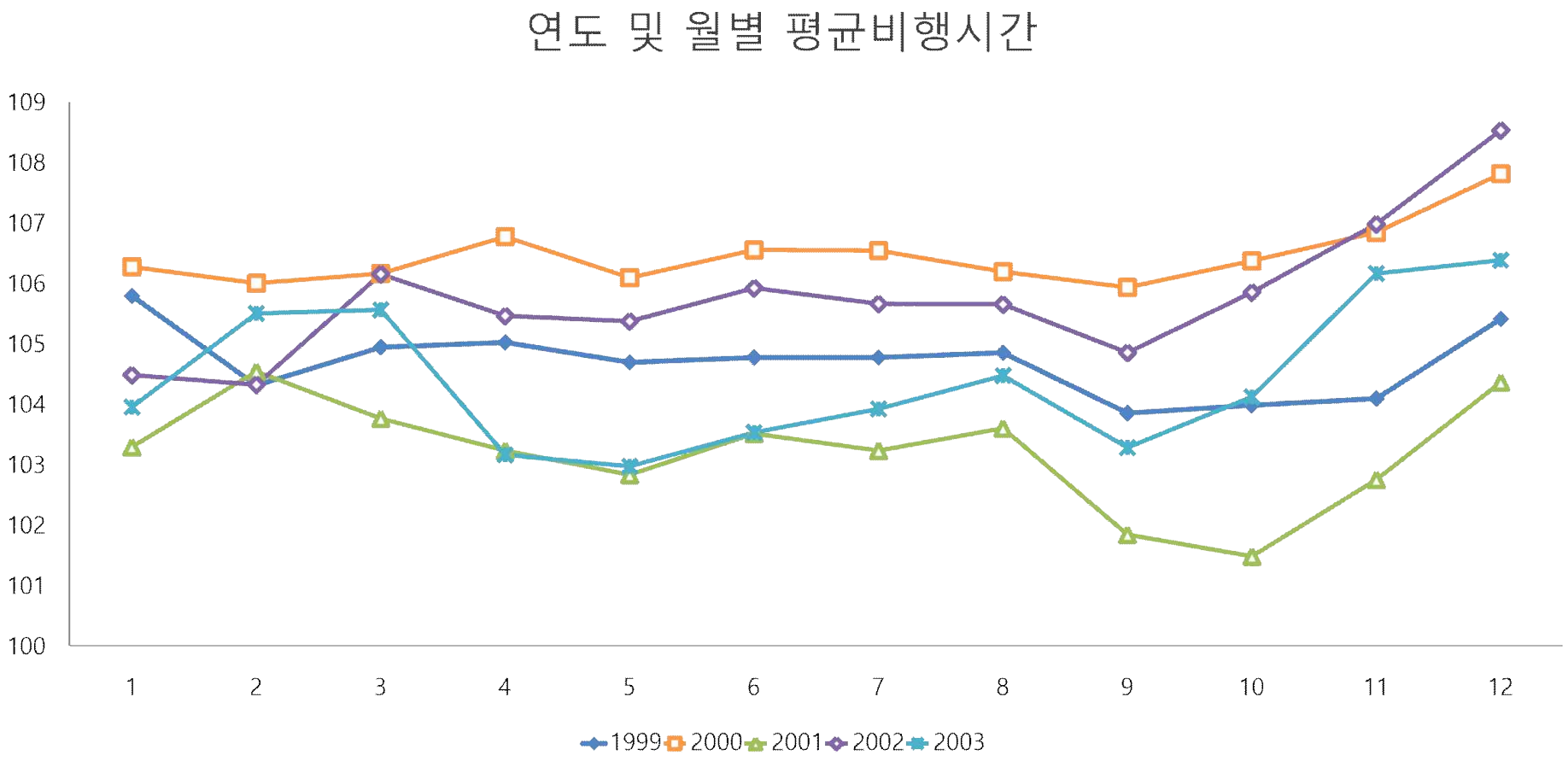
HIVE 예제문제

Airdelay Table

김영진

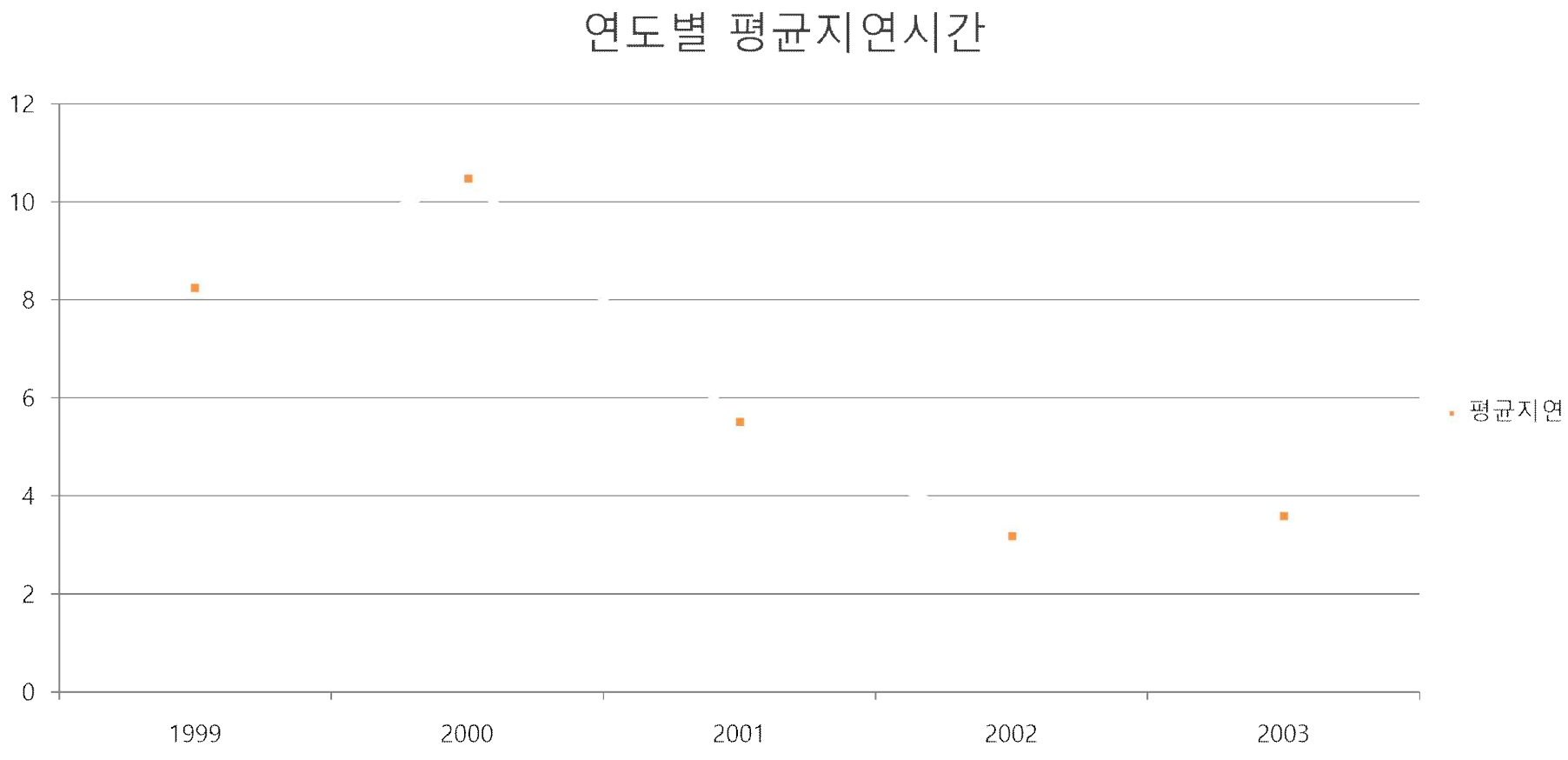
Q13. 5년간의 매월 비행시간 패턴을 구성시키시오.




```
insert overwrite directory '/user/hive/lwk'
  > select year, month, round(avg(airtime),2) as avg_airtime
  > from airdelay
  > group by year, month
  > order by year, month;
```

매년 8~9월에 비행시간이 감소하고
이후 매년 12월에 가장 많은 비행을 하는 경향

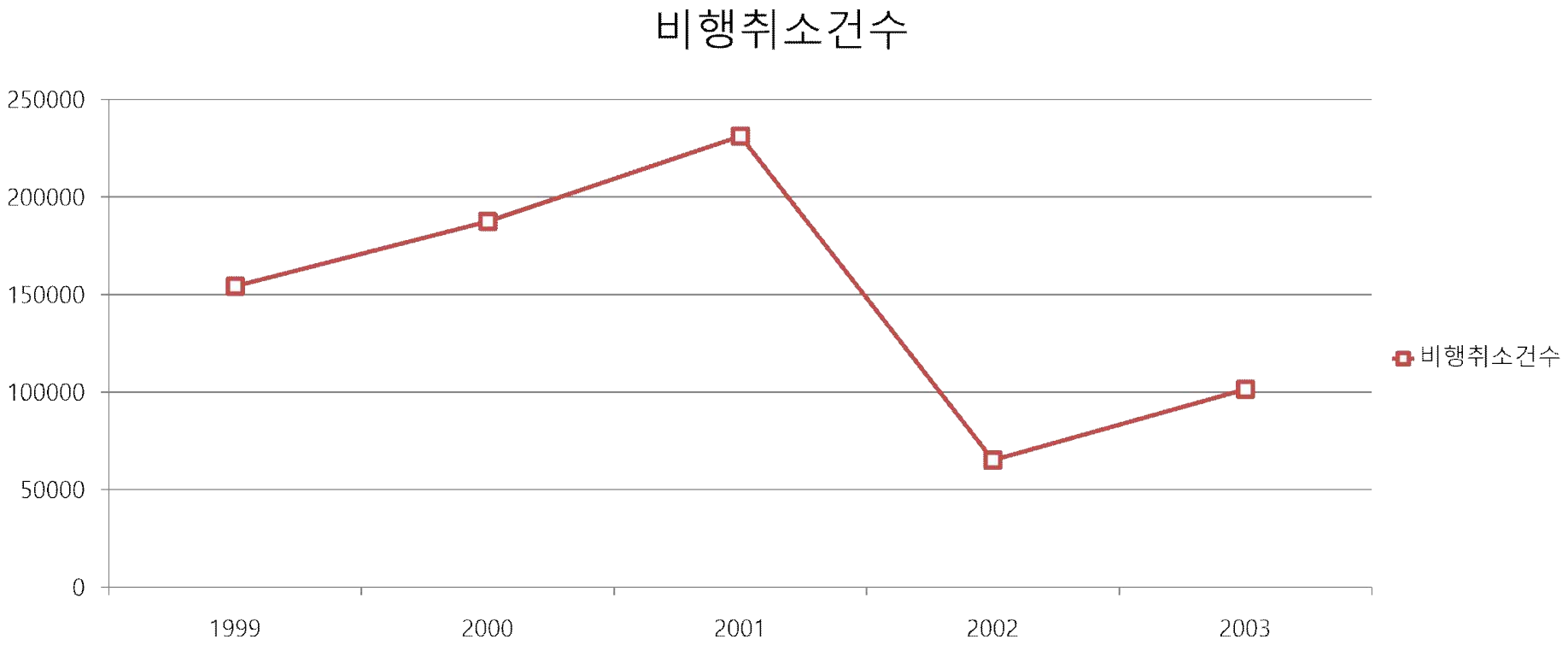
Q14. 5년간 연도별 지연 패턴을 보여라




```
insert overwrite directory '/user/hive/lwk'
  > select year, round(avg(arrdelay),2) as avg_arrdelay
  > from airdelay
  > group by year ;
```

2000년도에 평균지연시간이 가장 높음
이후에는 감소하는 경향이 뚜렷

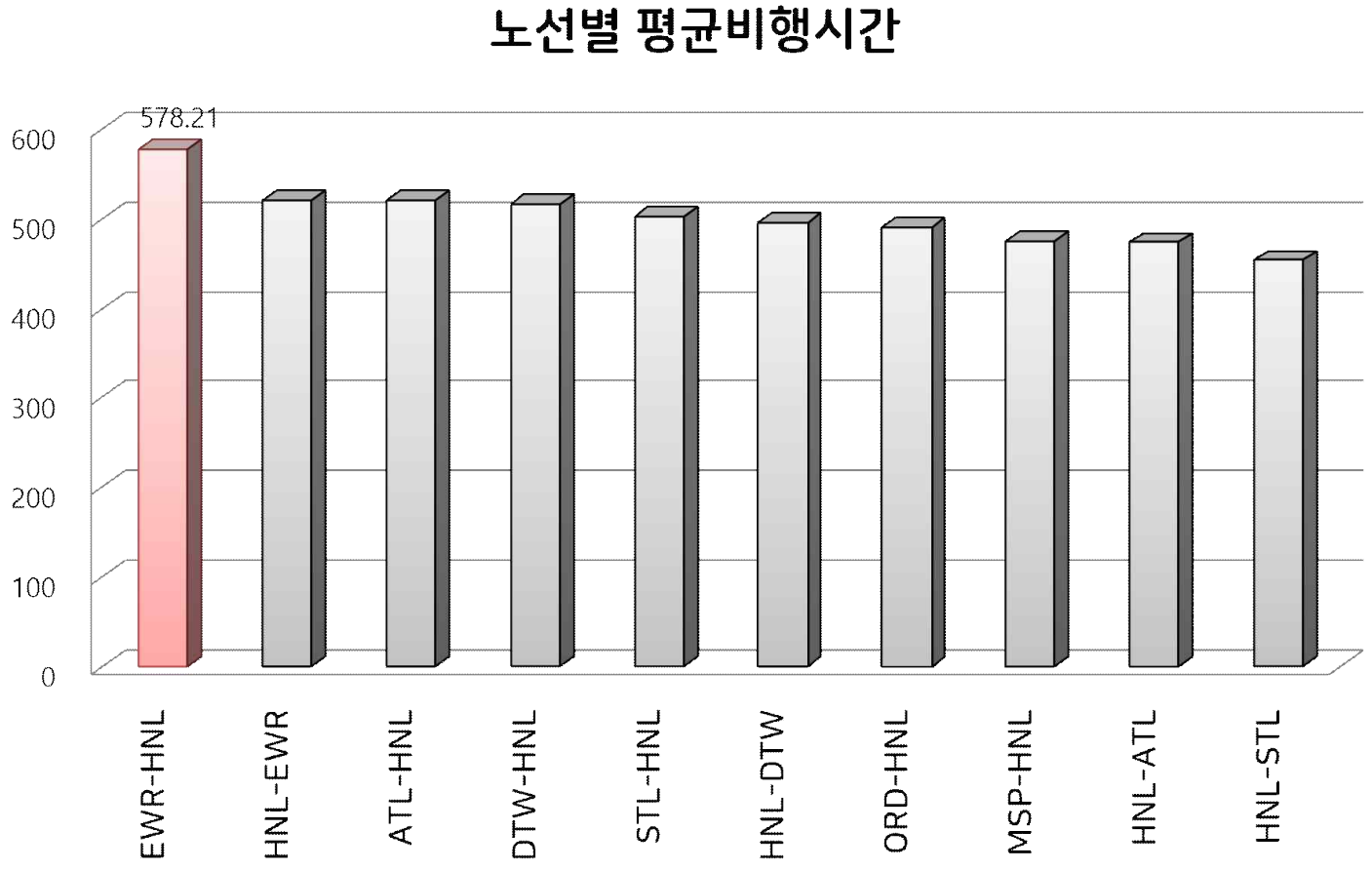
Q15. 연도별 비행취소 건수를 나타내라




```
insert overwrite directory '/user/hive/lwk'
  > select year,  sum(cancelled) as sum_cancel
  > from airdelay
  > group by year;
```

연도별 비행취소건수는 2001년까지 증가하다 이후 크게 감소

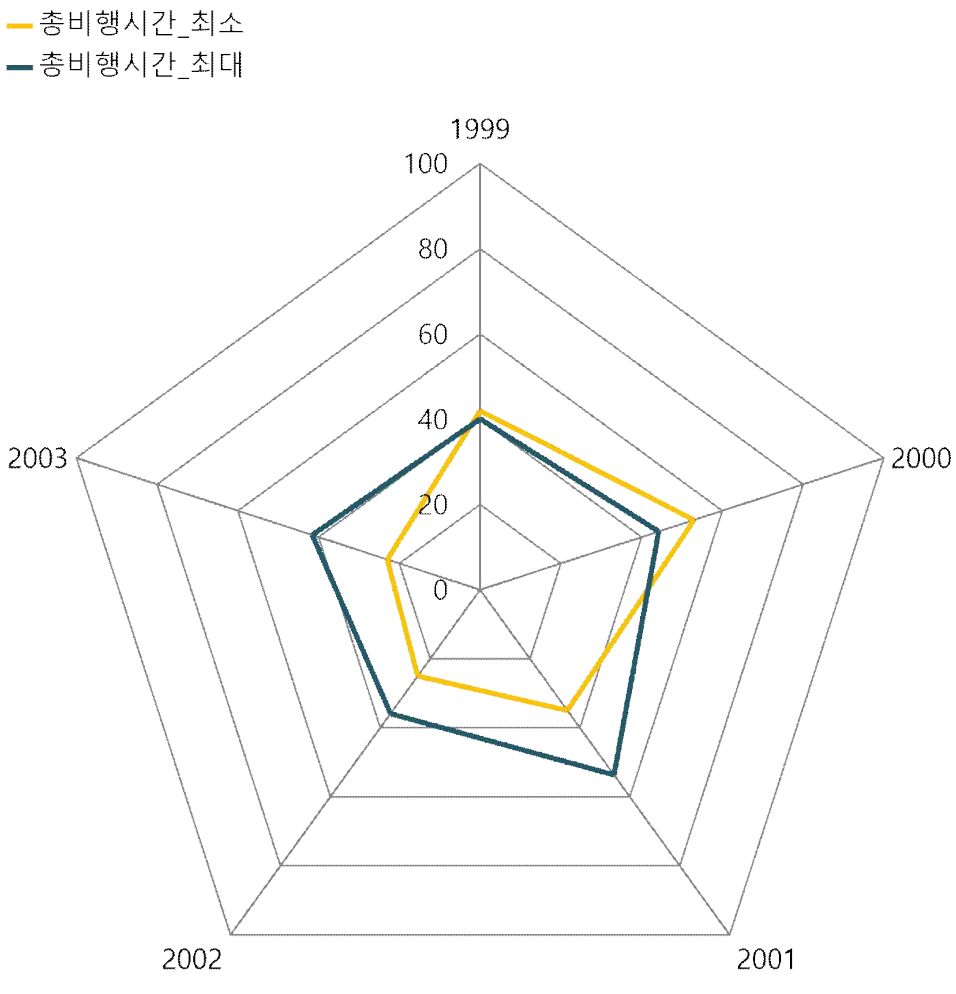
Q16. 5년간의 출발지-도착지의 평균비행시간이 큰 순서대로 나열하라




```
insert overwrite directory '/user/hive/lwk'
  > select year, origin, dest, avg(airtime) as avg_airtime
  > from airdelay
  > group by year, origin, dest
  > order by year, avg_airtime desc;
```

매년 8~9월에 비행시간이 감소하고
이후 12월에 가장 많은 비행을 하는 경향

Q17. 연도별 총비행시간이 최대인 비행편과 최소인 비행편을 나타내라



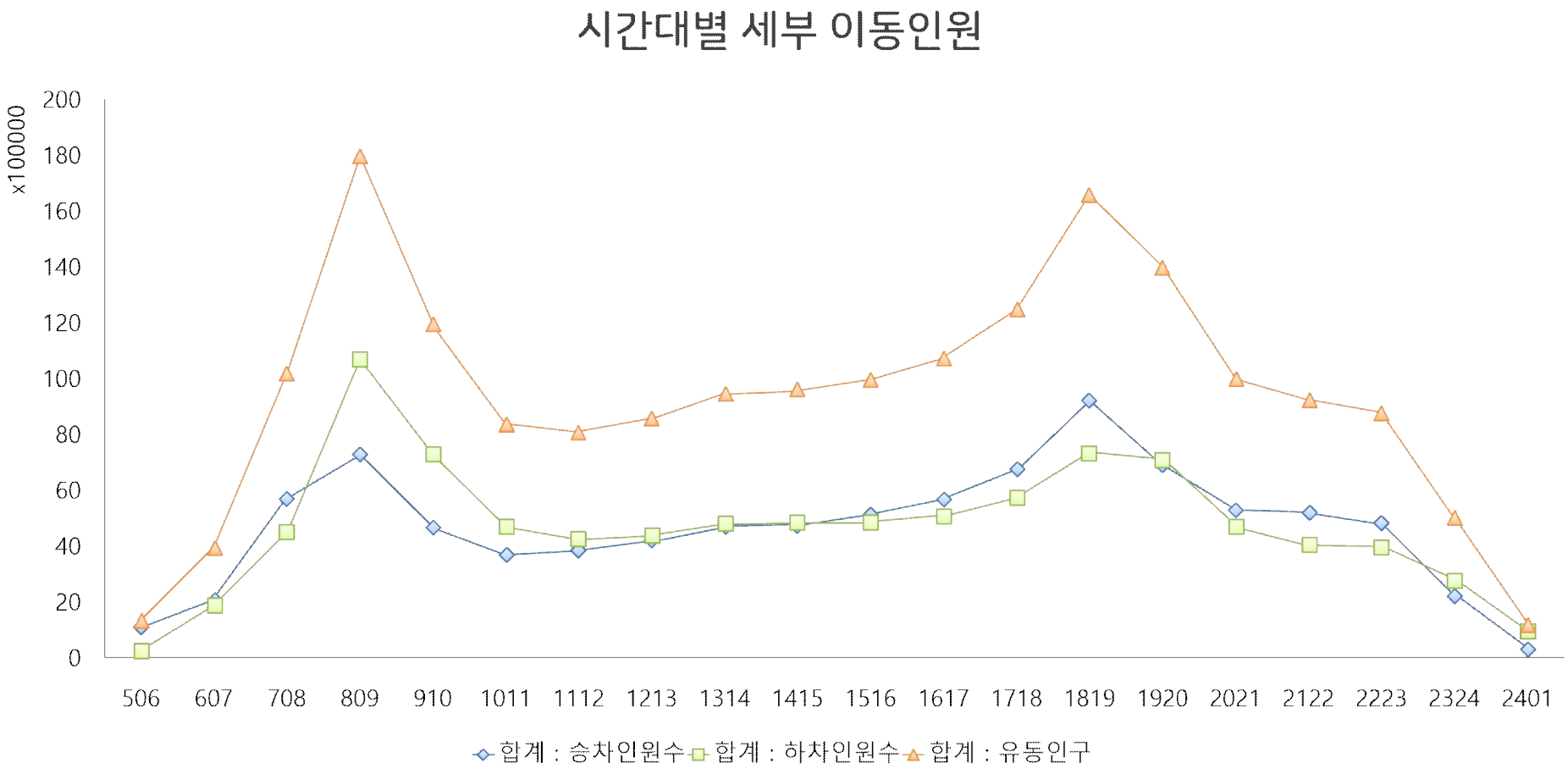
연도	비행기편	총비행시간_최대	배행기편	총비행시간_최소
1999	1	760250	2920	42
2000	1	776983	1325	53
2001	197	814649	2340, 5393	35
2002	73	743607	3963, 3961	25
2003	73	766332	6599	23

```
Insert overwrite directory '/user/hive/lwk'
 > select year, flightnum, sum(airtime) as sum_airtime
 > from airdelay
 > group by year, flughtnum
```

HIVE 예제문제

subway Table

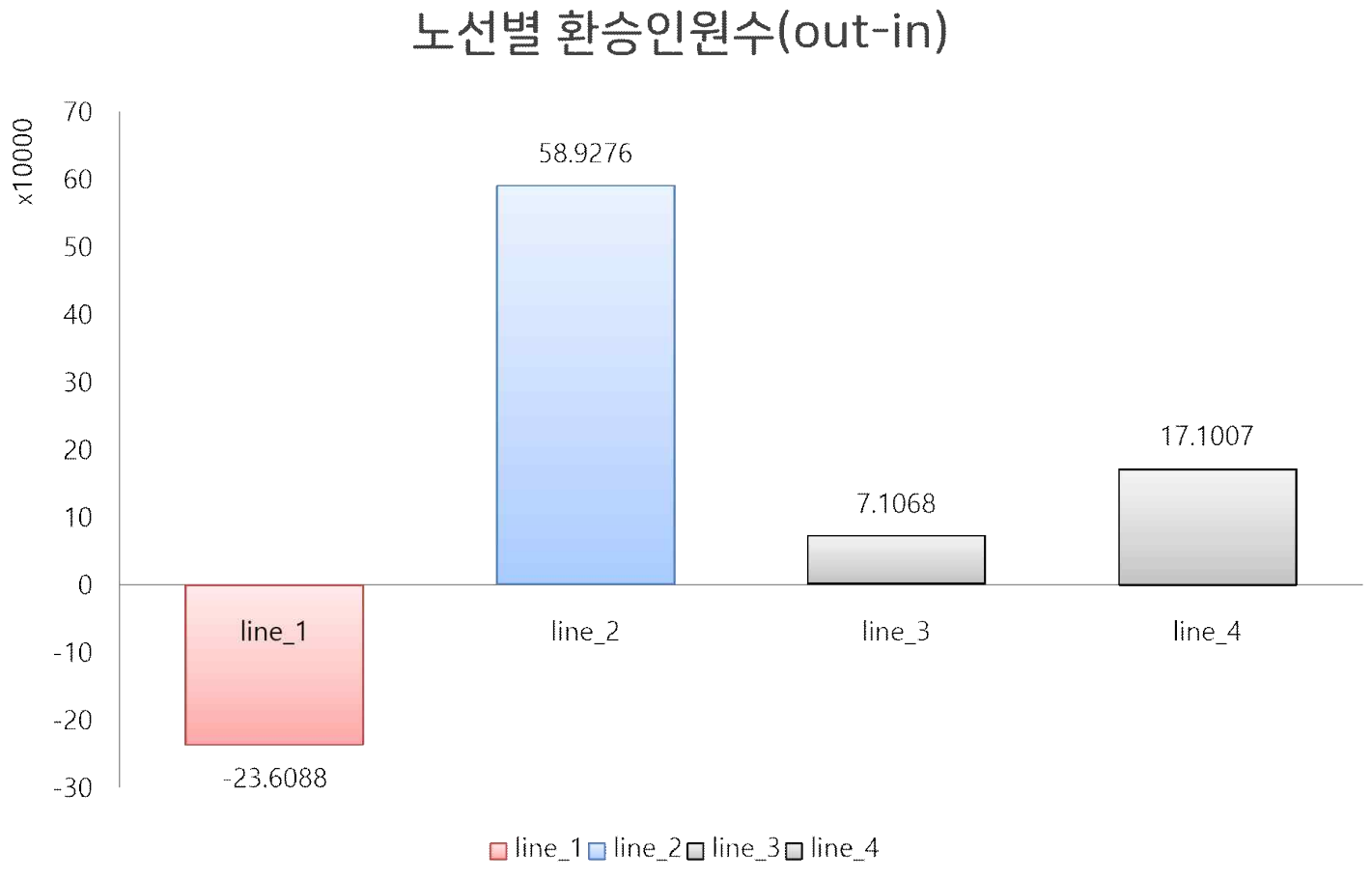
Q1. 시간대별 지하철 유동인구를 나타내라




```
insert overwrite directory '/user/hive/lwk'
  > select hour, sum(in_pass), sum(out_pass)
  > from sub_pass
  > group by hour
  > order by hour;
```

시간대별 지하철 유동인구는
출퇴근시간(0809, 1819))에 가장 크다

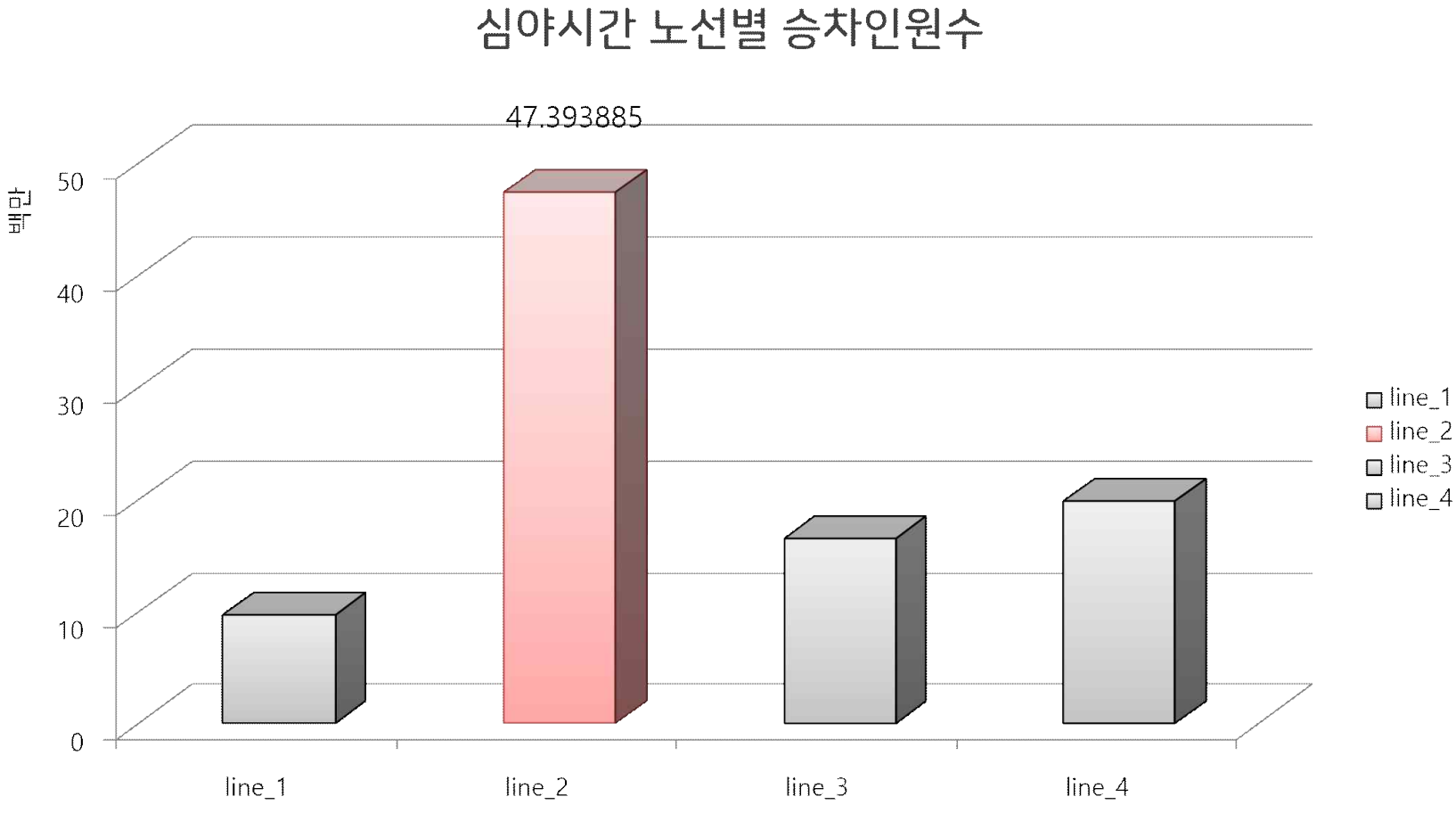
Q2. 환승인원이 가장 많은 호선은 어디인가?




```
insert overwrite directory '/user/hive/lwk'
  > select line_no, sum(out_pass - in_pass) as s_io
  > from sub_pass
  > group by line_no;
```

2~4호선 : 하차인원 > 승차인원
1호선 : 승차인원 > 하치인원

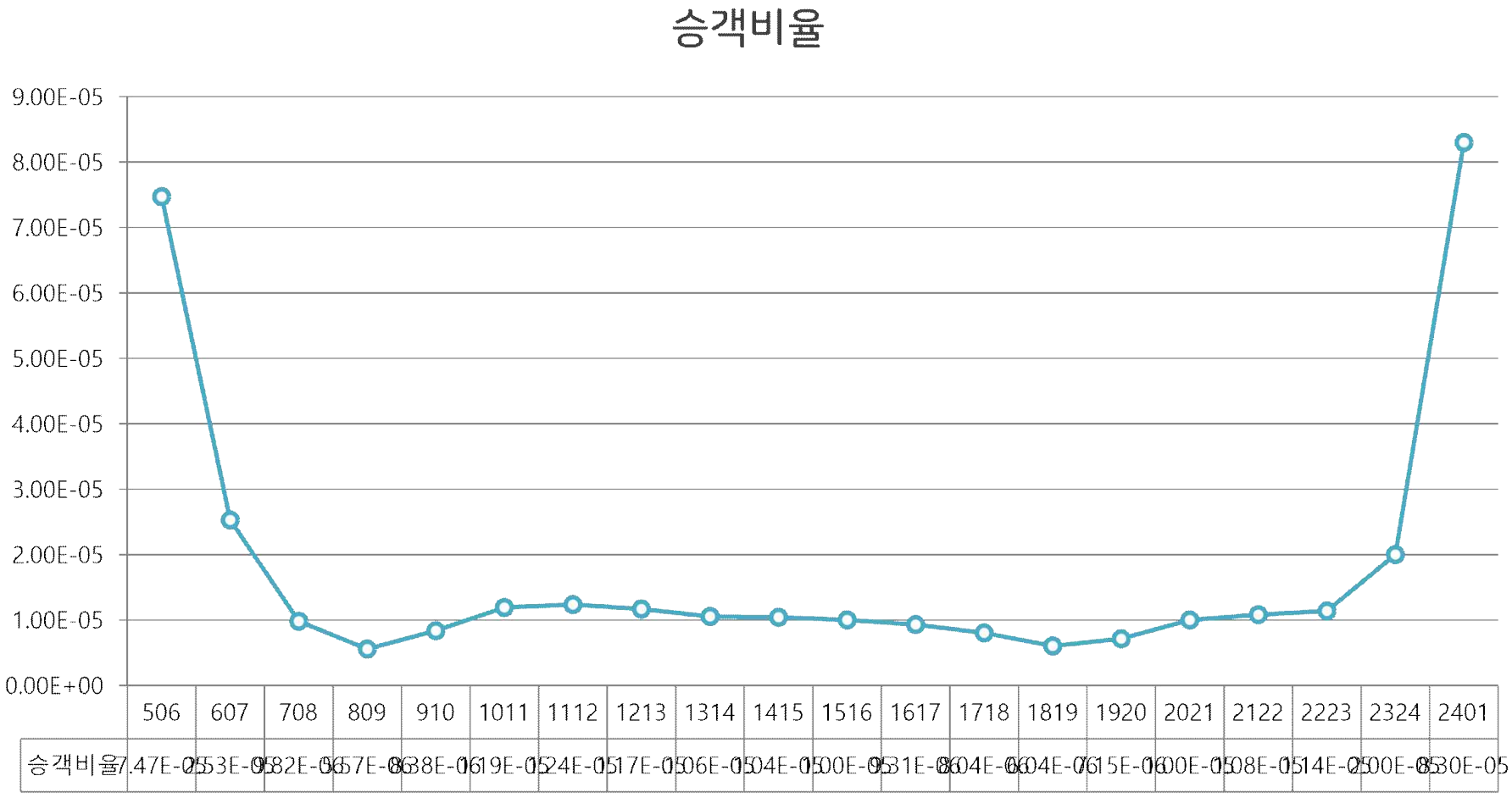
Q3. 심야(22~05)시간에 귀가객이 많은 호선은 무엇인가?




```
insert overwrite directory '/user/hive/lwk'
 > select line_no, sum(in_pass) s_i
 > from sub_pass
 > where hour>=22 or hour <=5
 > group by line_no;
```

심야시간에는 2호선의 승객이 가장 많다

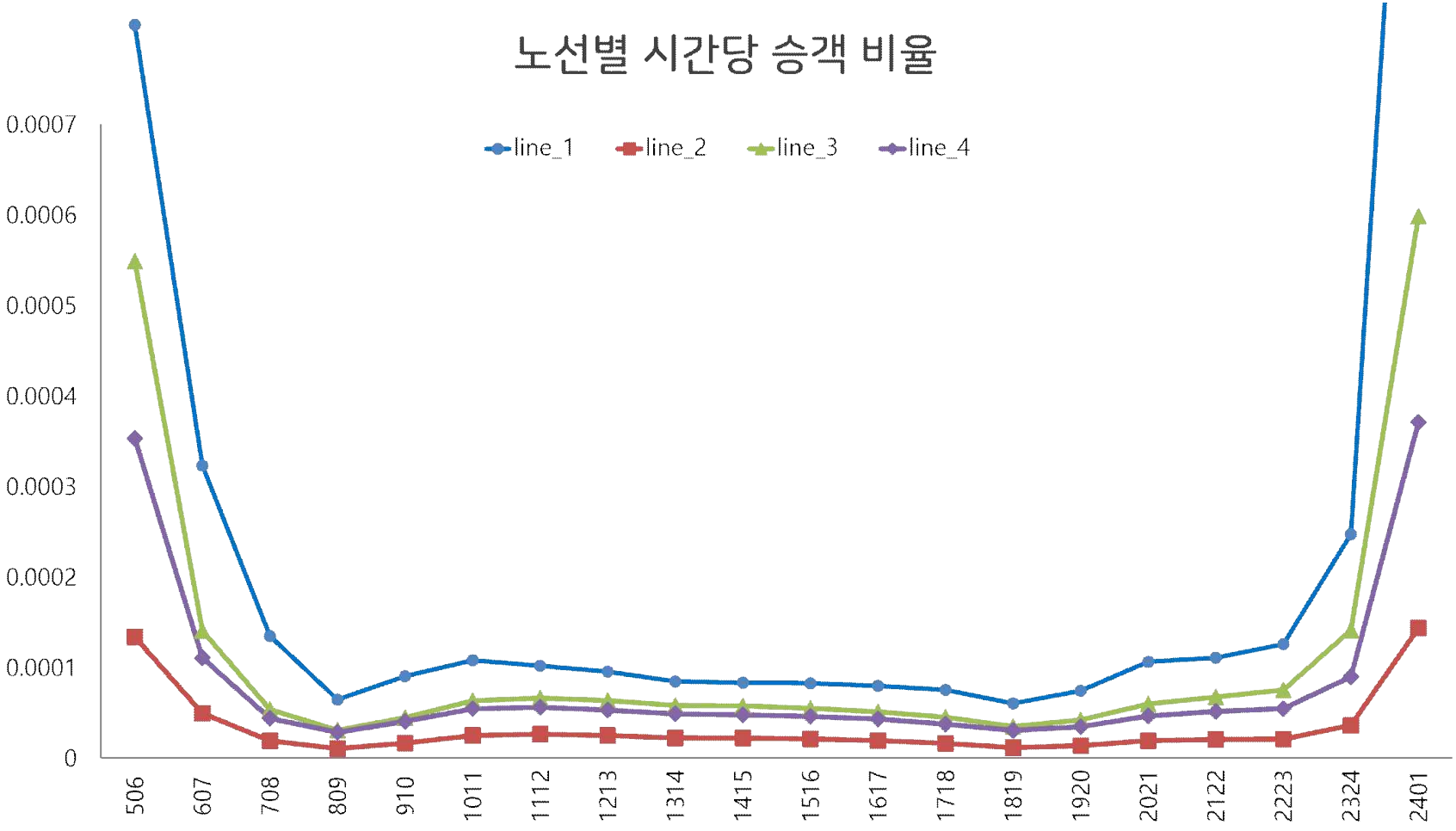
Q4. 시간당 승객비율을 나타내라




```
insert overwrite directory '/user/hive/lwk'
 > select hour, (100/sum(out_pass+in_pass)) s_oi
 > from sub_pass
 > group by hour
 > order by hour;
```

승객비율 : (100/승하차인원)
시간당 승객비율 또한 출퇴근시간과 유사한 경향

Q5. 노선별 시간당 승객비율을 나타내라



위와 마찬가지로 시간당 승객비율은
출퇴근시간과 유사한 패턴을 보이며,

1호선의 승객이 가장 적고
2호선의 승객이 가장 많음

```
insert overwrite directory'/USER/HIVE/LWK'
 > select line_no, hour, (100/sum(out_pass+In_pass)) s_oi
 > from sub_pass
 > group by line_no, hour
 > order by line_no, hour;
```