DOKUMENTASI TUGAS AKHIR DATA MINING A11.4408

Nama : Alfebrian Ivo Kurnia Adi  
NIM : A11.2021.13920

import numpy as np

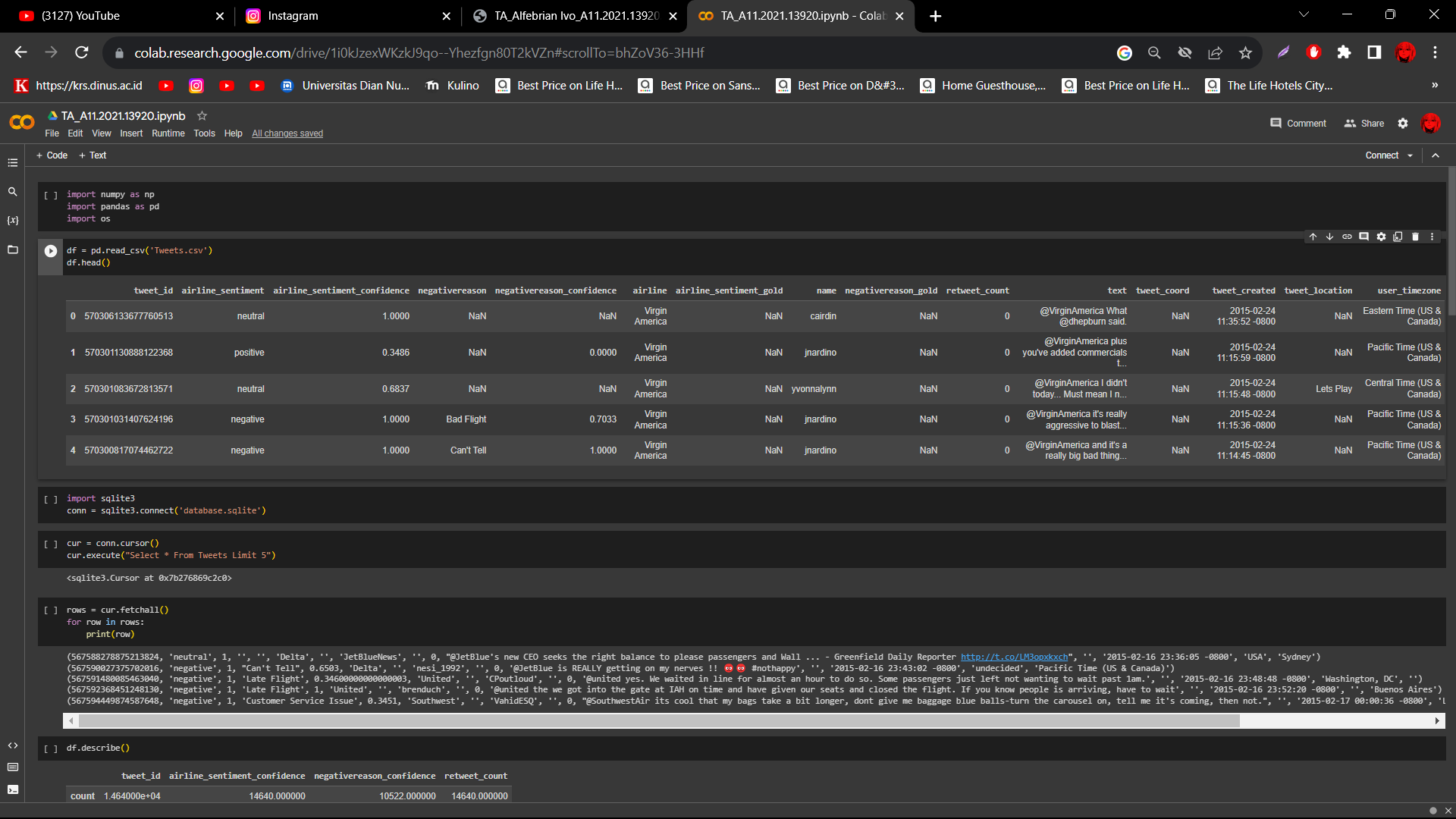
import pandas as pd

import os

df = pd.read\_csv('Tweets.csv')

df.head()

Output :



import sqlite3

conn = sqlite3.connect('database.sqlite')

cur = conn.cursor()

cur.execute("Select \* From Tweets Limit 5")

Output :

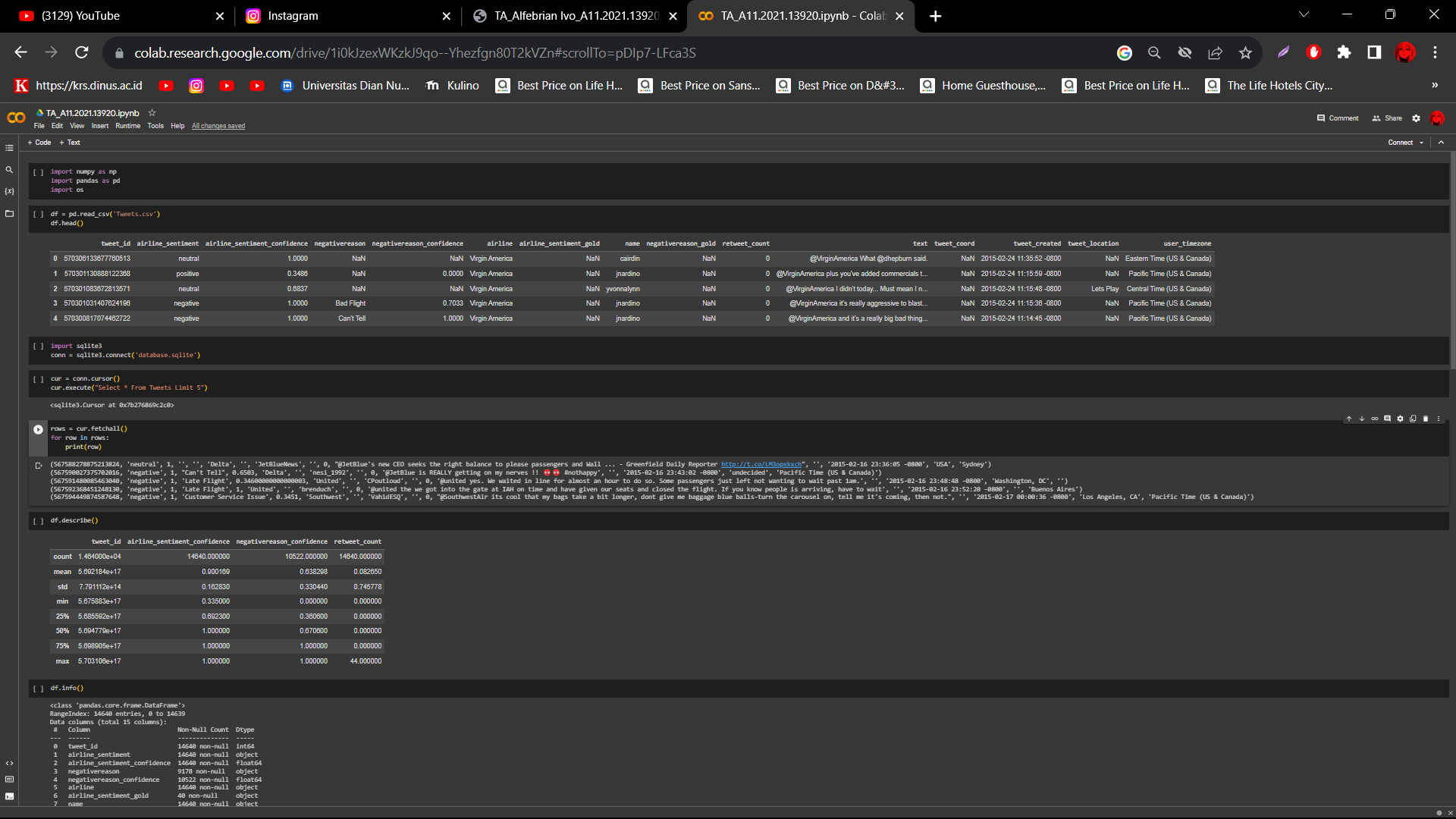
<sqlite3.Cursor at 0x7b276869c2c0>

rows = cur.fetchall()

for row in rows:

    print(row)

Output :



df.describe()

Output :

| **tweet\_id** | **airline\_sentiment\_confidence** | **negativereason\_confidence** | **retweet\_count** |
| --- | --- | --- | --- |
| **count** | 1.464000e+04 | 14640.000000 | 10522.000000 | 14640.000000 |
| **mean** | 5.692184e+17 | 0.900169 | 0.638298 | 0.082650 |
| **std** | 7.791112e+14 | 0.162830 | 0.330440 | 0.745778 |
| **min** | 5.675883e+17 | 0.335000 | 0.000000 | 0.000000 |
| **25%** | 5.685592e+17 | 0.692300 | 0.360600 | 0.000000 |
| **50%** | 5.694779e+17 | 1.000000 | 0.670600 | 0.000000 |
| **75%** | 5.698905e+17 | 1.000000 | 1.000000 | 0.000000 |
| **max** | 5.703106e+17 | 1.000000 | 1.000000 | 44.000000 |

df.info()

Output :

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 14640 entries, 0 to 14639

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 tweet\_id 14640 non-null int64

1 airline\_sentiment 14640 non-null object

2 airline\_sentiment\_confidence 14640 non-null float64

3 negativereason 9178 non-null object

4 negativereason\_confidence 10522 non-null float64

5 airline 14640 non-null object

6 airline\_sentiment\_gold 40 non-null object

7 name 14640 non-null object

8 negativereason\_gold 32 non-null object

9 retweet\_count 14640 non-null int64

10 text 14640 non-null object

11 tweet\_coord 1019 non-null object

12 tweet\_created 14640 non-null object

13 tweet\_location 9907 non-null object

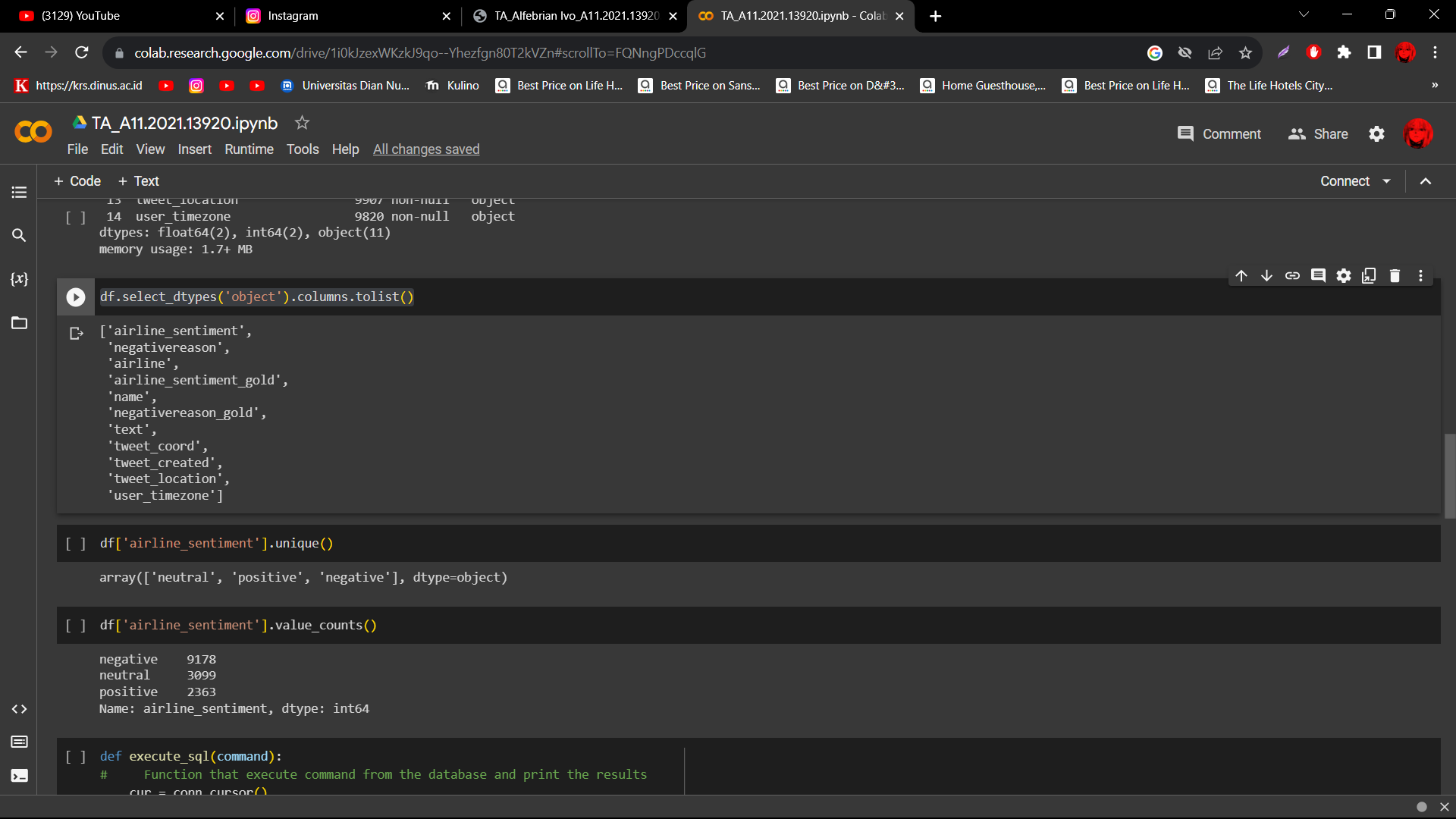
14 user\_timezone 9820 non-null object

dtypes: float64(2), int64(2), object(11)

memory usage: 1.7+ MB

df.select\_dtypes('object').columns.tolist()

Output :



df['airline\_sentiment'].unique()

Output :

array(['neutral', 'positive', 'negative'], dtype=object)

df['airline\_sentiment'].value\_counts()

Output :

negative 9178   
neutral 3099   
positive 2363   
Name: airline\_sentiment, dtype: int64

def execute\_sql(command):

#     Function that execute command from the database and print the results

    cur = conn.cursor()

    cur.execute(command)

    rows = cur.fetchall()

    for row in rows:

        print(row)

execute\_sql("Select DISTINCT(airline\_sentiment) From Tweets")

('neutral',)

('negative',)

('positive',)

Output :

('neutral',)

('negative',)

('positive',)

('positive',)

execute\_sql("Select airline\_sentiment, COUNT(airline\_sentiment) From Tweets GROUP BY airline\_sentiment")

('negative', 9082)

('neutral', 3069)

('positive', 2334)

Output :

('negative', 9082)

('neutral', 3069)

('positive', 2334)

('positive', 2334)

pd.read\_sql("Select DISTINCT(airline\_sentiment) From Tweets", conn)

Output :

|  | **airline\_sentiment** |
| --- | --- |
| **0** | neutral |
| **1** | negative |
| **2** | positive |

df\_neg = df.loc[(df.airline\_sentiment == 'negative') & (df.negativereason != "Can't Tell")].reset\_index(drop=True)

df\_neg.shape

Output :

(7988, 15)

df\_neg\_sql = pd.read\_sql('''Select \* From Tweets WHERE airline\_sentiment = 'negative' \

                        AND negativereason != "Can't Tell"''', conn)

df\_neg\_sql.shape

Output :

(7906, 15)

df\_neg.groupby('negativereason').negativereason\_confidence.mean()

Output :

negativereason   
Bad Flight 0.631731   
Cancelled Flight 0.783096   
Customer Service Issue 0.780054   
Damaged Luggage 0.733432   
Flight Attendant Complaints 0.659639   
Flight Booking Problems 0.606797   
Late Flight 0.768907   
Lost Luggage 0.813019   
longlines 0.594076   
Name: negativereason\_confidence, dtype: float64

pd.read\_sql('''Select negativereason, AVG(negativereason\_confidence) AS average\_confidence From Tweets \

            WHERE airline\_sentiment = 'negative' \

            AND negativereason != "Can't Tell" GROUP BY negativereason''', conn)

Output :

| **negativereason** | **average\_confidence** |
| --- | --- |
| **0** | Bad Flight | 0.630785 |
| **1** | Cancelled Flight | 0.783200 |
| **2** | Customer Service Issue | 0.779946 |
| **3** | Damaged Luggage | 0.734204 |
| **4** | Flight Attendant Complaints | 0.658255 |
| **5** | Flight Booking Problems | 0.607153 |
| **6** | Late Flight | 0.768978 |
| **7** | Lost Luggage | 0.812209 |
| **8** | longlines | 0.593856 |