Data Science Term Project

권영근 교수님

20182132 이진

20192149 신재희

20192119 김상훈

울산대학교
UNIVERSITY OF ULSAN

목차

Are Spring and Autumn Getting Shorter?

도입

누군가 '현대사회에서 가장 큰 화두가 무엇이라고 생각하나요?'라고 묻는다면 국내외를 불문하고 대다수가 '지구온난화'라고 대답 할 것입니다. 지구온난화로 인해 전 세계적으로 뚜렷한 연평균 기온의 상승이 보이는 가운데 저희는 '지구온난화로 인해 어떤 것이 변화하고 있을까?'에 대해서 의문점을 갖게 되었고, 최근 몇 년간 봄과 가을의 길이가 예전에 비해 짧아진 것 같은 느낌을 가지게 되었습니다. 이것이 단순히 체감상 이렇게 봄, 가을 길이가 짧아진다고 느껴지는 것인지 아니면 실제로 봄, 가을의 길이가 짧아지고 있는가? 하는 의문점이 들었습니다.

의문점을 해결하기 위해 우선 사람들이 흔히 생각하는 계절을 수치적 편의상 12개월을 3개월씩 4계절로 나눈, 즉 절대적인 기준에 의해 분류하는 것으로는 의미가 없다고 생각하였습니다. 이러한 이유 때문에 기존의 정량적 기준보다 직관적이고 현 기온변화의 동향을 잘 반영하는 사계절의 정의 및 분류 기준이 필요했고 아래 표의 실제 기상청에서 일평균기온을 이용해 분류한 사계절의 시작일 산출 기준을 이용했습니다.

기상청에 따르면 각 계절을 일정한 일평균 기온이 올라가거나 내려가지 않는 첫 날이라 정의를 했고, 저희는 이에 따라 일평균 기온 조건을 만족하는 기간을 9일로 정의하고 봄과 가을의 시작일을 구해 데이터를 분석해보았습니다.

	계절 시작일 산출 기준
봄 시작일	일평균기온이 5℃ 이상 올라간 후 다시 내려가지 않는 첫날
여름 시작일	일평균기온이 20℃ 이상 올라간 후 다시 내려가지 않는 첫날
가을 시작일	일평균기온이 20℃ 미만으로 내려간 후 다시 올라가지 않는 첫날
겨울 시작일	일평균기온 5℃ 미만으로 내려간 후 다시 올라가지 않는 첫날

가설

저희는 봄과 가을의 길이가 짧아짐을 알아보기 위해 아래와 같이 가설을 세우게 되었고, 이것을 토대로 실제 50년간 일평균기온 데이터를 분석을 해보기로 했습니다.

· Null hypothesis

- 봄과 가을의 길이는 예전에 비해 짧아지지 않았다.

· Alternative hypothesis

- 봄과 가을의 길이가 예전에 비해 짧아졌다.

데이터 설명

· Test statistic

- 1973~2021년의 대한민국 전국 평균기온, 최저기온, 최고기온

날짜	지점	평균기온(℃)	최저기온(℃)	최고기온(℃)
1973-01-01	전국	2.3	-1.3	6.2
1973-01-02	전국	-3.1	-7.7	2.6
1973-01-03	전국	-4.6	-9.5	0.3
1973-01-04	전국	-1.1	-5.8	4.2
1973-01-05	전국	0.1	-3.7	5.3
1973-01-06	전국	0.8	-4	5.4
1973-01-07	전국	2.2	0.9	4
1973-01-08	전국	2.1	-0.3	5.1
1973-01-09	전국	2.3	-1	6.3
1973-01-10	전국	0.9	-3.4	4

... (17887 rows omitted)

저희가 활용할 원본 데이터는 1973년 이후 전국의 하루 평균기온, 최저기온, 최고기온을 담고 있으나, 분석을 위해 필요한 데이터는 날짜와 평균기온이기 때문에 날짜와 평균기온 column만 가져와 데이터를 분석했습니다.

· Data Column별 설명

- **Date:** 날짜

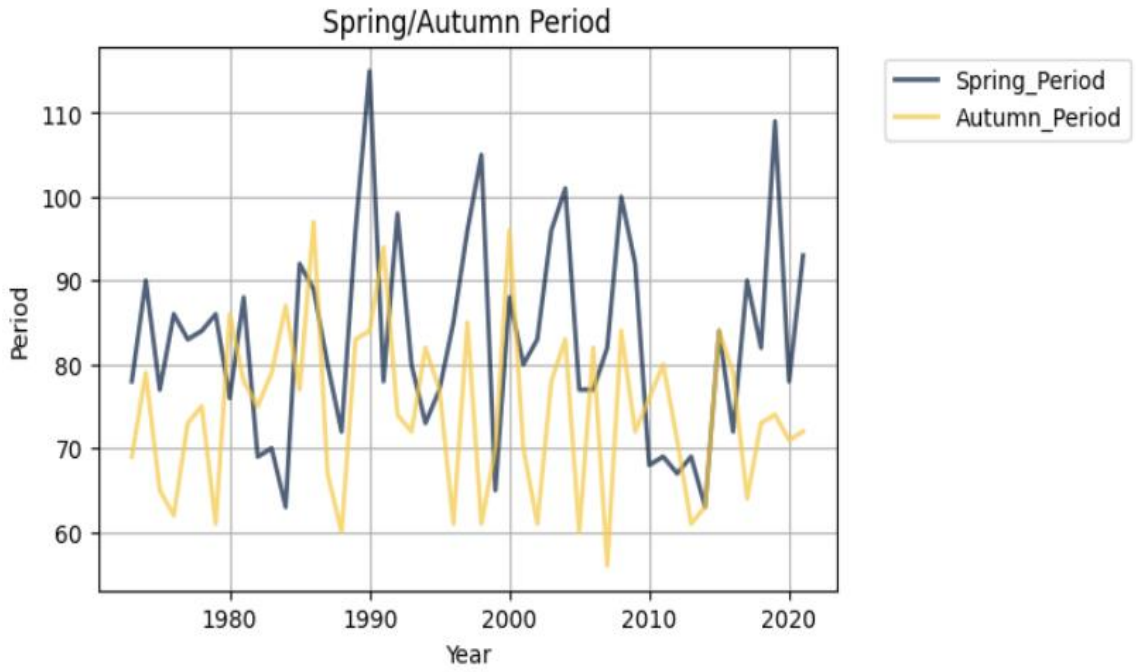
- **Temp:** 평균기온

Date	Temp
1973-01-01	2.3
1973-01-02	-3.1
1973-01-03	-4.6
1973-01-04	-1.1
1973-01-05	0.1
...	...
2021-12-27	-4.6
2021-12-28	0.0
2021-12-29	2.5
2021-12-30	-0.7
2021-12-31	-3.9

17897 rows × 1 columns

데이터 분석 방법 및 결과

· Period of Spring & Autumn

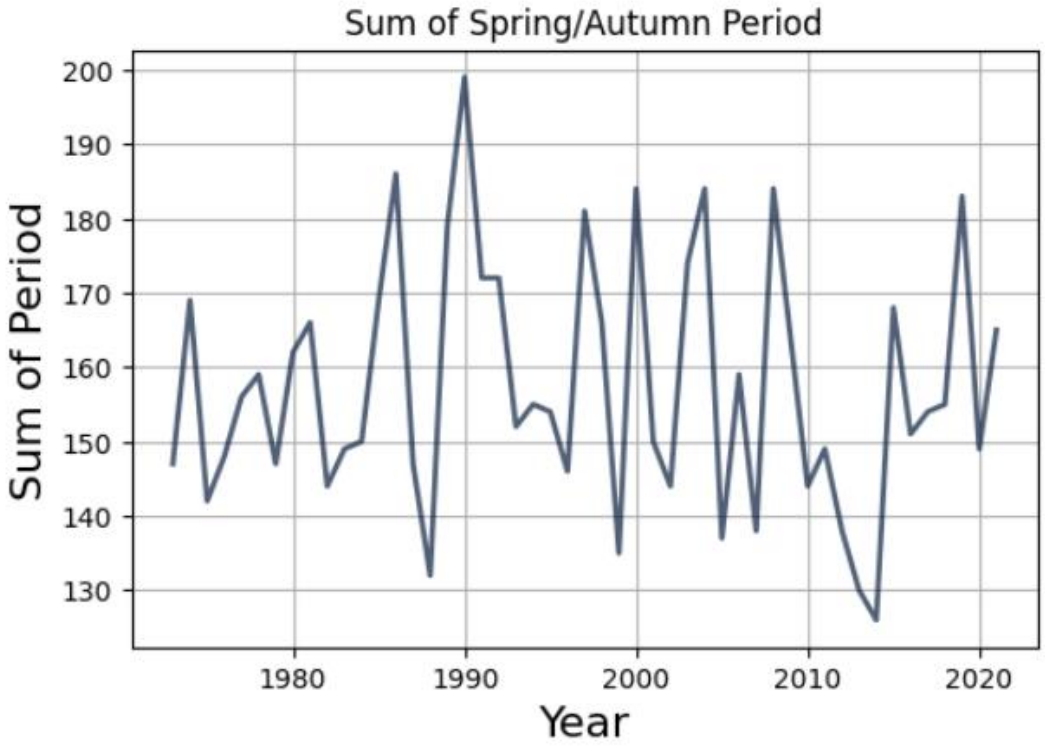


먼저 봄과 가을의 길이의 변화에 대해 알아보기 위해 저희가 활용할 일 평균기온 Table 에서 봄과 가을의 정의에 따라 계절의 시작일을 구해 [다음 계절 시작일 – 해당 계절 시작일]을 한 값인 계절의 길이를 구하여 Overlaid 하게 Plot Graph 로 시각화 하였습니다.

그래프의 x 축은 년도를, y 축은 봄과 가을의 길이를 의미합니다. 봄의 길이를 나타내는 파란색 Plot 과 가을의 길이를 나타내는 노란색 Plot 으로 1973 년에서 2021 년 사이의 봄과 가을의 길이의 동향을 살펴보았습니다.

봄과 가을의 길이가 크게 변화하지만 특정한 방향성이 보이지 않아 Plot Graph 만 보고 어떤 동향을 가지고 있는지 분석하기엔 무리가 있습니다.

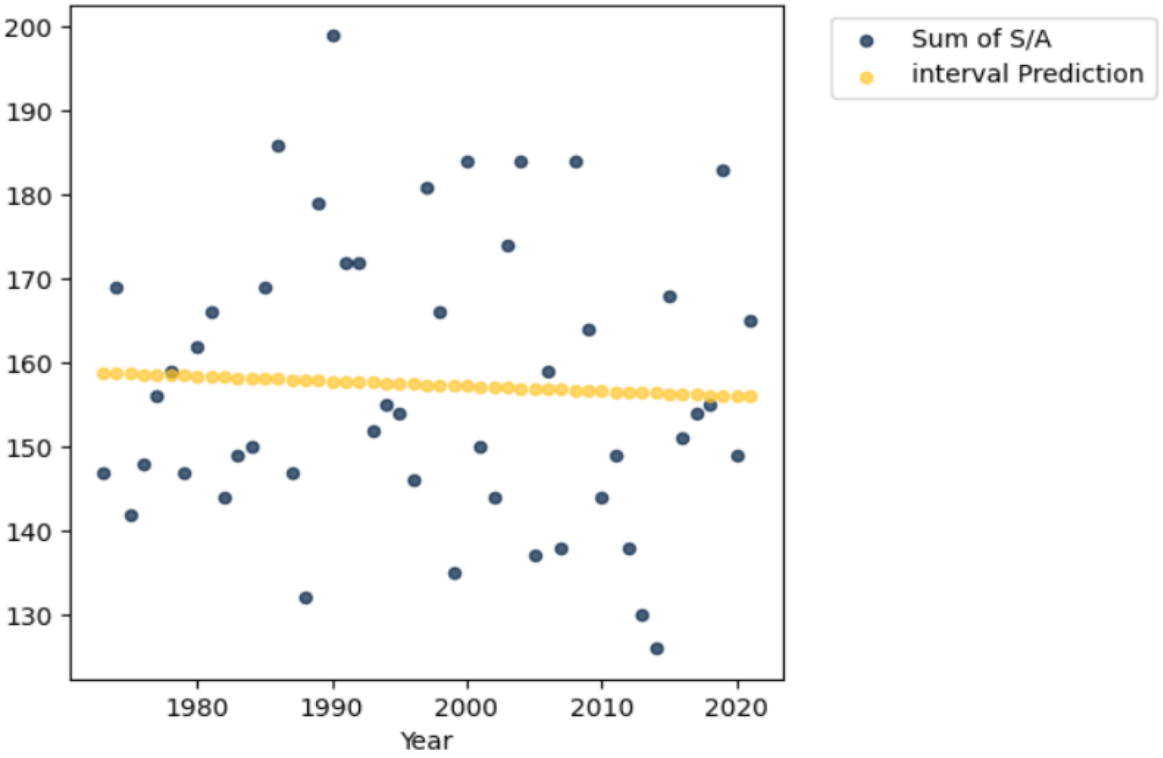
· Sum of the period of Spring & Autumn



저희는 봄과 가을의 길이 각각이 아닌 봄과 가을의 전체 길이를 알아야 하기에 봄과 가을의 길이의 합을 Plot Graph 로 표현하였습니다. x 축은 년도를, y 축은 [봄의 길이 + 가을의 길이]를 의미합니다.

봄,가을의 길이가 눈에 띄게 짧아지거나 길어지는 해가 드물지 않게 나타났지만, 전체적으로 봤을 때 최근의 봄,가을 길이 값이 짧아지지 않았다는 것을 확인할 수 있었습니다.

· Prediction of the period of Spring & Autumn




```
Correlation =  -0.05093180318720906
Slope =  -0.059795918367346934
```

이번에는 연도와 봄과 가을의 길이의 관계를 정량화하여 확인하기 위해 선형회귀 분석을 통해 Regression Line 과 이때의 기울기 및 상관계수를 나타냈습니다. 매 해 봄과 가을의 길이 모두 크게 변하는 현상이 나타나면서 회귀분석 결과 기울기가 미세하게 0 보다 적지만 상관계수 및 위의 Plot Graph 를 보았을 때, 변동이 커서 어떤 동향을 분석하기엔 역시 무리가 있었습니다.

저희는 위의 그래프들을 통해 봄과 가을의 자세한 동향을 알아내는 것에는 무리가 있다고 판단하였고, 최근 5 년과 그 이전의 데이터들을 비교 및 Shuffling 기법을 활용하여 분석을 해보았습니다.

· Shuffling

Shuffling 을 위해서 최근 5 년과 그 이전 년도로 데이터를 분류하였습니다. Years 는 old_year 과 recent_year 을 나타내는 column 이며 Sum_data 는 봄과 가을의 길이 합을 나타내는 column 입니다.

- recent_year
  - 최근 5 년의 데이터
  - 2017 년 ~ 2021 년의 봄/가을 길이
- old_year
  - 최근 5 년 이전의 데이터
  - 1973 년 ~ 2016 년의 봄/가을 길이

Years	Sum_data
recent_year	165
recent_year	149
recent_year	183
recent_year	155
recent_year	154
old_year	151
old_year	168
old_year	126
old_year	130
old_year	138

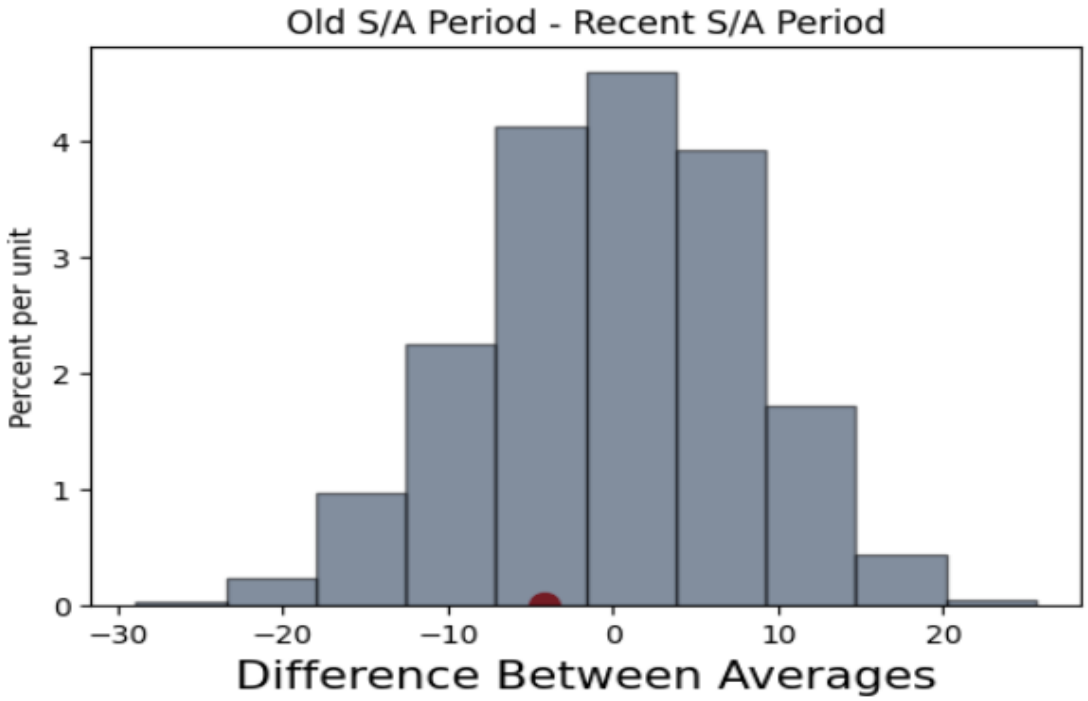
... (39 rows omitted)

Years	Sum_data mean
old_year	157
recent_year	161.2

-4.199999999999989

데이터를 최근 5년 및 이전 년도의 정보를 담은 Years column으로 grouping하여 평균값을 구하였고, [old_year의 봄/가을의 길이의 평균 – recent_year의 봄/가을의 길이의 평균]의 값을 관측값으로 설정 하였습니다.

이를 토대로 old_year과 recent_year의 봄/가을 길이의 합을 shuffling하여 [old_year의 봄/가을의 길이 합의 평균 – recent_year의 봄/가을의 길이 합의 평균]을 구하는 작업을 10000번 반복하여 이 값에 대한 Histogram을 그려 표현해보았습니다.



위 히스토그램의 x 축은 old_year 과 recent_year 의 봄과 가을의 길이 합을 random shuffling 하여 만들어진 shuffled data 의 [old_year 의 평균 봄/가을 길이 – recent_year 의 평균 봄/가을 길이]을 나타냅니다. 중앙에 위치한 붉은색 반원은 앞서 구한 관측값을 표시한 것입니다.

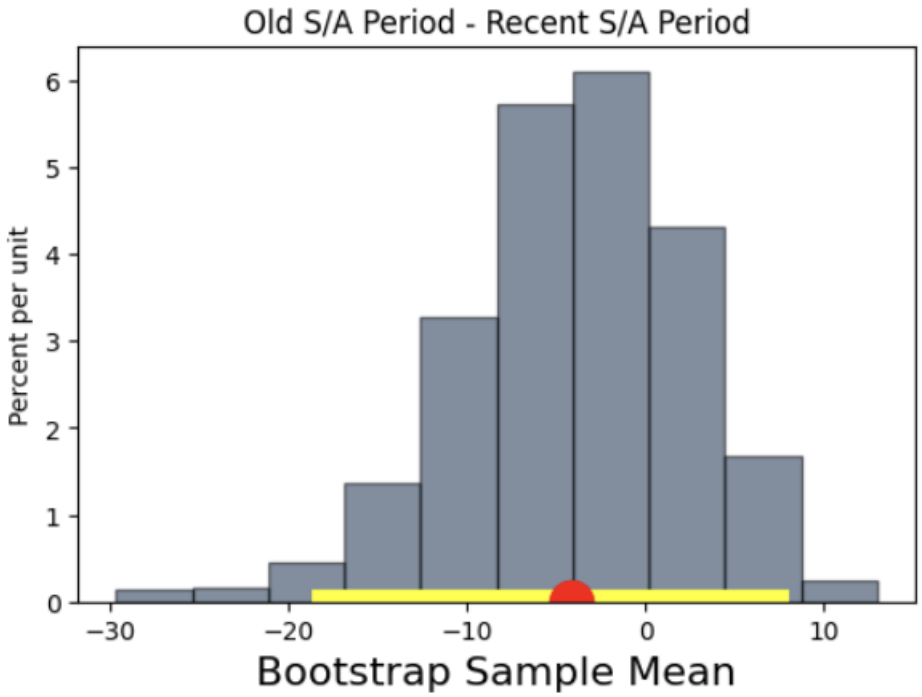
```
0.702
```

히스토그램 x 축의 대부분은 0 에 밀집하여 분포하였고 관측값은 중앙에서 살짝 왼쪽에 치우쳐 위치하고 있는 것을 알 수 있습니다. 저희는 이때의 p-value 값이 0.702 인 것을 보아 봄,가을의 길이가 짧아지지 않는다는 귀무가설을 기각하지 않고 유의한 차이가 없다는 결론을 내리게 됐습니다.

이 분석을 통하여 봄/가을의 길이가 짧아지고 있다는 Alternative Hypothesis 보다 봄/가을의 길이가 짧아지지 않았다는 Null Hypothesis 쪽에 힘이 실리게 됩니다.

하지만 이렇게 Null Hypothesis에 힘이 실리게 되는 관측값은 저희의 데이터에서만 두드러지는 값일 수 있으므로 모집단을 분석해볼 필요가 있습니다. 우리나라의 현재까지 모든 해의 일평균 기온 결과를 담은 데이터를 분석할 수 없으므로 저희는 Bootstrap방법을 이용하여 모집단의 데이터를 분석해보겠습니다.

· Bootstrap



랜덤샘플을 뽑아 (2017년 이전의 봄/가을 길이 - 2017년 이후의 봄/가을 길이)를 구하는 작업을 5000번 실시하였을 때의 히스토그램입니다.

그래프 하단의 노란색 선으로 95%의 신뢰구간을, 붉은색 반원으로 저희의 Shuffling 관측값을 표시하였습니다.

```
(-18.0, 7.30278)
```

위 값은 모집단을 추측해보았을 때 95%의 신뢰구간입니다. 높은 확률로 모집단의 관측값도 이 범위 안에 있을 것으로 추측됩니다. 물론 저희가 사용한 데이터의 관측값도 위의 그래프에 기재한 바와 같이 모집단의 신뢰구간 안에 포함됩니다.

Bootstrap 방법을 이용한 분석 이후에, 최근 5년간의 봄/가을 길이가 이전과 비교했을 때 짧아지지 않았다는 것이 더욱 확실해졌습니다.

· Why?

저희는 사람들이 봄, 가을의 길이가 짧아졌다고 느끼는 이유에 대해 최근 지구온난화로 인해 '여름의 길이가 길어지며 봄이 빨리 찾아오고, 가을은 늦춰지게 됐다'로 예상하였고 이에 대해 분석해보았습니다.

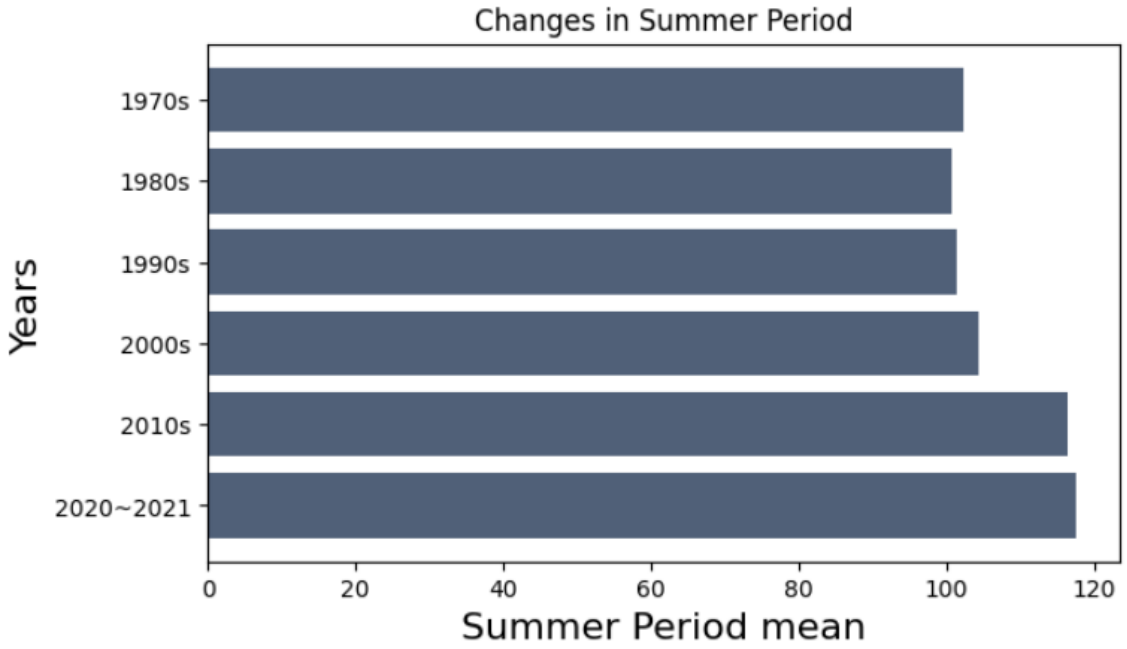
· New Table

Years	Spring_Start_Date	Summer_Start_Date	Autumn_Start_Date
1973	84	162	264
1974	82	172	256
1975	90	167	273
1976	83	169	255
1977	69	152	277
1978	80	164	275
1979	76	162	264
1980	84	160	259
1981	74	162	251
1982	84	153	262

... (39 rows omitted)

여름의 길이와 봄, 가을의 시작 날짜를 구하기 위해 각 계절마다 정의한 조건에 따라 1년 365일 기준 봄, 여름, 가을의 시작 날짜의 index를 담은 새로운 Table을 생성하였습니다. 예를 들어 1973년 봄의 시작 날짜 index 데이터가 84이면, 1973년의 84번째 날인 1973년 3월 25일을 뜻합니다.

· Change In Summer Period



위의 테이블에서 (가을 시작 날짜 – 여름 시작 날짜)를 하여 여름의 길이를 구한 다음, 년대별로 grouping 하여 평균값을 구한 데이터를 Bar Graph로 나타냈습니다. 위의 그래프에서 알 수 있듯이 2020년대에 들어서 1970년대에 비해 여름이 15일가량 늘어난 것을 확인할 수 있었습니다.

Years	Spring_start	Autumn_start
1970s	80.5714	266.286
2020s	68	271

또한, 1970년대와 2020년대의 평균 봄/가을 시작 날짜를 비교하여 1970년대에 비해 2020년대의 봄은 12일가량 빨라지고 가을은 4~5일 가량 늦춰진 것을 확인할 수 있었습니다.

결론

데이터를 다양한 방법으로 분석해본 결과 최종적으로 앞서 저희가 세운 Null Hypothesis 를 Reject 하는 근거를 찾지 못 하였고 이에 따라 '봄과 가을의 길이는 예전에 비해 짧아지지 않았다.'라는 귀무가설을 기각하지 못한다는 결론을 내었습니다.

따라서 최근 들어 봄,가을의 길이가 짧아지고 있다고 보긴 힘들다는 사실을 알게 되었습니다. 하지만 2010 년대에 50 년동안 봄,가을이 가장 짧았던 해가 있는 것으로 보았을 때 예전에 비해 봄과 가을이 짧아지고 있지 않긴 하나, 어떠한 다른 요인에 의해 봄,가을이 전년도에 비해 급작스레 짧아질 수 있다고 볼 수 있습니다.

그러나 이러한 극단적으로 짧아지는 경우 또한 포함한 데이터에서 수치상 유의미한 차이를 보이지 않았기 때문에 예전에 비해 봄과 가을의 길이가 짧아지지 않았다는 것은 확실합니다.

논의사항

2010 년대 초반에 유난히 봄/가을길이가 짧아지고 있다는 예측이 많았던 이유에 대해 논의를 해보았을 때, 당시에 근 50 년간 최단 기간의 봄, 가을을 기록한 해였기 때문이라 생각하였고 2010 년대 중반이후로 다시 봄,가을의 길이가 정상적으로 돌아오고 있다고 생각하였습니다.

미비점

데이터가 매년 변화량이 급격하게 줄거나 늘어나기 때문에 데이터 분석 결과값이 전체적으로 신뢰성이 떨어질 수 있다고 생각했습니다. 또한, 기상청의 자료인 1973 년 이후의 자료로 분석을 해보았는데 표본 집단의 크기가 상대적으로 작다고 생각했습니다.

이로 인해 Bootstrap 방법을 적용시 신뢰구간을 좀 더 줄이지 못하였고 만약 표본 데이터의 크기가 좀 더 컸더라면 부트스트랩의 신뢰성을 더할 수 있을 것이라 생각합니다.