미니 프로젝트

강수량 데이터 분석 및 시각화

안동대학교 정보통계학과

류승호

프로젝트 기간 : 2022.08.03 ~ 2022.08.05

CONTENTS

01. 데이터 선택

- 1960~2021년
  전국 강수량 데이터

02. 데이터 전처리

I. excel 원본 데이터 확인
II. excel 전처리
III. Python으로 데이터 확인
IV. Python 전처리

03. 데이터 분석

I. 지점별 데이터
II. 지점별 통계량
III. 지점별 최대 강수량
IV. 년도별 데이터
V. 년도별 통계량
VI. 년도별 최대 강수량

04. 데이터 시각화

I. 지점별 강수량(상위 30 and 하위 30)
II. 년도별 강수량(상위 30 and 하위 30)
III. 입력한 지역의 년도별 강수량 그래프
IV. 입력한 지역, 년도의 월별 그래프
V. 두개의 지역, 년도의 비교 그래프
VI. 입력한 지역, 년도의 계절별
강수량 평균 파이차트

데이터 선정이유

'낙동강 녹조' 부산에서도 심각 수준...강수량 감소·기온상승 영향

보리·마늘·양파 안보이네…강수량 부족에 생산량 급감

최근 강수량과 기온 상승으로 인하여 많은 피해가 일어나고 있다.
-> 지역별, 년도별 강수량 비교

02 데이터 전처리

I. excel 원본 데이터 확인
II. excel 전처리
III. Python으로 데이터 확인
IV. Python 전처리

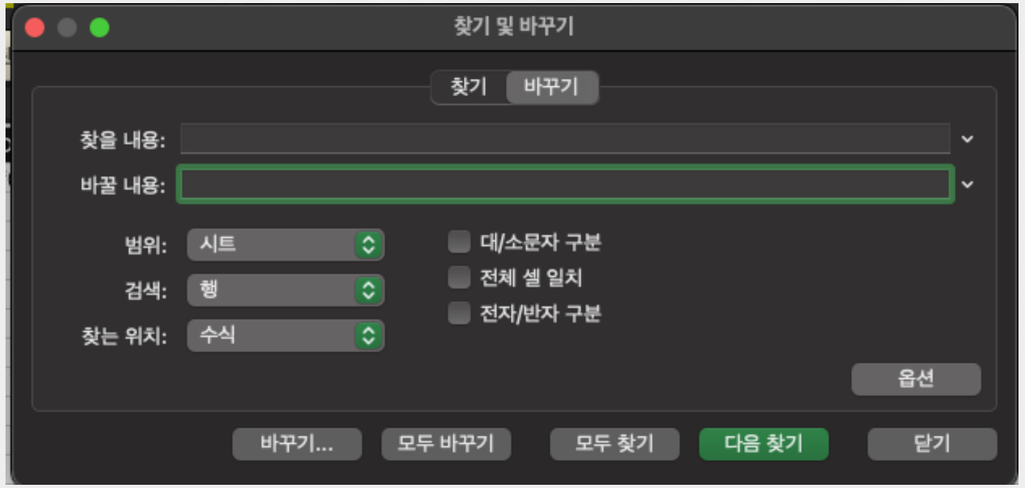
I. 데이터 확인

excel 전처리

지점번호

A
지점번호
90
90
90
90
90
90
90
90
90
90

A열 전체 데이터에 앞에 (Tab) 존재



A열 전체에 대하여 찾기 및 바꾸기 실행
찾을 내용 : (Tab)
바꿀 내용 : spacebar
-> 모두 바꾸기

지점번호

A
지점번호
90
90
90
90
90
90
90
90
90
90

A열 전체 데이터 (Tab) 제거

02 데이터 전처리

I. excel 원본 데이터 확인

excel 데이터 확인

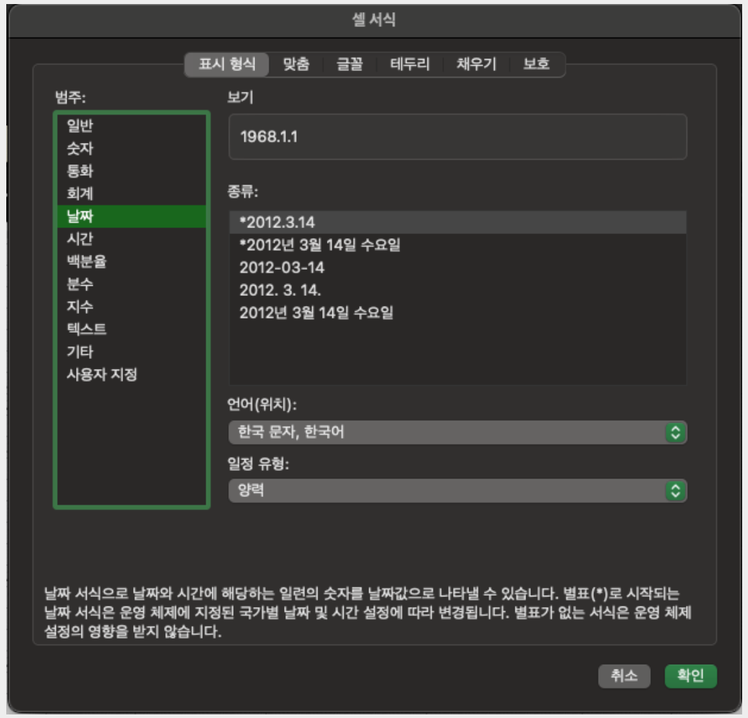
지점번호	지점명	일시	강수량(mm)	일최다강수량	일최다강수량	1시간최다강	1시간최다강수
90	속초	Jan.68	0	0	1968.1.24	0	1968.1.2
90	속초	Feb.68	3.3	2	1968.2.29	0	1968.2.10
90	속초	Mar.68	8.8	3.3	1968.3.24		
90	속초	Apr.68	6.9	3.4	1968.4.8	1.4	1968.4.5
90	속초	May.68	49.1	31.8	1968.5.18	5.7	1968.5.18
90	속초	Jun.68	48.3	14.5	1968.6.19	5.7	1968.6.19
90	속초	Jul.68	304.7	89.7	1968.7.16	20.4	1968.7.30
90	속초	Aug.68	299.8	155.2	1968.8.23	36.3	1968.8.23
90	속초	Sep.68	113.7	77.5	1968.9.5	9.6	1968.9.5
90	속초	Oct.68	342.1	170.3	1968.10.24	23	1968.10.24
90	속초	Nov.68	24.5	14.1	1968.11.27	4.7	1968.11.27
90	속초	Dec.68	60.8	14.8	1968.12.26	3.2	1968.12.12
90	속초	Jan.69	112.6	25.9	1969.1.28		
90	속초	Feb.69	139.1	29.9	1969.2.18		
90	속초	Mar.69	77.5	28.7	1969.3.16		
90	속초	Apr.69	254.3	139.7	1969.4.24	19.4	1969.4.23
90	속초	May.69	112.3	55.8	1969.5.27	10.1	1969.5.27
90	속초	Jun.69	59.6	22.5	1969.6.21	15.1	1969.6.21
90	속초	Jul.69	219	83.4	1969.7.30	25.2	1969.7.30
90	속초	Aug.69	236.4	84.2	1969.8.3	15.6	1969.8.6
90	속초	Sep.69	104.2	42.7	1969.9.30	6.8	1969.9.5
90	속초	Oct.69	3	1.5	1969.10.16	1.1	1969.10.16
90	속초	Nov.69	48.9	36.8	1969.11.16	7.9	1969.11.16
90	속초	Dec.69	23.2	22.5	1969.12.7	4.9	1969.12.7
90	속초	Jan.70	22.8	10.7	1970.1.4	0	1970.1.19
90	속초	Feb.70	83.4	42.7	1970.2.25	0	1970.2.19
90	속초	Mar.70	12.9	9.5	1970.3.9		

II. excel 전처리

excel 전처리

C
일시
Jan.68
Feb.68
Mar.68
Apr.68
May.68
Jun.68
Jul.68
Aug.68

C열 날짜데이터가 보기 불편합니다.



C열 날짜 데이터를 셀 서식을 이용하여 변경

C
일시
1968.1.1
1968.2.1
1968.3.1
1968.4.1
1968.5.1
1968.6.1
1968.7.1
1968.8.1
1968.9.1
1968.10.1
1968.11.1

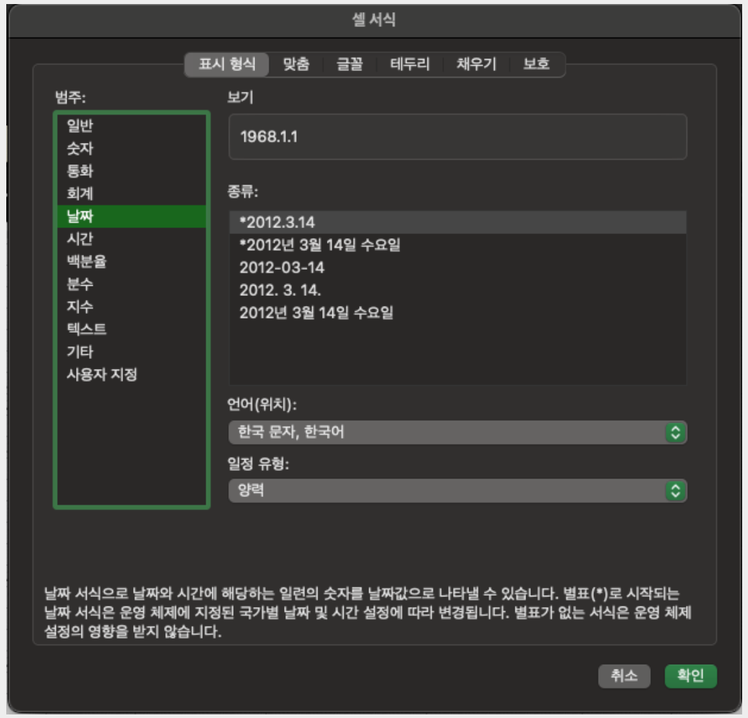
C열 날짜 데이터 형식 변경

II. excel 전처리

excel 전처리

C
일시
Jan.68
Feb.68
Mar.68
Apr.68
May.68
Jun.68
Jul.68
Aug.68

C열 날짜데이터가 보기 불편합니다.



C열 날짜 데이터를 셀 서식을 이용하여 변경

C
일시
1968.1.1
1968.2.1
1968.3.1
1968.4.1
1968.5.1
1968.6.1
1968.7.1
1968.8.1
1968.9.1
1968.10.1
1968.11.1

C열 날짜 데이터 형식 변경

III. Python으로 데이터 확인

Python 데이터 확인

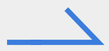
	지점번호	지점명	일시	강수량(mm)	일최다강수량(mm)	일최다강수량일자	1시간최다강수량(mm)	1시간최다강수량일자
0	90	속초	1968.1.1	0.0	0.0	1968.1.24	0.0	1968.1.2
1	90	속초	1968.2.1	3.3	2.0	1968.2.29	0.0	1968.2.10
2	90	속초	1968.3.1	8.8	3.3	1968.3.24	NaN	NaN
3	90	속초	1968.4.1	6.9	3.4	1968.4.8	1.4	1968.4.5
4	90	속초	1968.5.1	49.1	31.8	1968.5.18	5.7	1968.5.18
...	...	...	...	...	...	...	...	...
49396	295	남해	2022.4.1	142.1	72.8	2022.4.26	29.0	2022.4.26
49397	295	남해	2022.5.1	6.7	6.0	2022.5.21	6.0	2022.5.21
49398	295	남해	2022.6.1	137.1	47.7	2022.6.5	17.1	2022.6.24
49399	295	남해	2022.7.1	236.5	77.5	2022.7.18	31.4	2022.7.11
49400	295	남해	2022.8.1	59.9	29.5	2022.8.2	11.6	2022.8.2

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49401 entries, 0 to 49400
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0      지점번호        49401 non-null  int64
 1   지점명             49401 non-null  object
 2   일시               49401 non-null  object
 3   강수량(mm)         49021 non-null  float64
 4   일최다강수량(mm)     49021 non-null  float64
 5   일최다강수량일자       49021 non-null  object
 6   1시간최다강수량(mm)  30939 non-null  float64
 7   1시간최다강수량일자     30939 non-null  object
dtypes: float64(3), int64(1), object(4)
memory usage: 3.0+ MB
```

- 날짜 데이터(일시, 일최다강수량일자, 1시간최다강수량일자)를 object -> datetime64로 변경
- 날짜 데이터 (1960년 1월 ~ 2022년 8월까지) 확인
- 결측값 확인 -> 결측값 처리

Python 전처리

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49401 entries, 0 to 49400
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0      지점번호         49401 non-null  int64
 1   지점명             49401 non-null  object
 2   일시               49401 non-null  object
 3   강수량(mm)         49021 non-null  float64
 4   일최다강수량(mm)     49021 non-null  float64
 5   일최다강수량일자       49021 non-null  object
 6   1시간최다강수량(mm)   30939 non-null  float64
 7   1시간최다강수량일자     30939 non-null  object
dtypes: float64(3), int64(1), object(4)
memory usage: 3.0+ MB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49401 entries, 0 to 49400
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0      지점번호         49401 non-null  int64
 1   지점명             49401 non-null  object
 2   일시               49401 non-null  datetime64[ns]
 3   강수량(mm)         49021 non-null  float64
 4   일최다강수량(mm)     49021 non-null  float64
 5   일최다강수량일자       49021 non-null  datetime64[ns]
 6   1시간최다강수량(mm)   30939 non-null  float64
 7   1시간최다강수량일자     30939 non-null  datetime64[ns]
dtypes: datetime64[ns](3), float64(3), int64(1), object(1)
memory usage: 3.0+ MB
```

- 날짜 데이터를 분석하기 위하여 datetime64[ns]로 형(type) 변환

- 날짜 최소 데이터 `Timestamp('1960-01-01 00:00:00')`

- 날짜 최대 데이터 `Timestamp('2022-08-01 00:00:00')`

- 날짜 데이터가 원본과 같은지 확인 1960년 1월 부터 2022년 8월 까지

Python 전처리

- 결측치 확인 `True`

```
    지점번호      False
지점명          False
일시           False
강수량(mm)       True
일최다강수량(mm)     True
일최다강수량일자       True
1시간최다강수량(mm)    True
1시간최다강수량일자      True
dtype: bool
```

```
    지점번호      False
지점명          False
일시           False
강수량(mm)       False
일최다강수량(mm)     False
일최다강수량일자       True
1시간최다강수량(mm)    False
1시간최다강수량일자      True
dtype: bool
```

- 강수량, 일최다강수량, 1시간최다강수량 데이터 결측값 제거
- 일 최다 강수량일자, 1시간 최다 강수량 일자는 강수가 없을 수도 있으니 Null로 놔둔다.

Python 전처리

- 결측치 확인 `True`

```
    지점번호          False
지점명              False
일시               False
강수량(mm)          True
일최다강수량(mm)      True
일최다강수량일자       True
1시간최다강수량(mm)    True
1시간최다강수량일자     True
dtype: bool
```

```
    지점번호          False
지점명              False
일시               False
강수량(mm)          False
일최다강수량(mm)      False
일최다강수량일자       True
1시간최다강수량(mm)    False
1시간최다강수량일자     True
dtype: bool
```

- 강수량, 일최다강수량, 1시간최다강수량 데이터 결측값 제거

03 데이터 분석

I. 지점별 데이터

지역 입력(시군구) : 밀양

밀양을 입력하여 데이터 출력

지역 입력(시군구) : 밀양

	지점번호	지점명	일시	강수량(mm)	일최다강수량(mm)	일최다강수량일자	1시간최다강수량(mm)	1시간최다강수량일자	Month	Season
46982	288	밀양	1972-01-01	0.0	0.0	NaT	0.0	NaT	1	4
46983	288	밀양	1973-01-01	47.5	23.3	1973-01-24	4.0	1973-01-24	1	4
46984	288	밀양	1973-02-01	28.8	9.6	1973-02-22	2.5	1973-02-22	2	4
46985	288	밀양	1973-03-01	9.1	4.4	1973-03-09	1.5	1973-03-09	3	1
46986	288	밀양	1973-04-01	162.5	72.2	1973-04-24	19.0	1973-04-24	4	1
...	...	...	...	...	...	...	...	...	...	...
47574	288	밀양	2022-04-01	59.6	37.5	2022-04-26	21.0	2022-04-26	4	1
47575	288	밀양	2022-05-01	3.3	3.3	2022-05-02	2.6	2022-05-02	5	1
47576	288	밀양	2022-06-01	232.8	125.6	2022-06-27	38.4	2022-06-27	6	2
47577	288	밀양	2022-07-01	112.8	68.3	2022-07-18	19.2	2022-07-18	7	2
47578	288	밀양	2022-08-01	31.0	30.9	2022-08-02	19.8	2022-08-02	8	2

```
지역 입력(시군구): 밀양
```

밀양을 입력하여 밀영의 강수량 통계량 출력

```
지역 입력(시군구): 밀양

count    597.000000
mean     101.611055
std      109.346063
min        0.000000
25%       25.400000
50%       61.200000
75%      136.000000
max      695.000000
Name: 강수량(mm), dtype: float64
```

03 데이터 분석

IV. 년도별 데이터

년도 입력 : 2021

년도 입력 : 2021

	지점번호	지점명	일시	강수량(mm)	일최다강수량(mm)	일최다강수량일자	1시간최다강수량(mm)	1시간최다강수량일자	Month	Season
636	90	속초	2021-01-01	0.0	0.0	2021-01-31	0.0	NaT	1	4
637	90	속초	2021-02-01	5.0	4.8	2021-02-27	0.0	NaT	2	4
638	90	속초	2021-03-01	104.2	73.4	2021-03-01	0.0	NaT	3	1
639	90	속초	2021-04-01	97.2	40.5	2021-04-03	5.9	2021-04-03	4	1
640	90	속초	2021-05-01	154.9	67.0	2021-05-16	12.1	2021-05-16	5	1
...	...	...	...	...	...	...	...	...	...	...
49388	295	남해	2021-08-01	473.2	185.7	2021-08-21	73.5	2021-08-21	8	2
49389	295	남해	2021-09-01	164.6	56.2	2021-09-29	26.4	2021-09-29	9	3
49390	295	남해	2021-10-01	43.6	18.2	2021-10-10	11.4	2021-10-10	10	3
49391	295	남해	2021-11-01	53.5	31.3	2021-11-08	0.0	NaT	11	3
49392	295	남해	2021-12-01	2.4	2.4	2021-12-16	0.0	NaT	12	4

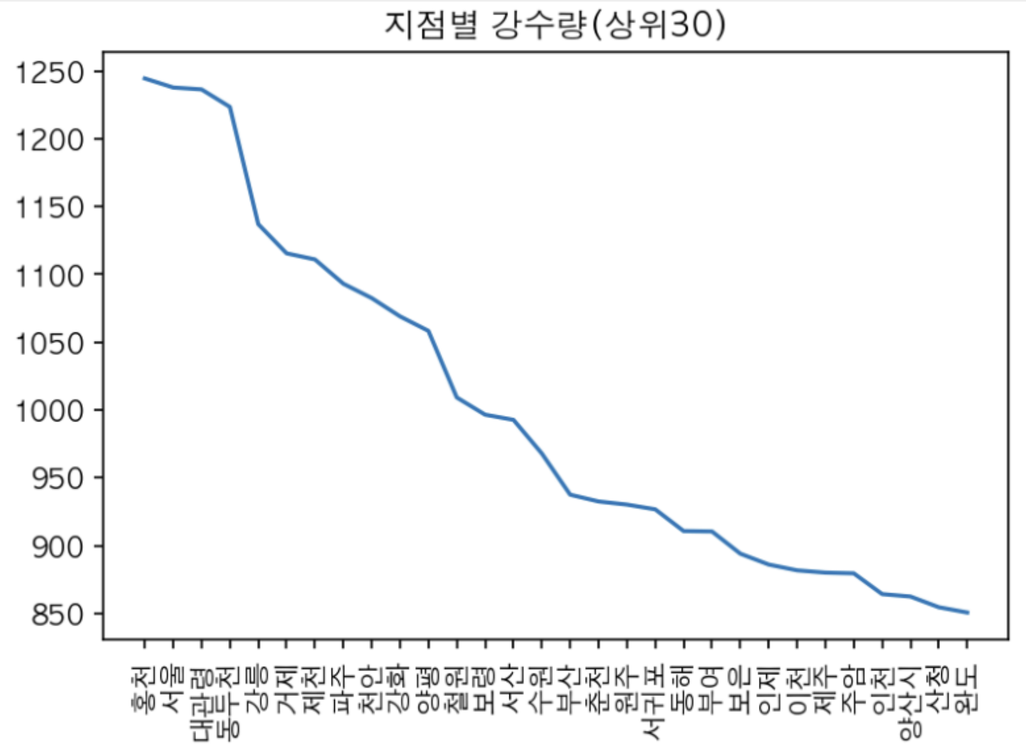
```
년도 입력 : 2021
```

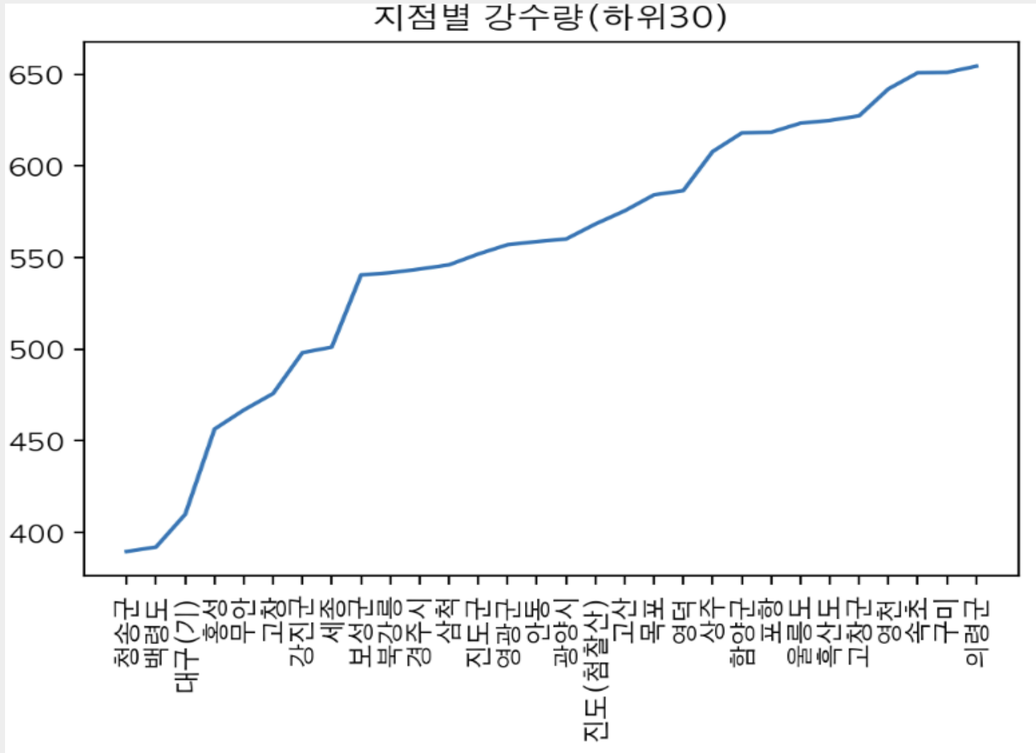
```
년도 입력 : 2021

count    1140.000000
mean      106.432632
std       101.368726
min         0.000000
25%        35.275000
50%        84.150000
75%       143.125000
max       692.400000
Name: 강수량(mm), dtype: float64
```

04. 데이터 시각화

Ⅰ.지점별 강수량(상위 30 and 하위 30)

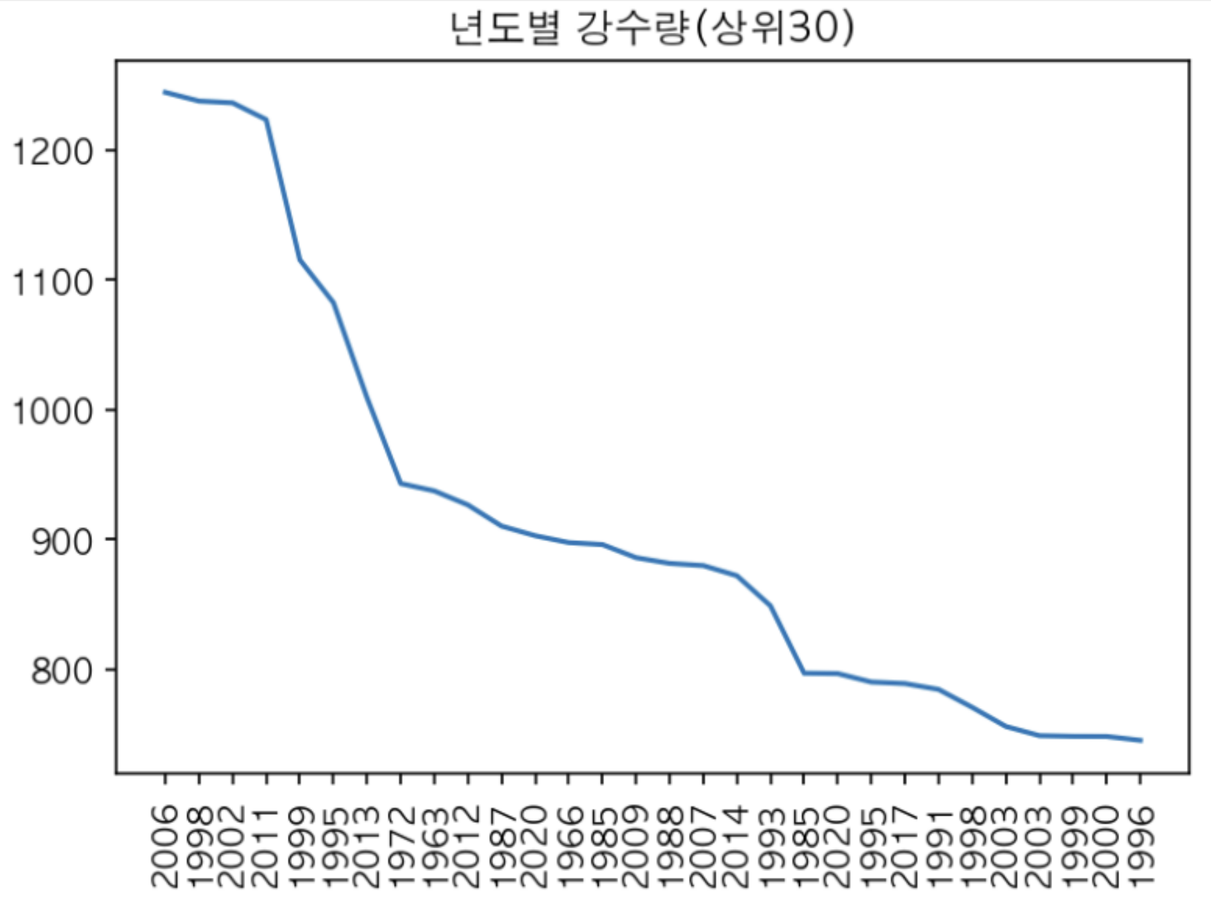


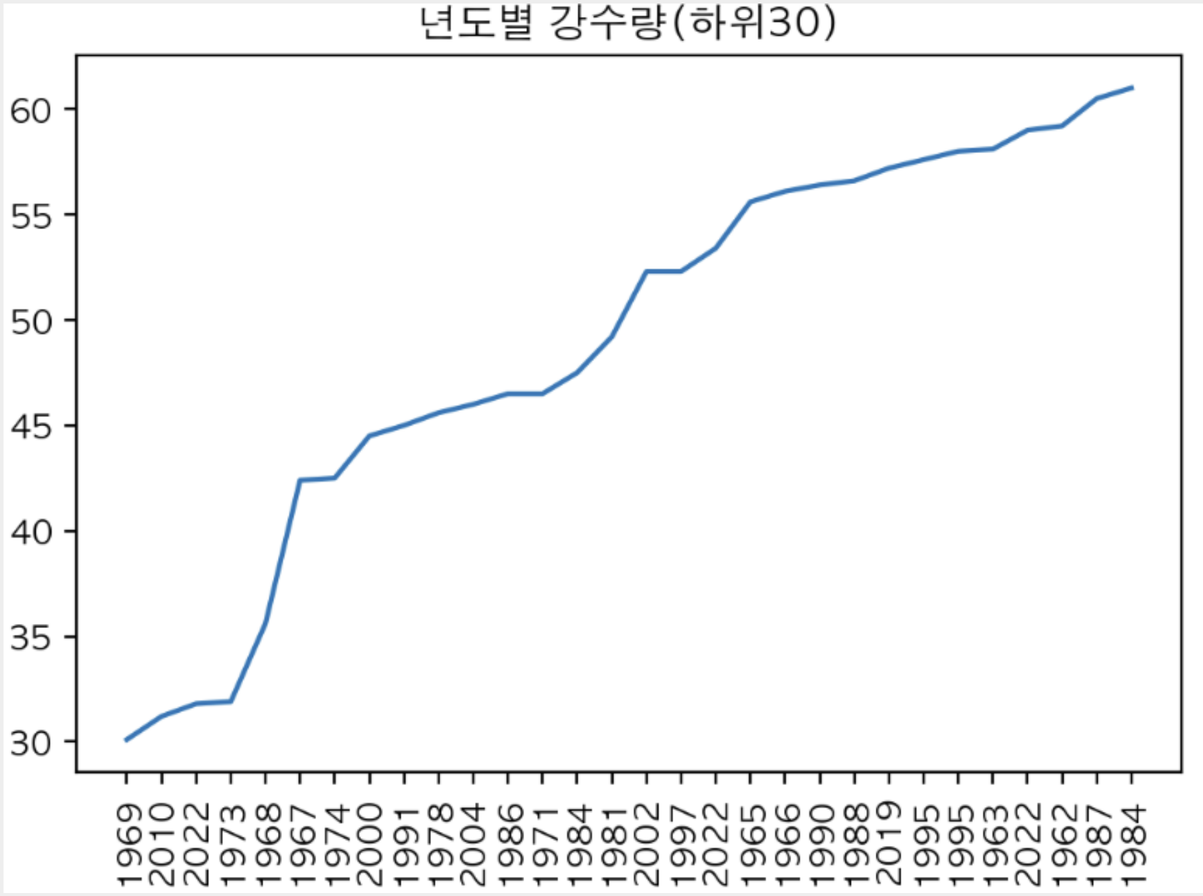


- 1960년 1월 ~ 2022년 8월 강수량 평균 데이터중 가장 많은 강수량을 가진 지역은 홍천이고 가장 적은곳은 청송이다.

04. 데이터 시각화

II. 년도별 강수량(상위 30 and 하위 30)



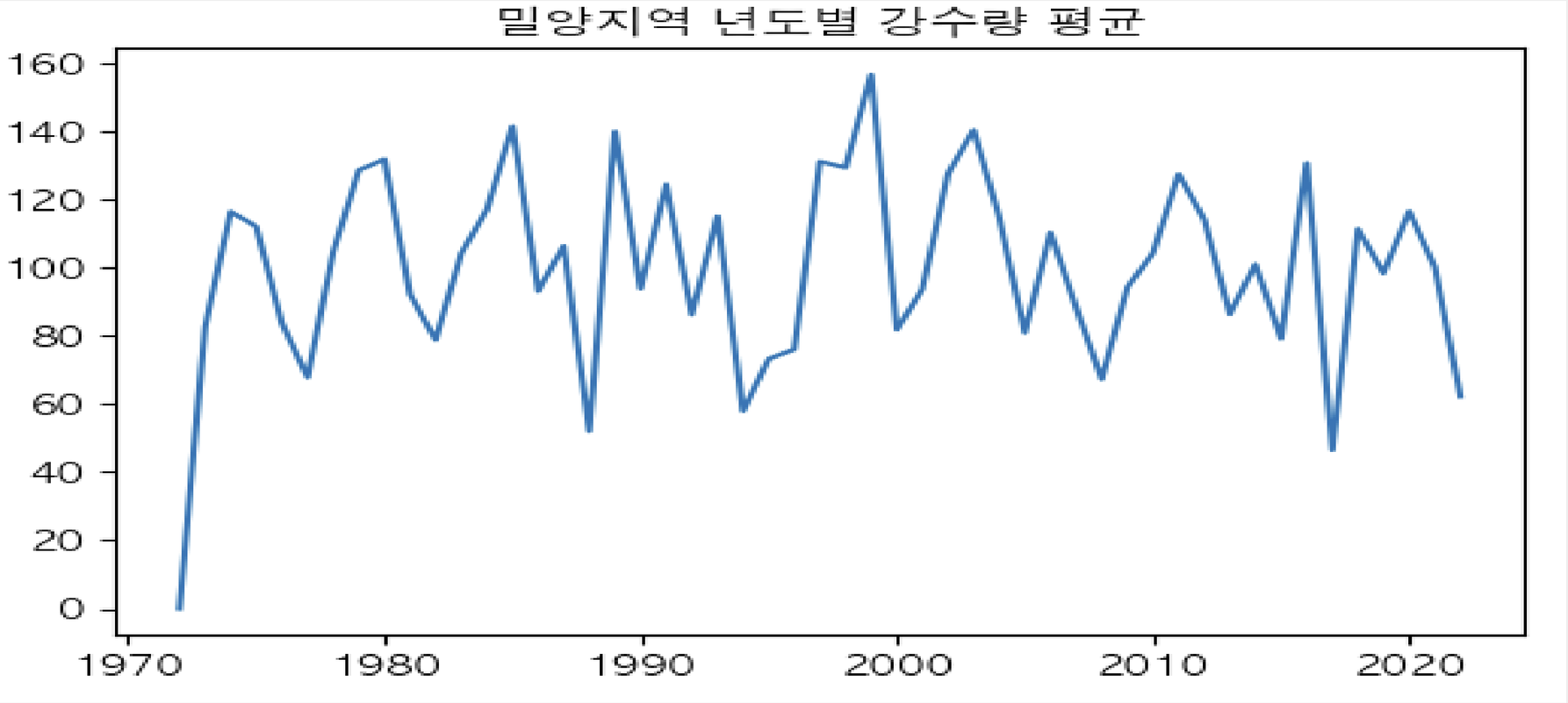


- 1960년 1월 ~ 2022년 8월 강수량 평균 데이터중 가장 많은 강수량을 가진 년도는 2006년이고 가장 적은것은 1969년이다.

04. 데이터 시각화

III. 입력한 지역의 년도별 강수량 그래프

지역 입력(시군구): 밀양

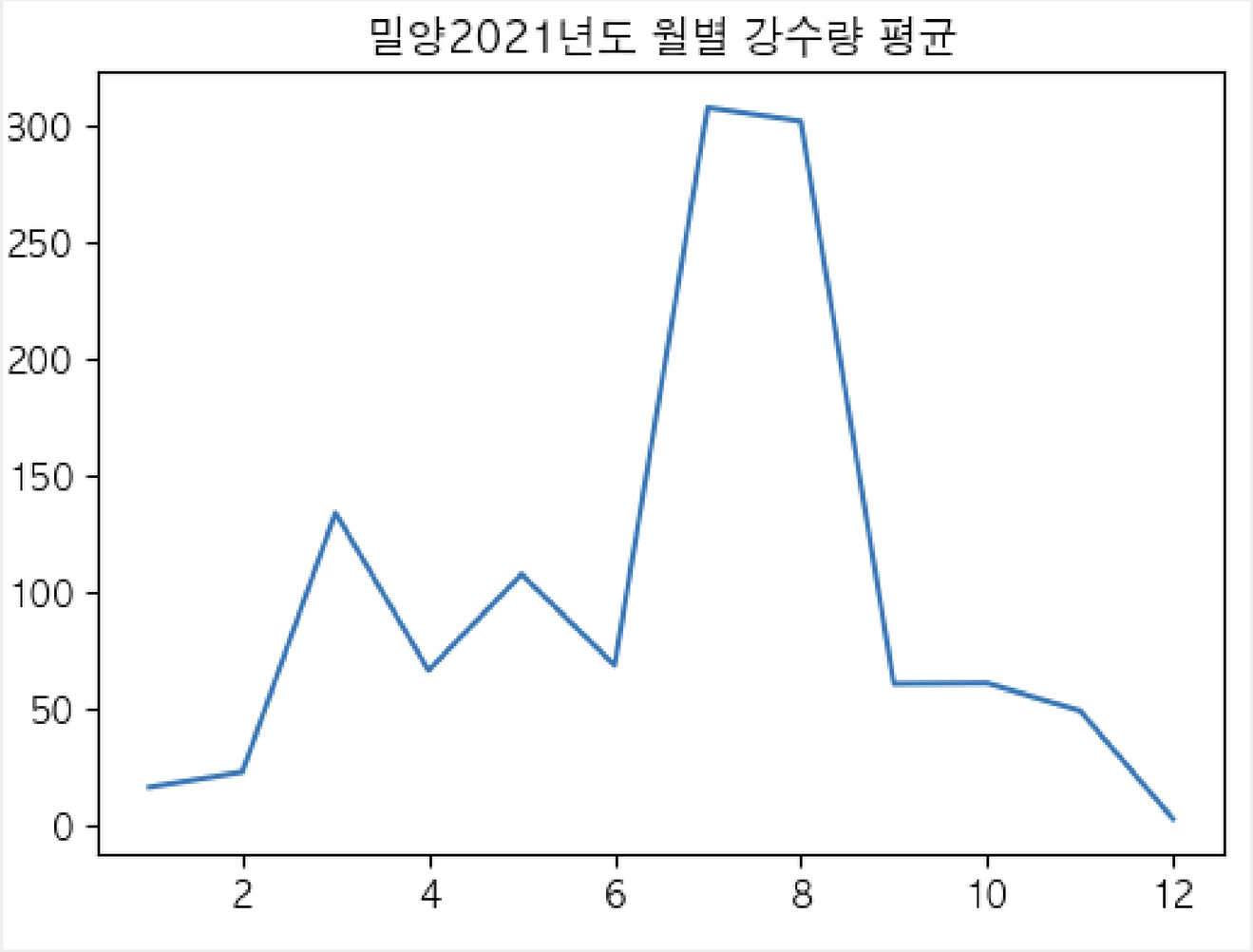


04. 데이터 시각화

IV. 입력한 지역, 년도의 월별 그래프

```
지역 입력(시군구): 밀양
날짜 입력 : 2021
```

	지점번호	지점명	일시	강수량(mm)	일최다강수량(mm)	일최다강수량일자	1시간최다강수량(mm)	1시간최다강수량일자	Month	Season
47559	288	밀양	2021-01-01	16.6	14.3	2021-01-26	0.0	NaT	1	4
47560	288	밀양	2021-02-01	23.1	14.7	2021-02-01	0.0	NaT	2	4
47561	288	밀양	2021-03-01	134.3	46.6	2021-03-01	0.0	NaT	3	1
47562	288	밀양	2021-04-01	66.7	28.2	2021-04-03	10.6	2021-04-04	4	1
47563	288	밀양	2021-05-01	107.9	28.5	2021-05-16	17.5	2021-05-28	5	1
47564	288	밀양	2021-06-01	68.9	24.3	2021-06-11	11.8	2021-06-23	6	2
47565	288	밀양	2021-07-01	308.2	87.9	2021-07-06	23.0	2021-07-07	7	2
47566	288	밀양	2021-08-01	302.3	121.6	2021-08-21	32.0	2021-08-21	8	2
47567	288	밀양	2021-09-01	60.9	19.1	2021-09-29	9.4	2021-09-29	9	3
47568	288	밀양	2021-10-01	61.2	40.7	2021-10-11	7.3	2021-10-11	10	3
47569	288	밀양	2021-11-01	49.4	29.0	2021-11-30	0.0	NaT	11	3
47570	288	밀양	2021-12-01	3.0	3.0	2021-12-16	0.0	NaT	12	4

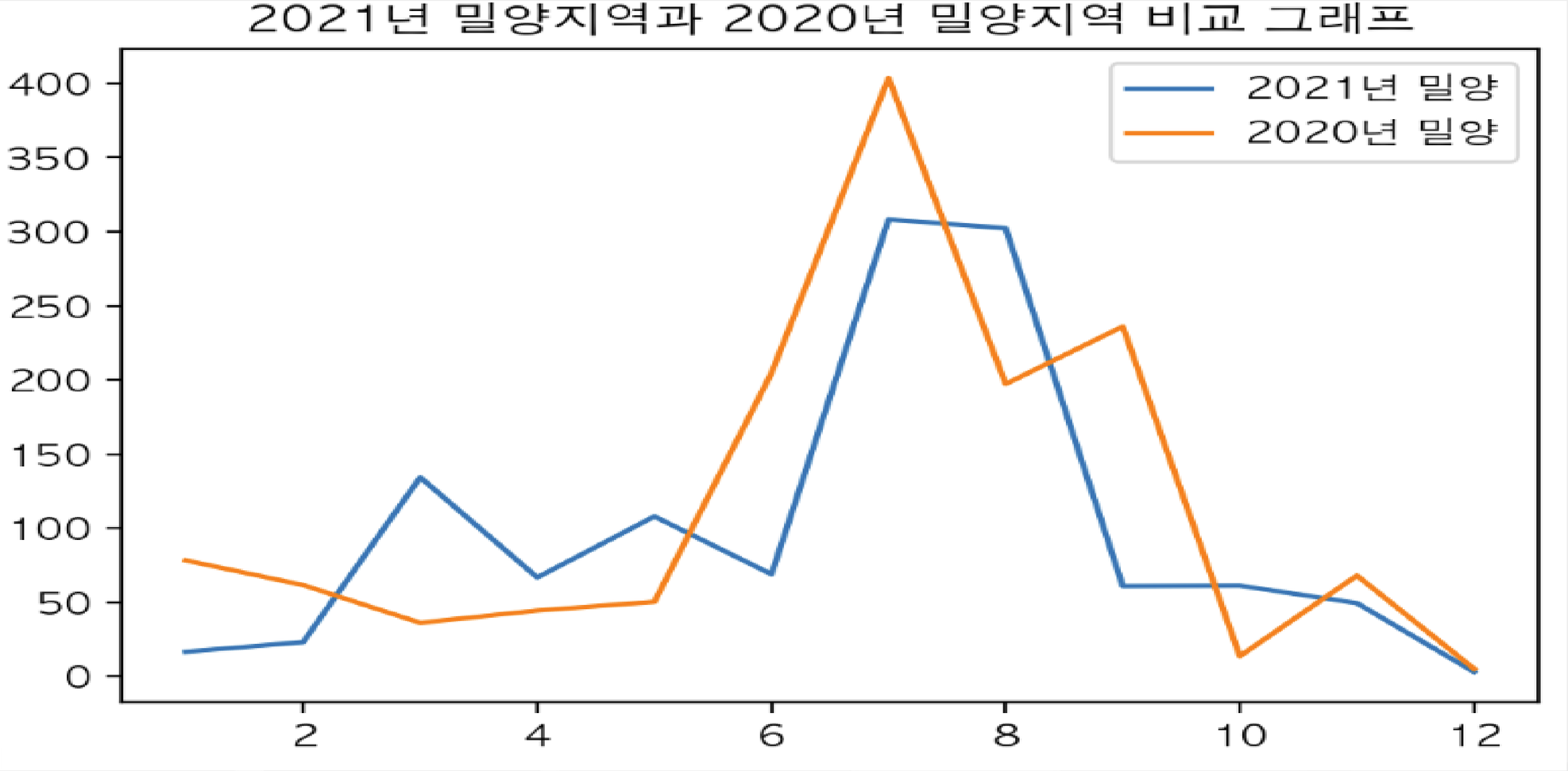


04. 데이터 시각화

V. 두개의 지역, 년도의 비교 그래프

```
지역 입력(시군구): 밀양
날짜입력(년도) : 2021
```

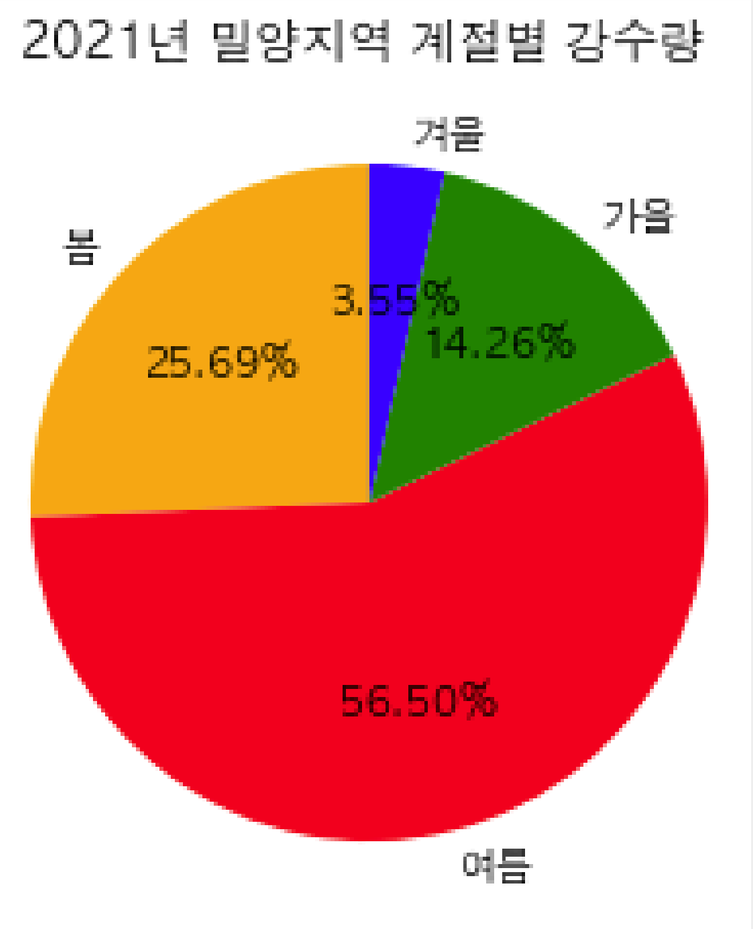
```
지역 입력(시군구): 밀양
비교할 날짜입력(년도) : 2020
```



04. 데이터 시각화

VII. 입력한 지역, 년도의 계절별 강수량 평균 파이차트

```
지역 입력(시군구) : 밀양
날짜 입력 : 2021
```



감사합니다.