EDS Theory Assignment

Mohammed Zaki Abdhusain Shahpure

202401080028

CS6-54

Twitter Sentiment Analysis - Problems and Commands

Used jupyter notebook for doing this activity, used Kaggle’s Text Classification Dataset of - Twitter Sentiment Analysis, which helps us understand the sentiments of the users, by their texts.

```
print("EDS Assignment-1")
```

```
print("Name: Mohammed Zaki Abidhusain Shahpure")
```

```
print("PRN: 202401080028")
```

```
print("Roll no: CS6-54")
```

Problem 1: Distribution of sentiment labels

```
sentiment_counts = df['target'].value_counts()
print(sentiment_counts)
print((sentiment_counts / len(df) * 100).round(2))
```

Problem 2: Tweet length by sentiment

```
df['tweet_length'] = df['text'].str.len()
length_by_sentiment = df.groupby('target')['tweet_length'].agg(['mean', 'median', 'min', 'max'])
print(length_by_sentiment)
```

Problem 3: Most active users

```
top_users = df['user'].value_counts().head(10)
print(top_users)
```

Problem 4: Tweets distribution across days

```
df['date'] = pd.to_datetime(df['date'], errors='coerce')
df.dropna(subset=['date'], inplace=True)
```

```
df['day_of_week'] = df['date'].dt.day_name()
day_counts = df['day_of_week'].value_counts()
print(day_counts)
```

Problem 5: Most common words

```
sampled_text = ' '.join(df['text'].sample(min(1000, len(df))))
words = re.findall(r'\b[a-zA-Z]{3,}\b', sampled_text.lower())
top_words = pd.Series(words).value_counts().head(15)
print(top_words)
```

Problem 6: Sentiment by time of day

```
df['hour'] = df['date'].dt.hour
hourly_sentiment = df.groupby('hour')['target'].mean()
print(hourly_sentiment)
```

Problem 7: Tweets containing specific keywords

```
keywords = ['happy', 'sad', 'love', 'hate', 'good', 'bad']
for word in keywords:
    count = df['text'].str.contains(fr'\b{word}\b', case=False, regex=True).sum()
    print(f"Tweets containing '{word}': {count}")
```

Problem 8: Average hashtag length

```
hashtag_pattern = r'#(\w+)'
df['hashtags'] = df['text'].str.findall(hashtag_pattern)
df['hashtag_count'] = df['hashtags'].apply(len)
df['avg_hashtag_length'] = df['hashtags'].apply(lambda x: np.mean([len(h) for h in x]) if x else 0)
print(df['hashtag_count'].mean())
print(df['avg_hashtag_length'][df['avg_hashtag_length'] > 0].mean())
```

Problem 9: Tweets activity by hour

```
hourly_counts = df['hour'].value_counts().sort_index()
print(hourly_counts)
```

Problem 10: Tweet length vs sentiment correlation

```
length_sentiment_corr = np.corrcoef(df['target'], df['tweet_length'])[0, 1]
print(length_sentiment_corr)
```

Problem 11: Users with mixed sentiments

```
user_sentiment_counts = df.groupby('user')['target'].nunique()
mixed_sentiment_users = (user_sentiment_counts > 1).sum()
print(mixed_sentiment_users)
```

Problem 12: Percentage of tweets containing URLs

```
url_pattern = r'http[s]?://(?:[a-zA-Z0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+'
df['contains_url'] = df['text'].str.contains(url_pattern, case=False)
url_percentage = df['contains_url'].mean() * 100
print(url_percentage)
```

Problem 13: Sentiment over months

```
df['month'] = df['date'].dt.month
monthly_sentiment = df.groupby('month')['target'].mean()
print(monthly_sentiment)
```

Problem 14: Distribution of tweet length

```
length_stats = df['tweet_length'].describe()
print(length_stats)
```

Problem 15: Highest average sentiment users

```
user_tweet_counts = df['user'].value_counts()
users_with_multiple_tweets = user_tweet_counts[user_tweet_counts >= 5].index
filtered_df = df[df['user'].isin(users_with_multiple_tweets)]
user_sentiment = filtered_df.groupby('user')['target'].mean().sort_values(ascending=False)
print(user_sentiment.head(10))
```

Problem 16: Correlation between mentions and sentiment

```
mention_pattern = r'@(\w+)'
df['mentions'] = df['text'].str.findall(mention_pattern)
df['mention_count'] = df['mentions'].apply(len)
mention_sentiment_corr = np.corrcoef(df['target'], df['mention_count'])[0, 1]
print(mention_sentiment_corr)
```

Problem 17: Sentiment by day of week

```
day_sentiment = df.groupby('day_of_week')['target'].mean().sort_values()
print(day_sentiment)
```

Problem 18: Tweets containing questions

```
question_pattern = r'\?'
question_tweets = df['text'].str.contains(question_pattern).sum()
print(question_tweets)
```

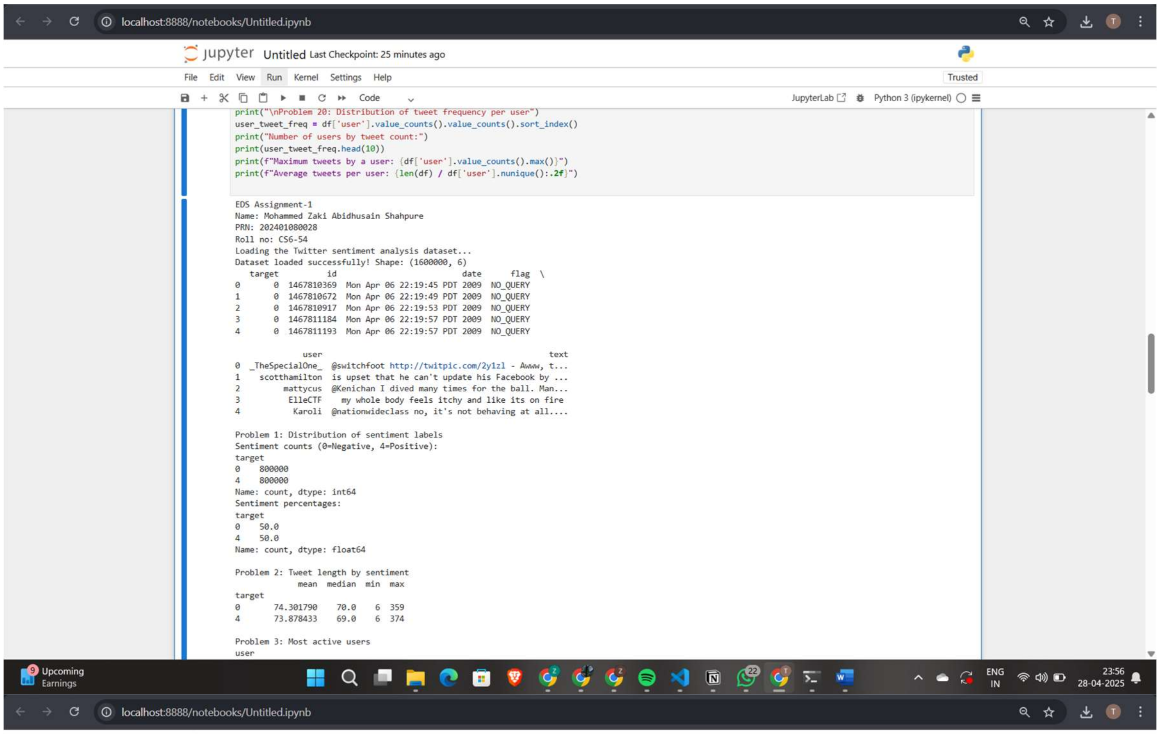
Problem 19: Sentiment comparison (hashtag vs no hashtag)

```
df['has_hashtag'] = df['hashtag_count'] > 0
hashtag_sentiment = df.groupby('has_hashtag')['target'].mean()
print(hashtag_sentiment)
```

Problem 20: Tweet frequency per user

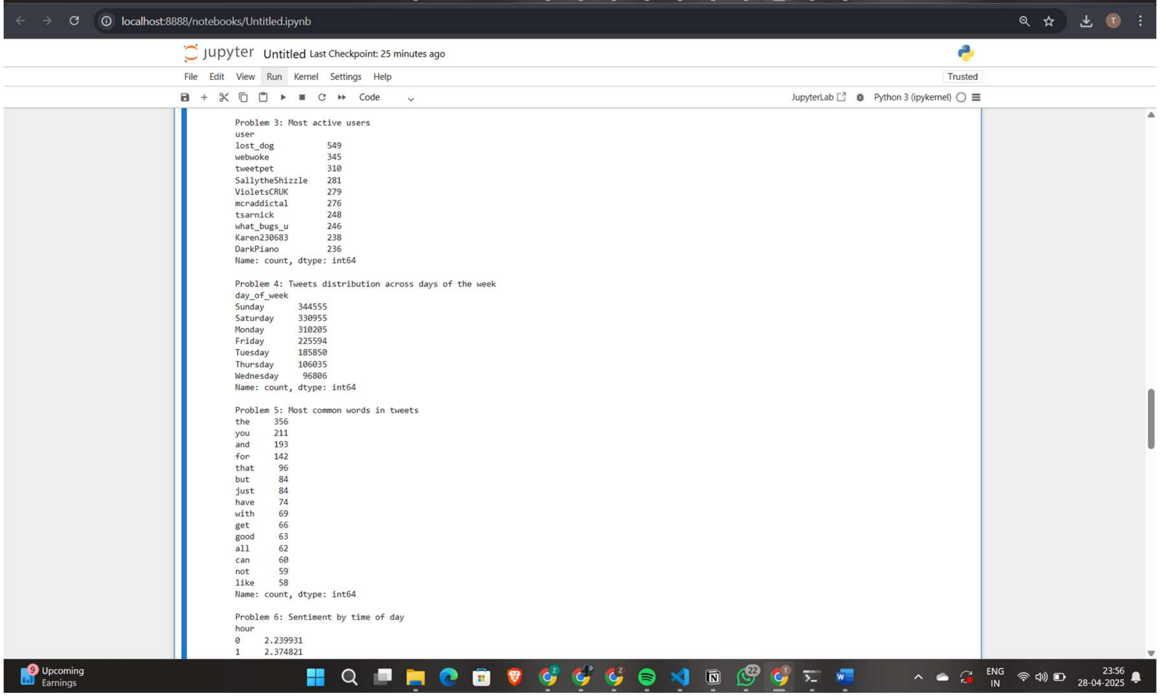
```
user_tweet_freq = df['user'].value_counts().value_counts().sort_index()
print(user_tweet_freq.head(10))
```

Below attached are the screenshots of the above commands run in the Jupyter Notebook, by importing the dataset locally and importing pandas and numpy, and then performing the required operations

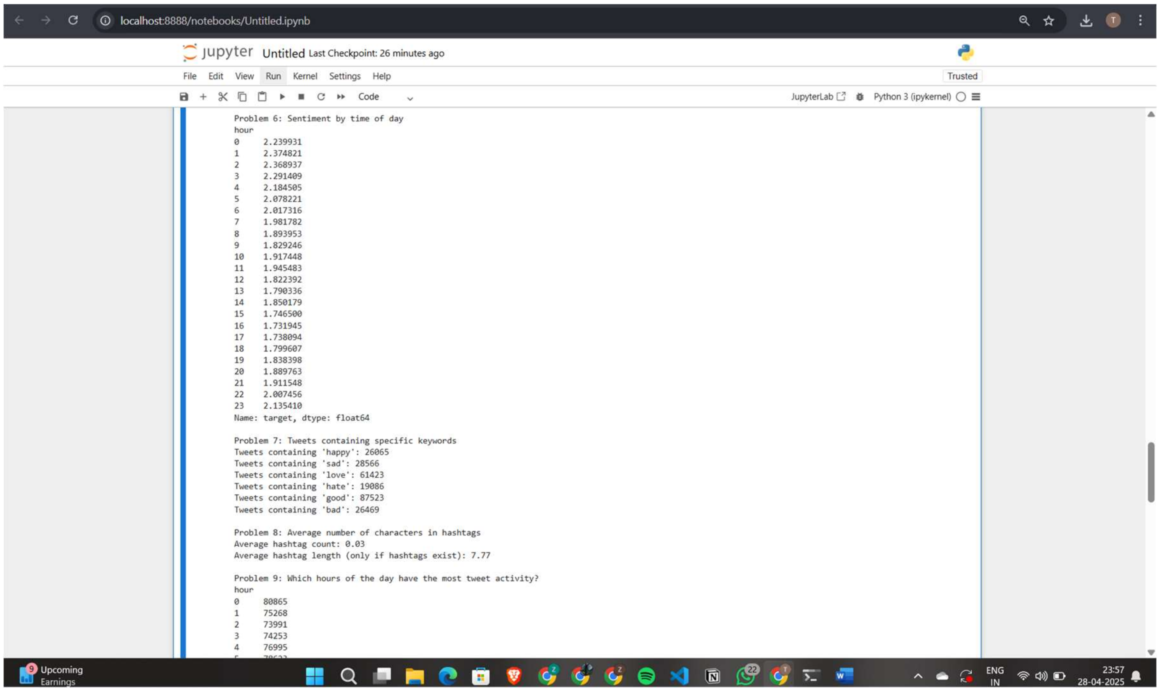

localhost:8888/notebooks/Untitled.ipynb
jupyter Untitled Last Checkpoint: 25 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Upcoming
Earnings
ENG
IN
23:56
28-04-2025

```
print("\nProblem 20: Distribution of tweet frequency per user")
user_tweet_freq = df['user'].value_counts().value_counts().sort_index()
print("Number of users by tweet count:")
print(user_tweet_freq.head(10))
print(f"Maximum tweets by a user: {df['user'].value_counts().max()}")
print(f"Average tweets per user: {len(df) / df['user'].nunique():.2f}")
```

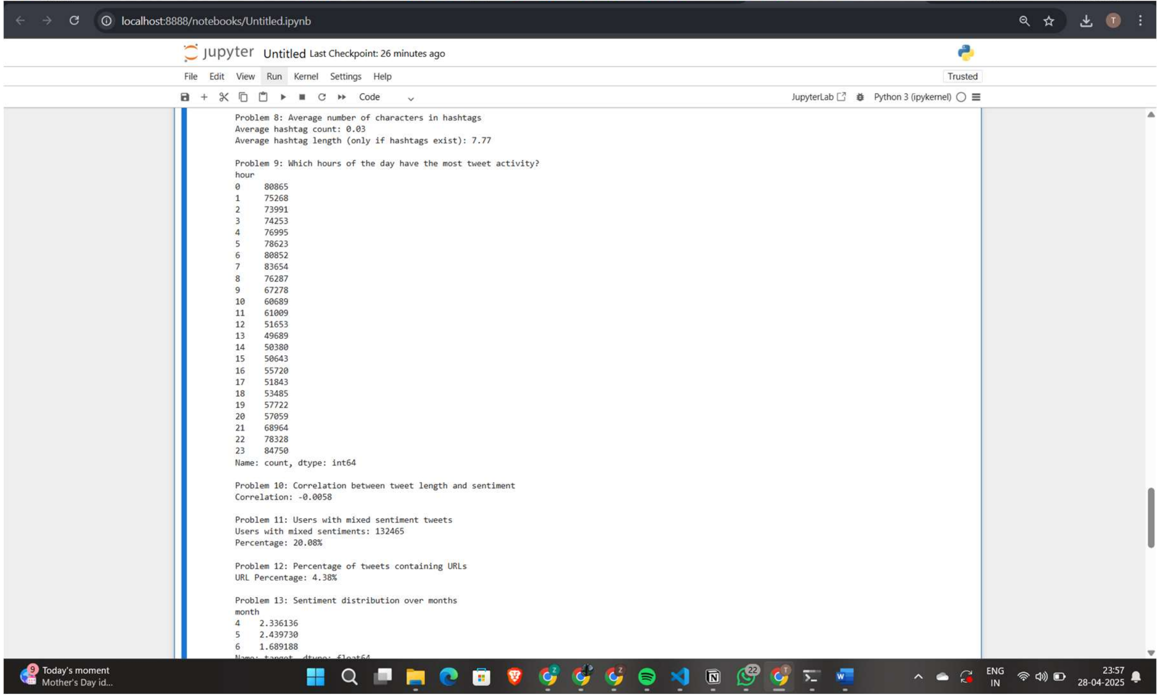
```
EDS Assignment-1
Name: Mohammed Zaki Abidhusain Shahpure
PRN: 202401080028
Roll no: CS6-54
Loading the Twitter sentiment analysis dataset...
Dataset loaded successfully! Shape: (1600000, 6)
   target          id                          date      flag  \
0       0  1467810369  Mon Apr 06 22:19:45 PDT 2009  NO_QUERY
1       0  1467810672  Mon Apr 06 22:19:49 PDT 2009  NO_QUERY
2       0  1467810917  Mon Apr 06 22:19:53 PDT 2009  NO_QUERY
3       0  1467811184  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY
4       0  1467811193  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY

              user                                               text
0  _TheSpecialOne_  @switchfoot http://twitpic.com/2y1zl - Awww, t...
1    scotthamilton  is upset that he can't update his Facebook by ...
2         mattycus  @Kenichan I dived many times for the ball. Man...
3          ElleCTF    my whole body feels itchy and like its on fire
4           Karoli  @nationwideclass no, it's not behaving at all....

Problem 1: Distribution of sentiment labels
Sentiment counts (0=Negative, 4=Positive):
target
0    800000
4    800000
Name: count, dtype: int64
Sentiment percentages:
target
0    50.0
4    50.0
Name: count, dtype: float64

Problem 2: Tweet length by sentiment
             mean  median  min  max
target
0       74.301790    70.0    6  359
4       73.878433    69.0    6  374

Problem 3: Most active users
user
```


localhost:8888/notebooks/Untitled.ipynb
jupyter Untitled Last Checkpoint: 25 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Upcoming
Earnings
ENG
IN
23:56
28-04-2025

```
Problem 3: Most active users
user
lost_dog           549
webwoke            345
tweetpet           310
SallytheShizzle    281
VioletsCRUK        279
mcraddictal        276
tsarnick           248
what_bugs_u        246
Karen230683        238
DarkPiano          236
Name: count, dtype: int64

Problem 4: Tweets distribution across days of the week
day_of_week
Sunday       344555
Saturday     330955
Monday       310205
Friday       225594
Tuesday      185850
Thursday     106035
Wednesday     96806
Name: count, dtype: int64

Problem 5: Most common words in tweets
the     356
you     211
and     193
for     142
that     96
but      84
just     84
have     74
with     69
get      66
good     63
all      62
can      60
not      59
like     58
Name: count, dtype: int64

Problem 6: Sentiment by time of day
hour
0     2.239931
1     2.374821
```


localhost:8888/notebooks/Untitled.ipynb
jupyter Untitled Last Checkpoint: 26 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Upcoming Earnings
ENG IN
23:57 28-04-2025

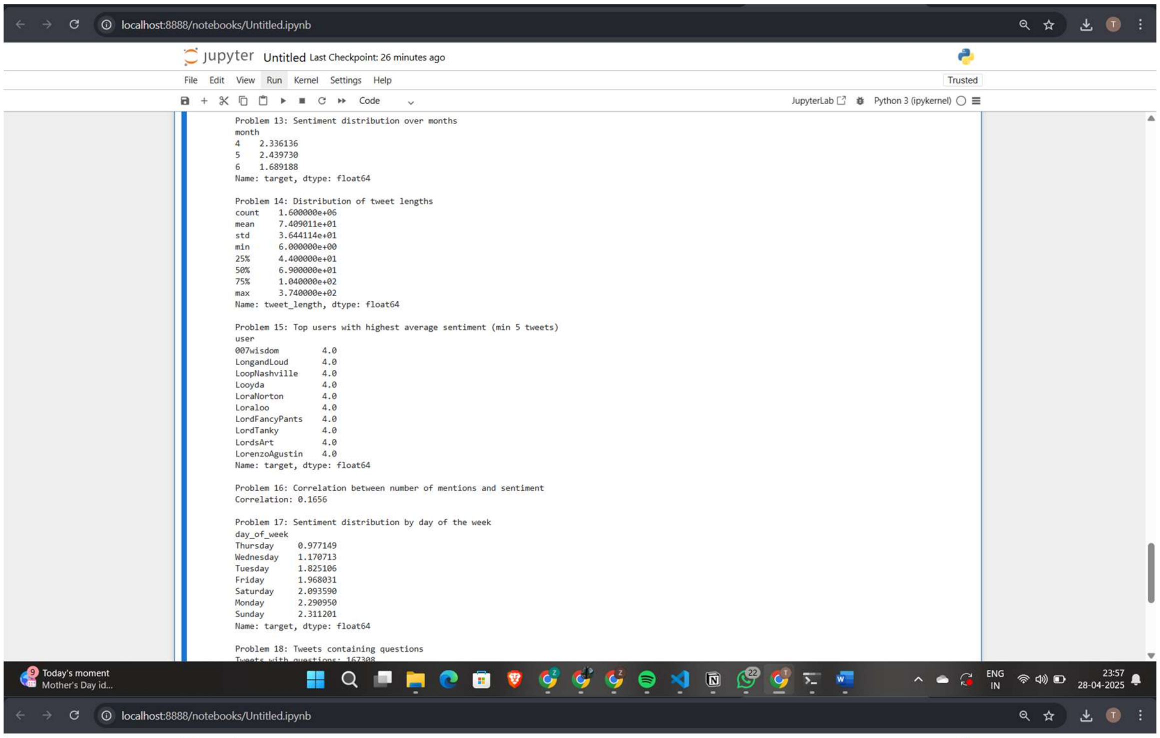
```
Problem 6: Sentiment by time of day
hour
0     2.239931
1     2.374821
2     2.368937
3     2.291409
4     2.184505
5     2.078221
6     2.017316
7     1.981782
8     1.893953
9     1.829246
10    1.917448
11    1.945483
12    1.822392
13    1.790336
14    1.850179
15    1.746500
16    1.731945
17    1.738094
18    1.799607
19    1.838398
20    1.889763
21    1.911548
22    2.007456
23    2.135410
Name: target, dtype: float64

Problem 7: Tweets containing specific keywords
Tweets containing 'happy': 26065
Tweets containing 'sad': 28566
Tweets containing 'love': 61423
Tweets containing 'hate': 19086
Tweets containing 'good': 87523
Tweets containing 'bad': 26469

Problem 8: Average number of characters in hashtags
Average hashtag count: 0.03
Average hashtag length (only if hashtags exist): 7.77

Problem 9: Which hours of the day have the most tweet activity?
hour
0     80865
1     75268
2     73991
3     74253
4     76995
```


localhost:8888/notebooks/Untitled.ipynb
jupyter Untitled Last Checkpoint: 26 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Today's moment Mother's Day id...
ENG IN
23:57 28-04-2025

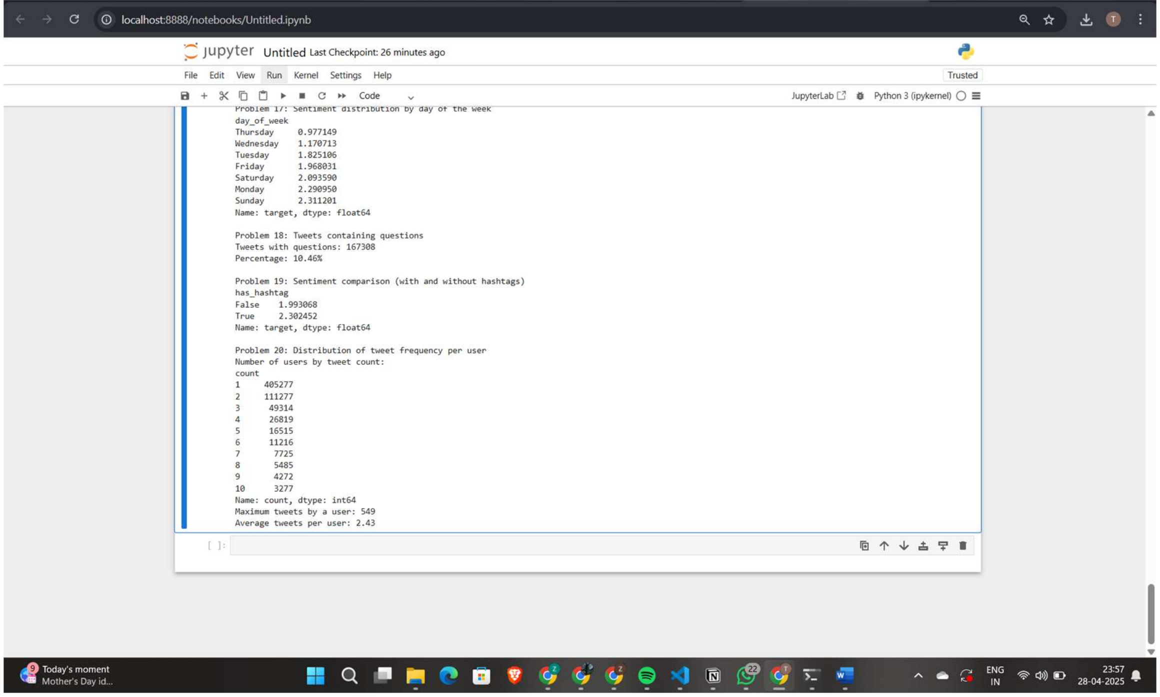
Problem 8: Average number of characters in hashtags
Average hashtag count: 0.03
Average hashtag length (only if hashtags exist): 7.77

Problem 9: Which hours of the day have the most tweet activity?
hour
0   80865
1   75268
2   73991
3   74253
4   76995

```
5     78623
6     80852
7     83654
8     76287
9     67278
10    60689
11    61009
12    51653
13    49689
14    50380
15    50643
16    55720
17    51843
18    53485
19    57722
20    57059
21    68964
22    78328
23    84750
Name: count, dtype: int64

Problem 10: Correlation between tweet length and sentiment
Correlation: -0.0058

Problem 11: Users with mixed sentiment tweets
Users with mixed sentiments: 132465
Percentage: 20.08%

Problem 12: Percentage of tweets containing URLs
URL Percentage: 4.38%

Problem 13: Sentiment distribution over months
month
4    2.336136
5    2.439730
6    1.689188
Name: target, dtype: float64
```



```
Problem 13: Sentiment distribution over months
month
4    2.336136
5    2.439730
6    1.689188
Name: target, dtype: float64

Problem 14: Distribution of tweet lengths
count    1.600000e+06
mean     7.409011e+01
std      3.644114e+01
min      6.000000e+00
25%      4.400000e+01
50%      6.900000e+01
75%      1.040000e+02
max      3.740000e+02
Name: tweet_length, dtype: float64

Problem 15: Top users with highest average sentiment (min 5 tweets)
user
007wisdom         4.0
LongandLoud       4.0
LoopNashville     4.0
Looyda            4.0
LoraNorton        4.0
Loraloo           4.0
LordFancyPants    4.0
LordTanky         4.0
LordsArt          4.0
LorenzoAgustin    4.0
Name: target, dtype: float64

Problem 16: Correlation between number of mentions and sentiment
Correlation: 0.1656

Problem 17: Sentiment distribution by day of the week
day_of_week
Thursday     0.977149
Wednesday    1.170713
Tuesday      1.825106
Friday       1.968031
Saturday     2.093590
Monday       2.290950
Sunday       2.311201
Name: target, dtype: float64

Problem 18: Tweets containing questions
Tweets with questions: 167308
```



```
Problem 17: Sentiment distribution by day of the week
day_of_week
Thursday     0.977149
Wednesday    1.170713
Tuesday      1.825106
Friday       1.968031
Saturday     2.093590
Monday       2.290950
Sunday       2.311201
Name: target, dtype: float64

Problem 18: Tweets containing questions
Tweets with questions: 167308
```

```
Percentage: 10.46%

Problem 19: Sentiment comparison (with and without hashtags)
has_hashtag
False    1.993068
True     2.302452
Name: target, dtype: float64

Problem 20: Distribution of tweet frequency per user
Number of users by tweet count:
count
1     405277
2     111277
3      49314
4      26819
5      16515
6      11216
7       7725
8       5485
9       4272
10      3277
Name: count, dtype: int64
Maximum tweets by a user: 549
Average tweets per user: 2.43
```