보고서 - SearchTrend Part

1. 자료 정리하기

SearchTrend

- 국내 최대 포털 중 하나인 NAVER에서 검색한 키워드의 트렌드 데이터 (자료 1642개, 변수 5개)

- 단위는 해당 기간의 가장 높은 검색량을 100으로 설정하여 상대값을 의미한다.

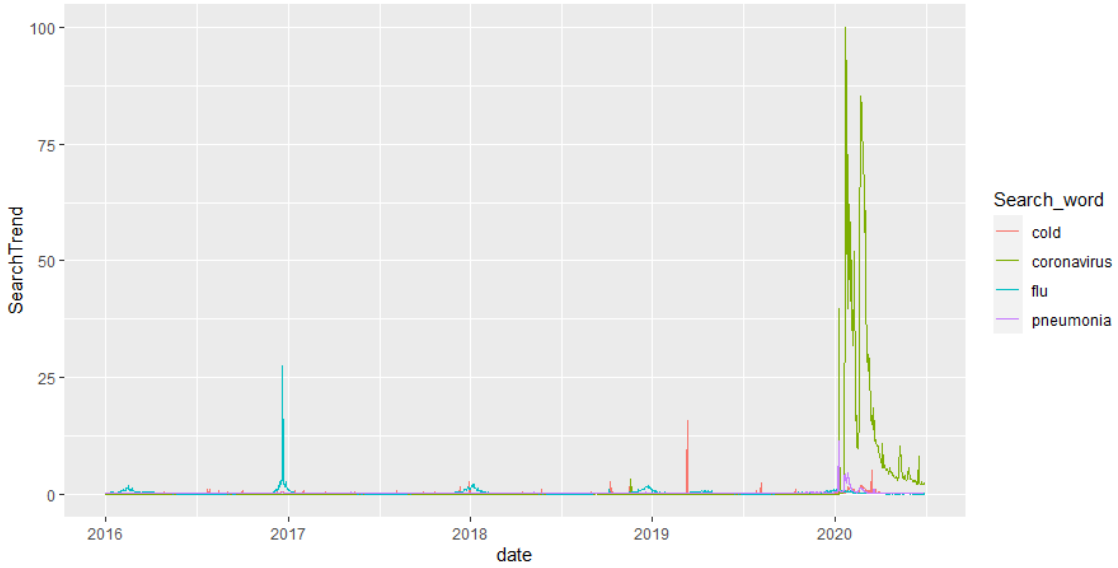
변수명	변수 설명	변수 종류
date	날짜 (2016-01-01부터 2020-06-29까지)	date
cold	키워드 '감기' 검색량.	float
flu	키워드 '독감' 검색량	float
pneumonia	키워드 '폐렴' 검색량	float
coronavirus	키워드 '코로나바이러스' 검색량	float

2. EDA를 통하여 자료를 살펴보기

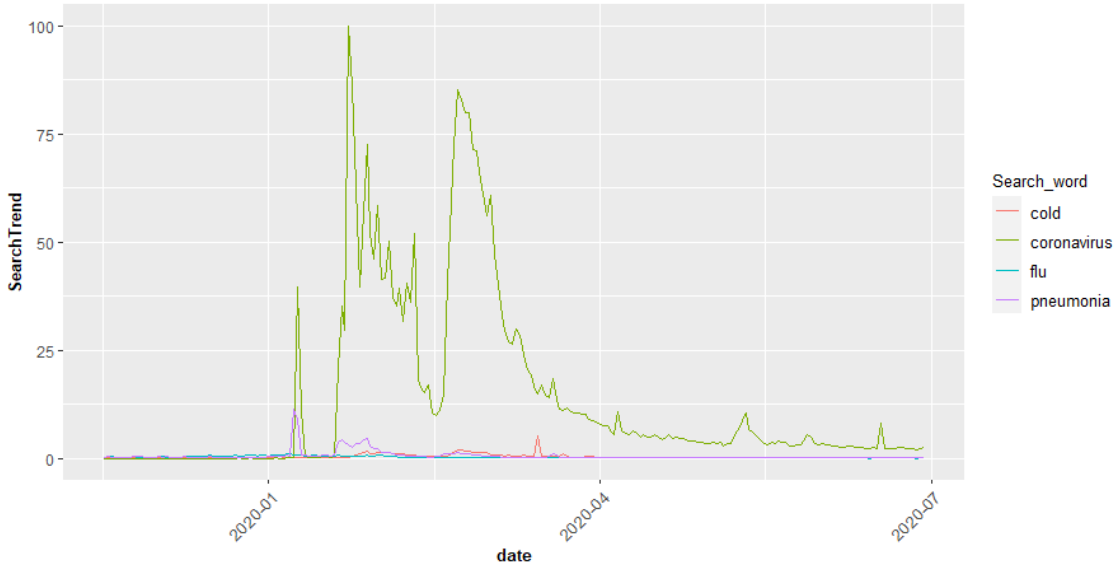
4가지 호흡기 질환(감기/독감/폐렴/코로나바이러스)에 대한 일별 검색량의 tibble 자료이다.

```
> SearchTrend
# A tibble: 1,642 x 5
   date        cold    flu pneumonia coronavirus
   <date>     <dbl>  <dbl>     <dbl>       <dbl>
 1 2016-01-01 0.117 0.0559     0.157     0.00736
 2 2016-01-02 0.134 0.171      0.208     0.0089
 3 2016-01-03 0.149 0.223      0.193     0.00845
 4 2016-01-04 0.175 0.186      0.290     0.0114
 5 2016-01-05 0.172 0.151      0.246     0.0138
 6 2016-01-06 0.173 0.144      0.251     0.0138
 7 2016-01-07 0.174 0.124      0.251     0.0119
 8 2016-01-08 0.167 0.125      0.290     0.0157
 9 2016-01-09 0.135 0.102      0.244     0.0104
10 2016-01-10 0.151 0.331      0.211     0.00699
# ... with 1,632 more rows
```

(2016-01-01부터 2020-06-29까지) 데이터 전체 기간의 시계열 그래프를 보면, 코로나바이러스가 나타나기 전에는 2017년 초에 flu(독감)과 2019년 중반에 cold(감기)가 반짝 검색되었던 것을 제외하면, 한국에서 coronavirus(코로나바이러스)만큼의 폭발적인 검색은 없었다.



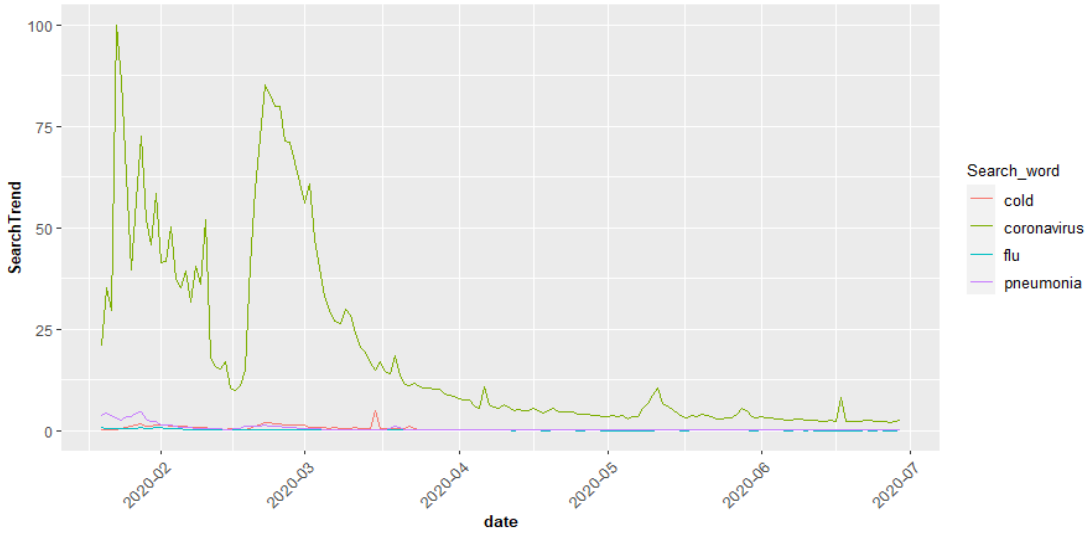
(2019-11-17부터 2020-06-29까지) 중국에서 처음 코로나가 발병한 이후 동안의 시계열 그래프를 보면, 첫 1개월 반 동안은 큰 변화가 나타나지 않았다. 눈에 띄는 점은 pneumonia(폐렴) 검색량 증가가 coronavirus(코로나바이러스)보다 먼저 일어난 것이다. 한국에서는 처음에 '우한성 폐렴'이라고 불렸기 때문이라 볼 수 있다. 이후에는 코로나바이러스가 계속해서 높은 검색량을 차지한다. 한국에서 첫 확진자가 발생한 후 3일 뒤인 2020-01-23에는 검색량이 100%인 최고조에 달했다.



```
> SearchTrend_corona <- SearchTrend %>% gather(key="Search_word", value="Sea
rchTrend", -date) %>% filter(Search_word=="coronavirus")
> SearchTrend_corona
# A tibble: 1,642 x 3
   date       Search_word SearchTrend
   <date>     <chr>             <dbl>
 1 2016-01-01 coronavirus     0.00736
 2 2016-01-02 coronavirus     0.0089
 3 2016-01-03 coronavirus     0.00845
 4 2016-01-04 coronavirus     0.0114
 5 2016-01-05 coronavirus     0.0138
 6 2016-01-06 coronavirus     0.0138
 7 2016-01-07 coronavirus     0.0119
 8 2016-01-08 coronavirus     0.0157
 9 2016-01-09 coronavirus     0.0104
10 2016-01-10 coronavirus     0.00699
# ... with 1,632 more rows

> SearchTrend_corona %>% arrange(desc(SearchTrend))
# A tibble: 1,642 x 3
   date       Search_word SearchTrend
   <date>     <chr>             <dbl>
 1 2020-01-23 coronavirus         100
 2 2020-01-24 coronavirus          86.1
 3 2020-02-22 coronavirus          85.2
 4 2020-02-23 coronavirus          82.9
 5 2020-02-25 coronavirus          80.0
 6 2020-02-24 coronavirus          79.9
 7 2020-01-28 coronavirus          72.6
 8 2020-02-21 coronavirus          72.2
 9 2020-02-26 coronavirus          71.3
10 2020-02-27 coronavirus          71.1
# ... with 1,632 more rows
```

(2020-01-20부터 2020-06-29까지) 첫 번째 피크 통과 후, 코로나바이러스 검색량은 약 한달 동안 감소한다. 그러나 2020-02-21 대구에서 신천지 교회 집단감염 사건이 발생하였고, 코로나바이러스 검색량은 그 다음 날인 2020-02-22에 정점을 찍은 것을 알 수 있다. 검색 경향은 이후로도 확진자 수와 큰 유사성을 보여주었기에, 이들 사이에 상관관계가 있을 수 있다고 생각하였다.



3. 자료로부터 궁금한 가설 세우기

1) '코로나 바이러스' 검색량과 확진자의 수는 유의미한 관계가 있다.

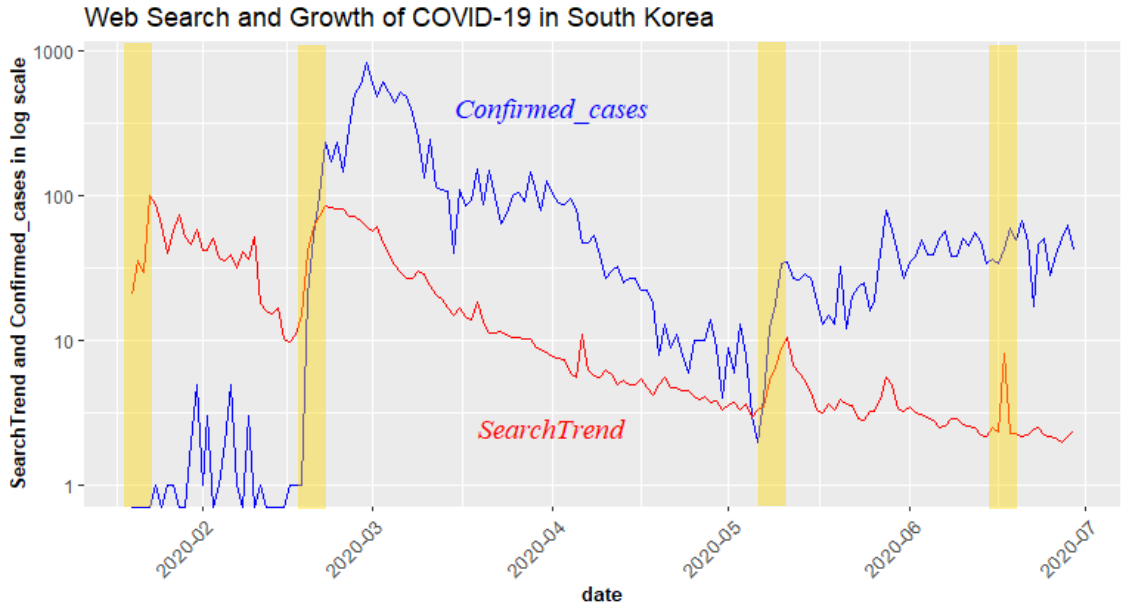
'코로나 바이러스' 검색량이 전일 대비 급증한 date에는 코로나 확진 판정수 또한 전일 대비 급증했을 것이다. 확진자 수가 증가하면 코로나 검색량도 증가하는지 상관관계를 알아보고자 한다.

2) '코로나 바이러스' 검색량은 바이러스가 등장했던 1월에 가장 많을 것이다.

3) '코로나 바이러스' 검색량은 기온이 낮을수록 증가하고, 특정 Policy가 시행된 date에 증가할 것이다.

4. 세운 가설을 알아내기 위한 그림들 그리기

1-1) 'coronavirus' SearchTrend(검색량, 1~100) 및 Confirmed_cases(확진자 수, log) 시계열 그래프



(2020-01-20부터 2020-06-29까지) SearchTrend.csv 자료에서 coronavirus의 SearchTrend(검색량)을 시간에 따라 나타낸 시계열 그래프이다(빨간색). 그와 동시에 Time.csv 자료의 누적 확진자 수로부터 Confirmed_cases(일별 확진자 수)를 구해 동일한 시계열 그래프상에 그 로그값을 나타냈다(파란색). 두 그래프를 비교한 결과, 확진자 수가 증가한 시기에 코로나바이러스 검색량 또한 어느 정도 증가한다는 것을 알 수 있다. 시계열 그래프 상 검색량의 4가지 피크는 다음과 같다.

(1) 2020-01-23(국내 코로나 환자 첫 발생 3일 뒤) 코로나바이러스 검색량은 최고조(100%)였다.

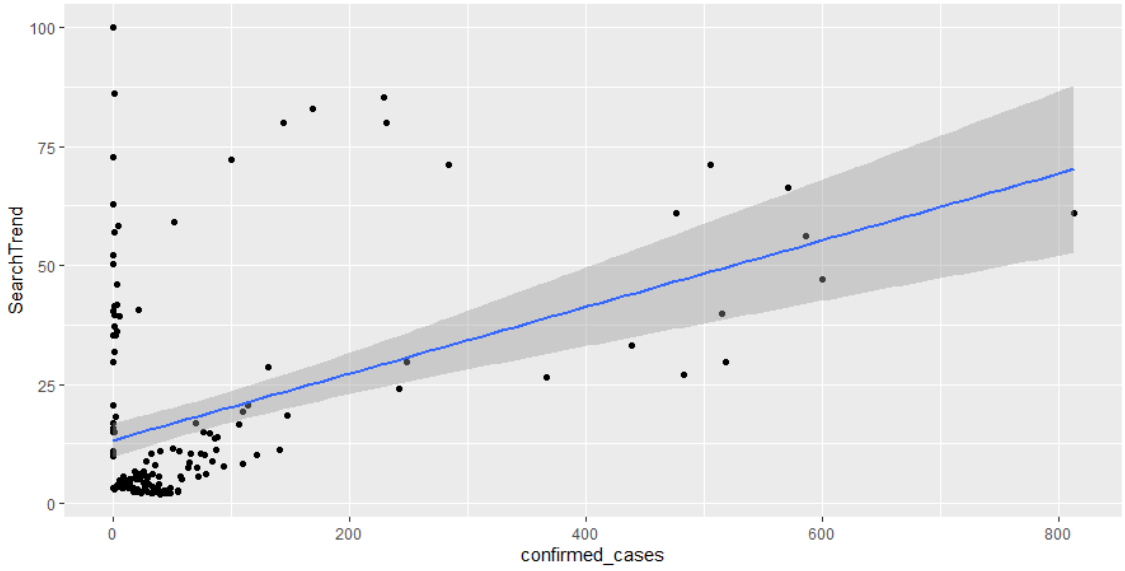
(2) 2020-02-22(신천지 교회의 집단감염 하루 뒤) 검색량 또한 급증하였다.

(3) 2020-05-11(연휴 이후, 이태원 클럽 집단감염 4일 뒤) 이태원 집단감염 사건을 시작으로 점차 줄고 있었던 확진자 수가 급증하는 추세를 보였으며, 검색량 역시 증가하였다.

(4) 2020-06-17(다단계 리치웨이 집단감염 일주일 뒤) 리치웨이발 집단감염이 교회, 콜센터, 학원 등으로 번지며 일일 확진자가 또다시 급증하였고, 검색량도 마찬가지로 늘어났다.

1-2) confirmed_cases(확진자 수)와 'coronavirus' SearchTrend(검색량)의 상관관계 비교

- 확진자 수에 따른 검색량의 산점도 및 회귀선



그래프 왼쪽에 산점도가 집중한 것을 보아, 확진자가 없을 때도 코로나바이러스 검색량은 많았던 것을 알 수 있다. 그러나 확진자가 증가할수록 코로나바이러스 검색량 또한 증가하는 추세임을 회귀선을 통해 확인할 수 있다.

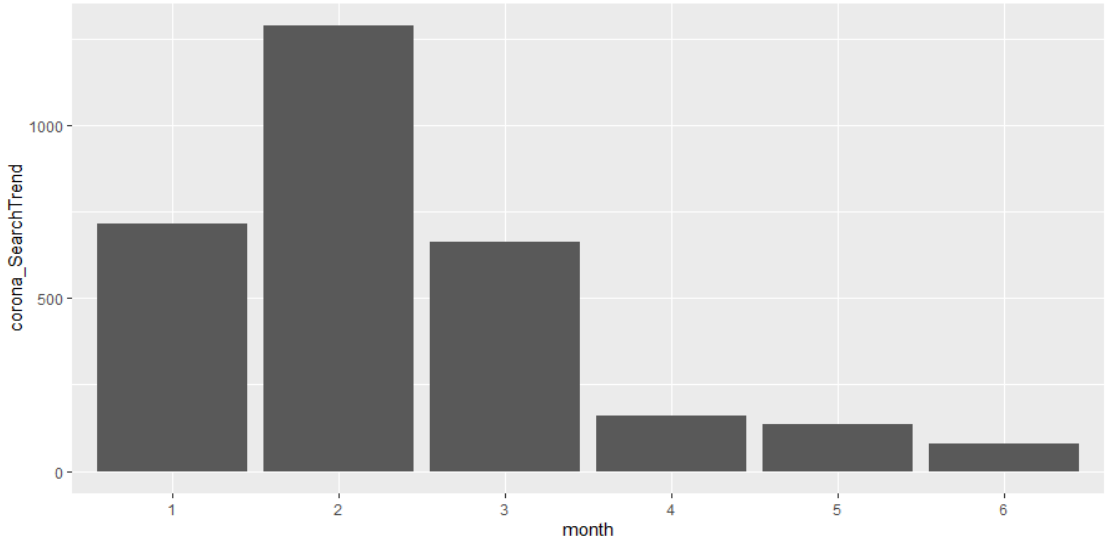
- 상관계수(상관관계의 크기를 나타내는 값)는 0.42로, 코로나바이러스 확진자 수와 검색량 간에 약한 양의 상관관계가 있음을 알 수 있다.

```
> cor(search_and_case_confirmed$SearchTrend,search_and_case_confirmed$diff)
[1] 0.4237677
```

① 상관계수가 양의 부호를 가지므로 두 변수 사이엔 양의 선형관계가 존재한다.

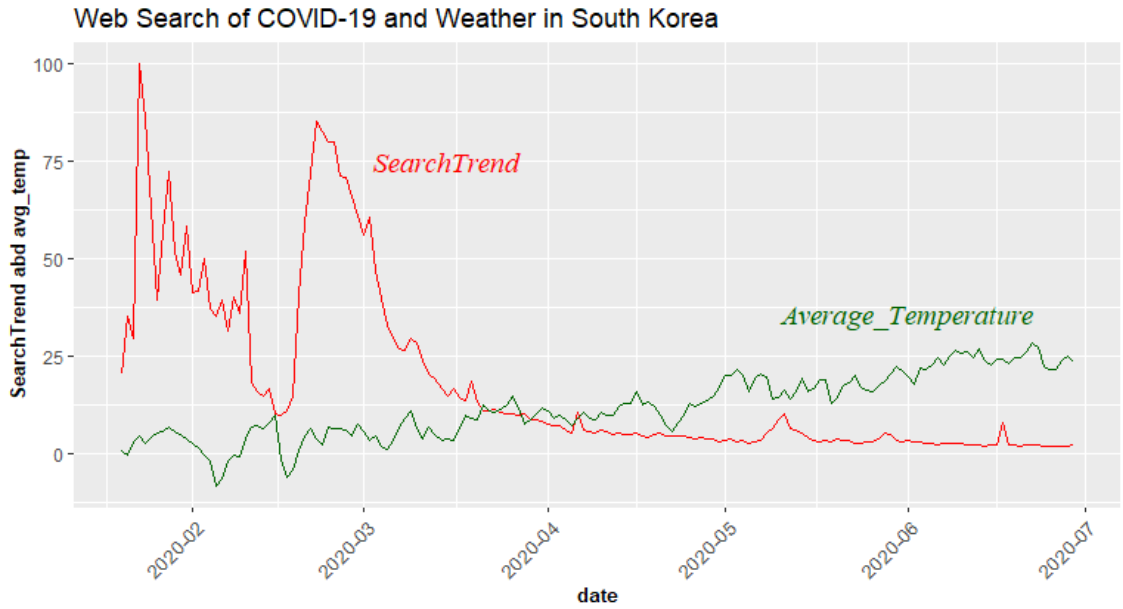
② 상관계수가 1보다는 0에 좀 더 가깝기 때문에 선형관계가 약하다.

2) 2020년 상반기 월별 코로나 검색량을 나타난 bar chart



'코로나바이러스' 검색량은 바이러스가 등장했던 1월이 아니라, 확진자 수가 급증하기 시작했던 2월에 가장 많은 것으로 드러난다. 1월 20일에 국내 첫 코로나 확진자가 등장하면서 1월 23일 검색량 최고조를 찍었지만, 아무래도 1월 중순 이후부터 이슈되었기 때문에 1월 검색량은 3월과 비슷한 수준이었다. 반면 집단감염으로 인해 확진자가 급증했던 2월, 코로나바이러스 검색량이 가장 두드러진다.

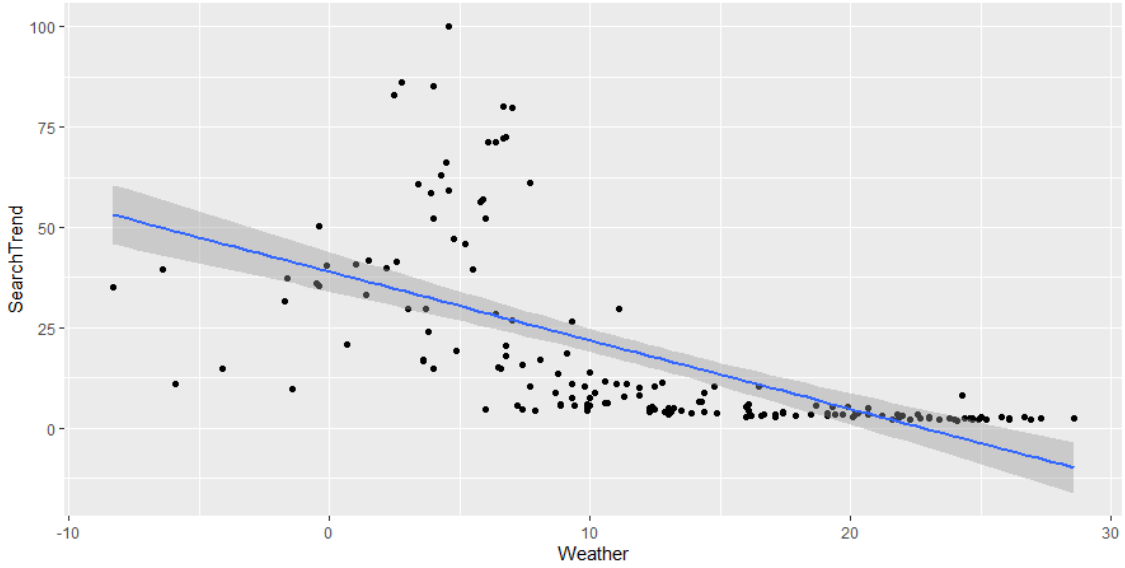
3-1) 'coronavirus' SearchTrend(검색량, 1~100) 및 Avg_Temp(평균 기온, ℃) 시계열 그래프



상대적으로 기온이 낮은 2,3월에 비해 따뜻한 6,7월에는 검색량이 감소하는 양상을 찾을 수 있다.

3-2) Avg_Temp(평균 기온)과 'coronavirus' SearchTrend(검색량)의 상관관계 비교

- 기온에 따른 검색량의 산점도 및 회귀선



기온이 높아질수록 '코로나 바이러스' 검색량은 감소하여 점점 0에 가까워지는 것을 확인할 수 있다.

- 상관계수(상관관계의 크기를 나타내는 값)는 -0.618로, 유의한 상관관계가 있음을 알 수 있다.

```
> cor(ww$avg_temp,search_and_case_confirmed$SearchTrend)
[1] -0.6181275
```

① 상관계수가 음의 부호를 가지므로 두 변수 사이엔 음의 선형관계가 존재한다.

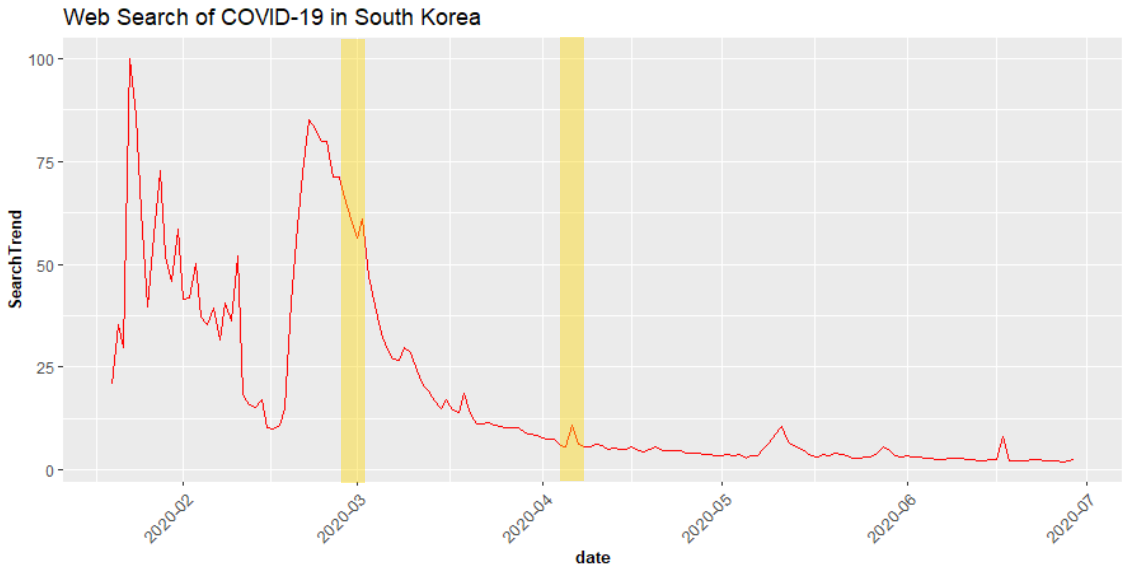
② 상관계수가 0보다는 -1에 좀 더 가깝기 때문에 선형관계가 유의하다.

* 이 자료는 2020 상반기 우리나라에서만 조사된 자료로, 시공간이 한정되어 있다는 점에 주목하고자 한다. 유럽과 북미 등 해외 수많은 곳에서는 여름이 되어서도 코로나바이러스의 확산세는 좀처럼 위축되지 않았기 때문이다.

한국에서 코로나바이러스가 발병한 시기가 겨울인 1월이었기 때문에, 이때 미지의 바이러스에 대한 검색량이 늘었을 수밖에 없다. 그러므로 날씨가 검색량에 영향을 미치는 주요한 요인이라고는 볼 수 없다.

3-3) 코로나바이러스 검색량 시계열 그래프

특정 Policy가 시행된 date에 검색량이 증가했는지 확인해보고자 한다.



(1) 2020-02-29(사회적 거리두기 강화 시작) 코로나바이러스 검색량이 증가하지는 않지만, 여전히 큰 비율(75%)의 검색량을 보인다.

(2) 2020-04-09(온라인 수업 개교 시행) 코로나바이러스 검색량의 미세한 증가를 확인할 수 있다.

5. 결론 도출하기

1) '코로나 바이러스' 검색량과 확진자의 수는 상관관계가 있다.

2) '코로나 바이러스' 검색량은 바이러스가 등장했던 1월이 아니라, 확진자 수가 급증하기 시작했던 2월에 가장 많다.

3) 기온이 낮을 때 '코로나 바이러스' 검색량 또한 증가하였다. 단 기온이 확진자 수와 유의미한 관계가 없는 것으로 보아, 기온이 검색량에 영향을 미치는 주요한 요인이라고 할 수 없다.

일부 정책이 시행된 날 '코로나 바이러스' 검색량은 증가한다.