EDS Theory Assignment 1

Name: Adarsh Bhagwat Shinde

Div.: CS2 Roll No.: 82

PRN: 202401040006

1. What is the shape of the dataset?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 1. Shape of the Dataset
print("1. Dataset shape:", df.shape)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"
1. Dataset shape: (1525, 10)
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

2. What are the column data types?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 2. Column Data Types
print("\n2. Column data types:\n", df.dtypes)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

2. Column data types:
 _unit_id                        int64
_golden                           bool
_unit_state                     object
_trusted_judgments               int64
_last_judgment_at               object
policies_violated               object
policies_violated:confidence    object
city                            object
policies_violated_gold         float64
review                          object
dtype: object
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

3. How many null values exist per column?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 3. Null Values per Column
print("\n3. Null values per column:\n", df.isnull().sum())
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

3. Null values per column:
 _unit_id                        0
_golden                          0
_unit_state                      0
_trusted_judgments               0
_last_judgment_at                0
policies_violated                0
policies_violated:confidence    54
city                            87
policies_violated_gold        1525
review                           0
dtype: int64
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

4. What are the top 5 most common policy violations?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 4. Top 5 Most Common Policy Violations
top_violations = df['policies_violated'].value_counts().head(5)
print("\n4. Top 5 most common policy violations:\n", top_violations)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

4. Top 5 most common policy violations:
 policies_violated
na              304
RudeService     186
SlowService     135
OrderProblem    121
BadFood         107
Name: count, dtype: int64
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

5. Which cities appear most frequently in the reviews?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 5. Most Frequent Cities in Reviews
top_cities = df['city'].value_counts().head(5)
print("\n5. Top 5 cities appearing in reviews:\n", top_cities)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

5. Top 5 cities appearing in reviews:
 city
Las Vegas      409
Chicago        219
Los Angeles    167
New York       165
Atlanta        130
Name: count, dtype: int64
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

6. Which city has the highest number of RudeService violations?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 6. City with Highest RudeService Violations
if 'RudeService' in df['policies_violated'].unique():
    rude_service_city = df[df['policies_violated'] == 'RudeService']['city'].value_counts().idxmax()
    print("\n6. City with highest number of RudeService violations:", rude_service_city)
else:
    print("\n6. No 'RudeService' violations found in dataset.")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

6. City with highest number of RudeService violations: Las Vegas
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

7. How many unique cities are there?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 7. Number of Unique Cities
unique_cities = df['city'].nunique()
print("\n7. Number of unique cities:", unique_cities)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

7. Number of unique cities: 9
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

8. What percentage of reviews have multiple policy violations?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 8. Percentage of Reviews with Multiple Policy Violations
multiple_violations = df['policies_violated'].str.split(',').str.len() > 1
percentage_multiple = (multiple_violations.sum() / len(df)) * 100
print(f"\n8. Percentage of reviews with multiple policy violations: {percentage_multiple:.2f}%")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

8. Percentage of reviews with multiple policy violations: 0.00%
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

9. What is the average number of policy violations per review?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 9. Average Number of Policy Violations per Review
average_violations = df['policies_violated'].str.split(',').str.len().mean()
print(f"\n9. Average number of policy violations per review: {average_violations:.2f}")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

9. Average number of policy violations per review: 1.00
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

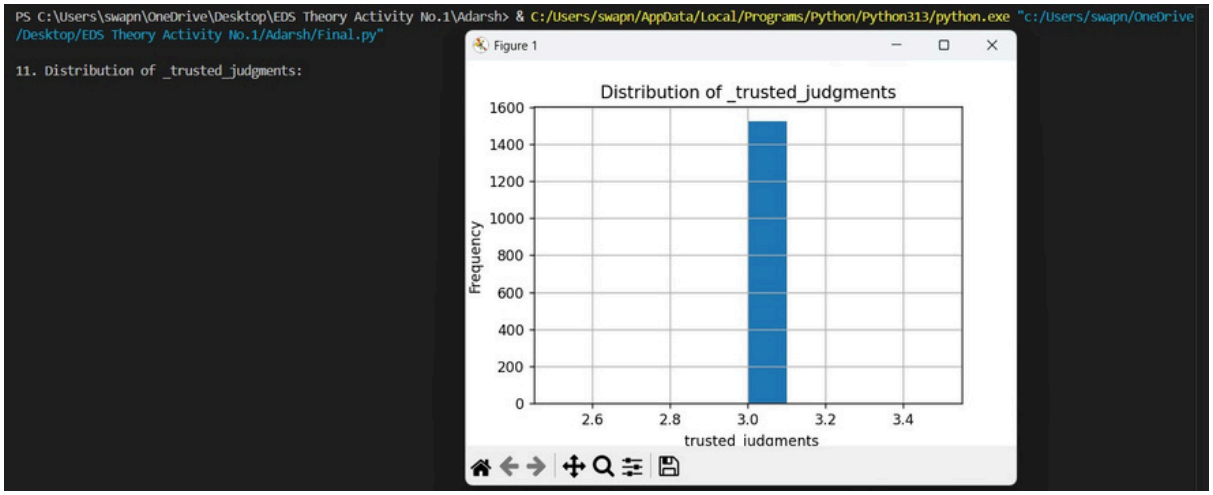
10. What is the average number of words per review?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 10. Average Number of Words per Review
average_words = df['review'].dropna().str.split().str.len().mean()
print(f"\n10. Average number of words per review: {average_words:.2f}")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

10. Average number of words per review: 96.56
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

11. What is the distribution of _trusted_judgments?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 11. Distribution of _trusted_judgments
print("\n11. Distribution of _trusted_judgments:")
df['_trusted_judgments'].hist()
plt.xlabel('_trusted_judgments')
plt.ylabel('Frequency')
plt.title('Distribution of _trusted_judgments')
plt.show()
```



12. How many reviews were finalized (_unit_state == 'finalized')?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 12. Number of Finalized Reviews (_unit_state == 'finalized')
finalized_reviews = df[df['_unit_state'] == 'finalized'].shape[0]
print("\n12. Number of finalized reviews:", finalized_reviews)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

12. Number of finalized reviews: 1525
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

13. How many reviews have no policy violations?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 13. How Many Reviews Have No Policy Violations?
# Assuming reviews with empty string or NaN in 'policies_violated' have no violations
no_violations_count = df['policies_violated'].fillna('').str.strip().eq('').sum()
print(f"\n13. Number of reviews with no policy violations: {no_violations_count}")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

13. Number of reviews with no policy violations: 0
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

14. Which policy is most commonly violated in each city?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 14. Most Commonly Violated Policy in Each City
def most_common_policy(violations):
    return violations.str.split(',').explode().value_counts().idxmax()

common_violations_per_city = df.groupby('city')['policies_violated'].apply(most_common_policy)
print("\n14. Most commonly violated policy per city:")
print(common_violations_per_city)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

14. Most commonly violated policy per city:
city
Atlanta                 na
Chicago                 na
Cleveland      SlowService
Dallas                  na
Houston                 na
Las Vegas               na
Los Angeles             na
New York                na
Portland                na
Name: policies_violated, dtype: object
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

15. What is the average confidence score for each type of violation?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 15. Average Confidence Score for Each Type of Violation
# Explode the policy violations and confidence scores into long format for averaging
df_long = df[['policies_violated', 'policies_violated:confidence']].copy()
df_long = df_long.dropna(subset=['policies_violated', 'policies_violated:confidence'])
df_long['policies_violated'] = df_long['policies_violated'].str.split(',')
df_long = df_long.explode('policies_violated')
df_long['policies_violated:confidence'] = pd.to_numeric(df_long['policies_violated:confidence'], errors='coerce')
average_confidence = df_long.groupby('policies_violated')['policies_violated:confidence'].mean()
print("\n15. Average confidence score for each type of violation:")
print(average_confidence)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

15. Average confidence score for each type of violation:
policies_violated
BadFood                                0.876383
BadFood\rCost                               NaN
BadFood\rFilthy                             NaN
BadFood\rFilthy\rRudeService                NaN
BadFood\rMissingFood                        NaN
                                          ...
SlowService\rScaryMcDs                      NaN
SlowService\rScaryMcDs\rBadFood             NaN
SlowService\rScaryMcDs\rFilthy              NaN
SlowService\rScaryMcDs\rRudeService         NaN
na                                     0.909448
Name: policies_violated:confidence, Length: 140, dtype: float64
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

16. Find the longest review by word count.

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 16. Longest Review by Word Count
longest_review_index = df['review'].dropna().str.split().str.len().idxmax()
longest_review = df.loc[longest_review_index, 'review']
longest_review_length = len(longest_review.split())
print(f"\n16. Longest review by word count ({longest_review_length} words):")
print(longest_review)
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

16. Longest review by word count (914 words):
It's 2 for 1 review day! Î¼I can't review the Walmart without reviewing the McDonald's that resides inside. Î¼And boy is it some tantalizing stuff.I try not to
give too many 1-star reviews but this McDonald's so deserves it. Î¼I hardly frequent a McDonald's at all (although this is my second McDonald's review in less t
han a week) but I need to let the peeps in on this one. Î¼No Yelper should go unwarned. Î¼Of course I came to McDonald's since I have those coupons for their ne
w contraption, the McWrap. Î¼Which when I pronounce very fast and in my Scottish accent sounds much like 'McCrap'. Î¼After my second go round with this thing, I
 have made up my mind it is decidedly so. Î¼I need to write McDonald's execs to change the name on this thing immediately.My coupon for the day was buy a medium
 fries and a medium drink, get a McWrap of your choice for FREE. Î¼I have a hard time resisting free. Î¼I'm here to tell you...WORST ALMOST FREE MEAL OF MY LIFE
!!! Î¼Seriously.It was the tail end of lunchtime so the line was still a bit busy. Î¼I patiently waited for almost 10 minutes...which is like forever in McDonal
d's time. Î¼At least two folks left the line while I was there. Î¼I know it's not McDonald's fault, but the couple in front of me took up majority of the line s
pace because they were in motorized wheelchairs. Î¼And they were very loud. Î¼This doesn't bother me much. Î¼But...they were also very stinky. Î¼Yes, stinky. Î¼
I have no problem with stinky in most cases (I occasionally help the homeless so I know) but the guy in front of me hadn't washed in DAYS if I were to guess. Î¼
To each his own, and I know it's almost summer so I know we all want to save water...but DAMN. Î¼If I get a whiff and I have to take a step back? Î¼You might be
 S-T-I-N-K-Y. Î¼But who am I to judge? Î¼I took a big step back and waited my turn. Î¼Free is the motivation here. Î¼I finally get up there to the McDonald's ca
shier and present my coupon to the guy. Î¼He takes a look at it and says 'We don't usually take the promo coupons at this location but we'll let you use it." Î¼
Then he proceeds to put the coupon onto a stack of...more coupons! Î¼Mind boggling to say the least. Î¼He takes my order of a medium fries and medium drink, ask
s me what type of McWrap I want (Grilled Chicken & Bacon) and rings me up. Î¼$2.91 later I'm given my receipt and asked to wait for my number to be called.I wai
t. Î¼And wait. Î¼#1! Î¼#2! Î¼#3! Î¼#5! Î¼#6! Î¼#7! Î¼#8! Î¼Obviously I'm #4 and 10 minutes later I am still waiting. Î¼Honestly, what the Hell are you guys doin
g back there? Î¼Mixing the masa for my fresh pressed tortilla? Î¼Feathering a chicken? Î¼Perhaps ripening a tomato? Î¼I ask the McDonald's runner and he tells m
e, "We're waiting for your McWrap." Î¼Of course.At this point I'm almost running late to be back at work, so the minute I get my order I bolt back to the office
 and have lunch at my desk. Î¼Thankfully I'm only three minutes away. Î¼I really, really should have just gone to the vending machine or eaten my stash of Reese
's chocolate peanut butter eggs from last Easter. Î¼Talk about disappointment! Î¼I take out my McWrap from my bag and lay down the box of fries on my desk. Î¼I
continue working while munching away on what is supposed to be McDonald's claim to fame. Î¼Their golden fries. Î¼I bite into one, eh a bit stale. Î¼I bit into a
nother one, eh not enough salt. Î¼I bite into yet another one, eh cold?? Î¼I look down, and whaddaya know? Î¼My damn fry is not even barely cooked! Î¼White like
 Olivia Wilde in the winter. Î¼I look through the entire box and find at least half a dozen fries not cooked properly! Î¼Not golden brown! Î¼Not delicious! Î¼Ba
```

17. How many reviews mention "clean" or "dirty" in the text?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 17. Number of Reviews Mentioning "clean" or "dirty"
mentions_clean_dirty = df['review'].dropna().str.contains('clean|dirty', case=False).sum()
print(f"\n17. Number of reviews mentioning 'clean' or 'dirty': {mentions_clean_dirty}")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

17. Number of reviews mentioning 'clean' or 'dirty': 185
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

18. Are there duplicate reviews in the dataset?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 18. Count of Duplicate Reviews
duplicates = df.duplicated().sum()
print(f"\n18. Number of duplicate reviews: {duplicates}")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

18. Number of duplicate reviews: 0
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

19. What is the median review length in characters?

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 19. Median Review Length in Characters
median_length = df['review'].dropna().str.len().median()
print(f"\n19. Median review length (characters): {median_length}")
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

19. Median review length (characters): 388.0
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh>
```

20. Plot the top 5 cities with the highest average number of violations per review.

```python
import pandas as pd
import matplotlib.pyplot as plt

filename = 'McDonalds-Yelp-Sentiment-DFE.csv'
df = pd.read_csv(filename, encoding='ISO-8859-1')

# 20. Plot Top 5 Cities with Highest Average Number of Violations per Review
avg_violations_per_city = df.groupby('city')['policies_violated'].apply(
    lambda x: x.str.split(',').apply(len).mean()
).nlargest(5)

print("\n20. Top 5 cities with highest average violations per review:")
print(avg_violations_per_city)

avg_violations_per_city.plot(kind='bar', color='skyblue')
plt.xlabel('City')
plt.ylabel('Average Number of Violations per Review')
plt.title('Top 5 Cities with Highest Average Number of Violations per Review')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
PS C:\Users\swapn\OneDrive\Desktop\EDS Theory Activity No.1\Adarsh> & C:/Users/swapn/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/swapn/OneDrive
/Desktop/EDS Theory Activity No.1/Adarsh/Final.py"

20. Top 5 cities with highest average violations per review:
city
Atlanta      1.0
Chicago      1.0
Cleveland    1.0
Dallas       1.0
Houston      1.0
Name: policies_violated, dtype: float64
```