2022 Spring AI·Big Data MBA

데이터 사이언티스트 인력 수요에 대한 텍스트 마이닝 분석 연구

최 태 림

August 2022

목차

표목차

그림목차

I. 서론

1. 연구 배경

“21세기 가장 섹시한 직업” 이라는 강렬한 수식어로 하버드 비즈니스 리뷰는 2012년에 ‘데이터 사이언티스트’ 라는 직업을 소개했다. 그로부터 10년이 넘게 지난 현재 구글 트렌드에서 ‘컴퓨터 사이언스’와 ‘데이터 사이언스’의 관심도의 수치 차이는 2005년에 비해 줄었다.[1]

또한, 위와 같은 관심도와 더불어 ‘데이터 사이언스’에 대한 교육적 수요도 증가했다. 465개의 미국 데이터 사이언스 프로그램의 데이터 분석을 하면 다음과 같은 결론을 도출할 수 있다. 데이터 사이언스 학문의 석사 학위가 전체 데이터 사이언스 교육 분야의 63.6%을 차지할 정도로 절대 다수이다. 이 결과는 구글 트렌드로 분석한 전 세계 데이터 와도 일치한다.[2]

2. 연구 목적

하지만, ‘데이터 사이언스’를 연구 혹은 실무적으로 다루고 있는 ‘데이터 사이언티스트’ 라는 직업에는 다양한 종류가 존재한다. 늘어나는 데이터 사이언스 학위 속에서 구직자는 자신에게 가장 필요한 역량이 무엇인지 파악하기 쉽지 않을 것이다. 그 중에서도 추가적으로 어떤 학문을 공부해야 하는지, 혹은 어떤 기술 스택 및 프로그래밍 툴을 공부해야 하는지 알고 준비하는 것이 어려울 것이다. 물론 매년 Stackoverflow에서는 각 분야의 유망한 프로그래밍 툴을 소개해준다. 하지만, ‘데이터 사이언티스트’ 라는 직군의 채용공고를 기준으로 ‘텍스트 마이닝’을 통해 분석했을 때 넓게는 현재 데이터 사이언티스트를 구인하는 기업에서 그들에게 가장 많이 요구하는 기술 스택 및 프로그래밍 툴이 무엇인지, 좁게는 기준을 더 추가하여 국가별로는 어떤 차이를 갖는지, 직급별로는 어떤 차이를 갖는지를 알아보고자 한다. 이를 통해, 데이터 사이언티스트가 되기 위한 구직자 혹은 기존 데이터 사이언티스트가 이직을 통해 산업군을 바꾸고자 할 때, 승진에 있어서 자신의 배경과 학습 목적에 맞는 데이터 프로그램을 공부할 수 있을 것이다. 그리고 커리어 패스를 설정하거나 본인에게 필요한 데이터 사이언스 관련 석사 학위를 찾게 될 때 큰 도움이 될 수 있을 것이다. 본 연구는 데이터 사이언티스트가 되기 위해 필요한 데이터 관련 프로그램 종류를 시장 수요를 바탕으로 분류하여 구직자에게 유용한 정보를 제공하고자 한다.

---

[1] Norita Ahmad, Areeba Hamid, Vian Ahmed. *Data Science: Hype and Reality.* IEEE, 2022

[2] Norita Ahmad, Areeba Hamid, Vian Ahmed. *Data Science: Hype and Reality.* IEEE, 2022

II. 이론적 배경

1. 데이터 사이언티스트의 개념[3]

데이터 사이언티스트의 개념에 앞서 우선 '데이터 사이언스'의 정의에 대해 알 필요가 있다. 데이터 사이언스는 수학적 지식, 컴퓨터 사이언스, 그리고 도메인 지식 세 가지 분야의 교집합으로 이루어져 있다. (참고 <그림 1>) 즉, 수학 및 통계학적 지식과 컴퓨터 사이언스 지식뿐만 아니라 비즈니스 도메인 지식까지 갖고 활용을 해야 데이터 사이언스라고 불리는 것이다.

<그림 1> 데이터 사이언스의 정의[4]

1) 데이터 제너럴리스트

그렇다면, 데이터 사이언스를 직업으로 하는 데이터 사이언티스트는 어떻게 정의할 수

[3] Renata Rawling-Goss. *Data Science Careers, Training, and Hiring*. Springer, 2019

[4] DataMixi, 데이터 사이언스 정의 이미지, http://datamixi.com/datascience

있을까? 미국의 저명한 학자 르나타 롤링 고스(Renata Rawlings Goss)는 '데이터 사이언스'를 데이터를 다루는 방식으로 크게 2가지로 나눴다. 하나는 '데이터 제너럴리스트(Data Generalist)'고, 다른 하나는 '데이터 스페셜리스트(Data Specialist)'이다. '데이터 제너럴리스트'는 좀 더 폭넓은 분야에서 데이터를 다루는 사람들에 대한 총칭이다. 르나타 롤링 고스는 여기에 그치지 않고 데이터를 다루는 방법에 따라 데이터 제너럴리스트를 다음과 같이 7개로 나눴다.

- 데이터 사이언티스트(Data Scientists)
- 데이터 분석가(Data Analysts)
- 데이터 설계자(Data Architects)
- 데이터 엔지니어(Data Engineers)
- 데이터베이스 관리자(Data Adminstrators)
- 비즈니스 분석가(Business Analysts)
- 데이터 분석 관리자(Data and Analytics Managers)

데이터 사이언티스트

데이터 사이언티스트는 문제를 진단하고, 문제를 해결하는 데 필요한 방법과 도구를 결정하고, 무엇이 효과가 있는지 탐색하고, 다시 시도하기 위해 높은 속도의 과학적 방법을 사용하기 때문에 진정한 과학자 또는 문제 해결사이다.

데이터 분석가

데이터 분석가는 통계 정보를 일상적인 사람들이 이해하고 실질적인 결정을 내리기 위한 기준으로 사용할 수 있는 언어로 번역한다. 이렇게 분석된 데이터는 더 합리적인 가격으로 원료를 조달하거나, 비즈니스 운영에 수반되는 운송 비용을 줄이거나, 회사에 너무 많은 비용이 드는 문제를 추적하는 데 사용될 수 있다.

데이터 설계자

빌딩 설계자와 마찬가지로 데이터 설계자는 데이터를 수용하고 데이터를 효과적으로 수집, 저장 및 관리하는 데 필요한 구조를 설계한다. 데이터 설계자는 디지털 비즈니스가 성공하는 데 필요한 정보를 해석하는 도구를 구축하기 위해 분석 기술을 사용하여 무한해 보이는

수많은 경쟁업체에서 기업이 확고한 명성을 쌓을 수 있도록 지원하는 데 필요한 인프라를 설계한다.

데이터 엔지니어

데이터 엔지니어는 데이터 설계자가 설계한 솔루션을 테스트, 평가 및 개선한다. 데이터 엔지니어는 강력하고 사용 가능한 솔루션을 만들기 위해 코딩 경험과 결합된 소프트웨어 엔지니어링 및 테스트 패턴에 대한 정확한 지식과 이해를 가지고 있다.

데이터베이스 관리자

데이터베이스 관리자는 물리적 및 가상 공간에 이 정보를 저장하고 백업하는 중요한 작업을 담당한다.

비즈니스 분석가

비즈니스 분석가는 일반적으로 기술적인 지식이 다소 떨어지지만 성공적인 비즈니스 운영에 수반되는 프로세스에 대한 깊은 지식을 제공한다. 그들은 이러한 통찰력을 기업이 더 성공할 수 있도록 지원하는 실제 세계 전략과 연결시키는 데 능한 사람들이다. 조직의 기술적 측면을 비즈니스 임무와 통합하는 중개자로 간주하고, 이 두 가지를 결합하여 성공적인 비즈니스 운영을 목표로 하는 솔루션 지향 전략을 제공한다.

데이터 및 분석 관리자

데이터 및 분석 관리자는 모든 당사자들에게 올바른 목표와 우선순위가 정해져 있을 뿐만 아니라 적절한 인재를 고용해야 할 책임이 있다. 데이터 및 분석 관리자는 선임 엔지니어, 분석가 또는 사이언티스트는 다른 유형의 강력한 사회적 기술을 요구합니다. 이들은 거의 위원회나 부서의 의장과 같은 독립적인 개인들로 구성된 팀을 이끌어야 하며, 또한 데이터 결과를 분석하고 검증할 수 있는 기술적 노하우를 가지고 있다. 성공적인 데이터 팀을 운영하고 장애물을 극복하는 것은 다양한 전문가의 전문적인 입력이 필요한 복잡한 일련의 이벤트다. 따라서 데이터 및 분석 관리자는 이러한 모든 배경에 대처하고, 결속력을 제공하며, 기술적으로 성향이 있는 사람들을 위해 흥미롭고 솔루션 지향적인 다양한 직업 선택을 제공해야 한다.

2) 데이터 스페셜리스트

데이터 제너럴리스트와는 다르게 데이터 스페셜리스트는 데이터 분야에 좀 더 세부적으로 전문적인 배경이 있는 사람이라 할 수 있다. 기업이 비즈니스와 관련된 특정 사용 사례에 집중할 수 있게 됨에 따라 전문화 경향이 점점 커지고 있다. 데이터 스페셜리스트는

기계학습, 자연어처리(NLP), 컴퓨터 비전 및 이미징 그리고 사물 인터넷(IoT)에 전문적인 성격을 가진 사람들이다.

2. 데이터 사이언티스트 특징[5]

데이터 사이언티스트가 갖는 특징을 데이터 사이언티스트가 갖는 스킬로 나타내 보았다. 르나타 골링 로스는 데이터 사이언티스트에게 필요한 스킬을 크게 2가지로 나눴다. 이노베이션 스킬(Innovation Skills)과 테크니컬 스킬(Technical Skills)이 그것이다. 본 연구에서는 테크니컬 스킬 그 중에서도 프로그램 스킬에 대해서 보다 깊게 다룰 예정이다.

1) 이노베이션 스킬(Innovation Skills)

혹자는 이 스킬을 '소프트 스킬 (Soft Skills)' 라고 한다. 하지만 르나타는 '소프트' 라는 단어가 주는 어감을 '도움은 되지만, 필수적이지 않은'으로 해석하여, 이를 좀 더 필수적인 느낌을 주기 위하여 이름을 바꿨다. 데이터 사이언스에서 혁신적인 마인드셋은 항상 가치 있게 쓰인다. 혁신은 산업 또는 프로세스의 변형으로부터 온다. 후술할 이노베이션 스킬의 4단계는 테크니컬 스킬을 갖추기 이전에 필수적으로 갖춰야할 과정이며, 이를 통해 데이터 사이언티스트의 시간 및 돈을 절약할 수 있다.

명확한 문제 인지

데이터 사이언티스트는 훌륭한 문제 해결사이다. 그래서 그들의 업무는 문제를 명확히 인지하는 데서 시작한다. 예를 들어, 단순히 '회사 콜센터의 효율성을 높이기 위해서 데이터를 통해 해결하자' 라는 말 보다는 콜센터의 운영 방식이 어떤 식으로 진행되고 있는지 파악하는 게 우선이다. 전부 기록되고 있는지, 사람들이 가장 빈번하게 묻는 것은 무엇인지, 데이터는 어떻게 저장되는지, 어디로부터 가장 많은 콜을 받는지 등과 같은 보다 명확한 프로세스 정의 및 문제 인지 능력을 보여야 한다.

도메인 전문가의 도움받기

배경 정보를 얻는 가장 좋은 또는 유일한 방법은 다양한 사람들과 이야기하고 동의를 얻고 의견을 묻는 것이다. 기술 담당자나 고위 경영진은 대부분의 경우 상황이 어떻게 돌아가는지 알고 있다고 잘못 생각하지만, 도메인 전문가가 작동 방식에 대해 뭐라고 말하든지 간에 항상 그들과 기본적인 수준의 차이가 있다. 데이터 사이언스 관리자나 번역가와 같은 중개자가 없는 한 이는 분석가의 몫이다. 그런 사람이 있더라도 당사자 간의 의사소통은 완벽하

[5] Renata Rawling-Goss. *Data Science Careers, Training, and Hiring*. Springer, 2019

게 이뤄지지 않기 때문이다. 이것을 조금이라도 해소할 수 있는 방법은 도메인 전문가와 자주 얘기하고 문제점을 함께 해결하는 것이다.

커뮤니케이션 능력

위에 언급한 도메인 전문가에게 도움을 받기 위해서는 원활한 의사소통 능력을 갖춰야 한다. 사람들은 데이터 사이언티스트가 하는 일을 좀 더 알기 쉬운 단어로 이해하고, 전달하고 싶어한다.

호기심과 지속적인 학습

데이터 사이언티스트는 호기심과 지속적인 학습을 유지해야 해당 분야에 대한 동향을 파악할 수 있고, 자신이 파악한 데이터를 통해서 미래를 예측할 수 있다. 이제 최상위 관리직의 경우에도 데이터 프로젝트를 지속적으로 관리하는 것이 핵심이라는 것이 분명해졌다. 관리직에서 데이터 사이언티스트가 되려는 경우 최신의 지식과 기술을 유지하는 것이 그 어느 때보다 중요하다. 많은 기업들이 리더십뿐만 아니라 팀을 지도할 수 있을 만큼 충분히 실무적으로 임할 수 있다는 점에서 이른바 '플레잉 코치'가 될 수 있는 지도자를 찾고 있다. 데이터 사이언스가 새로운 산업으로 빠르게 확장되고 리더가 팀에 처음 고용되는 경우가 많기 때문에 특히 그렇다.

2) 테크니컬 스킬 (Technical Skills)

앞서 언급한 이노베이션 스킬을 갖춘 뒤에 후술할 테크니컬 스킬을 얻는다면, 훌륭한 데이터 사이언티스트가 될 수 있다. 본 연구가 데이터 사이언티스트가 되기 위해서 어떤 프로그램을 주로 사용하는지 확인해보는 연구이고, 데이터 사이언티스트가 많이 쓰는 프로그램을 다루는 능력이 바로 이 테크니컬 스킬에 해당한다.

데이터 전처리 작업

손상되거나 불완전한 데이터를 찾아서 수정하거나 보완하는 작업은 데이터 사이언티스트가 전체 업무시간의 80%를 할애하는 작업이다.

기초통계, 미적분 및 대수학에 대한 올바른 이해

대부분의 데이터 사이언스 도구를 이해하고 정확하게 사용하기 위해서는 몇 가지 기본적인 통계 기술이 필요하다.

데이터 시각화 기술

데이터를 기반으로 의사 결정을 내리는 기업에게 특히 중요하다. 스토리를 말하고 미래 예측을 사용할 수 있는 기반을 마련하기 위한 접근 방법으로 쓰인다.

3. 데이터 사이언티스트 인력현황

1) 인력 현황[6]

앞서 설명한 것처럼, 데이터 사이언티스트에 대한 관심의 증가와 더불어 이것이 인력시장에는 어떠한 영향을 끼치는지 아는 것은 중요하다. 대한민국 기준으로 과학기술정보통신부에서 2020년에 작성한 '데이터 산업현황 조사' 통계정보보고서에서 데이터 산업에 종사하는 근로자 현황을 알 수 있다. (참고 <표 1>)

구 분		종사자 규모					
대분류	중분류	1-9인	10-49인	50-99인	100-299인	300인 이상	합계
데이터 솔루션	데이터 수집	257	165	56	24	2	504
	DBMS	87	48	6	1	4	146
	데이터 분석	148	145	19	9	1	322
	데이터 관리	519	392	72	67	9	1,059
	데이터 보안	67	70	36	21	5	199
	데이터 플랫폼	68	54	1	2	1	126
	소계	**1,146**	**874**	**190**	**124**	**22**	**2,356**
데이터 구축/ 컨설팅	데이터구축	1,016	718	242	119	45	2,140
	데이터컨설팅	265	203	34	24	8	534
	소계	**1,281**	**921**	**276**	**143**	**53**	**2,674**
데이터 서비스	데이터 거래	19	25	53	16	12	125
	정보제공	1,014	320	819	41	28	2,222
	데이터분석제공	100	61	8	64	9	242
	소계	**1,133**	**406**	**880**	**121**	**49**	**2,589**
합 계		**3,560**	**2,201**	**1,346**	**388**	**124**	**7,619**

<표 1> 대한민국 데이터 산업 종사자 현황 (단위: 개)

2) 데이터 산업 직접매출 시장규모[7]

[6] 과학기술정보통신부. *데이터 산업현황 조사*, 2020 년

[7] 과학기술정보통신부. *데이터 산업현황 조사*, 2020 년

(단위 : 억 원)

구분	2015년	2016년	2017년	2018년	2019년(E)	증감률 '17-'18	CAGR '17-'19(E)
데이터 솔루션	14,124	15,720	16,457	18,617	20,409	13.1%	11.4%
데이터 구축/컨설팅	26,698	27,875	30,847	37,009	38,971	20.0%	12.4%
데이터 서비스	16,128	16,928	18,339	30,102	32,714	64.1%	33.6%
합계	56,950	60,523	65,642	85,728	92,094	30.6%	18.4%

<표 2> 대한민국 데이터 산업 직접매출 시장규모

<표 2>에서 보는 것처럼 대한민국의 데이터 산업의 직접매출 시장규모도 매해 10%가 넘는 신장율을 보인다.

3) 연봉 정보

연봉 정보는 Stackoverflow에서 매년 발간하는 'Developer Survey'에서 확인해봤을 때, 아래 그래프와 같다. 미국, 영국, 독일, 인도, 캐나다의 개발자 대상이며, 한 해동안 받는 연봉의 중간값을 나타낸다. (단위: USD)

<표 3> Stackoverflow Developer Survey 2022 – Salary by developer type[8]

<표 4> Stackoverflow Developer Survey 2021 – Salary by developer type[9]

<표 2>와 <표 3>을 보면 Data Scientists or machine learning specialist 직군의 연봉 중간값이 약 18% 증가했고, Data or business analyst 직군은 약 14% 증가했다. 이는 이 직군에 대한 수요가 갈수록 높아지고 있다는 것이다.

[8] 'Developer Survey'. Stackoverflow. 2022. https://survey.stackoverflow.co/2022/

[9] 'Developer Survey'. Stackoverflow. 2021. Stack Overflow Developer Survey 2021

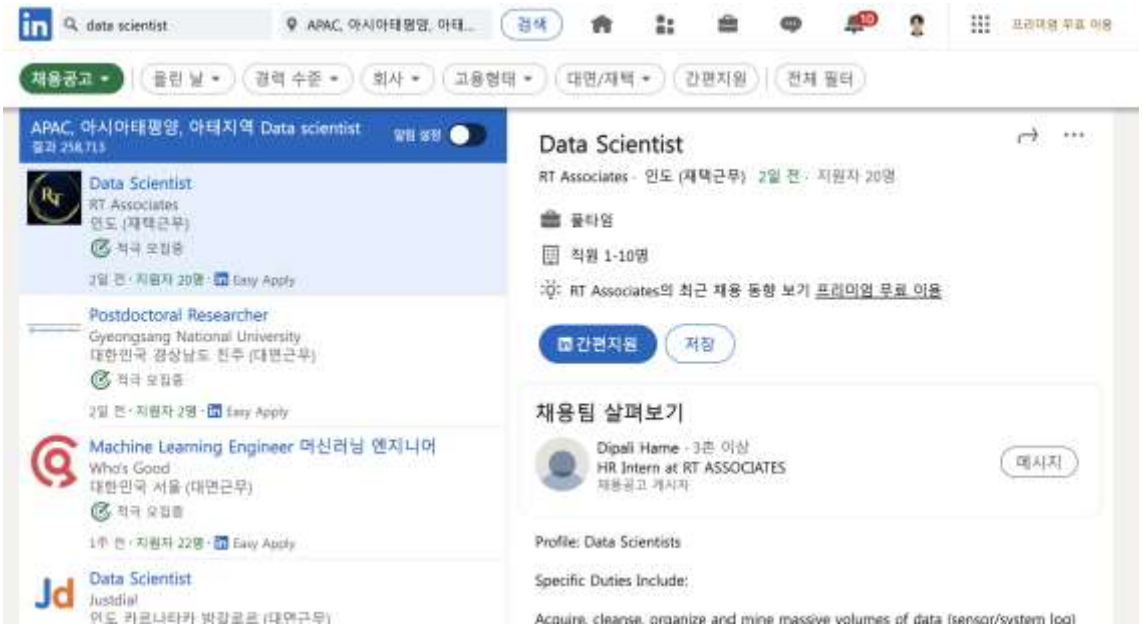
III. 연구 방법

1. 데이터 수집 및 정리

1) 크롤링

데이터 수집은 세계 최대 구인, 구직 사이트 '링크드인(Linkedin)'에서 진행했다. 링크드인을 데이터 수집 대상 서비스로 선정한 이유는 세계 최대 비즈니스 소셜 미디어 업체이기 때문이다. 이 소셜 미디어는 특정 업계 사람들이 서로 구인 및 구직에 대해서 동종 업계 사람을 파악할 수 있는 서비스를 제공한다. 2021년 기준으로 전세계 7억명이 이 서비스를 이용중이다. 그만큼 대량의 데이터를 수집할 수 있다고 판단하였기 때문에 이 소셜미디어를 데이터 수집 대상으로 선정하였다. 데이터 수집 방법은 '크롤링(Crawling)' 이라는 방법으로 진행하였는데, 크롤링의 정의는 '웹 페이지의 정보를 그대로 가져와서 거기서 데이터를 추출해내는 행위' 이다. 크롤링을 통해서 수집한 데이터를 바탕으로 .csv 파일을 만든 뒤 그 파일을 Python과 같은 프로그램으로 정리할 것이다.

우선 크롤링을 하기 위해서 아래 <그림 2>와 같이 링크드인 페이지에서 검색에 'data scientst'라는 글자를 입력 후 채용공고를 눌러보았다.



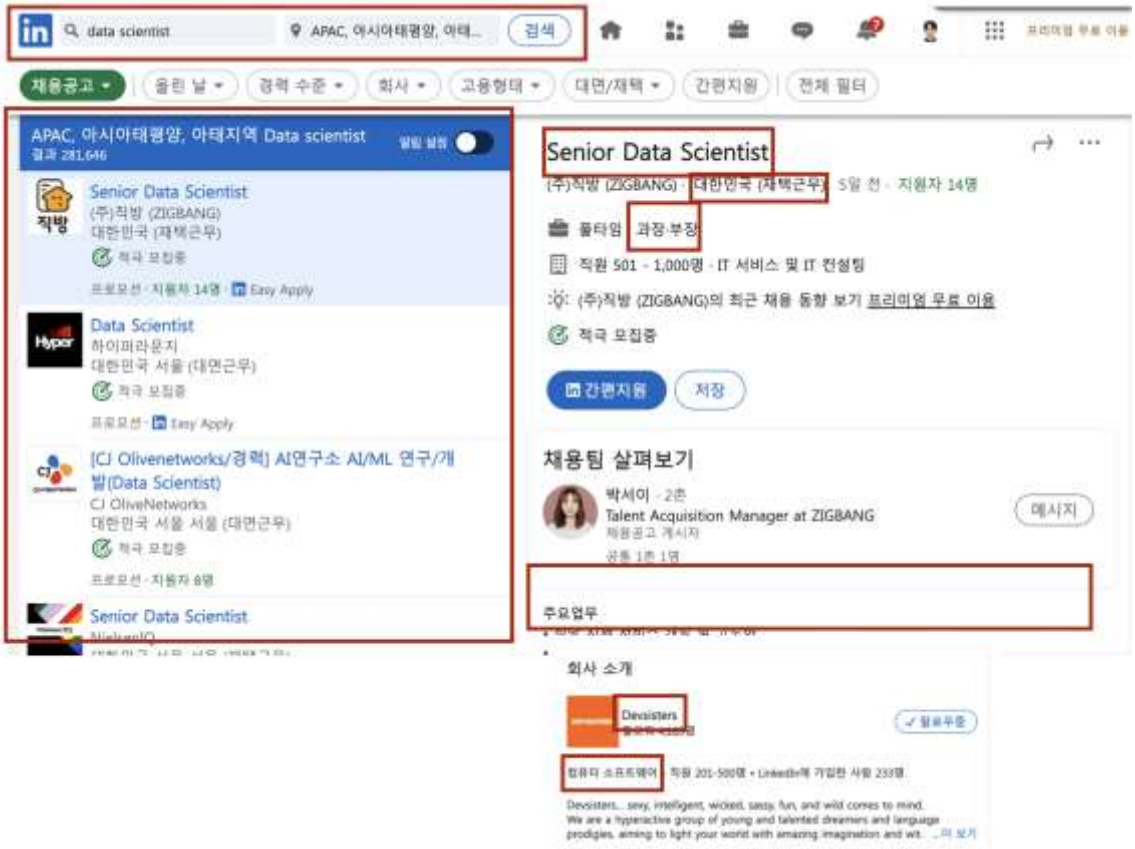
<그림 2> 링크드인 검색 페이지 화면

처음에 지역설정을 대한민국으로만 했을 때 'data scientist'의 채용공고가 약 1,300개 정도였

기 때문에, 이를 바탕으로 데이터를 수집하기엔 빅데이터 기준에 미치지 못한다고 판단했다. 따라서 지역 기준을 APAC(아시아 태평양)으로 넓혀서 한국, 일본, 미국, 중국, 대만, 호주, 뉴질랜드, 캐나다, 인도, 싱가폴, 베트남, 태국, 인도네시아, 필리핀의 채용공고로 추출하여 약 12,000개 데이터를 수집했다.

크롤링을 하기 위해서 준비한 툴은 크롤링 하는 툴은 AWS의 ec2에서 Rselenium이었고, 실제로 링크드인 페이지에서 검색결과는 한 챕터당 25개씩 40개 총 1,000개 가량만 보여줬기 때문에 위에 언급한 국가 14개를 검색어에 일일이 넣고 검색하여 URL을 확보한 뒤, 크롤링을 진행했다.

수집하는 데이터는 다음 <그림 3>에서 붉은 색 박스 친 부분이고, 그 명칭은 다음과 같다.



<그림 3> 링크드인 데이터 수집 대상 확인

- 검색어 'data scientist'
- 지역 'APAC(아시아 태평양)'
- 직무명

- 지역명
- 직급명
- 채용공고 상세내용 (e.g. 주요 업무 소개, 상세 내용)
- 회사명
- 산업군명

만약 과거의 채용공고부터 현재까지의 시계열 데이터를 볼 수 있으면 좋았겠지만, 채용공고의 특성상 지나간 공고를 확인할 수는 없었다. 따라서 채용공고만으로 구분될 수 있는 특징 위주로 위의 수집 대상 데이터를 선정하였다. 크롤링한 데이터를 csv 파일로 만들어 놓으면 다음 <그림 4>와 같은 결과를 얻을 수 있다.

	idx	recruit_name	company_name	location	industry	occupation	employment_form	career_level	content	target_url	eng_flag
0	1	Junior Data Scientist	Tridge	서울, 대한민국	도매 수입 및 수출	IT, 에널리스트 및 과학	풀타임	과장 부장	최고의 인재들과 최고의 문제를 풀어 최대 다수에 사람들과 그 혜택을 나눌 수 있는 ...	https://kr.linkedin.com/jobs/view/junior-data-...	0
1	2	Data Scientist	원티레이스	서울, 대한민국	데이터처리 제조	공학 및 IT	풀타임	대리	[이미지처리 기반 데이터 사이언스]■ 담당 업무1) 이미지 데이터의 효율적 관리 개발...	https://kr.linkedin.com/jobs/view/data-scientist...	0
2	3	Data Scientist	BAE Systems, Inc.	서울 인천 지역	국방 및 우주 제조	공학 및 IT	풀타임	신입	Job Description Join BAE Systems' Intelligence...	https://kr.linkedin.com/jobs/view/data-scientist...	0
3	4	Data Scientist	Miso, Inc.	서울	IT 서비스 및 IT 컨설팅	공학 및 IT	풀타임	신입	Services Productmiso의 Mission은 "We build techn...	https://kr.linkedin.com/jobs/view/data-scientist...	0
4	5	Machine Learning Engineer	포항공과대학교	포항	NaN	NaN	Contract	NaN	POSTECH(포항공과대학교)에서는 전국의 중소기업을 대상으로 최신 스마트 기술 및...	https://kr.linkedin.com/jobs/view/machine-lear...	0

<그림 4> Python으로 .csv파일을 읽었을 때 예시

2) 한글 워드카운트

앞서 csv 파일을 읽었을 때, 제일 먼저 전처리가 필요한 부분은 content 칼럼이다. 이것은 채용공고 상세 내용이 들어가 있는데, 원하는 기술 스택 및 프로그램 툴을 뽑기 전에 먼저 단어 단위로 바꾸어서 단어의 빈도수를 확인하는 것이 첫번째 과제다. 우선 단어의 형태로 바꾸기 전에, 이 파일의 문제점을 짚고 넘어가야 한다. 워드 카운트를 하기 위해서는 각 언어의 맞는 라이브러리를 활용할 줄 알아야 하고, 여러 언어를 동시에 적용하기엔 무리가 있다. 따라서 크롤링을 진행하기 전에 미리 확인 후 하나의 언어로만 분석에 들어갔어야 했다. 이 파일의 분석을 위해서 본 연구에서는 한글과 영어로 이루어진 채용 공고에 대한 워드 카운트를 할 예정이다. (일본어, 중국어, 인니어, 태국어 등 미분석)

한글 채용공고에서 워드 카운트를 하기 전에, 우선 한글로만 이루어진 csv 파일을 만들어야 한다. 뒤에서 언급하겠지만, location 칼럼에서 도시나 지역 이름은 지운 뒤 국가이름으로만 country라는 칼럼을 만들었다. 워드 카운트 준비를 하기위해서 다음과 같은 과정을 거친다.

- Location 칼럼에서 국가이름이 Korea라고 돼 있는 채용공고만 모두 모은 뒤, csv 파일을 만든다.
- 필요한 라이브러리를 파이썬에서 실행시킨다. (참고 <그림 5>)
- 한글을 제외한 모든 형태의 글자를 공백으로 만들고, 새로운 칼럼으로 만든다. (참고 <그림 6>)
- 새로운 칼럼의 모든 글자를 한 문장으로 만든다. (참고 <그림 7>)
- 불용어*는 제거 처리를 한다. (* 쓰지 않는 단어 예: 을/를, 이/가 등) (참고 <그림 8>)
- 단어 리스트를 만든다. (참고 <그림 9>)

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline
        import seaborn as sns
        from collections import Counter
        from konlpy.tag import Okt

In [59]: df = pd.read_csv("C:\Users\godlo\Desktop\최신태림\linkedin2.csv")

         df.head()
```

<그림 5> Python의 필요한 라이브러리 실행 후 csv 파일 읽기

```
In [54]: # 한글제외한 문자 스페이스 처리
         df['re_content'] = df['content'].str.replace(r'[^ㄱ-ㅣ가-힣]',' ', regex=True)
```

<그림 6> Python pandas로 데이터 읽기

```
In [57]: content_all = ''
         for i in range(len(df['re_content'])):
             content_all = content_all + ' ' + df['re_content'].loc[i]

In [58]: content_all = str(content_all)
```

<그림 7> 채용공고를 모두 한 문장으로 만들기

```
In [61]: rank_text = dict(rank_text)
         count_len = 10
         temp_dic = {}
         for key, value in rank_text.items():
             if value > count_len:
                 temp_dic[key] = value
         rank_text = temp_dic

         k_stopword = pd.read_csv("C:\Users\godlo\Desktop\최신태림\korean_stopword.csv")
         k_stopword = list(k_stopword['불용어'])

         temp_dic = {}
         for key, value in rank_text.items():
             if key not in k_stopword:
                 temp_dic[key] = value
```

<그림 8> 불용어 처리하기

```
'경험': 3712,
'개발': 3395,
'서비스': 2406,
'지원': 2193,
'데이터': 1705,
'기술': 1654,
'업무': 1611,
'운영': 1460,
'팀': 1342,
'채용': 1079,
'제공': 970,
'사용': 933,
'기반': 919,
'시스템': 868,
'서류': 864,
'관련': 824,
'분석': 773,
'플랫폼': 771,
'환경': 757,
```

<그림 9> 단어 세기

채용공고에 대한 단어 빈도를 세어 보면 <그림 9> 처럼 나온다. 이를 워드 클라우드 형태로 전환시키면 <그림 10>과 같이 한 눈에 알아볼 수 있다.

<그림 10> 한글 채용공고 word cloud

(워드 클라우드에서는 빈도가 높은 단어일수록 글씨 크기가 크게 나온다)

3) 영어 워드카운트

영어 워드 카운트 역시 한글 워드 카운트와 마찬가지의 과정을 진행한다. 하지만 한글 워드 카운트 시 쓰는 라이브러리와 차이가 있다. 한글은 konlpy 라는 라이브러리로 진행하지만, 영어는 nltk라는 라이브러리를 활용한다는 차이점이 있기에 코드도 약간의 차이가 난다. (참고 <그림 11>)

```
word_tokens = nltk.word_tokenize(content_all)

tokens_pos = nltk.pos_tag(word_tokens)

NN_words = []

for word, pos in tokens_pos:
    if 'NN' in pos:          ## Noun 종류 4개는 모두 'NN'을 포함하고 있어서 (NN, NNS, NNP, NNPS)
        NN_words.append(word)

# nltk.WordNetLemmatizer() 사용

wlem = nltk.WordNetLemmatizer()
lemmatized_words = []

for word in NN_words:
    new_word = wlem.lemmatize(word)
    lemmatized_words.append(new_word)

# 1차적으로 nltk에서 제공하는 불용어사전을 이용해서 불용어를 제거
from nltk.corpus import stopwords
stopwords_list = stopwords.words('english') # nltk에서 제공하는 영어 불용어사전

unique_NN_words = set(lemmatized_words)  ## 중복을 제거하기 위해 set(집합형)으로 변환
final_NN_words = lemmatized_words

# 불용어 제거
for word in unique_NN_words:
    if word in stopwords_list:
        while word in final_NN_words: final_NN_words.remove(word)

from collections import Counter  # import해서 사용

c = Counter(final_NN_words)  ## 단어 개수를 세어준다
print(c)
```

<그림 11> 영어 워드 카운트 시 Python 코드

<그림 11> 처럼 코드를 작성하면, 불용어를 제외한 단어 수를 아래 <그림 12>처럼 셀 수 있다.

word	counts
data	18512
experience	8411
team	6689
business	6026
science	4811
machine	4277
model	4152
product	3600
solution	3252
work	3161
analytics	3033
customer	2933
opportunity	2901
skill	2807
analysis	2718
development	2688
scientist	2652
technology	2646

<그림 12> 영어 채용공고의 워드 카운트

영어 역시 워드 클라우드의 형태로 만들면 <그림 13>과 같은 형태로 나타낼 수 있다.

<그림 13> 영어 워드 클라우드

4) 프로그래밍 툴 키워드 임의 설정

앞서 진행한 한글 및 영어 단어 빈도수 확인하는 과정에서 본 연구의 주제인 프로그래

밍 툴을 확인할 수 있는 단어는 상위권에 나오지 않았다. 한글 분석은 아예 추출이 안됐고, 영어 단어 빈도수 확인 작업에서는 30위에 'python'이 나온다. 따라서 채용공고 내용을 담고 있는 content 컬럼에서 뽑고 싶은 키워드를 임의로 설정한 후, 키워드를 임의로 세어주는 방법을 통해 csv 파일을 재정리했다. 프로그래밍 툴과 관련된 키워드는 Stackoverflow에서 매년 발표하는 'Developer Survey' 2022년판에서 가장 인기 있는 기술 스택 및 프로그래밍 툴을 확인 후 키워드 리스트를 만들었다. 키워드 리스트는 다음과 같다.[10]

python	**SAS**	**Databricks**	**Julia**
R	**Hadoop**	**airflow**	**MATLAB**
SQL	**Spark**	**PowerBI**	**Google Analytics**
Torch/Pytorch	**Tableau**	**ElasticSearch**	
Keras	**Scikit-learn**	**kafka**	
Tensorflow	**Superset**	**Numpy**	
Presto	**Scala**	**Pandas**	
AWS	**C++**	**Spring**	
EMR	**Hive**	**Azure**	
Kinesis	**Excel**	**Google Cloud**	

<표 5> 기술 스택 및 프로그래밍 툴 키워드 설정

<표 5>에 나온 리스트를 바탕으로 다음과 같은 과정을 통해서 새로운 데이터 프레임을 만든다.

- 키워드를 기준으로 새로운 리스트를 생성한다. (참고 <그림 14>)
- Content 컬럼에서 해당 키워드의 빈도수를 센 새로운 컬럼을 만든다. (참고 <그림 15>)
- 데이터 프레임을 다시 분석한다.

[10] 'Developer Survey'. Stackoverflow. 2022. https://survey.stackoverflow.co/2022/

```
In [227]: klist = {'python': ['python', '파이썬', '파이선'],
                   'R': ['R', '알'],
                   'SQL':['sql', '에스큐엘', 'MySQL', 'mysql', 'MS SQL', 'MS sql', 'ms sql', 'mssql'],
                   'Torch/Pytorch':['torch', 'Torch', 'pytorch', 'Pytorch'],
                   'Keras':['keras'],
                   'Tensorflow':['tensorflow', 'Tensorflow', 'TensorFlow'],
                   'Presto':['presto', 'Presto'],
                   'AWS':['Aws', 'aws'],
                   'EMR':['Emr', 'emr'],
                   'Kinesis':['kinesis', 'Kinesis'],
                   'SAS':['sas', 'Sas'],
                   'Hadoop':['HADOOP', 'Hadoop', 'hadoop'],
                   'Spark':['SPARK', 'Spark', 'Apache Spark', 'spark', 'apache spark'],
                   'Tableu':['태블로', 'tableau', 'TABLEAU', 'Tableau'],
                   'Scikit-learn':['scikit-learn', 'Scikitlearn', 'scikitlearn', 'Scikit-learn', 'scikit learn', 'scikit'],
                   'Superset':['superset', 'Superset'],
                   'Scala':['scala', 'Scala'],
                   'PySpark':['Pyspark', 'pyspark', 'PySpark'],
                   'C++':['C++', 'c++'],
                   'Hive':['hive', 'Hive'],
                   'Excel':['MS Excel', 'excel', '엑셀', 'Excel', 'ms excel'],
                   'Databricks':['databricks', 'Databricks'],
                   'airflow':['Airflow', 'airflow'],
                   'PowerBI':['powerBI', 'PowerBI', 'powerbi', 'power bi'],
                   'ElasticSearch':['elasticsearch', 'Elasticsearch', 'ElasticSearch', 'elastic search'],
                   'kafka':['kafka', 'Kafka', 'Apache kafka', 'Apache Kafka', 'apache kafka'],
                   'Numpy':['numpy', 'Numpy'],
                   'Pandas':['Pandas', 'pandas'],
                   'Spring':['spring', 'Spring'],
                   'Azure':['azure', 'Azure', 'Microsoft Azure', 'MS Azure', 'ms azure'],
                   'Google Cloud':['구글 클라우드', 'Google cloud', 'Google Cloud', 'google cloud'],
                   'Julia':['julia', 'Julia'],
                   'MATLAB':['matlab', 'MATLAB'],
                   'Google Analytics':['google analytics', 'Google Analytics']
                  }
```

<그림 14> <표 3>의 키워드를 바탕으로 키워드 재정리 리스트

```
In [229]: %%time
          for idx in df.index:
              print(df.loc[idx,'idx'], '----------')

              counts = Counter(df.loc[idx,'re_content'].split())

              for key, val in klist.items():
                  cnt = 0
                  for word in val:
                      cnt += counts[word.lower()]
                  print(f'key = [{key}], cnt = [{cnt}]')
                  df.loc[idx,key] = cnt

          print('\n[End of job] ----------\n')
```

```
1 ----------
key = [python], cnt = [1]
key = [R], cnt = [1]
key = [SQL], cnt = [1]
key = [Torch/Pytorch], cnt = [0]
key = [Keras], cnt = [0]
key = [Tensorflow], cnt = [0]
key = [Presto], cnt = [0]
key = [AWS], cnt = [0]
key = [EMR], cnt = [0]
key = [Kinesis], cnt = [0]
key = [SAS], cnt = [2]
key = [Hadoop], cnt = [0]
key = [Spark], cnt = [0]
key = [Tableu], cnt = [0]
key = [Scikit-learn], cnt = [0]
key = [Superset], cnt = [0]
key = [Scala], cnt = [0]
key = [PySpark], cnt = [0]
```

<그림 15> 설정된 키워드를 바탕으로 content 컬럼의 키워드 세기

2. 데이터 분석 방법

본 연구의 주제는 채용공고를 텍스트 마이닝을 통해서 채용공고에서 가장 많이 쓰이는

기술 스택 혹은 프로그램 툴을 확인하는 것이다. 따라서 이 파일에서 content 부분에 채용 공고 상세내용이 들어가 있기 때문에 여기서 필요한 기술 스택 및 프로그램 툴을 추출해야 한다. 아래와 같은 과정을 통해서 csv 파일을 분석할 것이다.

- 키워드 그룹에서 나오는 기술 스택 혹은 프로그래밍 툴을 확인한다.
- 그룹화 시킬 수 있는 컬럼의 데이터 전처리를 한다. (예시: location 컬럼에서 country 정보만 추출)
- 확인된 기술 스택 혹은 프로그래밍 툴을 앞의 정리된 여러 컬럼으로 그룹화 한다.
- 그룹별 특징이나 차이점을 파악한다.

IV. 분석 결과

1. 종합

앞서서 csv 파일 속 키워드를 새로운 컬럼으로 만들면 아래 <그림 16>과 같은 표를 얻을 수 있다.

In [230]: df.head()

Out[230]:

career_level	content	target_url	...	ElasticSearch	kafka	Numpy	Pandas	Spring	Azure	Google Cloud	Julia	MATLAB	Google Analytics
과장 부장	최고의 인재들과 최고의 문제를 풀어 최대 다수의 사람들과 그 혜택을 나눌 수 있는 ...	https://kr.linkedin.com/jobs/view/junior-data-...	...	0	0	0	0	0	0	0	0	0	0
대리	[이미지처리 기반 데이터 사이언스] ■ 담당업무1) 이미지 데이터의 효율적 관리 개발 ...	https://kr.linkedin.com/jobs/view/data-scienti...	...	0	0	0	0	0	0	0	0	0	0
신입	Job Description Join BAE Systems' Intelligence ...	https://kr.linkedin.com/jobs/view/data-scienti...	...	0	0	0	0	0	0	0	0	0	0

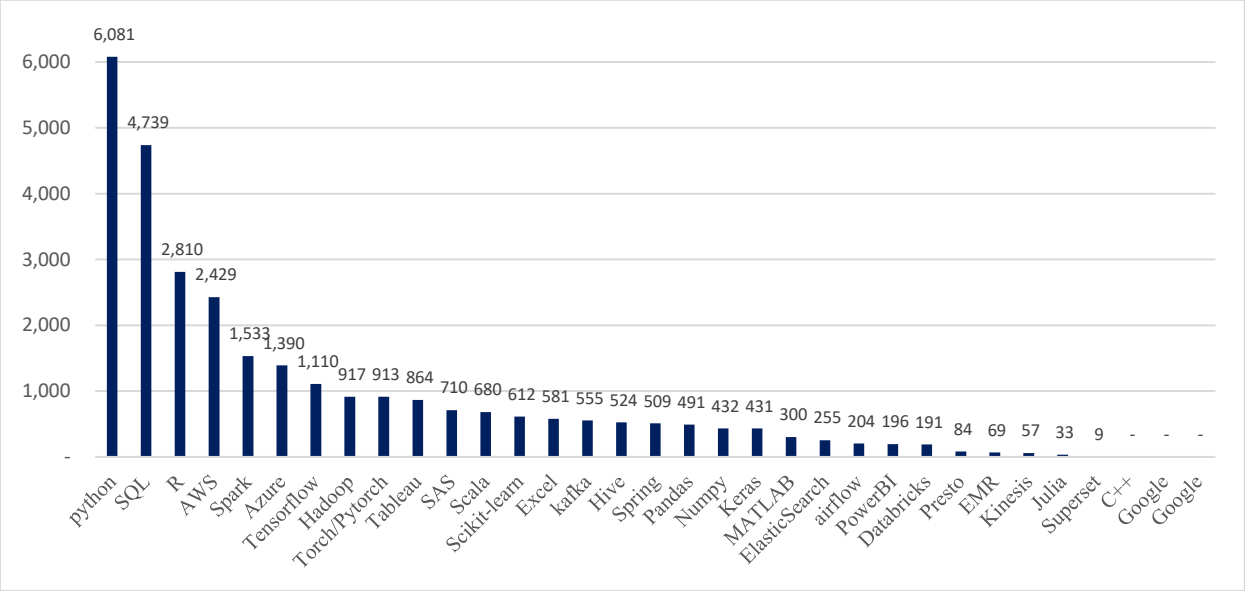
<그림 16> 기술 스택 및 프로그래밍 툴의 새로운 컬럼화

위 csv 파일을 바탕으로 기술 스택 및 프로그래밍 툴을 전체적으로 분석해 본 결과 <표 6>처럼 1위 python이 6,081 2위 SQL 4,739건, 3위 R 2,810건 순서로 나타난다. 이 상위 3개의 결과치는 데이터 사이언티스트 채용공고에서 다뤄야 할 기술 스택 및 프로그래밍 툴로 가장 중요한 것이 python이라는 것을 알 수 있다. 이 3가지 언어에 대해서 짧게 언급하자면 다음과 같다.

Python은 Stackoverflow의 'Developer Survey 2022'에서 언어 랭킹 4위로 현재 데이터 사이언티스트 사이에서 가장 널리 쓰이는 프로그래밍 언어이다. 데이터를 쉽게 다룰 수 있는 pandas, numpy와 같은 라이브러리를 활용할 수 있고, 머신러닝, 인공지능 분야에서도 쓸 수 있는 다양한 종류의 라이브러리를 보유하고 있다.

SQL은 'Structured Query Language'의 약자로 데이터베이스 관리 시스템 (RDBMS)에서 데이터를 처리하기 위해 설계된 특수 목적의 프로그래밍 언어이며, 질의 언어라고 불리기도 한다. Python과 마찬가지로 데이터 사이언티스트 사이에서 널리 쓰이는 언어이며, Stackoverflow의 'Developer Survey 2022'에서 python보다 1단계 위인 언어 랭킹 3위로 나온다. 회사별로 언어의 차이가 있으며 대표적인 언어로 'MySQL', 'MSSQL, 'Oracle SQL'

등이 있다.

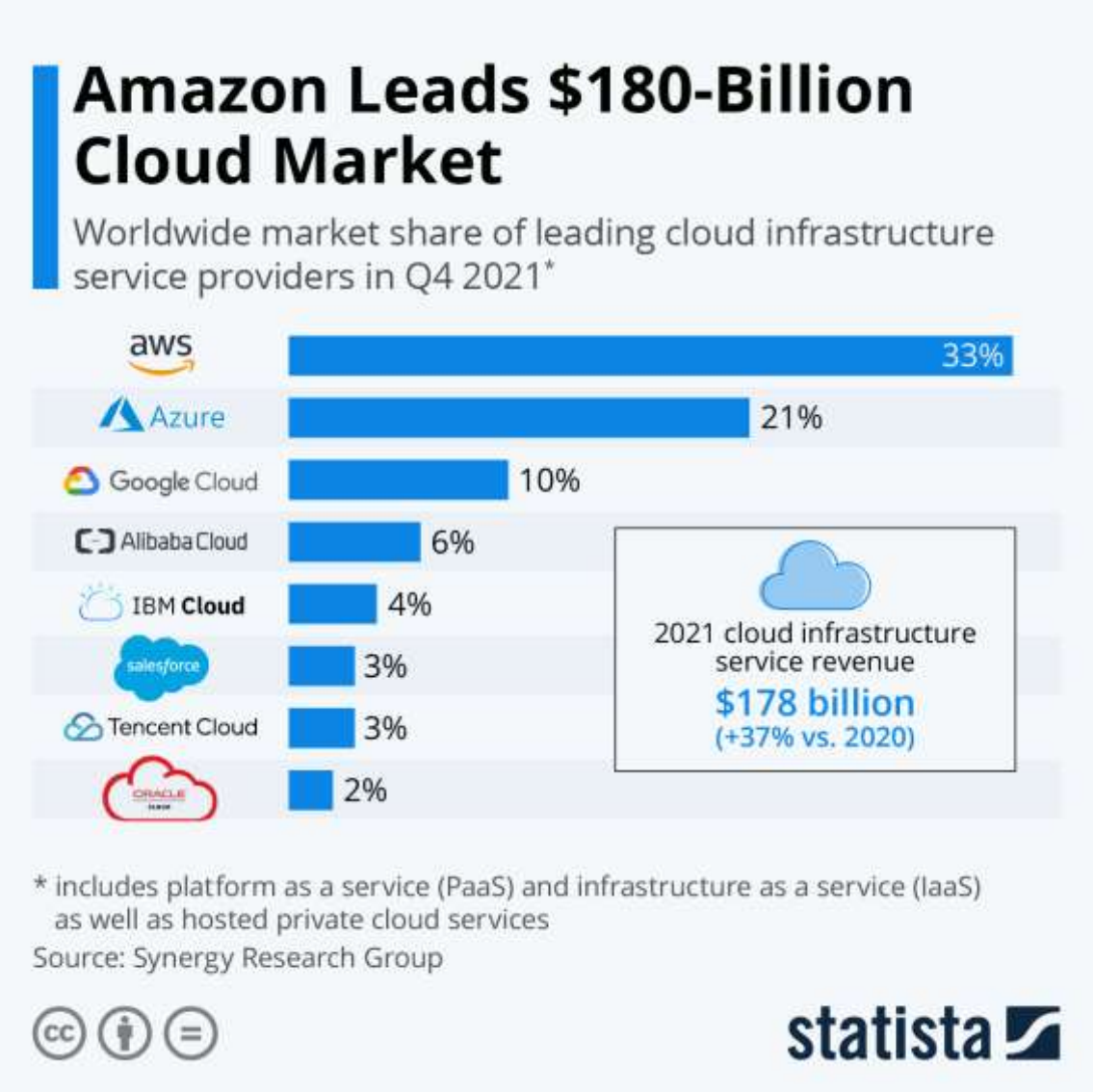


<표 6> content 컬럼의 기술 스택 및 프로그래밍 툴 현황

R은 원래 통계 계산에서 주로 쓰이던 언어이다. 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경이다. 오픈소스로 쓰였으며 무료로 사용이 가능하다는 특징으로 많은 사람들이 즐겨 쓰고 있는 통계 데이터 프로그램이다.

위의 세가지 프로그램과 다르게 4위에 위치한 AWS와 6위에 위치한 Azure는 클라우드 컴퓨팅을 다루는 프로그램이다. 클라우드 컴퓨팅이란[11], 인터넷 기반의 컴퓨팅을 의미한다. 인터넷 상의 가상화 된 서버에 프로그램을 두고 필요할 때마다 컴퓨터나 스마트폰 등에 불러와 사용하는 서비스이다. 클라우드(cloud)라는 단어가 말해주듯, 인터넷 통신망 어딘 가에서 구름에 싸여 보이지 않는 컴퓨팅 자원 (CPU, 메모리, 디스크 등)을 원하는 대로 가져다 쓸 수 있다. 인터넷에 연결된 어느 곳에서나 이 자원을 보장받을 수 있다는 의미다. 클라우드 컴퓨팅 중에서 AWS와 Azure는 IaaS (Infrastructure as a Service)라 불린다. IaaS는 서비스로써의 인프라를 뜻하는데, 이는 사용자가 관리할 수 있는 범위가 가장 넓은 클라우딩 컴퓨팅 서비스다. 서버 OS부터 미들웨어, 런타임, 그리고 데이터와 어플리케이션까지 직접 구성하고 관리할 수 있다. 위 표에서 보듯 데이터 사이언티스트가 되기 위해서 이 클라우드 컴퓨팅에 대한 이해도를 갖는 것이 중요하다는 것을 알 수 있다. (참고 <그림 17>)

[11] 가비아 라이브러리 – 클라우드 컴퓨팅이란 무엇인가



<그림 17> 전세계 클라우드 컴퓨팅 시장 점유율 (출처: statista)[12]

이 외에도 5위 Spark, 7위 Tensorflow, 8위 Hadoop, 9위 Torch/Pytorch 그리고 13위 Scikit-learn과 같은 그룹으로 묶일 수 있는데, 이들은 모두 빅데이터를 처리하기 위한 프로그래밍 툴이다. Spark와 Hadoop은 주로 대규모 클러스터 컴퓨팅을 도와주는 라이브러리이다. 클러스 컴퓨팅이란, 분산 컴퓨팅이라고 불리며 인터넷에 연결된 다량의 컴퓨터를 통해서 거대한 계산 문제를 해결하기 위한 즉, 빅데이터 처리를 위한 컴퓨팅이다. 또한, Tensorflow, Torch/Pytorch 그리고 Scikit-learn은 딥러닝과 기계학습 분야에서 주로 활용되는 라이브러리이다. 이 프로그램에 대한 요구는 단순히 작은 데이터를 다루는 것을 넘어서 빅데이터를 처리하고 계산할 수 있는 능력까지 필요하다는 것을 알 수 있다.

[12] 'Amazon Leads $180-Billion Cloud Market'. Statista. 2022. https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/

2. 유형별 그룹화

데이터 전처리 하는 과정에서 기본 컬럼을 바탕으로 총 4가지 유형을 나눠보았다. Location 컬럼에서 국가/지역별 구분을 얻을 수 있었고, employment_form컬럼에서 고용형태별, career_level 컬럼에서 직급별, 마지막으로 industry 컬럼에서 산업군별 유형을 지었다.

1) 국가/지역별

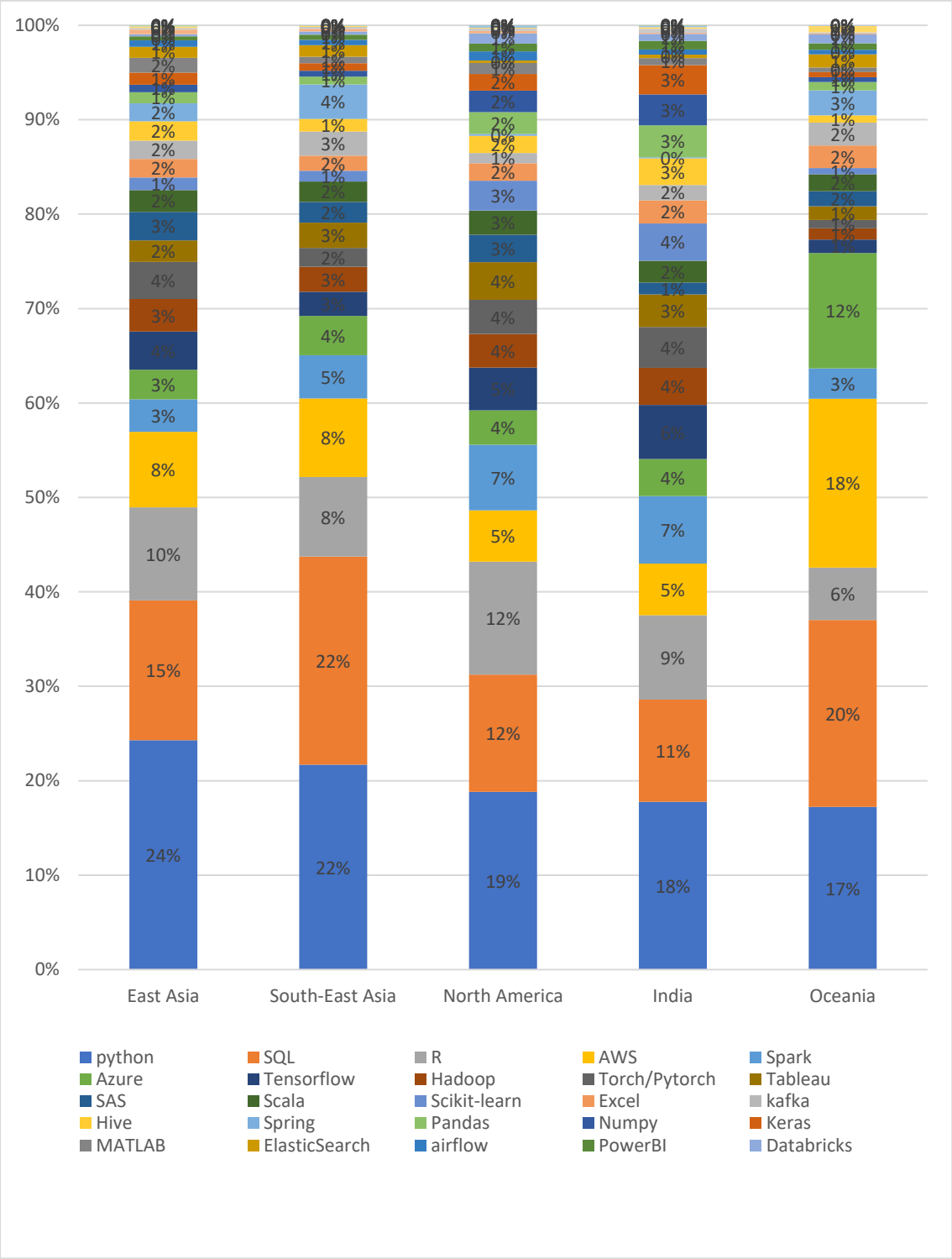
Location 컬럼을 전처리하는 과정에서 도시 이름이 있는 데이터와 없는 데이터가 공존했기 때문에 도시 이름을 지우고 국가 이름만으로 새로운 컬럼인 country 컬럼을 만들었다. Country 컬럼은 Korea, Japan, USA, China, Taiwan, Australia, New Zealand, Canada, India, Singapore, Vietnam, Thailand, Indonesia, Philippine으로 정리했고, 지역별로 region 컬럼을 만들었는데, East Asia (Korea, Japan, Taiwan, China), North America(USA, Canada), Oceania(Australia, New Zealand), India, South-east Asia(Vietnam, Thailand, Indonesia, Philippine)로 나눴다. 각 국가별, 지역별 건수는 아래 <표 7>와 같다.

지역구분	건수
계	12,780
South-East Asia	4,337
East Asia	3,498
North America	1,993
Oceania	1,974
India	978

국가구분	건수
계	4,595
Australia	989
Canada	996
China	796
India	978
Indonesia	836
Japan	843
Korea	950
New Zealand	985
Philippine	955
Singapore	863
Taiwan	909
Thailand	762
USA	997
Vietnam	921

<표 7> 지역별, 국가별 건수 (총 12,780건)

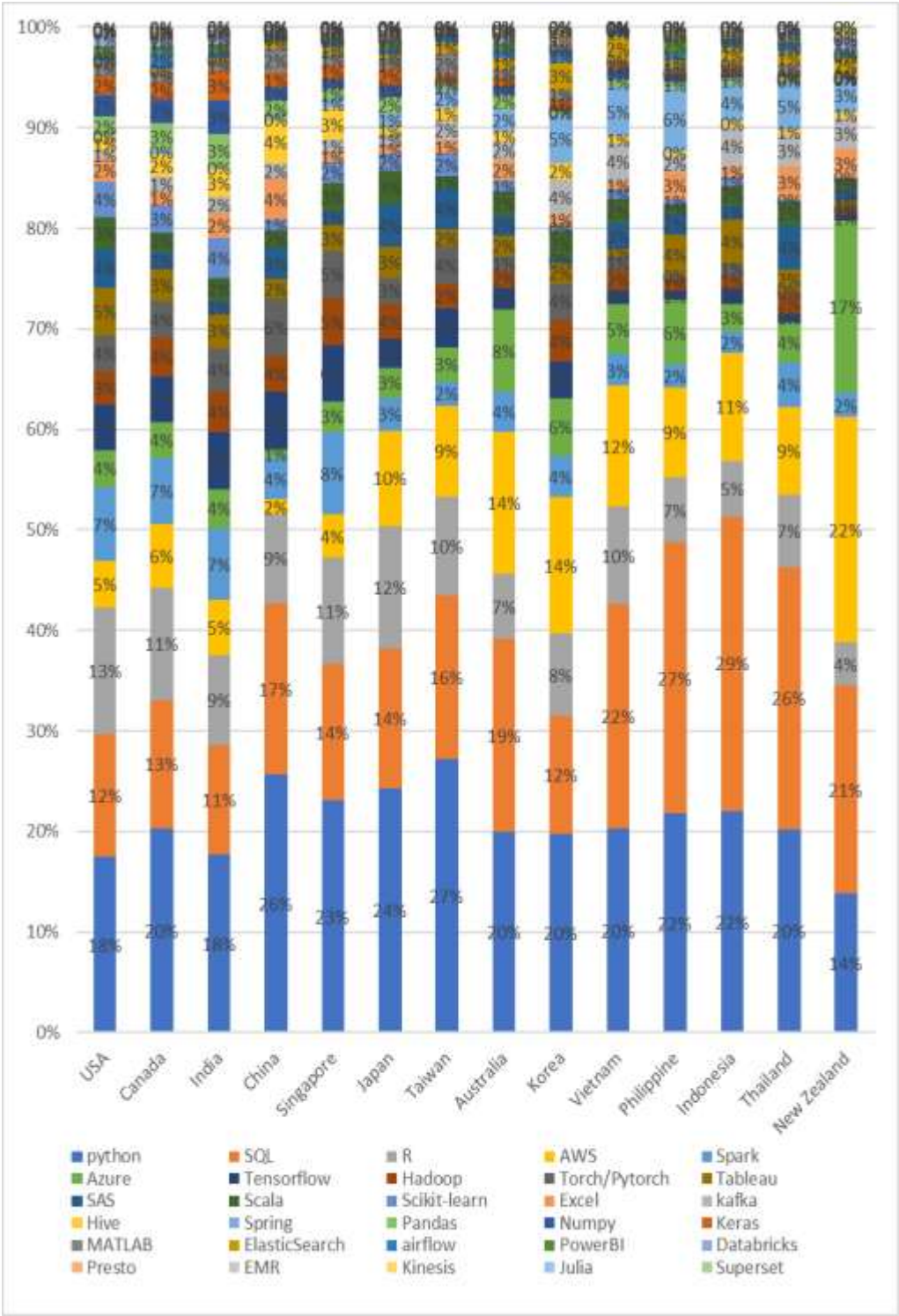
먼저 지역별로 나눠서 그래프를 그리면 <표 6>처럼 나타낼 수 있다. 위에서 종합으로 살펴보면 python이 가장 많이 언급되는 기술 스택 및 프로그래밍 툴이지만, 지역별로 보면 동남아시아에서는 SQL이 python과 동률이고, 오세아니아에서는 SQL이 python보다 더 많이 언급되는 것을 알 수 있다. (동남아시아 SQL 22%, python 22%/오세아니아 SQL 20%, python 17%) 오세아니아 지역은 AWS, Azure등 클라우드 컴퓨팅에 대한 언급이 3위, 5위에 언급될 정도로 많은 수요가 있는 것으로 보인다.



<표 8> 지역별 기술 스택 및 프로그래밍 툴 요구 비중 현황

<표 8>의 내용을 국가별로 나타내면 아래 <표 9>처럼 나타낼 수 있다. SQL이 python보다 많은 비중을 차지하는 국가는 동남아시아 4개국 (Singapore, Vietnam, Philippine, Thailand)과 오

세아니아의 New Zealand다. 그 특성이 지역에 반영된 것으로 보인다. 또한 오세아니아의 뉴질랜드와 호주 모두 AWS에 대한 수요가 조사한 국가 중 각각 1위, 2위에 해당하는 비중을 나타낸다.



<표 9> 국가별 기술 스택 및 프로그래밍 툴 비중 현황

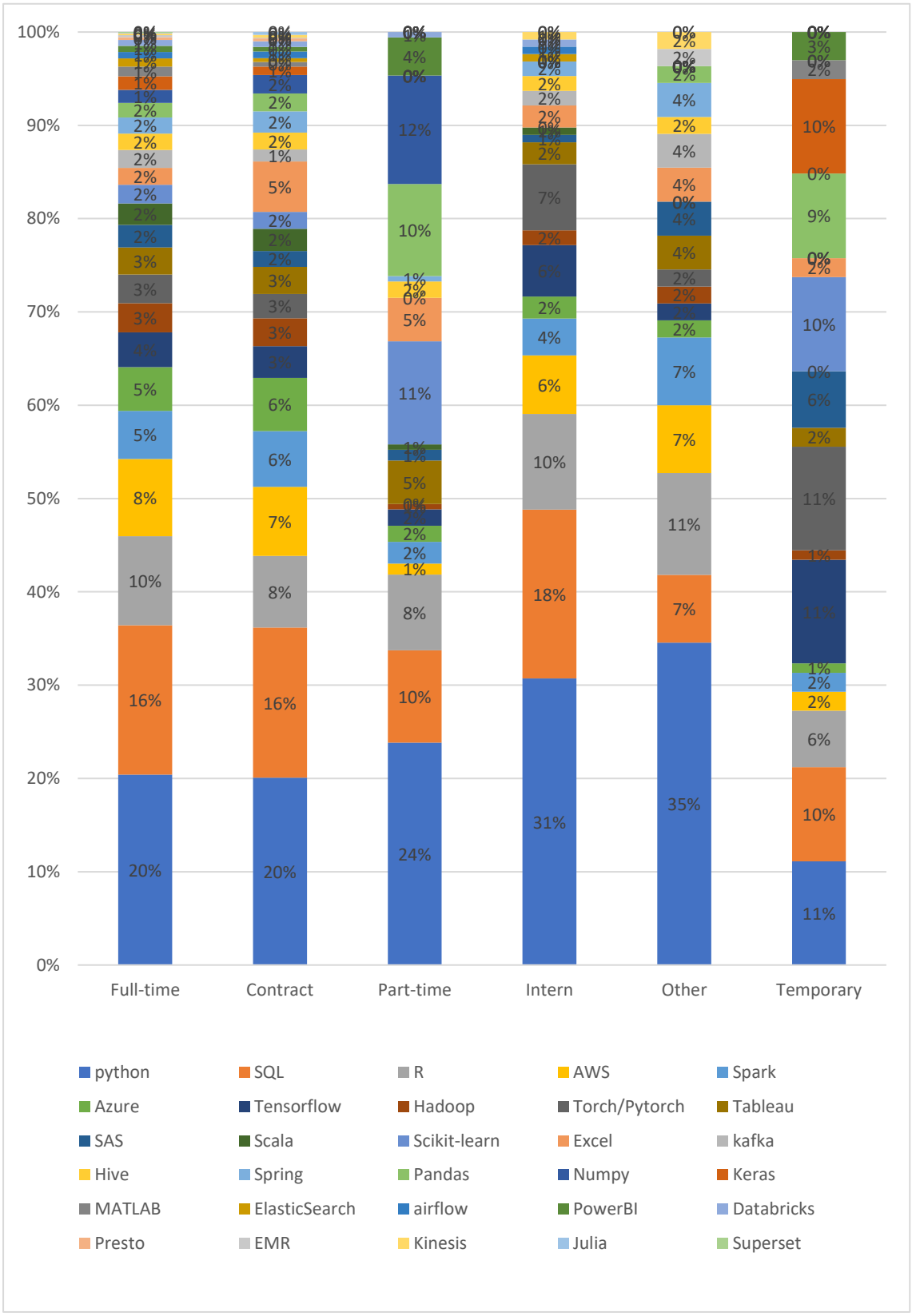
2) 고용형태별

Employment_form 컬럼을 전처리 하면서 고용형태를 다음과 같이 총 7가지로 나눴다. 각 단어의 정의를 말하면 Full-time은 정규직, Contract는 계약직, Intern은 인턴, Part-time은 아르바이트, Ohter는 기타, Temporary는 임시직, Volunteer는 자원봉사의 의미로 볼 수 있다. 이 중에서 1건이 나온 Volunteer를 제외하고 특징을 살피기로 했다. 각 고용형태별 건수 표는 <표 10>에서 알 수 있다.

고용형태	건수
계	**12,780**
Full-time	12,108
Contract	455
Intern	83
Part-time	59
Other	41
Temporary	33
Volunteer	1

<표 10> 고용형태별 채용공고 건수 현황

그리고 고용형태별 특징을 국가/지역별 형태처럼 비중 표로 만들면 <표 11> 처럼 나타낼 수 있다. 종합으로 봤을 때, 상위 3개 기술 스택 및 프로그래밍 툴인 Python, SQL, R의 비중합이 Full-time, Contract, Intern, Other에서 약 50%의 비중을 차지한다. 이것은 상위 3개 프로그램만 잘 다뤄도 구직할 때 선택의 폭이 넓어진다는 것을 알 수 있다. 이와는 대조적으로 Part-time과 Temporary에서는 상위 3개의 python, SQL, R이 아닌 Scikit-learn에 대한 언급이 각각 11%, 10%에 해당하는 것으로 보아 머신러닝 및 딥러닝에 대한 기술자 활용을 해당 고용형태로 한다는 것을 알 수 있다. 이는 이런 기술을 보유한 기술자의 공급이 부족해서 나타날 수도 있고, 이런 기술은 매번 필요한 기술이기 보다는 몇몇의 프로젝트로 해결하는 과제일수도 있기 때문에 이런 결과가 나왔을 것이라 추측한다.



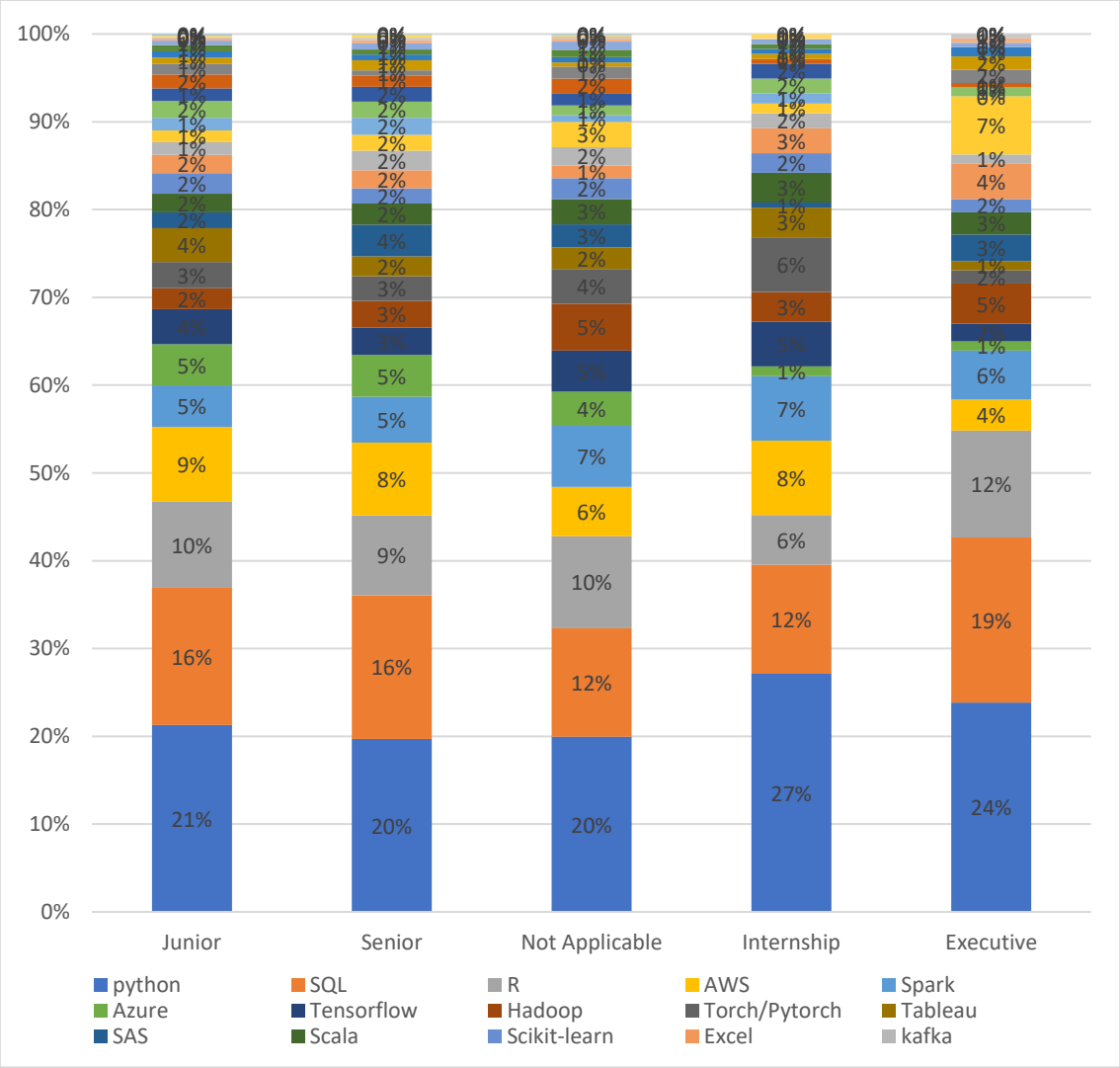
<표 11> 고용형태별 기술 스택 및 프로그래밍 툴 비중 현황

3) 직급별

Career-level 컬럼을 데이터 전처리 하면서 null값을 제외한 5개 직급 형태로 나눴다. 각 단어의 정의를 살펴보면, Junior 신입-대리 Senior 과장-부장 Executive 임원 Internship 인턴 Not Applicable 해당 없음으로 나눴다. 직급별 언급 건수를 보면 다음과 같다.

직급구분	건수
계	**11,193**
Junior	5,472
Senior	3,461
Not Applicable	2,055
Executive	103
Internship	102

<표 12> 직급별 채용공고 수 현황 (단위: 건)



<표 13> 직급별 기술 스택 및 프로그래밍 툴 비중 현황

직급에 따른 기술 스택 및 프로그래밍 툴 비중에 대한 차이는 거의 없는 것으로 나타났다. 직급은 기술 스택 및 프로그래밍 툴에 끼치는 영향이 적은 것으로 보인다.

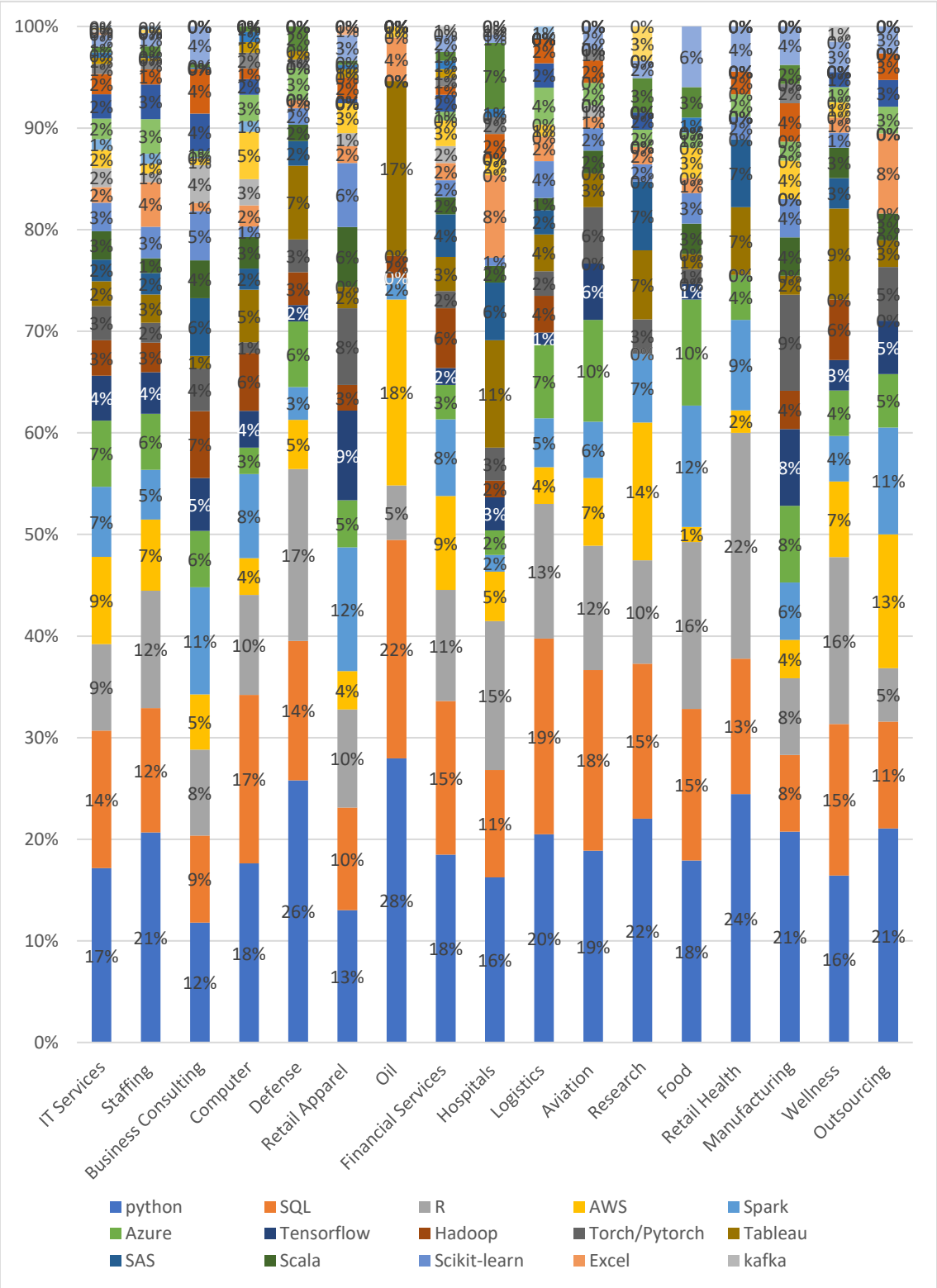
4) 산업군별

Industry 컬럼을 데이터 전처리 하는 과정에서 영어를 제외한 언어의 데이터를 전처리 하는 과정에서 산업 구분이 영어로 돼 있는 자료로만 구분했다. 산업군 중 10개 미만의 건수가 나오는 산업군은 제외하고 총 17개의 산업군으로 구분했다. (참고 <표 14>)

구 분	건수
총계	**8,312**
IT Services	5,442
Business Consulting	864
Staffing	614
Retail Apparel	238
Computer	193
Defense	124
Hospitals	123
Financial Services	119
Oil	93
Aviation	90
Logistics	83
Food	67
Wellness	67
Research	59
Manufacturing	53
Retail Health	45
Outsourcing	38

<표 14> 산업군별 채용공고 개수 표 (단위: 건)

산업군을 기술 스택 및 프로그래밍 툴 비중으로 나타내면 <표 15>의 결과를 얻을 수 있다. 대부분의 산업군에서 상위 3개 프로그래밍 툴인 python, SQL, R의 비중이 높게 나오지만, 몇 산업군에서는 빅데이터 처리를 요구하는 기술 스택 및 프로그래밍 툴을 요구하는 것으로 나타난다. 대표적으로 빅데이터를 처리할 때 필요한 Spark의 비중이 10%가 넘게 나타난다. Retail Apparel, Food (각 12% 비중)/Business consulting, Outsourcing (각 11% 비중). 또한, Oil 산업군에서는 AWS와 같은 클라우드 컴퓨팅 역량의 비중이 18%로 전체 산업군 중 가장 높았고, 데이터의 시각화 서비스를 도와주는 Tableau의 비중이 Oil(17%), Hospital(11%), Wellness(9%)로 타 산업군에 비해 높게 나왔는데 이에 대한 수요가 있는 것으로 보인다.



<표 15> 산업군별 기술 스택 및 프로그래밍 툴 비중 현황

V. 결론

1. 요약 및 결론

본 연구에서는 '링크드인'이라는 세계 최대 구인구직 사이트에서 APAC(아시아-태평양) 기준 데이터 사이언티스트의 12,780개 채용공고를 크롤링을 통해 자료를 모아 기업에서 데이터 사이언티스트 혹은 예비 데이터 사이언티스트에게 요구하는 기술 스택 및 프로그래밍 툴을 조사해 보았다. 조사 방법은 크롤링을 통해 모은 자료를 간단한 텍스트 마이닝을 통해 기술 스택 및 프로그래밍 툴과 관련된 단어의 빈도수를 확인하고, 채용 공고에 나온 국가/지역, 고용형태, 직급, 산업군에 따라 분류하여 파악했다.

기업에서 가장 많이 요구하는 기술 스택 및 프로그래밍 툴은 넓게는 python, SQL, R 순으로 나오고, 좁게는 유형별로 국가간, 고용형태 등에 따라 다르기는 하지만 python이 제일 많이 언급되는 점은 어떠한 유형에서도 공통적으로 나온다. 이 다음으로 기업에서 많이 언급하고 요구하는 역량으로는 AWS, Azure와 같은 클라우드 컴퓨팅을 다루는 능력 또한 수요가 있다는 점이다. 또한 python에서도 딥러닝, 머신러닝과 같은 빅데이터를 처리할 수 있는 라이브러리 예를 들어 Spark, Tensorflow, Hadoop, Scikit-learn 등을 익힌다면, 선택의 폭은 조금 더 넓어질 것으로 보인다.

국가/지역으로 보면 동남아시아 4개국 (Vietnam, Indonesia, Philippine, Singapore)에서 python 보다 SQL에 대한 수요가 높은 것이 특징이었고, 오세아니아 2개국은 AWS와 같은 클라우드 컴퓨팅에 대한 수요가 높은 것으로 나왔다.

고용형태로 보면 Part-time과 Temporary에서 머신러닝, 딥러닝과 관련된 Scikit-learn에 대한 수요가 높은 것으로 나타나 이에 대한 역량이 있는 구직자라면, 프리랜서나 개인사업자로 활동을 하는 것도 생각해 볼만하다.

직급으로 나눴을 때는 특별한 구분점이 나오지는 않았다. 아무래도 직급에 따라서 요구하는 기술 스택 및 프로그래밍 툴이 달라지지는 않는 것으로 보인다.

산업군으로는 대부분의 산업군에서 python, SQL, R이 높은 비중을 차지하는 것으로 나오지만, Retail Apparel, Food, Business consulting, Outsourcing과 같은 분야에서는 빅데이터 처리를 돕는 Spark의 비중이 높게 나오고, Oil산업군에서는 클라우드 컴퓨팅에 대한 수요, Oil, Hospital, Wellness에서는 데이터 시각화를 돕는 Tableau에 대한 수요가 높게 나타났다.

2. 시사점

본 연구의 가장 큰 시사점은 데이터 사이언티스트가 되고 싶은 구직자에게 다음과 같은 유용한 정보를 제공했다는 것에 있다. 첫번째로 본 연구를 통해 데이터 사이언티스트가 되기 위한 구직자가 해외 취업을 생각한다면 해당 국가에서 가장 많이 언급된 기술 스택 및 프로그래밍 툴 위주로 준비할 수 있다. 두번째로는 현직 데이터 사이언티스트가 이직을 준비하는 단계에서 업종을 변경하고자 할 때, 자신이 가진 역량에 맞는 업종을 선택할 수 있게 도움을 얻을 수 있다. 마지막으로 머신러닝 및 딥러닝에 대한 기술이 있는 데이터 사이언티스트에게 부업으로 Part-time이나 Temporary의 기회가 있다는 것을 알려줄 수 있다.

향후 연구에서는 본 연구에서 나눴던 유형간 관계를 파악하여 좀 더 고차원으로 분석하는 것이 필요할 것이다. 예를 들어, 국가/지역별 + 고용형태별에 어떤 상관관계가 있는지 혹은 국가/지역별 + 산업군별에 어떤 상관관계가 있는 지처럼 유형간 관계를 서로 엮어서 파악해 보면 기술 스택 및 프로그래밍 툴에 대한 조사를 깊게 할 수 있을 것이다. 또한, 본 연구 자료 조사에는 연봉정보가 없었기 때문에 데이터 사이언티스트에게 직접 연봉 정보 혹은 근무할 때 활용되는 기술 스택 및 프로그래밍 툴을 조사하여 같이 연결 지어 연구하는 것이 좀 더 다양한 시각으로 정보를 전달할 수 있을 것이다. 또한 유형별 그룹을 갖고 통계적 검정을 하여, 유형별로 나눈 기준점이 선호하는 기술스택 및 프로그래밍 툴에 영향을 끼치는지, 끼치지 않는지를 알아 보는 것이 좋을 것이다. 통계적 검정 방법의 예로 카이제곱 독립성 검정을 들 수 있다. 카이제곱 독립성 검정이란, 두 범주형 변수가 독립적으로 분포하는지를 테스트 하는 검정이다.[13] 예를 들어, 위의 관찰 값에서는 직급별로 기술스택 및 프로그래밍 툴의 차이가 없다고 보여지지만, 이것이 통계적으로도 상관이 있는지 없는지를 확인해볼 수 있을 것이다.

[13] JMP Statistical Discovery – 카이제곱 독립성 검정 예제

참고문헌

<국내문헌>

과학기술정보통신부. *데이터 산업현황 조사*, 2020 년

<외국문헌>

Norita Ahmad, Areeba Hamid, Vian Ahmed. *Data Science: Hype and Reality.* IEEE, 2022

Renata Rawling-Goss. *Data Science Careers, Training, and Hiring*. Springer, 2019

<참고사이트>

'Developer Survey'. Stackoverflow. 2022. https://survey.stackoverflow.co/2022/

'Developer Survey'. Stackoverflow. 2021. Stack Overflow Developer Survey 2021

'클라우드란 무엇인가'. gabia 라이브러리. https://library.gabia.com/contents/infrahosting/9114/

'카이제곱 독립성 검정 예제'. JMP Statistical Discovery. https://www.jmp.com/ko_kr/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html