Read CSV files, select columns, extract category and save to new CSVs

In [1]:

```
import os
import numpy as np
import pandas as pd
import json
```

In [2]:

```
base_path = '../data/raw/kickstarter_csvs'
target_path = '../data/interim/kickstarter_csvs'
```

In [3]:

```
# list of columns to select and save
properties = ['id', 'name', 'goal', 'pledged', 'usd_pledged', 'state', 'slug', 'disable_communication', 'country', 'currency', 'deadline', 'state_changed_at', 'created_at', 'launched_at', 'staff_pick', 'backers_count', 'blurb', 'category', 'spotlight']
properties = ['id', 'name', 'goal', 'pledged', 'usd_pledged', 'state', 'slug', 'disable_communication', 'country', 'currency', 'deadline', 'state_changed_at', 'created_at', 'launched_at', 'staff_pick', 'backers_count', 'blurb', 'spotlight']
len(properties)
```

Out[3]:

```
18
```

In [5]:

```
# read all csvs, select desired columns, and save as new csvs in the same format
# extract category and location from JSON strings
new_dfn = [] # list of new dataframes for concatenation
for folderName, subfolders, filenames in os.walk(base_path):
    _, dname = os.path.split(folderName)
    dname = os.path.join(target_path, dname)
    #os.mkdir(dname)
    for filename in filenames:
        if filename.endswith('.csv'):
            csv_fname = os.path.join(folderName, filename)
            #print("File", csv_fname)
            dfs = pd.read_csv(csv_fname)                       # read into DataFrame
            y = dfs['category'].map(lambda x: json.loads(x)['slug'])  # parse JSON

            dfn = dfs.reindex(columns=properties, copy=True)  # create a new dataframe
            new_dfn.append(dfn.assign(category=y.values))     # add a parsed category
            #print('New DF object:', new_dfn[-1].head(1))
            # save into separate csv files
            #dfn = dfs.reindex(columns=properties, copy=True) # create a new dataframe
            #dfn = dfn.assign(category=y.values)              # add a parsed category
            #dfn.to_csv(os.path.join(dname, filename))        # save new dataframe
            #print("Newfile: ", os.path.join(dname, filename))
            #print(dfs.columns)
```

In [5]:

```
len(new_dfn)
```

Out[5]:

```
961
```

In [6]:

```
df_single = pd.concat(new_dfn, ignore_index=True)
```

In [7]:

```
df_single.shape
```

Out[7]:

```
(3935527, 19)
```

```
In [9]: df_single.head()
```

Out[9]:

	id	name	goal	pledged	usd_pledged	state	slug	disable_communicatio
0	64486721	Along The Lines Of...	300.0	300.0	460.241994	successful	along-the-lines-of	False
1	755137951	Portrait of #NOW	500.0	595.0	595.000000	successful	portrait-of-now	False
2	796895846	A Dollar and a Dream	300.0	1071.0	1071.000000	successful	a-dollar-and-a-dream-0	False
3	2136864323	Correspondences: The Exhibition	1600.0	1735.0	1735.000000	successful	correspondences-the-exhibition	False
4	989395377	Abstraction of Utopia	750.0	760.0	760.000000	successful	abstraction-of-utopia	False

```
In [33]: df_single.to_csv(os.path.join(target_path, 'kick_all.csv'))

In [10]: dup = df_single.duplicated()

In [12]: from collections import Counter
         Counter(dup)

Out[12]: Counter({False: 325967, True: 3609560})

In [25]: df_alldup = df_single.drop_duplicates()

In [35]: df_alldup.to_csv(os.path.join(target_path, 'kick_nodup.csv'))

In [36]: df_alldup.shape

Out[36]: (325967, 19)

In [41]: df_iddup = df_single.drop_duplicates(['id'], keep='last')

In [42]: df_iddup.shape

Out[42]: (263765, 19)

In [43]: df_iddup.to_csv(os.path.join(target_path, 'kick_noiddup_last.csv'))        # save new dataframe

In [54]: df_id = df_iddup.set_index('id').sort_index()
```

In [58]:
```
df_id.head(2)
```

Out[58]:

	name	goal	pledged	usd_pledged	state	slug	disable_communication	country	currency
id									
18520	Grandma's are Life	15000.0	62.0	62.000000	failed	grandmas-are-life	False	US	USD
21109	Meta	150.0	173.0	258.036032	successful	meta	False	GB	GBP

In [56]:
```
df_id.shape
```

Out[56]:
```
(263765, 18)
```

In [57]:
```
df_id.to_csv(os.path.join(target_path, 'kick_id.csv'))        # save new dataframe
```

- The above is the processed single CSV file containing unique data over the history of kickstarter