

스타트업 정보 빅데이터 챌린지

기술개발서

팀명 :Wonder Women (WW)

목차

I. 주제 및 현황

1. 주제 ----- 3p

2. 개발 선정 이유 ----- 3p

II. 작업 과정

1. 사용 시스템 ----- 4p

2. 데이터 설명 ----- 5p

3. 전처리 ----- 6p

4. 시각화 ----- 8p

5.

III. 최종 및 의견 도출

1. 최종 결론 및 아이디어 ----- 16p

I. 주제

1. 주제

“

증가하는 창업자 수. 국가에서 지향하고 지원하는 스타트업.

그러나, 회사들의 생존율과 성장률은 지원하는 만큼 성과가 있는가?

대한민국의 스타트업 현황을 살펴보고, 앞으로 나아가야 할 방향성이 무엇인지 알아보자.

”

2. 개발 선정 이유

1) 프로젝트 배경

신생기업(New firm)은 신규 고용을 창출하고 국가 경제성장에 기여할 수 있는 잠재력을 가진 반면, 대한민국 스타트업은 낮은 생존율과 성장률 문제에 직면했다. 또한, '코로나'라는 변수에 의해 예측하지 못한 사태를 직면한 스타트업 기업의 파산 및 폐업은 더욱 가파르게 증가했다.

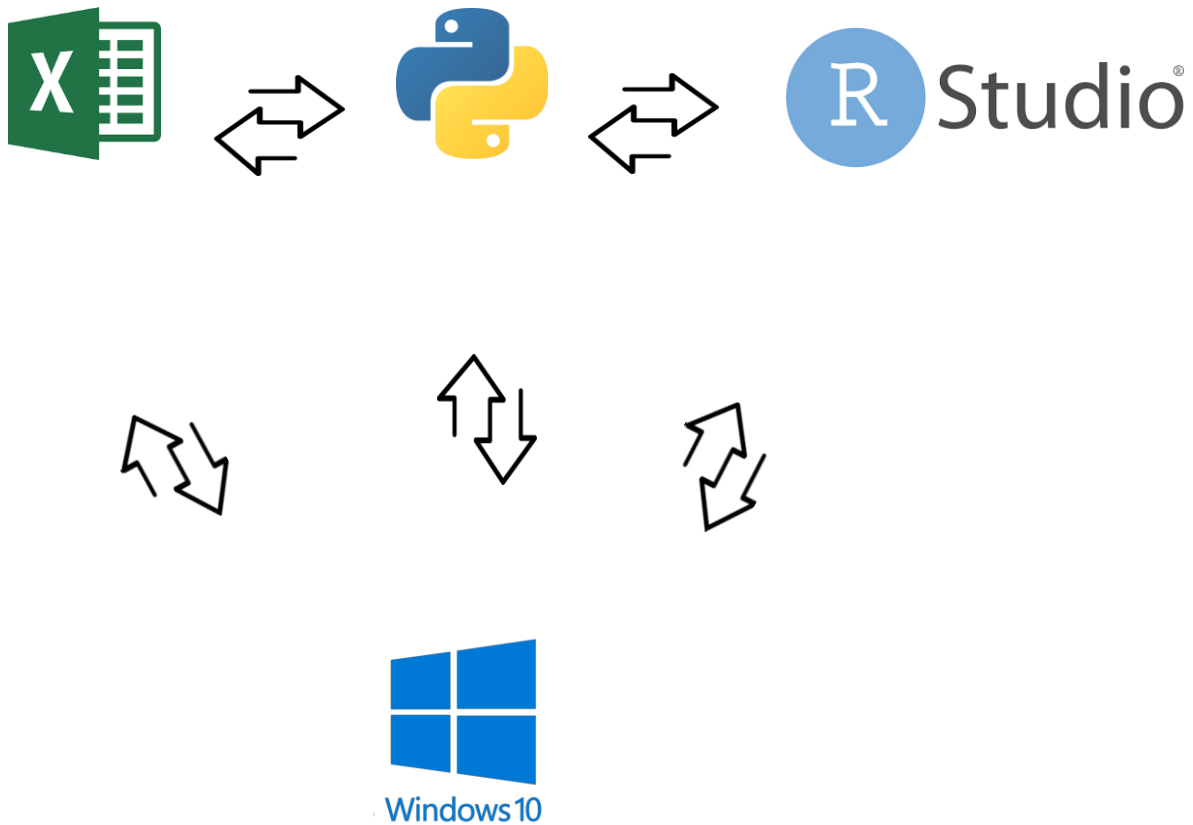
2) 프로젝트 목적

스타트업 기업을 대상으로 투자 생태계 관점에서 신생기업의 생존에 대한 영향요인을 분석하고 시사점을 도출한다. 또한, 스타트업의 현황을 바탕으로 '코로나'라는 특수 상황에 따른 변동사항을 체크한다.

최종적으로 대한민국에서의 스타트업 생존과 높은 투자단계 진입 장벽 및 기간 감축을 위한 데이터 분석을 진행하고자 한다.

II. 작업 과정

1. 사용 시스템



1) 프로젝트 개발 환경

- Python버전 : 3.9.7 (전체적인 분석에 활용)
- R 버전 : 4.2.1 / R studio 버전 : 2022.07.2-576 (전처리 및 시각화)
- Excel 버전 : 2016 (전처리)
- OS(운영체제) 버전 : Windows 11, windows 10 (검색)
- wordcloud.com : Windows 10 (워드 클라0우드 제작)

2.데이터 설명

1) 제공받은 데이터 설명(알리콘 주식회사 제공)

- 기업정보 : 기업에 대한 다양한 정보 기재
- 지역별 스타트업 현황 : 지역별 스타트업 업종 등 정보 기재
- 투자유치정보 : 기업별 투자 정보 기재
- 지역별 스타트업 투자 유치 현황 : 기업별 투자유치단계 등 기재

2) 별도 조사 데이터 설명(통계청, KOSIS 제공)

- 2020년 _직종별_인력구성_부족인력_수_및_비율 : 2020 인력에 대한 전반적인 정보 기재
- 2019년 _직종별_인력구성_부족인력_수_및_비율 : 2019 인력에 대한 전반적인 정보 기재
- 2021년_하반기_및_2022년_신규_채용계획이_없는_이유 : 설문조사 형식의 신규채용 없는 이유 기재

3.전처리

1) 메인 데이터 선정 : 기업정보.csv 데이터

2) 산업분류 코드 통일(한국 표준 산업 분류 기준)

- 기업이 지정한 내용이 그대로 입력, 통일되지 못한 기준
 - 한국 표준 산업분류 기준 카테고리를 새로 생성
 - 기업을 사람인과 잡코리아 중심으로 검색, 기업정보를 확인
 - 기재된 소분류 업종을 한국 표준 산업 대분류에 맞게 재검색하여 작성
 - 검색값이 없는 기업은 기업소개를 바탕으로 유사 유명 기업을 조사
 - 유사 기업의 업종을 가져와 기재
- 이후 시각화 작업 및 분류 작업의 원활한 진행을 위해 작업.

3) 중복, 폐업 회사 삭제

- 회사명(필수), 기업소개, 설립일자, 기업일련번호(필수) 등 겹치는 부분이 3개 이상인 항목 우선 삭제
- 언론매체에서 폐업 및 인수된 회사 중심으로 추가조사 후 사업자 등록번호
말소된 기업 삭제

4) 투자금액, 설립년차 조인

- 메인 데이터에 필요항목인 투자금액 합치기
- 작업일자 기준 설립일자를 대입, 설립년차를 계산. 컬럼 추가

5) 결측값 처리 (투자단계별 투자금액 평균)

- 우리가 사용할 때 결측값이 있으면 안되는 항목 선정

- 투자단계, 투자금액

- 투자단계 미공개 및 해당없음, 투자단계 항목 Null값 항목 모두 삭제
- 투자자마다 차이가 큰 Angel 투자 제외, 통상적인 규칙을 가진 나머지

투자단계는 해당 규칙에 따라 평균금액을 산정하여 null값에 기입

The screenshot shows the RStudio interface with a data table loaded. The table has columns for '투자단계' (Investment Stage), '투자금액' (Investment Amount), and '기업명' (Company Name). The data is filtered to show only rows where the investment stage is not null and the amount is not null. The table is sorted by investment stage, and the first few rows are visible.

- 데이터 전처리 종료

- 데이터 속 항목 중 사용한 컬럼

- 기업일련번호

- 투자단계명

- 설립연차

- 산업구분명

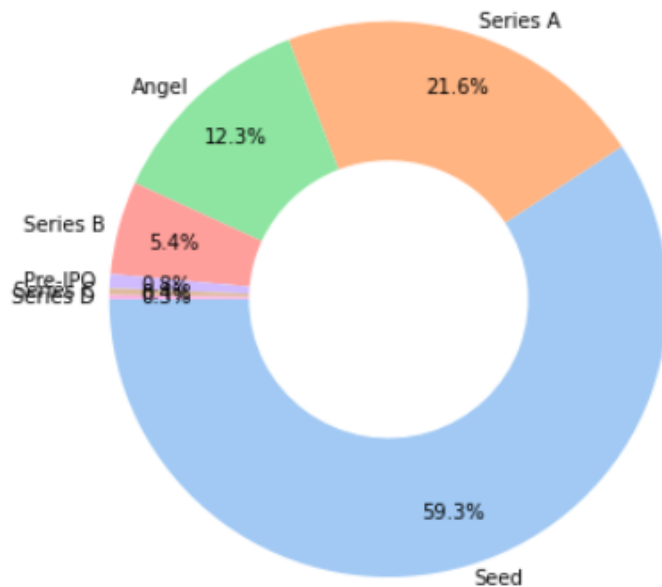
- 한국표준산업분류코드

- 설립연도

- 누적투자가치금액

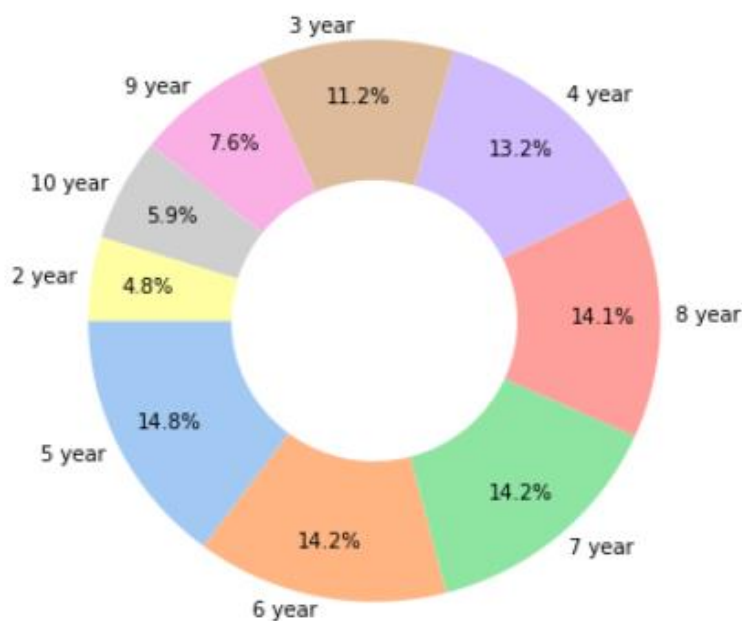
4. 사용 시각화

1) 스타트업 투자단계 분포 현황(원그래프)



- 1위 : 씨드
- 2위 : Series A
- 3위 : Angel

2) 스타트업 연차 분포 현황



- 1위 : 5년차
- 2위 : 6년차
- 2위 : 7년차

3) 스타트업 산업분류 분포 현황(그래프)

업종코드	개수
A	0
B	0
C	129
D	1
E	1
F	1
G	136
H	8
I	3
J	1097

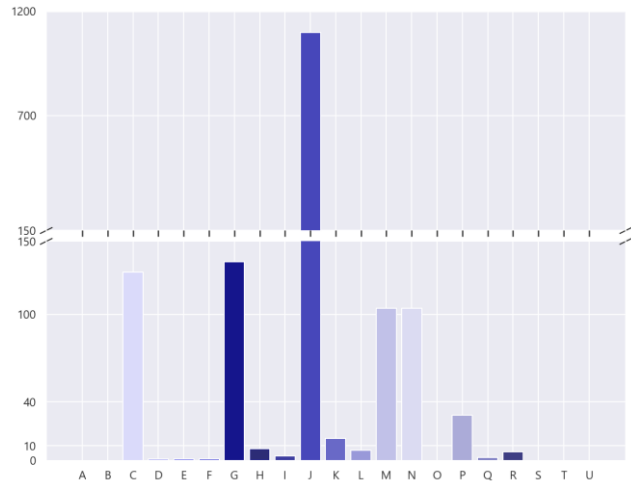
K	15
L	7
M	104
N	104
O	0
P	31
Q	2
R	6
S	6

● 엑셀 활용, COUNTIF함수를 사용하여 각 업종별 갯수 추출

● CSV로 저장, 파이썬에서 시각화 작업

- 값이 압도적으로 높은 J로 인해 보통의 시각화로는 한 눈에 들어오기 힘들
- 그래프를 2개 동시에 그려서 각각 y축 범위를 다르게 조절
- 이후 임의로 끊었음을 표시, 웨이브 삽입

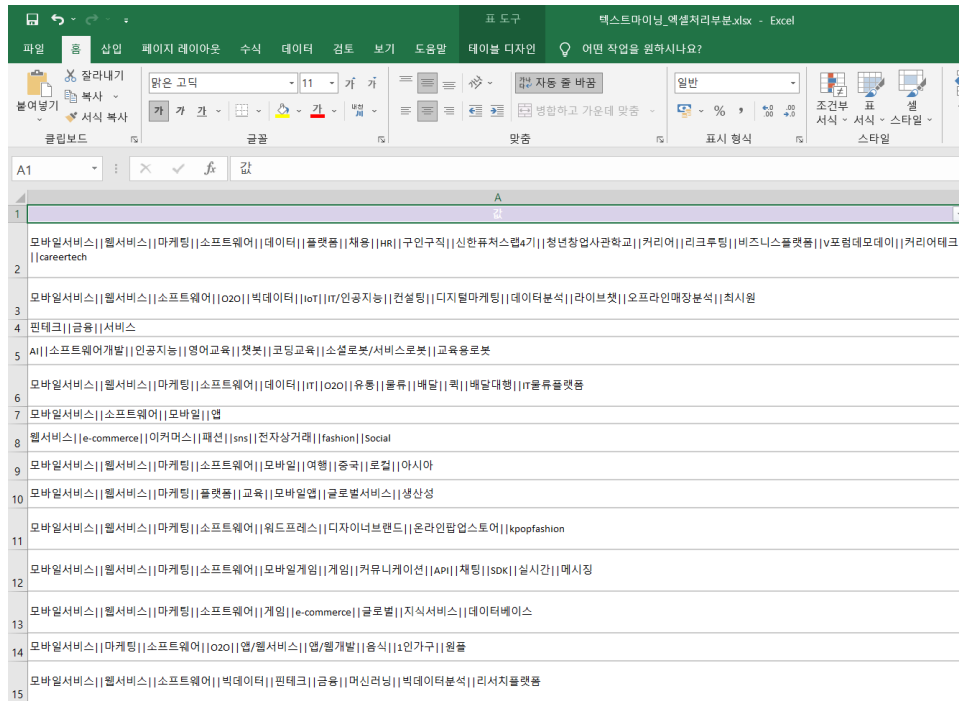
```
In [6]:
1 fig, (ax1, ax2) = plt.subplots(2, 1, sharex=True) # 그래프 두 개를 한 figure 내에 그리기
2 fig.subplots_adjust(hspace = 0.05) # 두 그래프 사이의 상하 간격 설정
3
4 # 그래프의 색 지정
5
6 colors = ['salmon', 'tomato', 'darksalmon', 'coral', 'lightcoral', 'lightsalmon', 'ivory', 'linen', 'b
7         'darkorange', 'burlywood', 'antiquewhite', 'tan', 'navajowhite', 'blanchedalmond', 'papay
8         'orange', 'wheat', 'oldlace']
9
10 # 각각의 그래프 그리기
11 ax1.bar(x, y, color = colors)
12 ax2.bar(x, y, color = colors)
13
14 ax1.set_ylim(150, 1200) # 첫 부분 y축 범위 설정
15 ax2.set_ylim(0, 150) # 아랫 부분 y축 범위 설정
16
17 # 두 그래프 사이의 경계선 제거
18 ax1.spines['bottom'].set_visible(False)
19 ax2.spines['top'].set_visible(False)
20 ax1.xaxis.tick_top()
21 ax2.tick_params(labeltop=False)
22 ax1.xaxis.tick_bottom()
23
24 # 두 그래프 y축 간격 조절
25 ax2.set_yticks([0, 10, 40, 100, 150])
26 ax1.set_yticks([1200, 700, 150])
27
28 # 두 그래프 사이의 y축에 물결선 효과 마커 표시
29 kwargs = dict(marker=[(-1, -0.5), (1, 0.5)], markersize=12,
30               linestyle='none', color='k', mec='k', mew=1, clip_on=False)
31 ax1.plot([0, 1], [0, 0], transform=ax1.transAxes, **kwargs)
32 ax2.plot([0, 1], [1, 1], transform=ax2.transAxes, **kwargs)
33
34 plt.show()
```



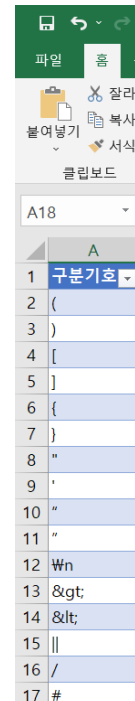
- 1위 : J 정보통신업
- 2위 : G 도매 및 소매업
- 3위 : C 제조업

4) 키워드 텍스트 마이닝

- 제공받은 데이터 내 '산업구분명'을 활용
- 해당 부분만 엑셀파일에 업로드
- 텍스트 구분 특수기호 및 제거 특수기호를 별도 지정

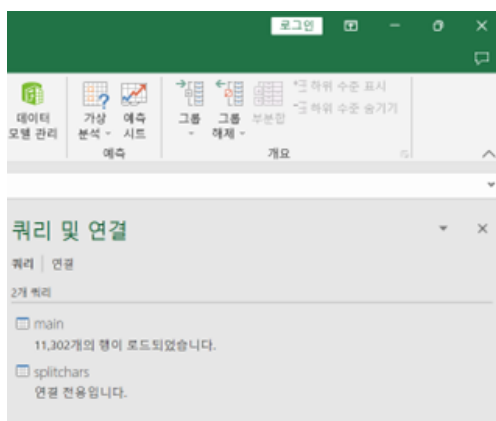


구분	서비스명	플랫폼	산업구분명
1	모바일서비스 웹서비스 마케팅 소프트웨어 데이터 플랫폼 채용 HR 구인구직 신한퓨처스텝4기 청년창업사관학교 커리어 리크루팅 비즈니스플랫폼 V포럼데모데이 커리어테크 careertech		
2	모바일서비스 웹서비스 소프트웨어 IoT 빅데이터 IoT 인공지능 컨설팅 디지털마케팅 데이터분석 라이브챗 오프라인매장분석 최시원		
3	핀테크 금융 서비스		
4	AI 소프트웨어개발 인공지능 영어교육 챗봇 코딩교육 소셜로봇/서비스로봇 교육용로봇		
5	모바일서비스 웹서비스 마케팅 소프트웨어 데이터 IT O2O 유통 유통 배달 퀵 배달대행 IT유통플랫폼		
6	모바일서비스 웹서비스 소프트웨어 모바일 앱		
7	웹서비스 e-commerce 이커머스 재선 SNS 전자상거래 fashion Social		
8	모바일서비스 웹서비스 마케팅 소프트웨어 모바일 여행 중국 로컬 아시아		
9	모바일서비스 웹서비스 마케팅 플랫폼 교육 모바일 글로벌서비스 생산성		
10	모바일서비스 웹서비스 마케팅 소프트웨어 워드프레스 디자이너브랜딩 온라인팝업스토어 kpopfashion		
11	모바일서비스 웹서비스 마케팅 소프트웨어 모바일 게임 커뮤니케이션 API 채팅 SDK 실시간 메시징		
12	모바일서비스 웹서비스 마케팅 소프트웨어 게임 e-commerce 글로벌 지식서비스 데이터베이스		
13	모바일서비스 마케팅 소프트웨어 O2O 앱/웹서비스 앱/웹개발 음식 1인가구 원룸		
14	모바일서비스 웹서비스 소프트웨어 빅데이터 핀테크 금융 머신러닝 빅데이터분석 리서치플랫폼		

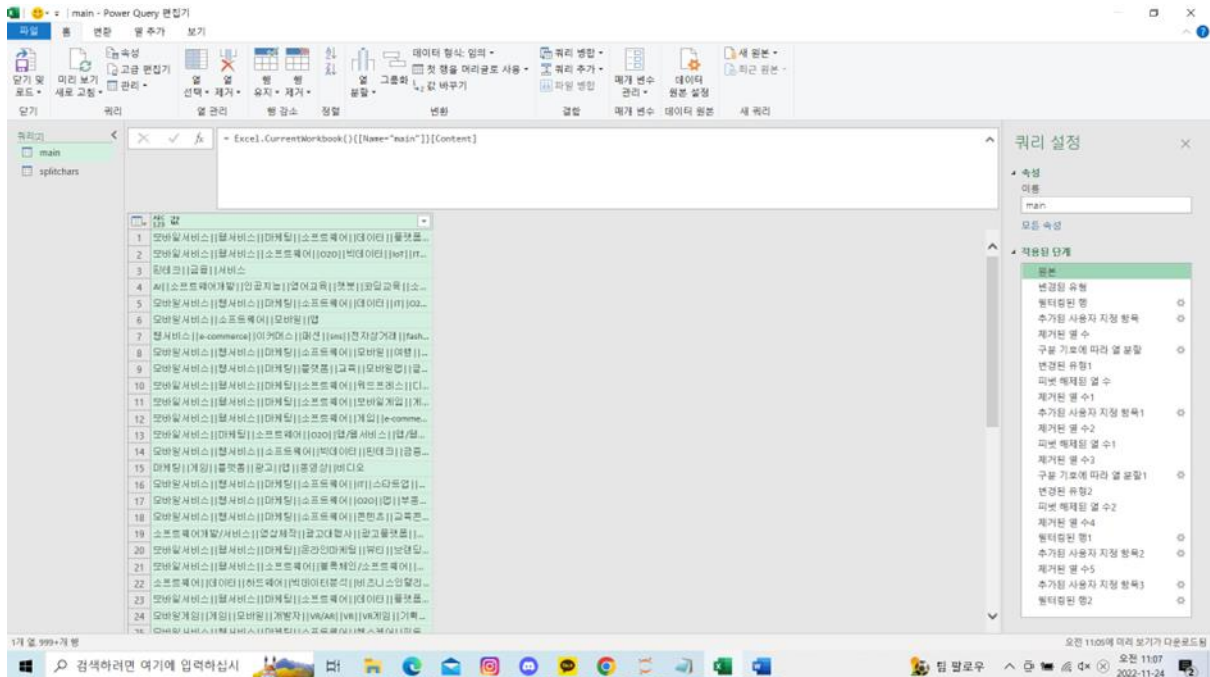


구분기호
1
2 (
3)
4 [
5]
6 {
7 }
8 "
9 '
10 "
11 "
12 Wn
13 >
14 <
15
16 /
17 #

- 각 데이터를 표로 변환, 연결전용으로 쿼리 저장



● 쿼리 편집기 사용, 텍스트 분류 및 특수문자 제거



- 사용된 코드

- 텍스트 구분짓기(특수문자로 텍스트를 구분짓기 전 과정)

```
List.Accumulate(
    List.Numbers(0, Table.RowCount(splitchars)-1),
    [값],
    (string, row) =>
        Text.Replace(string, splitchars{row}[구분기호], " ")
    )
```

- 각 행별 특수문자 제거

```
Text.Select([값],{"가".. "힉", "a".. "z", "A".. "Z", "0".. "9", " "})
```

- 출력된 쿼리를 표로 저장
- 피벗테이블로 제작
- 각 단어별 총 사용 개수 계산 완료, 표로 재정리

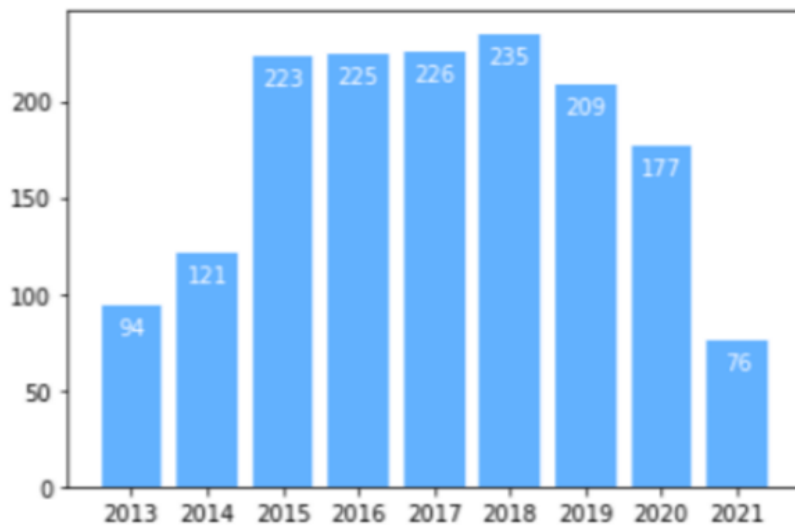
Category	Keyword	Count
1	웹서비스	416
2	소프트웨어	375
3	모바일서비스	356
4	IT	265
5	마케팅	256
6	플랫폼	159
7	인공지능	149
8	시	123
9	서비스	123
10	빅데이터	120
11	모바일	118
12	O2O	116
13	데이터	105
14	핀테크	96
15	IT서비스	87
16	소프트웨어	86
17	교육	74
18	스타트업	70
19	이커머스	67
20	머신러닝	65
21	헬스케어	64
22	IoT	63
23	블록체인	62
24	콘텐츠	56

- CSV(UTF-8)로 저장
- <https://www.wordclouds.com/> 사이트 접속
- 저장한 CSV파일 업로드, 기타 디자인 조절



- 1위 : 웹서비스
- 2위 : 소프트웨어
- 3위 : 모바일서비스

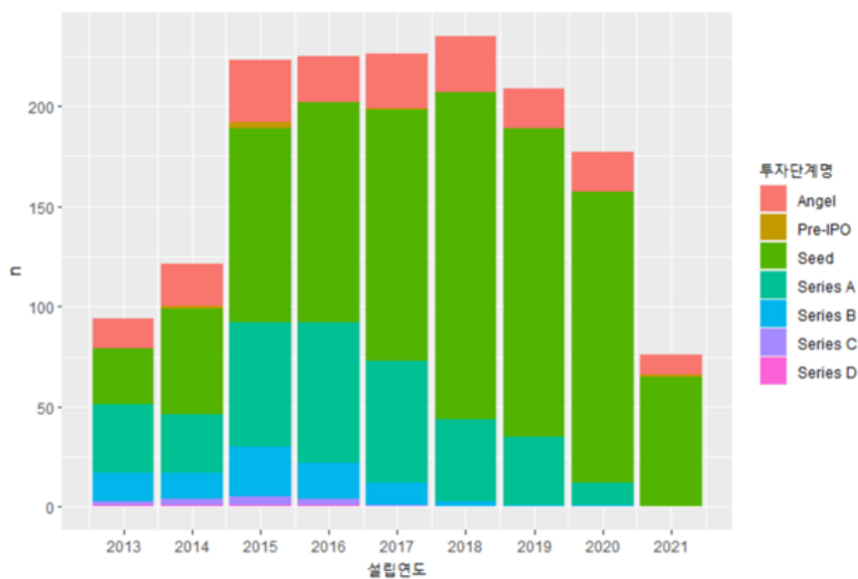
5) 설립연도별 설립 수 분포



- 19년도부터 스타트업의 신설 개수 하락, 21년도 급격히 감소
- 스타트업 활성화인 15년도~19년도 평균 223개
- 코로나 영향력의 20년도 ~ 21년도 평균 126개

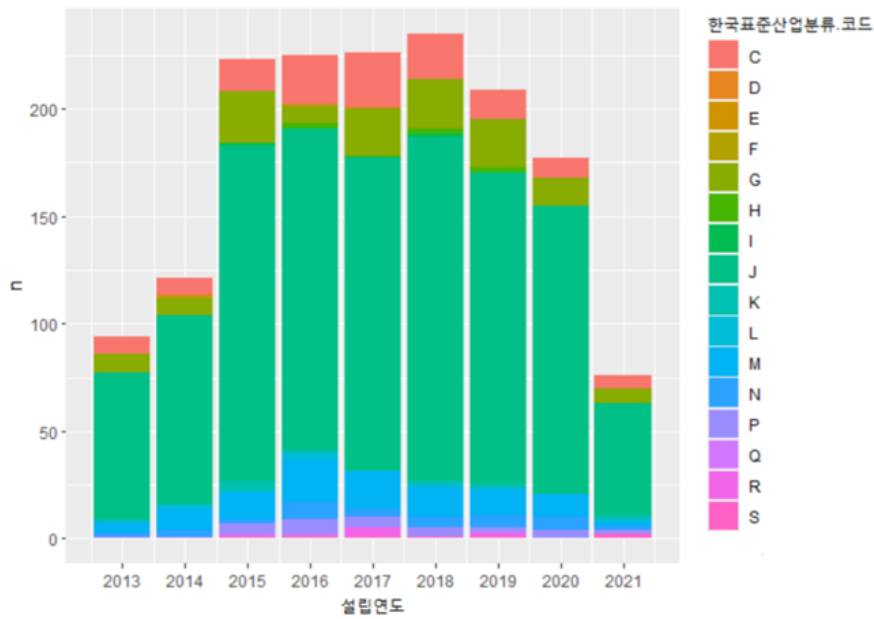
6) 코로나가 끼치는 신생기업의 영향력

● 설립연도별 투자단계 분포



- 2020년 Series B 1건 존재, C, D 0건
- 2021년 Series B, C, D 0건

● 설립연도별 산업분류 분포

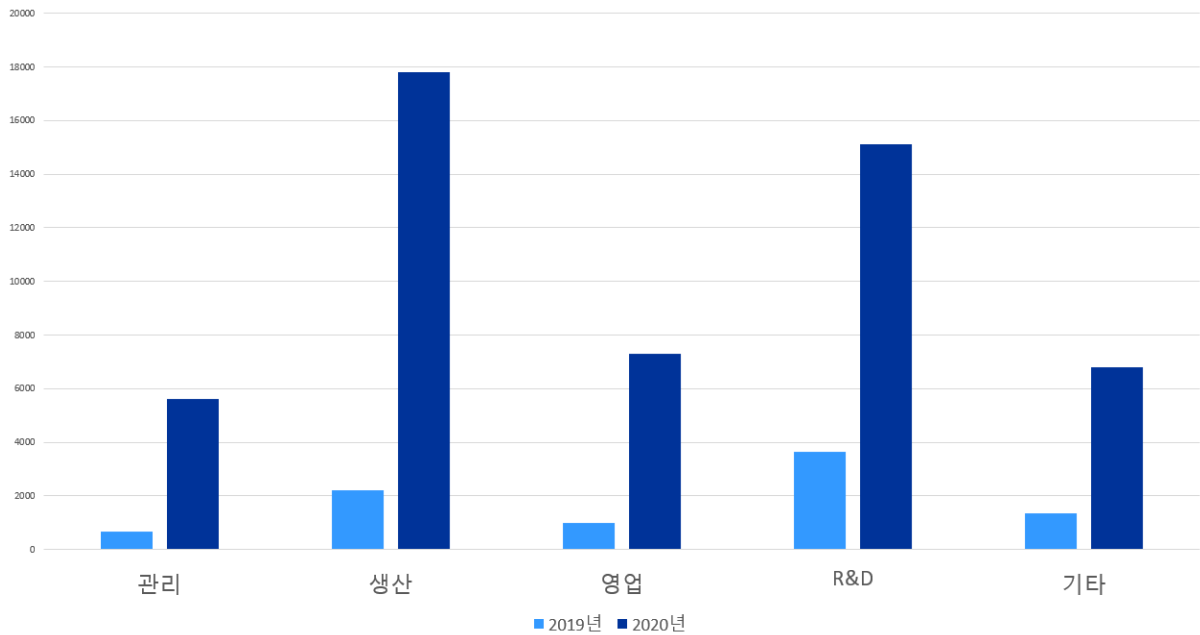


- 2019년 대비 2021년도 가장 적어진 산업코드(개수 순서)
- 1위 : J(정보통신업, -92건, 63% 하락)
- 2위 : G(도매 및 소매업, -15건, 68% 하락)
- 3위 : M(전문기술, -10건, 83% 하락)

2019년		2021년		년도별 개수 차이		증감률
산업코드	개수	산업코드	개수	산업코드	개수	
C	14	C	6	C	8	-57.1429
G	22	G	7	G	15	-68.1818
I	0	I	0	I	0	0
J	144	J	52	J	92	-63.8889
M	12	M	2	M	10	-83.3333
N	6	N	2	N	4	-66.6667
S	1	S	2	S	-1	100

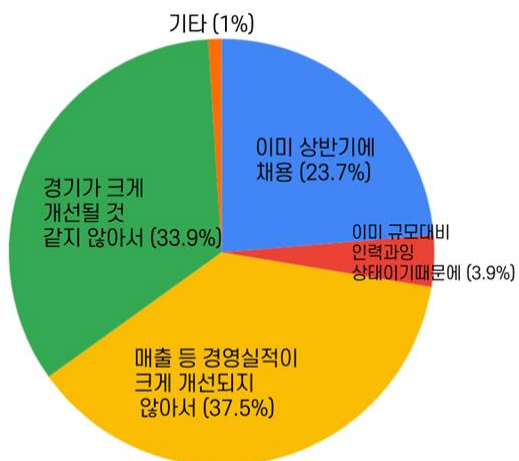
7) 코로나 영향에 따른 인력문제 현황

● 직군별 부족인력



- 전체적으로 모든 직군이 인력부족 현상을 보임
- 특히, 생산직군의 인력이 가장 도드라지게 부족함

● 신규채용이 없는 이유(하반기 ~ 내년 채용/설문조사 자료)



- 경기불황, 경영 실적이 주된 원인
 - 인력은 부족, 그러나 채용계획 없음
- = 기존 인원의 부담감이 상승가능성 존재

III. 최종의견 및 아이디어

1. 최종 의견 및 아이디어

1) 자료 기반 최종 결론

- 7년 이상 운영시 높은 투자단계유치 가능성 UP
- 현시점 '성장' 보다는 '생존'에 집중

2) 아이디어 도출

1. 자회사 AI 데이터 분석

[스타트업 생태계 파악 AI 프로그램 개발]

- 매출/영업이익금/부채/인력 등 자회사의 객관적인 점수를 통해 기회 및 위험요인 파악한다.

이를 통해 스타트업은 전략적인 운영 가능하다

2. 폐업/실패 데이터 활성화

[폐업에 대한 정보를 통한 실패요인 분석]

- 폐업 관련 데이터를 통해

문제가 되는 실패요인을 파악하여 대비한다.

폐업을 미리 방지함으로써 생존률을 높일 수 있다.

<참고 사례>

- 실패 데이터 적극 활용 사례 - 구글 묘지(<https://gcemetery.co/>)

↳ 구글에서는 실패에 대한 데이터를 따로 저장하고 있다.

실패요인에 대한 분석을 진행함으로써 위기 데이터를 기회로 활용하고 있다.

<참고 문헌>

김기만 부연구위원. [제 21-21호 KOSI 중소기업 포커스 - 스타트업 생태계 관점에서 바라본
신생기업 생존의 영향요인 : OECD 국가 비교 분석].중소벤처기업연구원, 2021

박건철·이승하.[서울시 벤처생태계 현황 및 성과분석].서울디지털재단, 2017

유정희 부소장[2020 벤처기업정밀실태조사].중소벤처기업부, (사)벤처기업협회, 2020

[2022년 주요업무 추진계획 : 위기를 넘어 혁신으로, 강한 중소·벤처·소상공인
육성].중소벤처기업부, 2021