

05

실습예제
(생년월일(YYMMDD) 데이터 분석)

ChatGPT 고급 데이터 분석 실습 예제

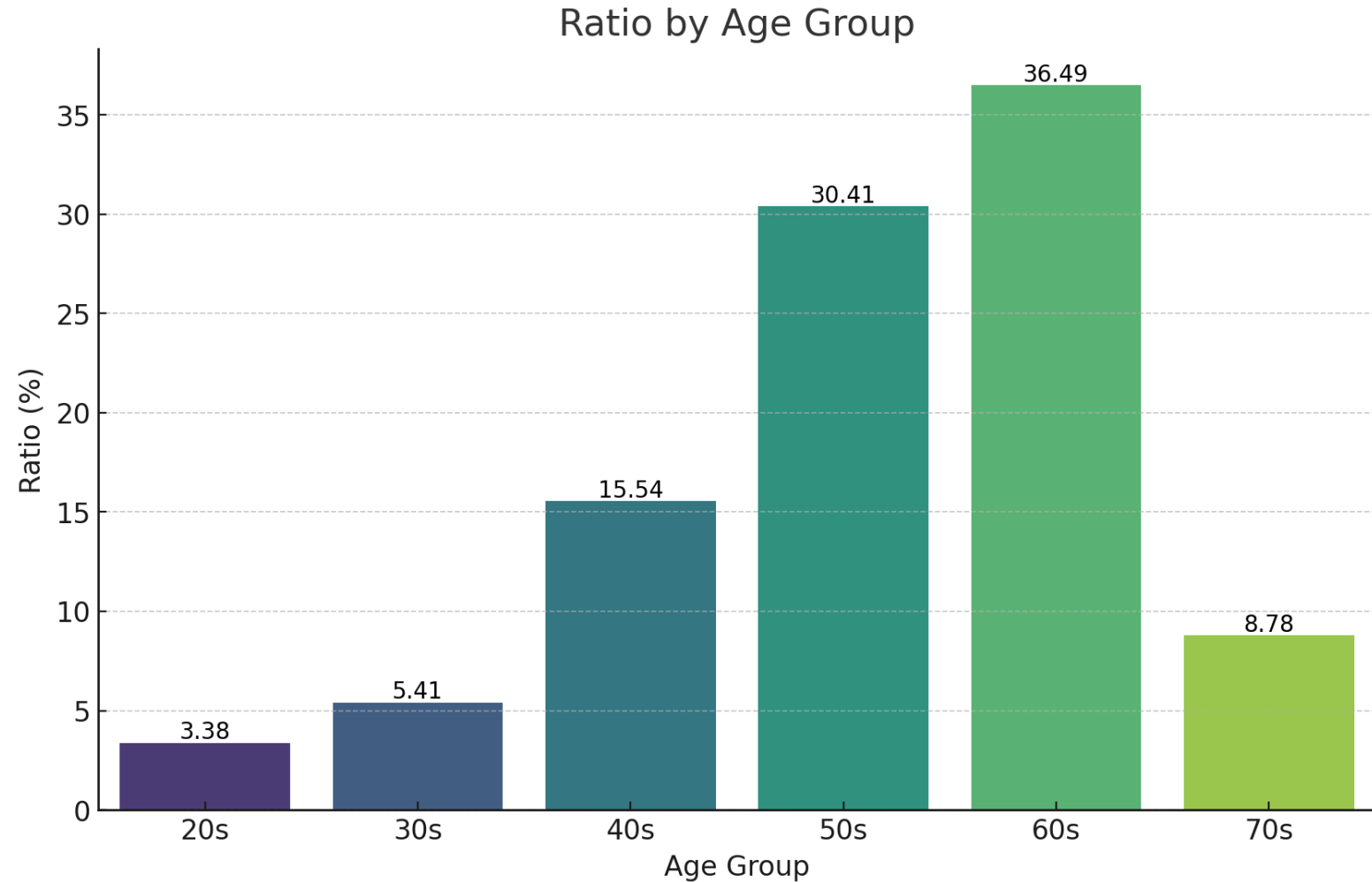
생년월일(YYMMDD) 데이터 분석

생성 AI 기술 발전으로 데이터 분석의 장벽이 낮아지고 있다.

생년월일 예제 데이터

YYMMDD 형식을 가지는 1 4 8 건의 데이터 샘플에서 연령대별(20대 ~ 70대) 비율을 보여주는 예제

| Date of Birth |
|---------------|
| 781224 |
| 561025 |
| 610423 |
| 680823 |
| 630517 |
| 630517 |
| 490525 |
| 671116 |
| 720925 |
| 610525 |
| 731117 |
| 731117 |
| 620625 |
| 620120 |
| 720324 |
| 581003 |
| 581003 |
| 591103 |
| 670219 |
| 570830 |
| 650305 |
| 580425 |
| 610302 |
| 660725 |
| 680310 |
| 680730 |
| 590219 |
| 730410 |



엑셀을 사용한다면 ?

매크로 활용 또는 데이터 가공 이후 시각화

=CountByAgeGroup(D\$6:D\$153,H6,I6)

| | D | E | F | G | H | J |
|--|--------|----|----|---|----|----|
| | | | | | 연령 | 인원 |
| | 781224 | 44 | 40 | | 20 | 5 |
| | 561025 | 66 | 60 | | 30 | 8 |
| | 610423 | 62 | 60 | | 40 | 21 |
| | 680823 | 55 | 50 | | 50 | 47 |
| | 630517 | 60 | 60 | | 60 | 54 |
| | 630517 | 60 | 60 | | 70 | 13 |

Function CountByAgeGroup

(rng As Range, lowerAge As Integer, upperAge As Integer) As Integer

Dim cell As Range

Dim count As Integer

Dim birthYear As Integer

Dim currentYear As Integer

Dim age As Integer

count = 0

currentYear = Year(Now)

For Each cell In rng

birthYear = 1900 + Left(cell.Value, 2)

age = currentYear - birthYear

If age >= lowerAge And age < upperAge Then

count = count + 1

End If

Next cell

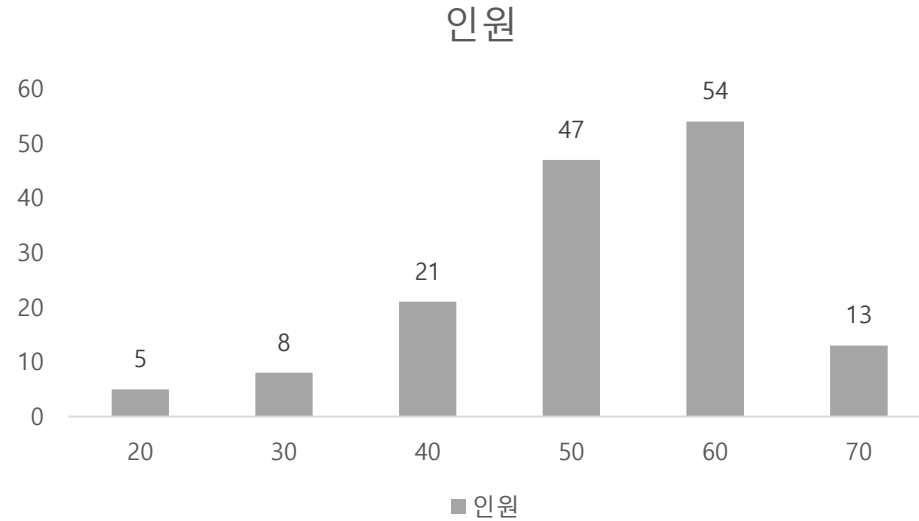
CountByAgeGroup = count

End Function

엑셀을 사용한다면 ?

직접적인 데이터 가공 이후 시각화 (FLOOR함수와 COUNTIF함수)

| =COUNTIF(F\$6:F\$153,H25) | | | | | |
|---------------------------|----|----|---|----|----|
| D | E | F | G | H | J |
| 670219 | 56 | 50 | | 연령 | 인원 |
| 570830 | 66 | 60 | | 20 | 5 |
| 650305 | 58 | 50 | | 30 | 8 |
| 580425 | 65 | 60 | | 40 | 23 |
| 610302 | 62 | 60 | | 50 | 45 |
| 660725 | 57 | 50 | | 60 | 54 |
| 680310 | 55 | 50 | | 70 | 13 |



CHATGPT를 활용하면?



birthday.xlsx
Spreadsheet

데이터를 요약해주세요.

2. Unnamed: 1:

- 평균: 56.31
- 표준 편차: 11.43
- 최솟값: 24
- 25% 분위수: 50
- 중앙값 (50% 분위수): 57
- 75% 분위수: 64
- 최댓값: 79

관찰 사항:

- "Date of Birth (yymmdd)" 열은 날짜 값을 숫자 형식 (yymmdd)으로 나타내고 있습니다. 데이터의 범위는 상당히 넓으며, 가장 오래된 것은 1944년 4월 25일 (440425)이고, 가장 최근의 것은 1998년 11월 30일 (981130)입니다.
- "Unnamed: 1" 열은 연령 또는 개인과 관련된 다른 숫자 지표일 수 있습니다. 값의 범위는 24에서 79까지이며, 평균은 약 56.31, 표준 편차는 약 11.43입니다.

다음 단계:

분석의 구체적인 목표에 따라 다음과 같은 단계를 진행할 수 있습니다:

1. "Date of Birth (yymmdd)"를 더 쉬운 분석을 위해 적절한 날짜-시간 형식으로 변환합니다.
2. "Unnamed: 1" 열의 이름을 그 목적에 따라 더 설명적인 것으로 변경합니다.
3. 누락된 값이나 이상치가 있다면 처리합니다.

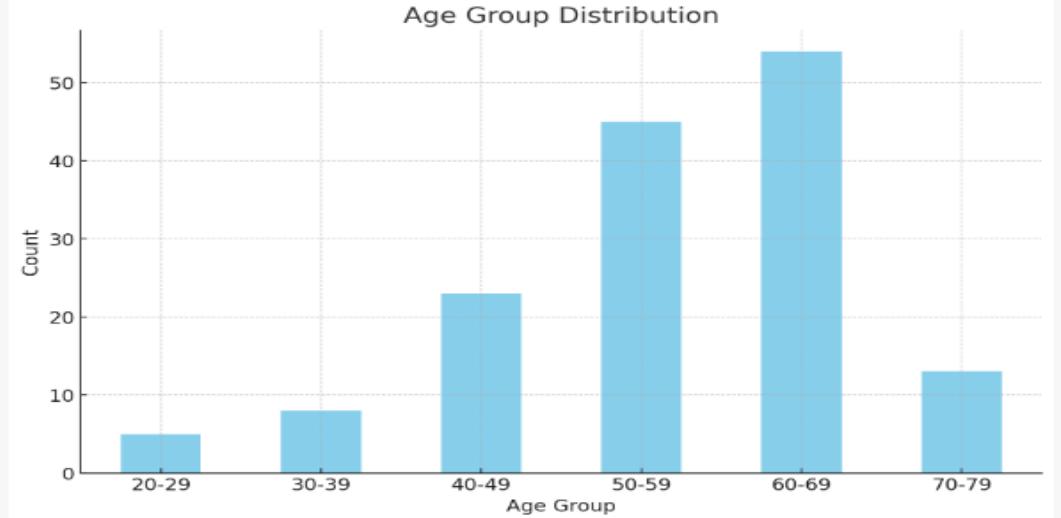


연령대별 분포를 막대그래프로 그려주세요.



Finished working

Show work



연령대별 분포

- **20-29:** 5명
- **30-39:** 8명
- **40-49:** 23명
- **50-59:** 45명
- **60-69:** 54명
- **70-79:** 13명

위의 막대그래프와 데이터를 보면, 가장 많은 사람들이 60-69 연령대에 속해 있으며, 그 다음으로는 50-59 연령대가 많습니다. 반면, 20-29 연령대와 30-39 연령대에는 상대적으로 적은 수의 사람들이 속해 있습니다.

이러한 연령 분포는 고령층이 많이 포함되어 있음을 나타냅니다, 이는 고령층 중심의 서비스나 제품을 고려할 때 유용한 인사이트가 될 수 있습니다.

다음 분석 단계나 특별한 요구사항이 있으신가요?

CHATGPT를 활용하면?

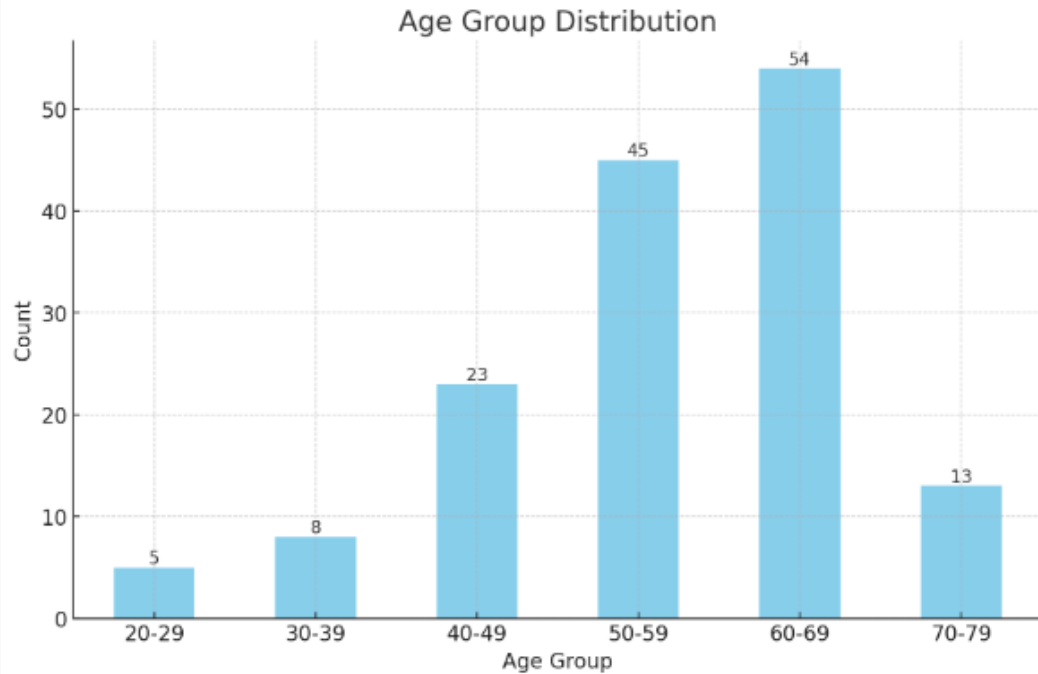


막대 그래프 위에 인원수를 넣어주세요.



Finished working

Show work



막대그래프 위에 각 연령대별 인원 수가 표시되어 있습니다. 이렇게 하면 그래프를 보면서 쉽게 각 연령대에 얼마나 많은 사람들이 있는지 파악할 수 있습니다.

이러한 시각화는 데이터의 분포를 빠르게 이해하고, 특정 연령대에 초점을 맞춘 전략을 구성할 때 도움이 될 수 있습니다.

다른 분석이나 질문이 있으신가요?

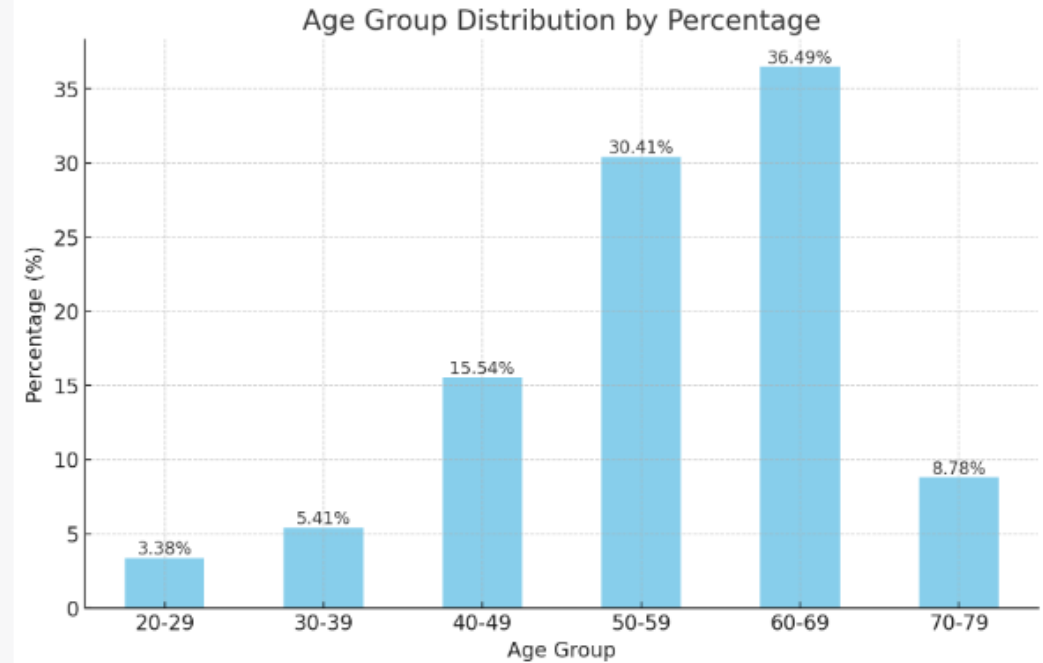


인원수 대신 비율로 변경해주세요.



Finished working

Show work



연령대별 분포 (비율)

- 20-29: 3.38%
- 30-39: 5.41%
- 40-49: 15.54%
- 50-59: 30.41%
- 60-69: 36.49%
- 70-79: 8.78%

막대그래프 위에 각 연령대별로 차지하는 비율이 표시되어 있습니다. 이를 통해 60-69 연령대가 전체의 약 36.49%로 가장 큰 비율을 차지하고 있으며, 그 다음으로는 50-59 연령대가 30.41%로 이어집니다. 반면, 20-29 연령대와 30-39 연령대는 각각 3.38%와 5.41%로

06

실습예제
(비정형 텍스트 데이터
쇼핑몰 리뷰 텍스트 데이터 분석)



Extract data efficiently with 리스틀리

A free easy-to-use web scraping tool.
Quickly solve your data problems today!

크롬확장프로그램 추가

사용법 안내 (영상)

★★★★☆ 전 세계 **209,400+** 사용자가 신뢰합니다

실시간으로 추출된 URL을 확인해보세요!

- https://new.land.naver.com/offices?ms... 1 hour ago
- https://besthomewarranties.inlinecrm.c... 1 hour ago
- https://www.google.com/search?sca_e... 1 hour ago



로보락 S8 Pro Ultra

최저 1,567,620원 무료

최저가 사러가기

판매처 216

제품정보

블로그리뷰

쇼핑몰리뷰 35,007

AITEMS추천

● **사용성** 어플 사용법도 어렵지 않고 보기 편하게 되어있습니다



랭킹순 | 최신순

포토&동영상리뷰만 보기 ☐ OFF

전체 (35,007)

5점 (32,884)

4점 (1,884)

3점 (173)

2점 (12)

1점 (54)

주제전체

만족도

가격

관리

소음

디자인

성능

조작성

사용성

기능

효과

향기

색상

품질



★★★★★ 5 로보락 공식 스토어 · ioja**** · 24.02.20.

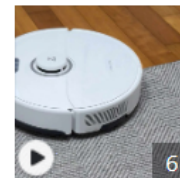
처음 구매하는 로봇청소기 이라 여러제품을 기능,AS 등을 비교하느라 10일정도 고민한것 같아요.기능이 좋고 가격이 저렴...

처음 구매하는 로봇청소기 이라 여러제품을 기능,AS 등을 비교하느라 10일정도 고민한것 같아요.

기능이 좋고 가격이 저렴해 해외 직구도 고민을 해 봤지만 AS 가 걸리고 해서 비싸더라도 AS,성능,판매량 등 종합적으로 로보락이 좋더라고요. 결정하고 바로 구매 하였어요.

설치 : 처음 제품 받고 무게가 엄청 무겁더라고요.

설치, 세팅이 어려울까 걱정도 있었는데




리뷰펼치기 ▼

★★★★★ 5 로보락 공식 스토어 · rons**** · 23.11.25.

포장이 꼼꼼하고 박스 개봉하면 크게 설치 설명서 있어서 초기 설치하기 편합니다. 무게감은 좀 있어서 박스 이동할 때 주...

포장이 꼼꼼하고 박스 개봉하면 크게 설치 설명서 있어서 초기 설치하기 편합니다. 무게감은 좀 있어서 박스 이동할 때 주의하셔야 할것 같습니다





로보락 S8 Pro Ultra

최저 1,567,620원 무료

최저가 사러가기

판매처 216

제품정보

블로그리뷰

쇼핑몰리뷰 35,007

AITEMS추천

사용성 어플 사용법도 어렵지 않고 보기 편하게 되어있습니다

랭킹순 | 최신순

포토&동영상리뷰만 보기

전체 (35,007) | 5점 (32,884) | 4점 (1,884) | 3점 (173) | 2점 (12) | 1점 (54)

주제전체 | 만족도 | 가격 | 관리 | 소음 | 디자인 | 성능 | 조작성 | 사용성 | 기능 | 효과 | 향기 | 색상 | 품질

★★★★★ 5 로보락 공식 스토어 · ioja**** · 24.02.20.

처음 구매하는 로봇청소기 이라 여러제품을 기능,AS 등을 비교하느라 10일정도 고민한것 같아요.기능이 좋고 가격이 저렴...

처음 구매하는 로봇청소기 이라 여러제품을 기능,AS 등을 비교하느라 10일정도 고민한것 같아요.

기능이 좋고 가격이 저렴해 해외 직구도 고민을 해 봤지만 AS 가 걸리고 해서 비싸더라도 AS,성능,판매량 등 종합적으로 로보락이 좋더라구요. 결정하고 바로 구매 하였어요.

설치 : 처음 제품 받고 무게가 엄청 무겁더라고요.

설치, 세팅이 어려울까 걱정도 있었는데

리뷰펼치기

★★★★★ 5 로보락 공식 스토어 · rons**** · 23.11.25.

포장이 꼼꼼하고 박스 개봉하면 크게 설치 설명서 있어서 초기 설치하기 편합니다. 무게감은 좀 있어서 박스 이동할 때 주...

< >

[-] [x]

CSS SELECTOR

#section_review

↑ ↓

EXTRACT

Default ▾

preview

RUN LISTLY

부분 + 기능 소개

TOP

07

실습예제

(자전거 대여 수요예측 kaggle 데이터)

<https://www.kaggle.com/competitions/bike-sharing-demand/overview>

Competitions

Results



Bike Sharing Demand

Forecast use of a city bikeshare system
Playground · 3242 Teams · 9 years ago



INF131 - 2019 Challenge

Build a regressor to predict bike sharing demand
Community · 69 Teams · 4 years ago



Bike Sharing Hourly Demand Forecasting

Forecast bike sharing system hourly demand
Community · 44 Teams · Invitation Only · 4 years ago



KAGGLE · PLAYGROUND PREDICTION COMPETITION · 9 YEARS AGO

Bike Sharing Demand

Forecast use of a city bikeshare system

Overview

Data

Code

Models

Discussion

Leaderboard

Rules

Overview

Start

May 29, 2014

Close

May 30, 2015

Competition Host

Kaggle



Prizes & Awards

Knowledge

Does not award Points or Medals

Participation

4,169 Entrants

3,559 Participants

3,242 Teams

32,809 Submissions

Description



```
In [1]: import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# 노트북 안에 그래프를 그리기 위해
%matplotlib inline

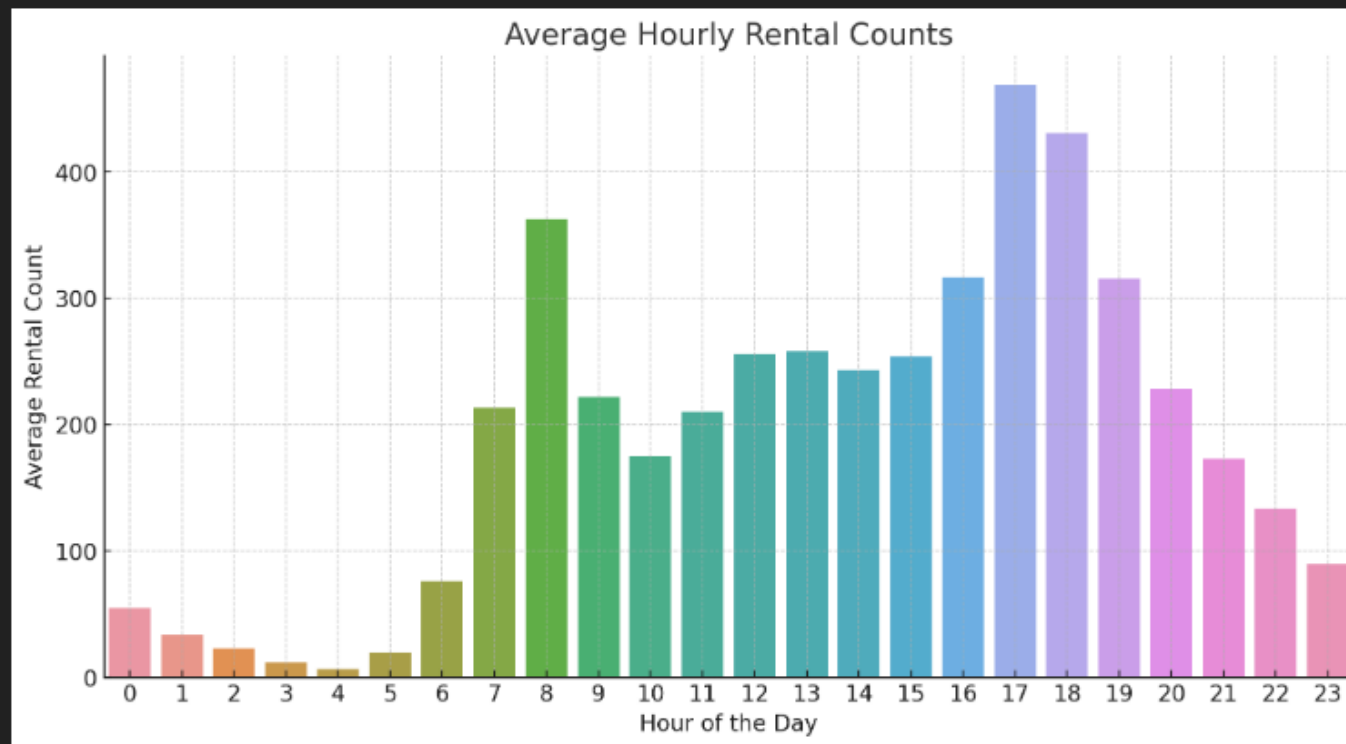
# 그래프에서 격자로 숫자 범위가 눈에 잘 띄도록 ggplot 스타일을 사용
plt.style.use('ggplot')

# 그래프에서 마이너스 폰트 깨지는 문제에 대한 대처
mpl.rcParams['axes.unicode_minus'] = False
```



ChatGPT

시간별로 총 대여량(`count`)을 시각화하여, 하루 중 각 시간대별 대여 패턴을 파악해 보겠습니다. 이를 위해 `datetime` 컬럼을 `datetime` 타입으로 변환하고, 시간 컬럼을 추출한 후, 시간별로 대여량의 평균을 계산하여 시각화하겠습니다.



위 그래프는 하루 중 각 시간대별 평균 대여량을 보여줍니다. 몇 가지 특징적인 패턴을 볼 수 있습니다: