>

* 1. WORKSITE\_STATE



<LCA ‘WORKSITE\_STATE’ histogram>

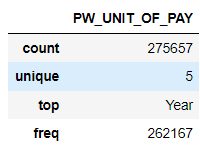
역시 캘리포니아 주가 가장 많은 것을 알 수 있다.

1. **PW**
   1. PREVAILING\_WAGE

위 경우 모슨 numerical data 앞에 $ 기호가 붙어 있어 string data로 인식된다. 이를 data processing을 통해 numerical type으로 바꿔준다면 유의미한 분석을 할 수 있을 것이다.

* 1. PW\_UNIT\_OF\_PAY

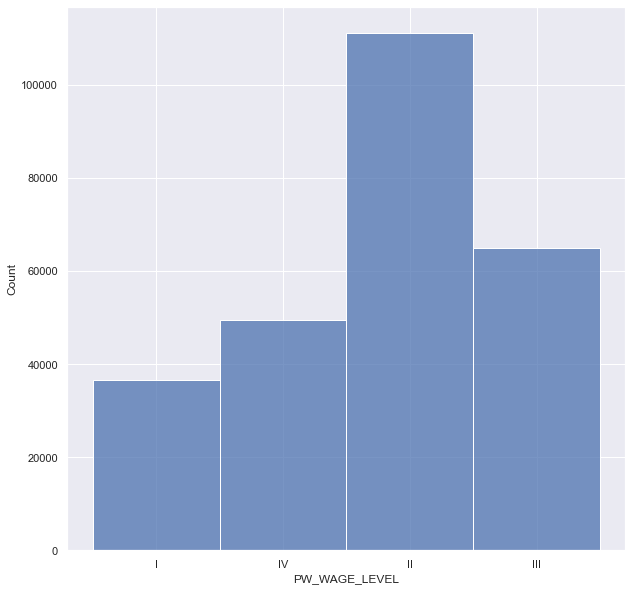
지불 단위에 대한 내용으로 거의 대부분이 연봉으로 임금을 지불하고 있다.



<LCA ‘PW\_UNIT\_OF\_PAY’ histogram>

* 1. PW\_WAGE\_LEVEL

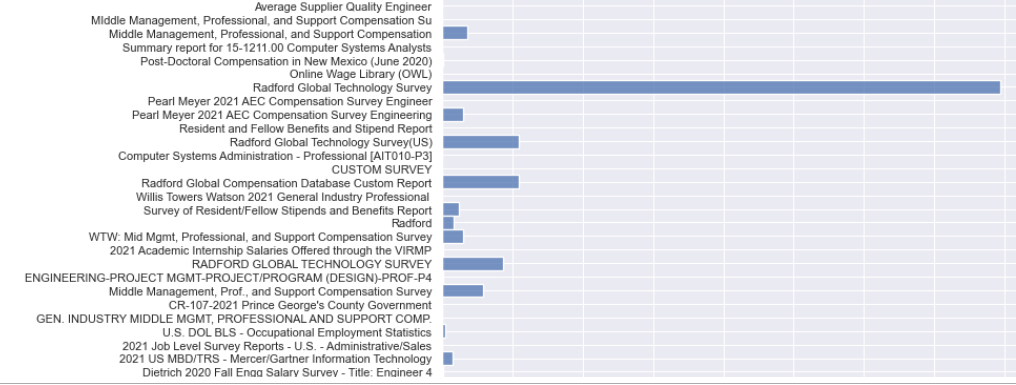
Level II가 가장 많았다. 아직까지는 유의미한 결과를 나타내지 못하지만 분포가 나름 고르기 때문에 PW LEVEL에 따른 추가적인 EDA 또한 가능 할 것이다.



<LCA ‘PW\_WAGE\_LEVEL’ histogram>

* 1. PW\_SURVEY \_PUBLISHER

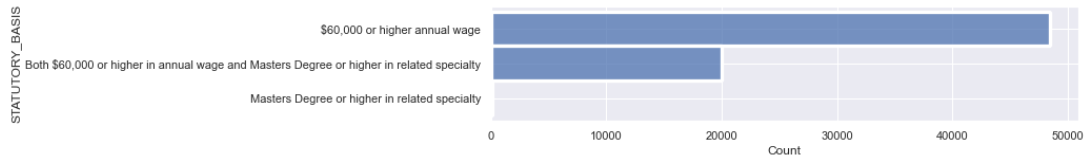
PW에 대한 survey를 진행한 publisher에 대한 내용으로 Radford Global Technology Survey가 압도적으로 많았다.



<LCA ‘PW\_SURVEY \_PUBLISHER’ histogram>

1. **STATUTORY\_BASIS**

의미적 분석 필요



<LCA ‘STATUTORY\_BASIS’ histogram>

LCA에 대한 EDA를 수행한 결과 대부분의 비자는 가장 기본적인 H-1B였으며 LCA를 통해 미국으로 입국하는 노동자들은 전반적으로 computer 분야의 직종의 노동자였다. 대부분이 캘리포니아 주로 유입되며 이는 미국의 첨단 산업 단지인 실리콘 밸리에 고용됐을 것이라 추정된다. PW LEVEL이 다른 지표에 비해 상대적으로 고른 편으로 이후 있을 PW에 대한 분석을 통해 유의미한 분석을 할 수 있을 것이라 고려된다.

PW와 PERM은 차주에 계속해서 분석을 진행 해 보도록하겠다.