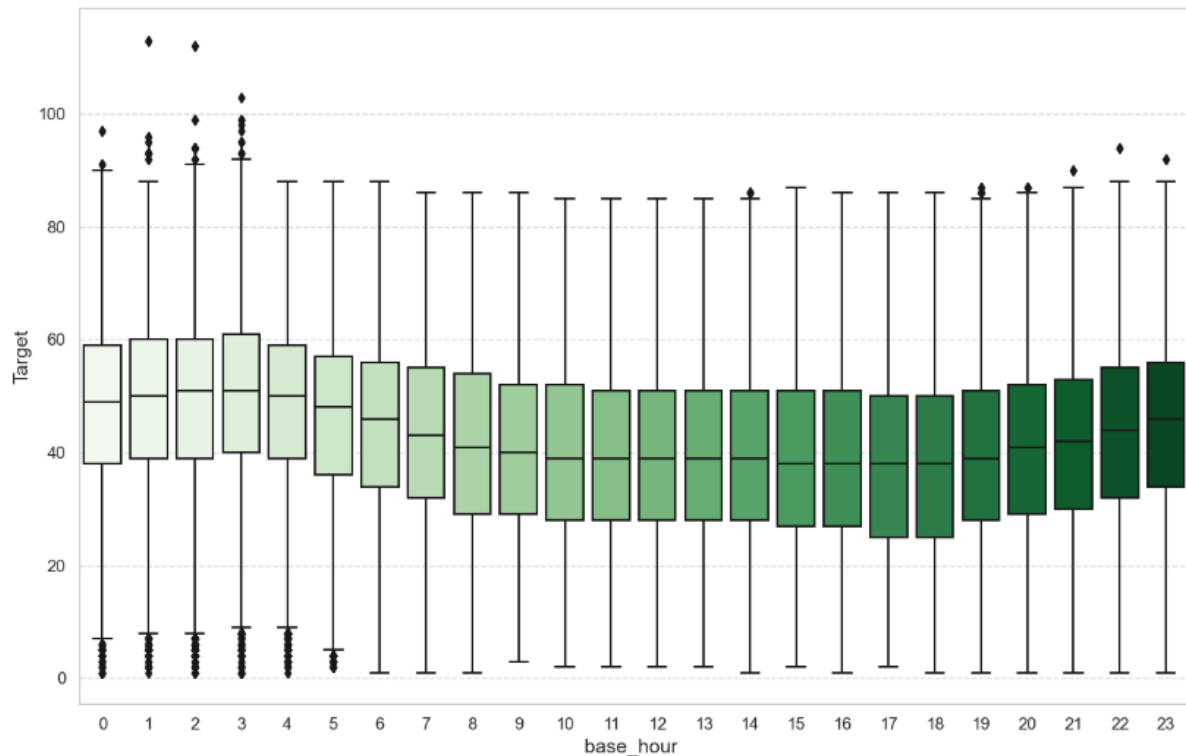


Q1) base_hour (시간)에 따른 교통량 분포를 알고 싶다. 시간에 따른 교통량 분포를 확인하기 위한 box plot을 그리고 시간에 따라 교통량이 차이가 있는지 본인 의견 작성



의견 – 중앙값을 보면 새벽시간인 00시부터 06시 보다 낮시간대인 9시부터 18까지 교통량이 활발하여 속도가 다소 줄어드는 것을 확인했다. 또한 새벽시간대에 유독 아웃라이어가 많이 나타나는 경향이 있는데 최대값을 넘은 outlier의 경우 그 도로상황이 차가 없고 빠르게 달려도 되겠다 싶은 환경 (예를 들어 가로등이 환하고 차선이 많고 시야가 잘 확보되는 직선도로)에서 나오는 값으로 추정되고, 최솟값보다 작은 outlier는 가로등이 열악하거나 도로가 구불구불하거나 밤이라서 안개 때문에 시야확보가 원활하지 않거나 차선이 좁고 커브가 많은 도로에서 나타나는 outlier로 추정했다. 나는 이 outlier가 무조건적으로 삭제해야되는 값이 아니라고 보고 시간에 영향을 받는 중요한 데이터라 판단하였다. 따라서 시간에 따라 교통량 차이는 현재 boxplot에서는 있다고 주장한다.

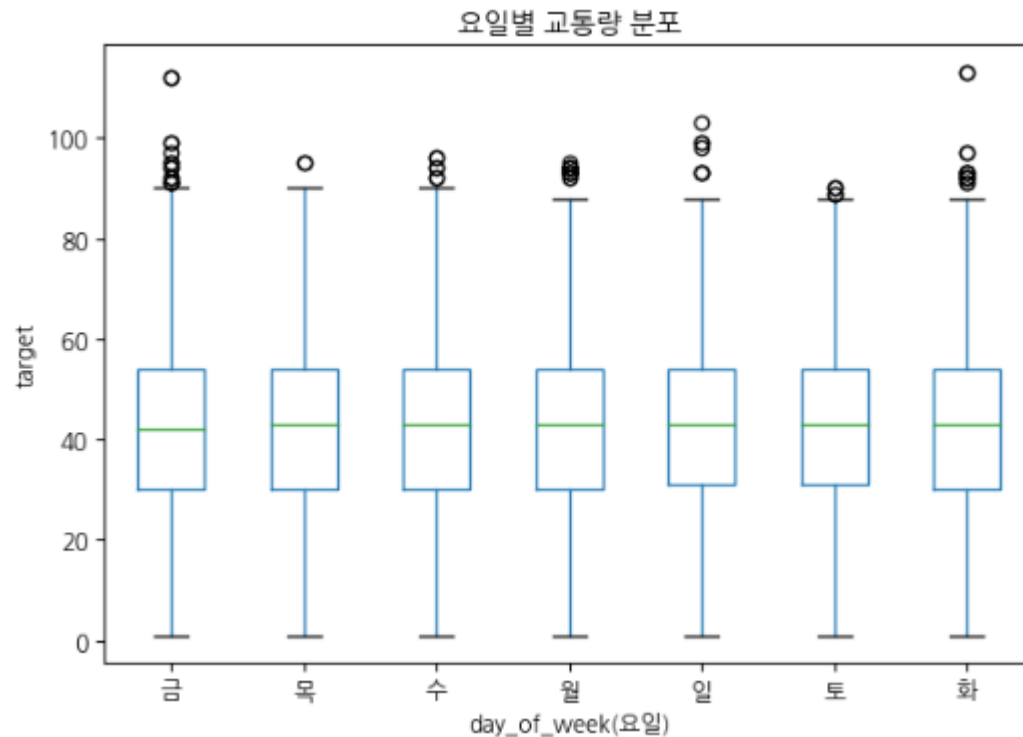
Q2) 요일별 교통량 분포를 box plot으로 그리고 요일별로 차이가 있는지 의견 작성

```
print(df['day_of_week'].value_counts())
```

executed in 447ms, finished 19:43:38 2024-11-19

```
금    684024
수    675583
목    674070
일    673632
토    669767
화    662498
월    661643
```

요일별 데이터는 비슷하다고 판단



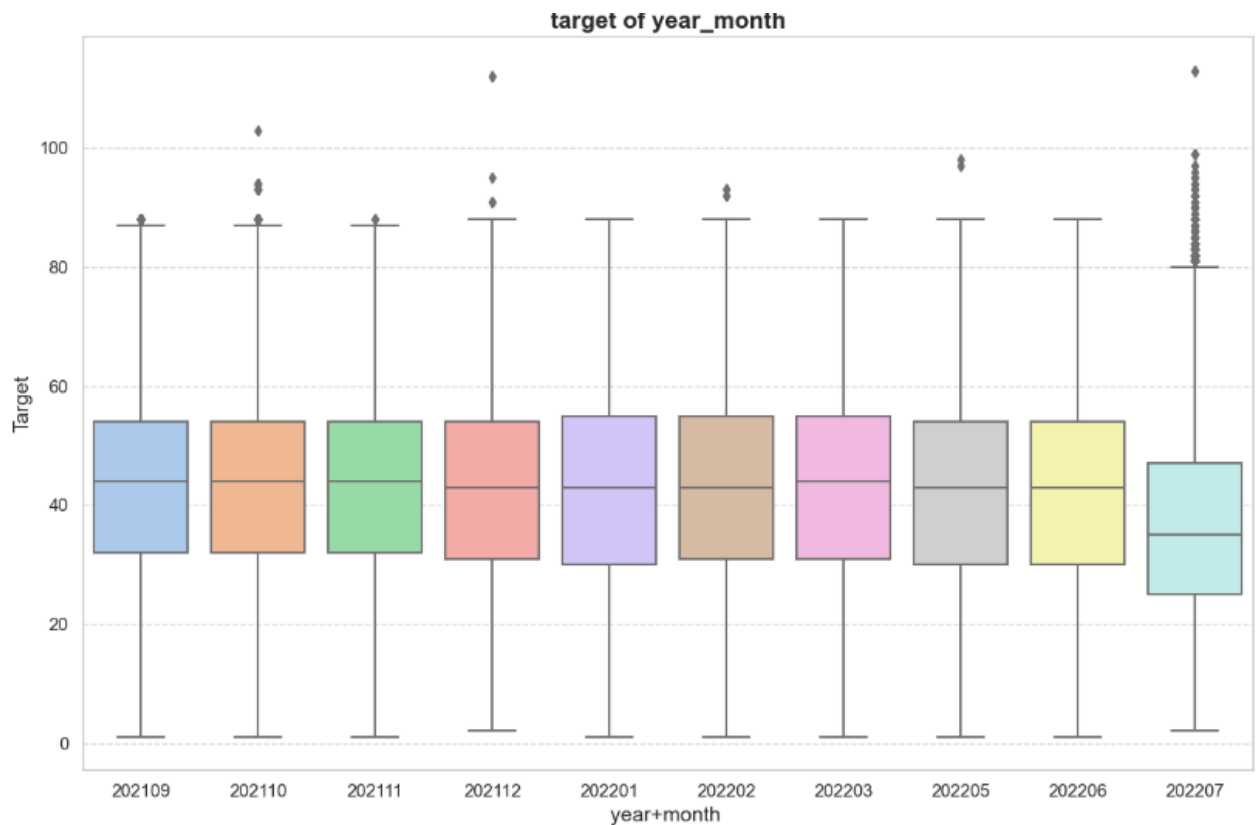
Boxplot을 보면 outlier가 있긴 하지만 중앙값을 비롯한 데이터의 분포가 비슷해보인다.

	Q1	Median (Q2)	Q3	
day_of_week				
월	30.0	43.0	54.0	실제로 큰 차이 없는 모습
화	30.0	43.0	54.0	
수	30.0	43.0	54.0	
목	30.0	43.0	54.0	
금	30.0	42.0	54.0	
토	31.0	43.0	54.0	
일	31.0	43.0	54.0	

주말에는 교통량이 많을 거라고 예상했지만 실제 데이터가 비슷하게 나왔다. 따라서 요일별 교통량의 차이는 없다고 판단했다.

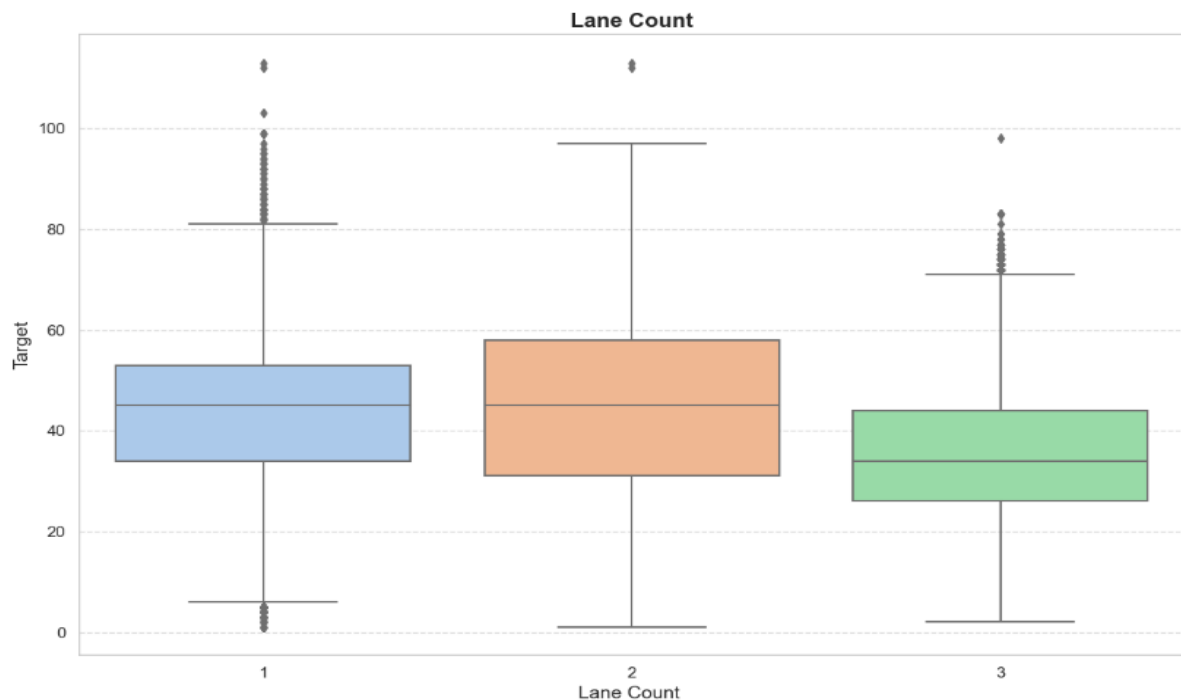
Q3) 연도+월별 교통량 분포를 box plot으로 그리고 시간이 흐름에 따라 교통량 차이가

있는지 의견 작성



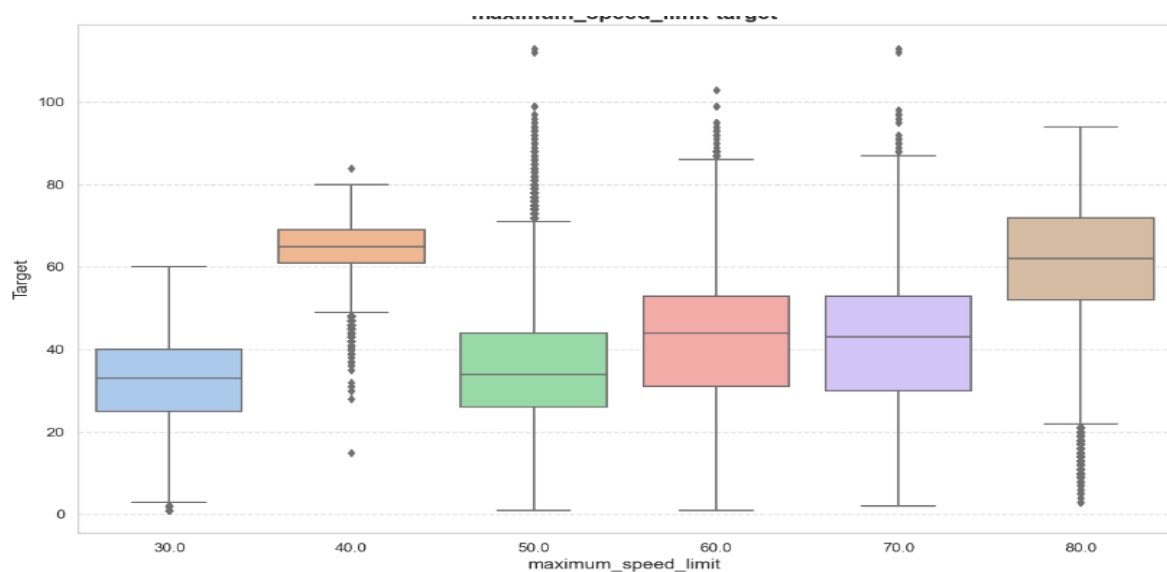
의견 - 2022년도 7월을 보면 다른 연도의 월과는 다르게 최댓값, 중앙값, q1, q2 등 전체적으로 outlier를 제외한 데이터들이 낮게 나온다. 이것은 7월은 여름휴가시즌이고 제주도는 우리나라의 대표적인 여름 여행지이기 때문에 여행객 증가로 인한 교통량 상승으로 보인다. 최대값을 넘는 아웃라이어도 다소 존재하는데 이는 원래 시즌이면 최대값안에 포함되는 속도지만 여행객들로 인해 여행명소 주변 도로는 교통량은 증가해 속도를 못내고 실제 도민들이 이용하거나 여행명소와 상관없는 도로들이 비교적 한적하여 속도가 높을때 나온 outlier라고 생각한다.

Q4) lane_count별 교통량 분포를 box plot으로 그리고 lane_count의 차이가 교통량 차이와 관련이 있을지 의견 작성



의견 - 2차선은 1차선, 3차선 보다 IQR의 범위가 넓고 최댓값도 더 크다 즉 교통량이 원할한 차선은 2차선으로 판단되고 3차선이 오히려 1차선보다 최댓값 중앙값 q1, q3 전부 낮은 경향을 보인다. 제주도의 1차선은 상대적으로 한적한 것에 비해 3차선은 도로는 넓지만 차량이 훨씬 더 많아 속도를 못내거나 신호교차로의 영향을 받는 등의 이유로 교통량이 늘어난 거라고 생각한다.

5. maximum speed limit 별 교통량 분포를 확인하면 상식과 위배되는 구간이 하나 나오는데 그게 어떤 구간인지 작성 - 왜 그럴까?



최고속도제한이 40인 구간을 보면 다른 데이터들과 다르게 Q1부터 Q3의 값이 모두 40을 훨씬 넘는 60이상에 위치해있다. 이는 대부분이 속도위반을 하는 것이라 상식과 위배된다.

```
filtered_df = df[df['maximum_speed_limit'] == 40]

# 도로명별 평균 차량 속도 계산
result = filtered_df.groupby('road_name')['target'].mean().reset_index()

# 컬럼 이름 변경 (보기 편하게)
result.rename(columns={'vehicle_speed': 'average_vehicle_speed'}, inplace=True)

# 결과 출력
print(result)
```

executed in 90ms, finished 09:52:58 2024-11-20

	road_name	target
0	일반국도12호선	66.903359
1	일반국도95호선	47.712526

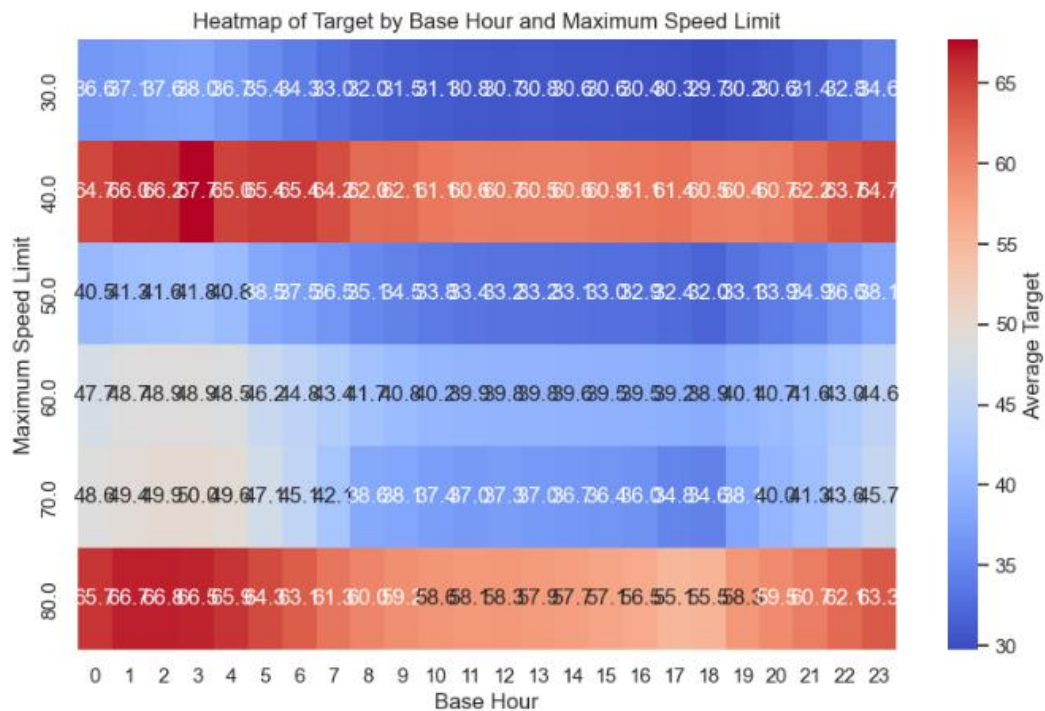
내가 생각하는 이유 - 최고속도 40km인 도로중 일반국도12호선에서 평균속도가 66.9km로 높게 나타나는데 이 국도는 제주특별자치도 제주시에서 해안선을 따라 서귀포를 지나 다시 제주시에 이르기까지 제주도를 완전히 일주하는 도로이다. 속도를 즐기며 제주도 바다를 볼 수 있는 바다이고 실제로 빠르게 주행하는 것으로 확인했다.

Q6) box plot으로 분포 확인이 불가능한 변수가 있나? 있나/없나 에 대한 이유를 작성

Box Plot을 통해 분포를 확인 하려면 x에 해당하는 변수가 범주형데이터형이어야 된다고 생각한다. 예를 들어 id와 같은 연속형 변수일경우 고유한 id의 값 하나당 target의 값이 1나 plot에 표현될 거고 id가 200개만 넘어가도 제대로 된 확인이 불가능할 분더러 id에 따른 target을 해석한다고 해서 얻어지는 것은 없다. 이와 같은 이유로 base_date도 마찬가지다.

Q7) Q4에서 가능한 변수들 중 시각화를 통해 교통량과 상관이 제일 큰 것 같은 두 변수를 선택하고 두 변수에 따른 교통 량 분포를 알기 위한 heatmap 그리고 어떤 특성이 있는지 작성 (추가 점수)

내가 고른 변수 - 시간과 최고속도제한이다. 이유는 plot을 눈으로 봤을 때 가장 차이가 두드러지게 나왔다고 생각했다.

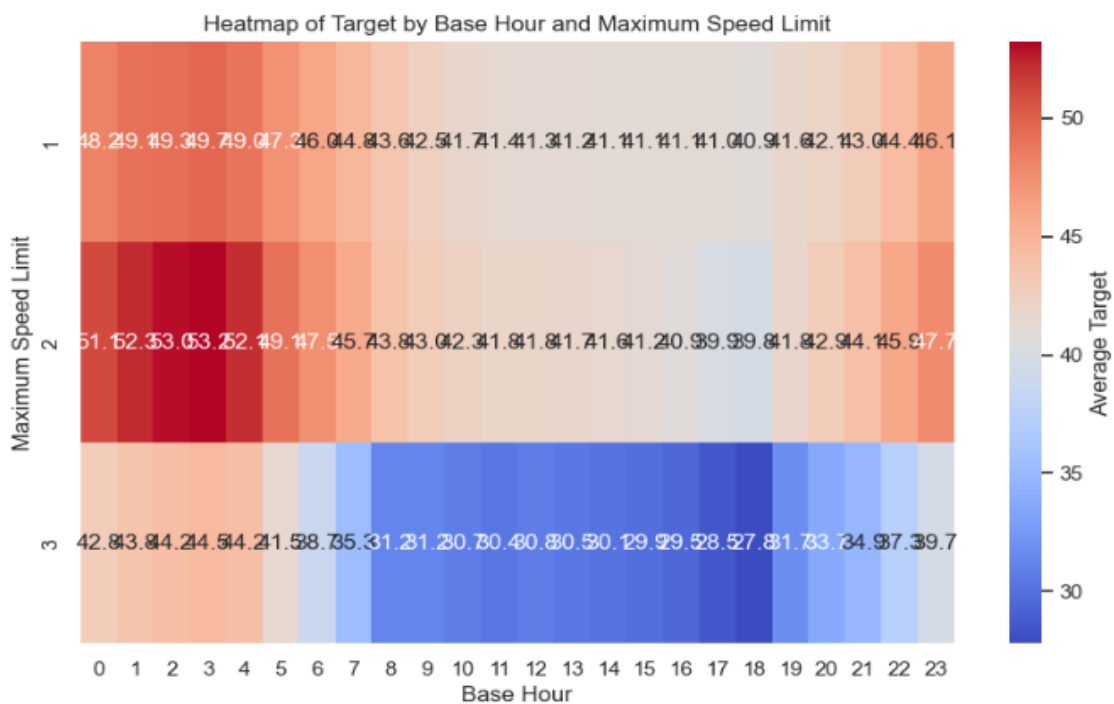


첫번째 특성 - 최고 속도 제한이 40km,80km인 도로는 모든 시간대에서 속도가 높은 즉, 교통량이 낮은 것으로 나왔다.

두번째 특성 - 최고 속도 제한이 60km와 70km도로는 는 새벽시간 00시부터 6시까지 교통량이 낮아지는 경향이 있다.(40,80을 제외한 나머지 4개 도로중 유독 색깔이 다름)

세번째 특성 - 모든 도로는 사람들이 많이 활동하는 낮 시간대에는 target이 낮아진다.

++ 추가 고른 변수 - 차선과 시간



첫 번째 특성 - 0시부터 5시까지 2차선에서 속도가 높아지는 , 교통량이 낮아지는 경향이 있다.

두 번째 특성 - 3차선은 사람들이 활동하는 낮시간대에 비교적 교통량이 높아지는 경향이 있다.

gpt레퍼런스 - plot를 더 시각적으로 그리는 방법을 물어봄



```
# Lane count별 교통량 분포를 box plot으로 시각화
plt.figure(figsize=(10, 6))
df.boxplot(column='target', by='lane_count', grid=False)
plt.title('lane_count별 교통량 분포')
plt.suptitle("") # Remove default subtitle
plt.xlabel('lane_count(차선)')
plt.ylabel('target')

plt.tight_layout()
plt.show() 위 코드도 좋은데 다른 느낌으로 옵션 더 추가해서 바꿔볼 수 있어?
```



아래는 기존 코드에 다양한 옵션을 추가하여 조금 더 시각적으로 향상된 방식으로 lane_count 별 교통량 분포를 box plot으로 그린 예제입니다:

python

코드 복사

```
import matplotlib.pyplot as plt
import seaborn as sns

# Seaborn 스타일 적용
sns.set(style="whitegrid")

# Box plot 그리기
plt.figure(figsize=(12, 8))
sns.boxplot(data=df, x='lane_count', y='target', palette='pastel')

# 타이틀과 레이블 추가
plt.title('Lane Count별 교통량 분포', fontsize=16, fontweight='bold')
plt.xlabel('Lane Count (차선 수)', fontsize=14)
plt.ylabel('Target (교통량)', fontsize=14)

# 눈금과 레이블 스타일 조정
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

# 격자 추가
plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.tight_layout()
plt.show()
```

