DHcode.org

resources | pandas methods

math methods

.value_counts( )

Counts a list of the same value, e.g., how many times does the word “refugee” appear?
Use this function if you want to know how many values exist.

Pro-tip: The **.count** with **.value_counts( )** method totals all of the tweets. In contrast, the **.contains** with **.sum( )** method sums the frequencies for each tweet.

```
df_tweet = df["text"]

ref_count = df_tweet.str.count("refugee").value_counts()

total_refs = df_tweet.str.contains("refugee").sum()

print("The total number of times 'refugee' appears in all of the tweets: " + str(total_refs))
print("\nThe word 'refugee' appears this number of times per tweet: \n" + str(ref_count))
```

```
The total number of times 'refugee' appears in all of the tweets: 491

The word 'refugee' appears this number of times per tweet:
0    509
1    442
2     47
4      1
3      1
Name: text, dtype: int64
```

math methods

.sum( )

Adds up a list of the same value, e.g., how many total followers and friends are in the sample data?
Use this function if you want to sum up one value in the column.

```
all = df["followers"].sum()
print(all)
```

```
14168083
```

```
all = df["friends"].sum()
print(all)
```

```
1317515
```

math methods

.describe( )

Gives a breakdown of basic statistics, e.g., count, mean, standard deviation, etc., of integer and float data types.

```
pals = df["followers"].describe()
pals.head(3)
```

```
count      1000.000000
mean      14168.083000
std      175082.133164
Name: followers, dtype: float64
```

```
pals = df["friends"].describe()
pals.head(3)
```

```
count    1000.000000
mean     1317.515000
std      1856.820347
Name: friends, dtype: float64
```

data analysis

.sort_values( )

Sorts a list of the same value, e.g., user locations, number of followers, etc.
Use this function if you want to know how many values exist across one column.

You will need to point Pandas to the column(s) you want to sort, and how you want to sort them

```
test = df.sort_values("followers", ascending = False)
test.head(3)
```

created_at	followers	friends	geolocation	id	original_text	text	tweet_type	user_id	user_location	user_name	user_screen_name	verified
2015-09-04 31:01+00:00	2774679	835	NaN	6.400000e+17	NaN	German Police Forced To Stop Accepting Refugee...	post	100077645	Los Angeles, California	PopWrapped	PopWrapped	True
2015-09-05 01:12+00:00	2772372	837	NaN	6.400000e+17	NaN	ICYMI Germans Forced To Stop Accepting Refugee...	post	100077645	Los Angeles, California	PopWrapped	PopWrapped	True
2015-09-06 00:47+00:00	2771642	845	NaN	6.410000e+17	NaN	Finnish Prime Minister Offers Private Home To ...	post	100077645	Los Angeles, California	PopWrapped	PopWrapped	True

Pro tip:
Sort the list by ascending order (True) or descending order (False)

data analysis

.sort_values( )

Sorts <u>two or more lists</u> of the same value, e.g., number of friends AND verified status etc.
Use this function if you want to know how many values exist across multiple columns.

```
test = df.sort_values(["verified", "friends"])
test.tail(3)
```

_at	followers	friends	geolocation	id	original_text	text	tweet_type	user_id	user_location	user_name	user_screen_name	verified
09 00	2314	1954	NaN	6.420000e+17	#Gevgelija needs support to face extraordinary...	RT @AFracassetti: #Gevgelija needs support to ...	share	1000093272	NaN	UNDP MK	UNDPMK	True
04 00	4680	1982	NaN	6.400000e+17	Hungarian citizens donated strollers to migran...	RT @TheLeadCNN: Hungarian citizens donated str...	share	1002161534	Denver, Colorado	Laura Keeney	LauraKeeney	True
05 00	4683	1982	NaN	6.400000e+17	Syrian refugee kids getting sweets from the ge...	RT @melissarfleming: Syrian refugee kids getti...	share	1002161534	Denver, Colorado	Laura Keeney	LauraKeeney	True

```
test = df.sort_values(["friends", "verified"])
test.tail(3)
```

followers	friends	geolocation	id	original_text	text	tweet_type	user_id	user_location	user_name	user_screen_name	verified
7257	7848	NaN	6.400000e+17	Migrant crisis: Egyptian billionaire offers to...	RT @ohboywhatashot: Migrant crisis: Egyptian b...	share	1002174228	Republic of Perthshire	ScotsVsAusterity	ScotIndyDebate	False
7262	7848	NaN	6.400000e+17	THE GUARDIAN FRONT PAGE: 'Cameron: we won't jo...	RT @SkyNews: THE GUARDIAN FRONT PAGE: 'Cameron...	share	1002174228	Republic of Perthshire	ScotsVsAusterity	ScotIndyDebate	False
8256	8284	NaN	6.390000e+17	NaN	lana banana https://t.co /ZE8XPbQ0uL	post	1000461469	Beacon Hills	«crybaby»	_owhstyles	False

Pro tip: The order of the columns matters here. See the difference? The first column is sorted, then the next, etc., which can return different results.

data analysis

.crosstab( )

Compares values across multiple columns, e.g., how many times did a verified user tweet (post) or retweet (share)? Use this method if you want to compare the frequency (how often) values occur between different columns of data

```
pd.crosstab(index=df["tweet_type"], columns=[df["verified"]], margins=True)
```

verified	False	True	All
tweet_type			
post	349	9	358
share	632	10	642
All	981	19	1000