Finding Reptile Photos using Social Engineering (and Python – the programming language, not the snake)" including the above codes of Python

Introduction:

In the realm of extraction and data analysis, the energy of social engineering and Python programming offers a powerful tool kit that plays a significant role in data analysis. This report searches for an application i.e., extraction of reptile photos and author's emails from bibliographic databases. The goal is to demonstrate the integration of programming skills with social engineering strategies process of photos and email retrieval.

Methods:

The data engineering and analysis of the bibliography dataset was done through the following process:

i. **Data Preparation**:

It is a starting step that involves preparing the data. The Python script begins with importing necessary libraries such as Pandas, NumPy, and Pandas, in data manipulation and re. The Panada library plays an important role in data engineering and data manipulation, NumPy for numerical operation, and re for regular expressions. The script then reads an Excel file data into DataFrame, creating a new copy and removing the duplicated columns.

Example: #
Import necessary libraries
import pandas as pd
import numpy as np
import re
# Set the pandas option to display full column width
pd.set_option('display.max_colwidth', None)
# Read an excelfile into Pandas DataFrame
df1=pd.read_csv('C:/Users/bhojr/Desktop/Reptile_bibliography/Reptile_final_filled_new_email.csv')
# ... (Continue with the code for data preparation)

ii. **Text cleaning**: The 'clean_corpus function, is defined to clean the text data by converting them into lowercase, removing non-alphanumeric characters, and eliminating the extra spaces. This function is used to email and 'author columns'.

Example: def clean_corpus(nlp):
words= nlp.lower()
mytext=re.sub(r'[^a-zA-Z0-9\,\@\.]', ' ',words)

```
 mytext=re.sub(r' +', ' ', mytext) return mytext.strip()
#Apply the 'clean_corpus' function to the email and author columns
df['email_clean'] = df.email.apply(clean_corpus)
df['author_clean'] = df.author.apply(clean_corpus)
```

iii. **Author and Email Parsing**: The script further splits the data by author name and email address components, creating separate columns for each. Social engineering comes into play when determining the likelihood of an author's email association by comparing components.

iv. **Author-Email Mapping**: A dictionary 'author_dict', is created to map authors to email addresses, enabling efficient replacement of author names with corresponding emails in DataFrame.

v. **Email Extraction and Concatenation**: Another function, 'find_email', is defined to extract email addresses from text using regular expressions. This function is applied to specified columns and the results are concatenated into a new 'email_final' column.

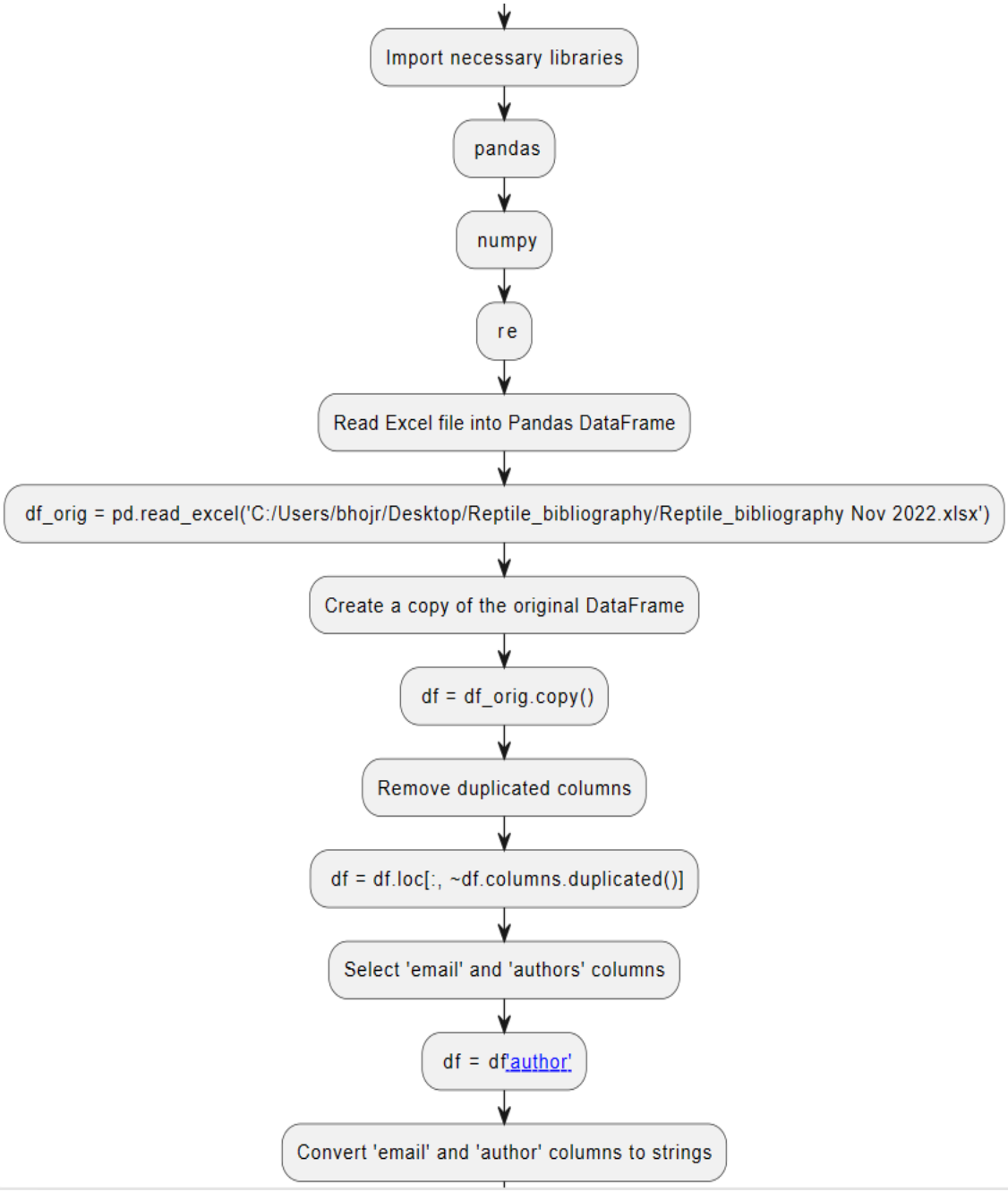
vi. **Data Merging**: The scripts demonstrate the merging of multiple DataFrame based on a common column, 'reference-numbers.' It also includes steps to handle problematic columns containing newline characters.

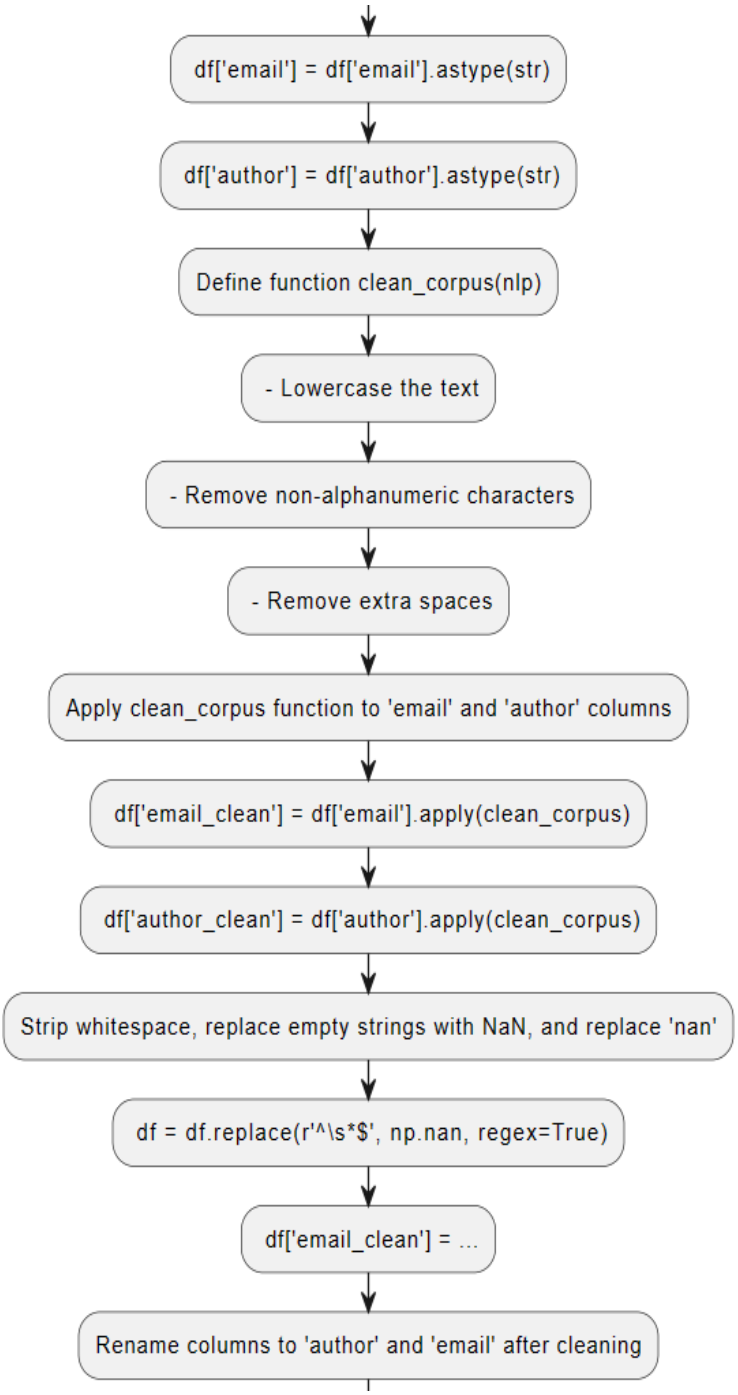
vii. **Data Filtering**: A key aspect of the script involves filtering the data based on specific criteria, such as the number of photos and related author's emails.

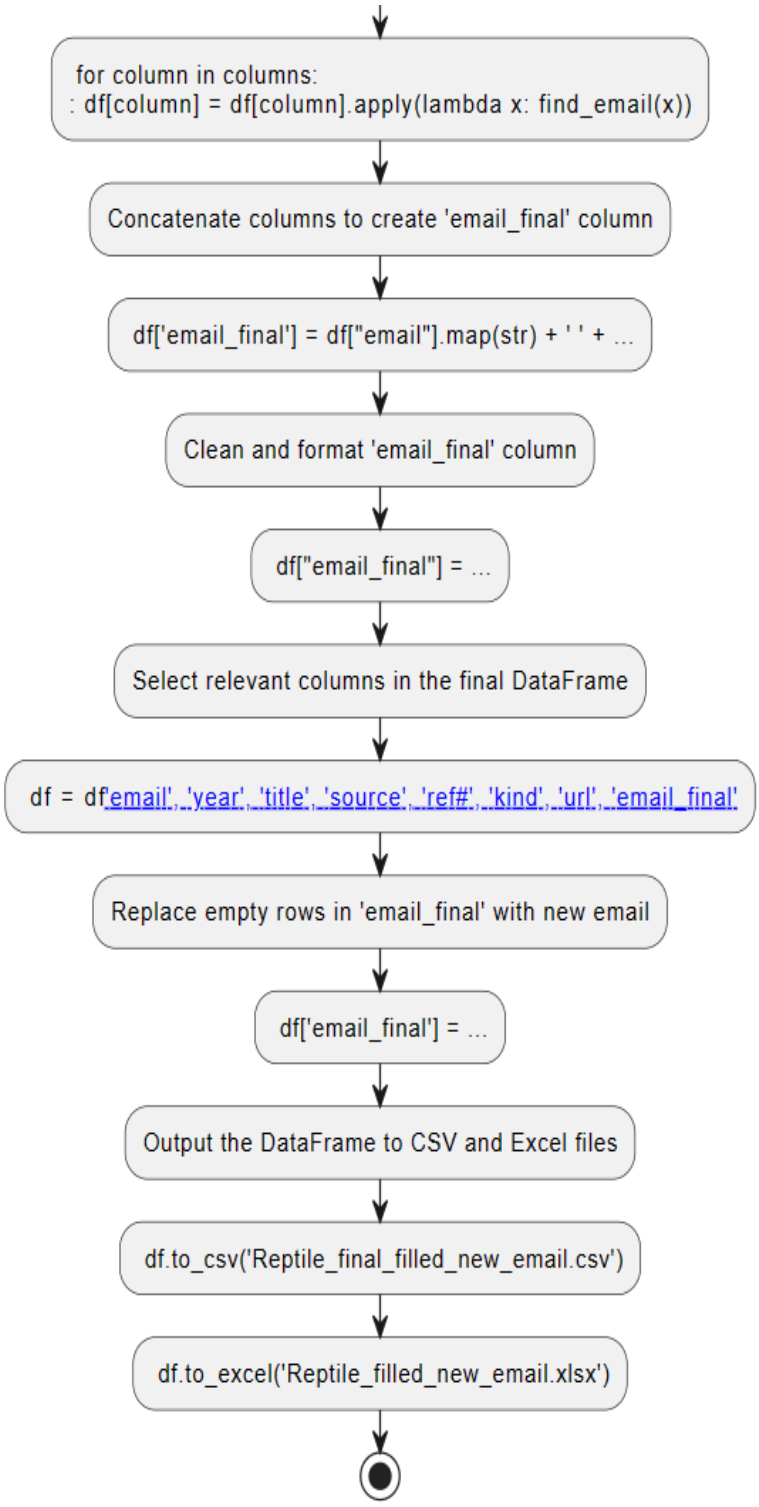
Results:

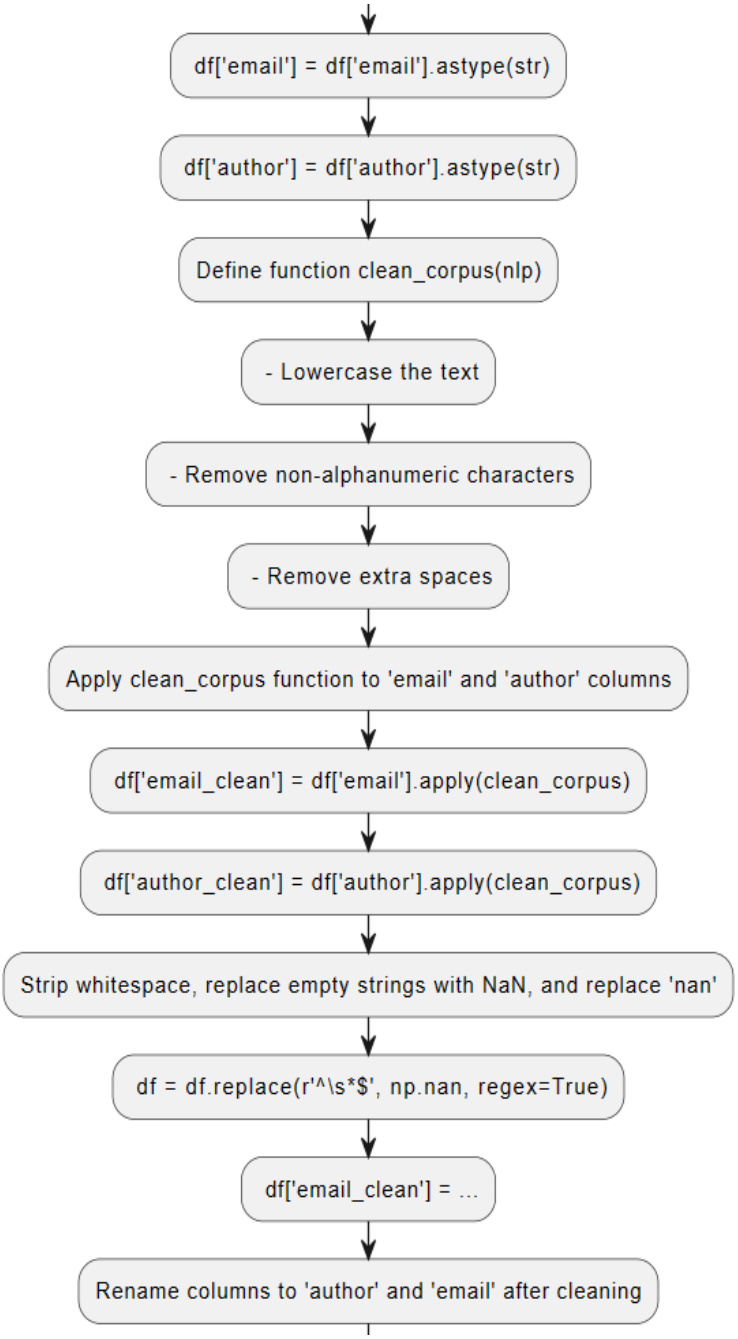
The Python script successfully executes the defined steps, showing the effectiveness of programming skills with social engineering strategies. The resulting DataFrame provides valuable insights into the relationship between authors, emails, photos, and references. This bibliography data analysis filled out 6386 emails in empty rows and mapped 319 images with their emails using the Python codes.

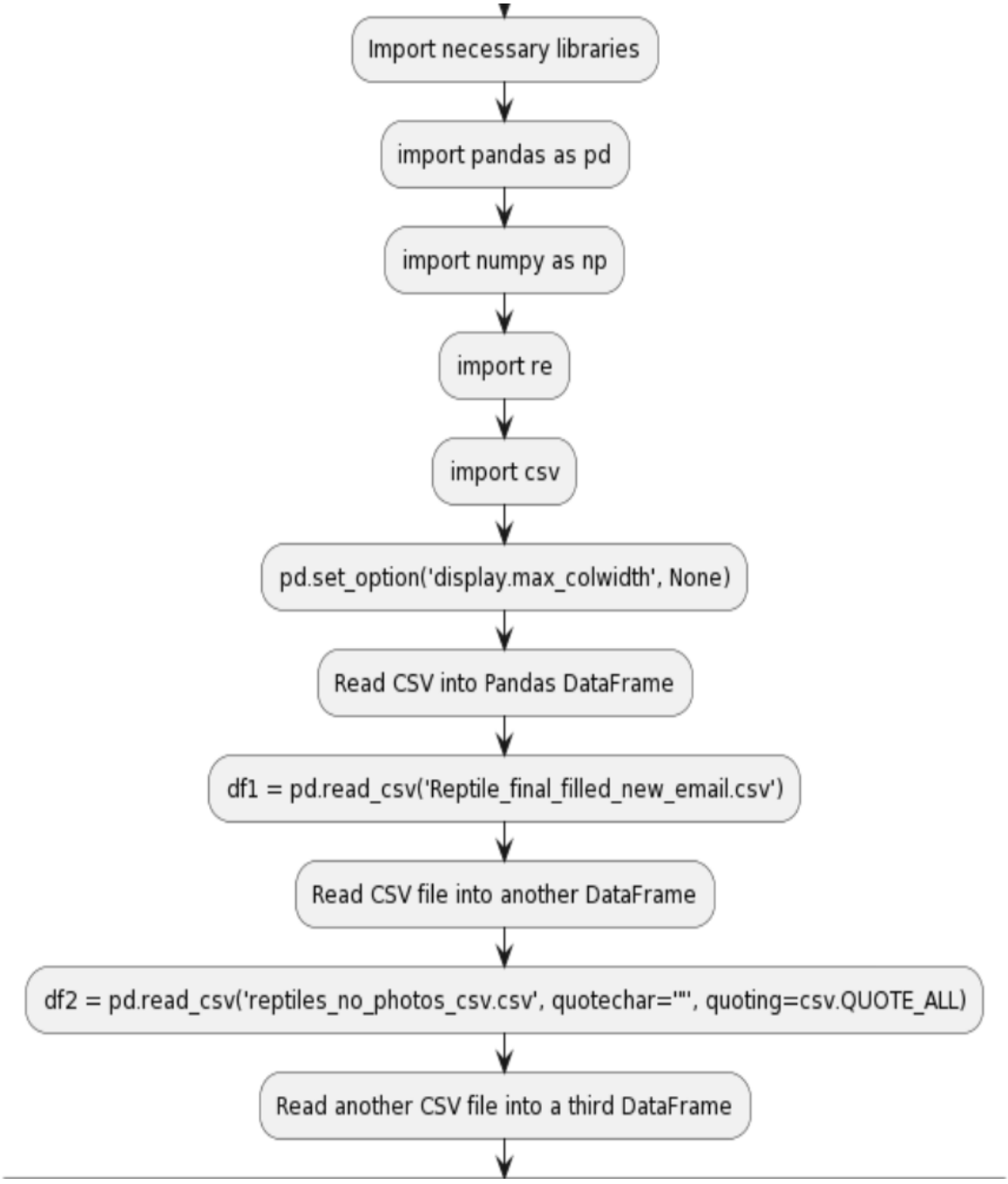
Flow Chart (Graphical Representation of the Codes)

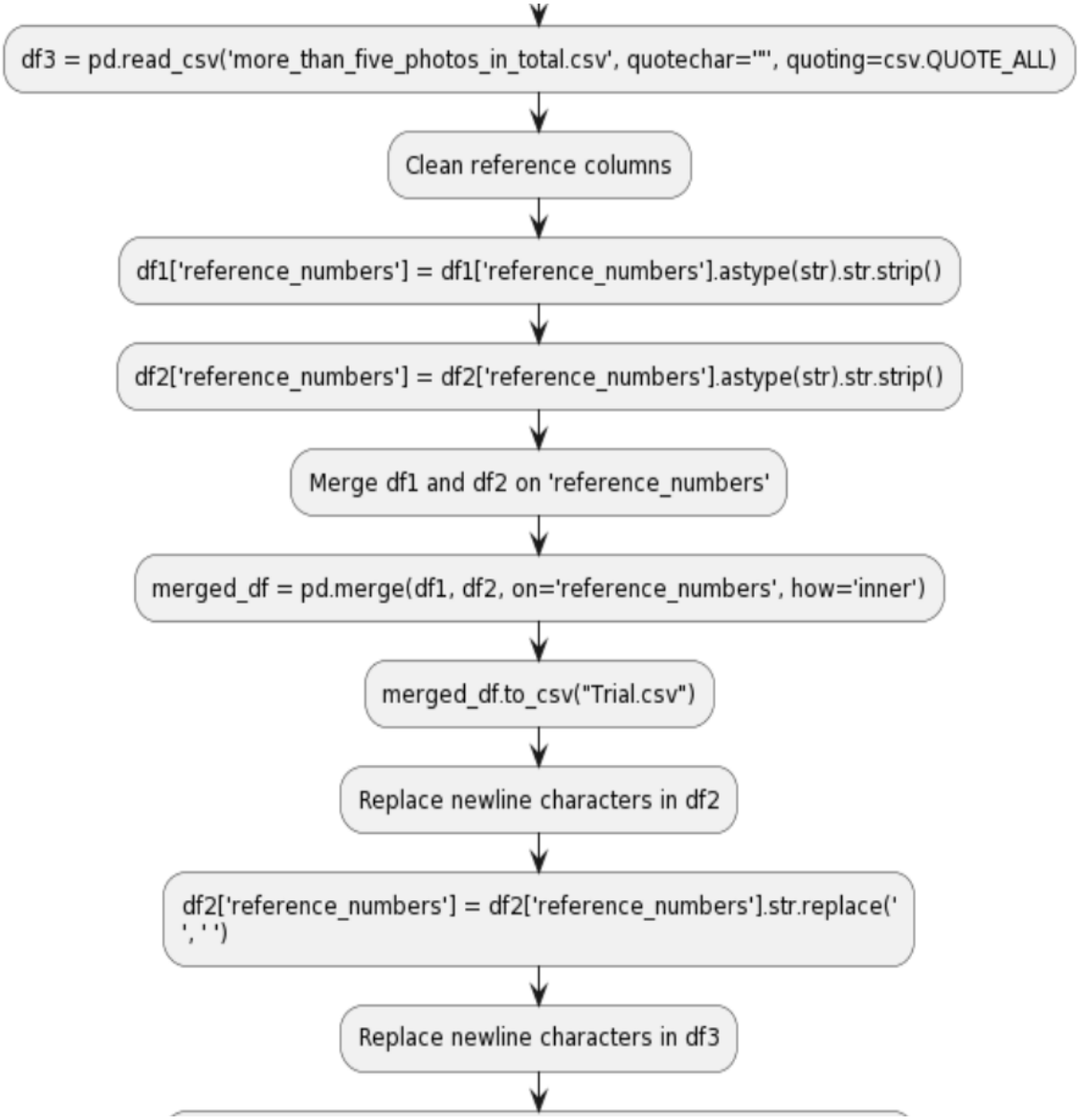


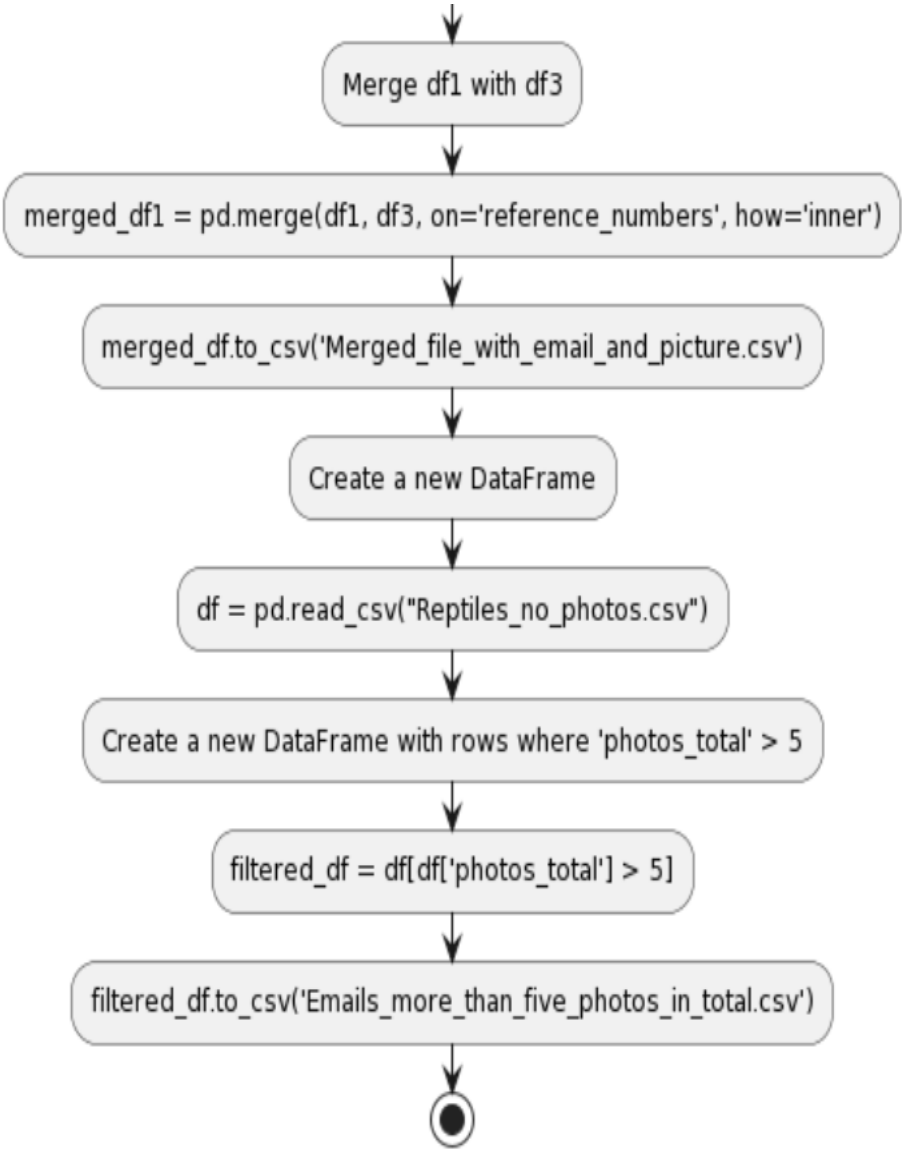

df['email'] = df['email'].astype(str)
df['author'] = df['author'].astype(str)
Define function clean_corpus(nlp)
- Lowercase the text
- Remove non-alphanumeric characters
- Remove extra spaces
Apply clean_corpus function to 'email' and 'author' columns
df['email_clean'] = df['email'].apply(clean_corpus)
df['author_clean'] = df['author'].apply(clean_corpus)
Strip whitespace, replace empty strings with NaN, and replace 'nan'
df = df.replace(r'^\s*$', np.nan, regex=True)
df['email_clean'] = ...
Rename columns to 'author' and 'email' after cleaning

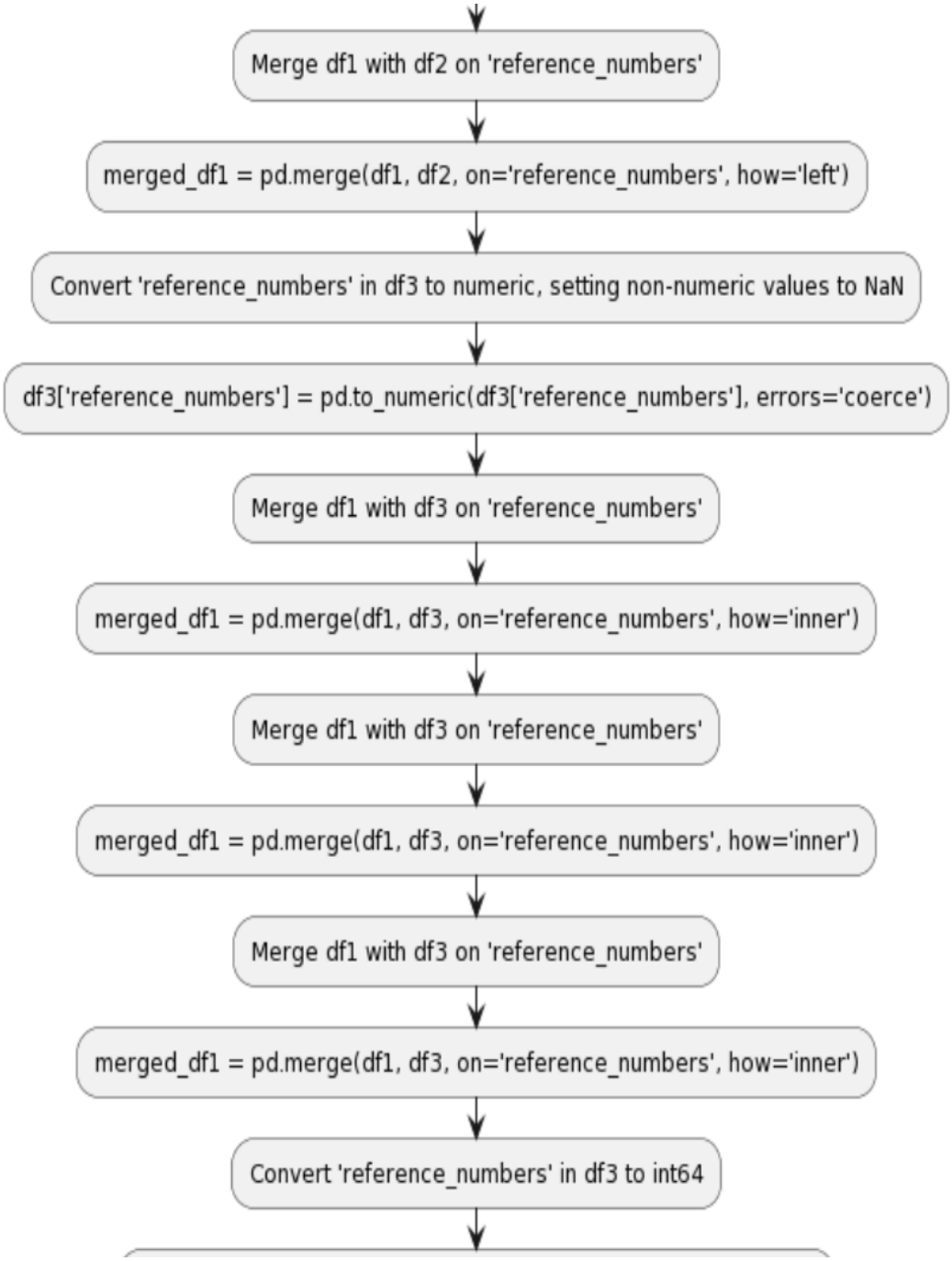

for column in columns:
: df[column] = df[column].apply(lambda x: find_email(x))
Concatenate columns to create 'email_final' column
df['email_final'] = df["email"].map(str) + ' ' + ...
Clean and format 'email_final' column
df["email_final"] = ...
Select relevant columns in the final DataFrame
df = df['email', 'year', 'title', 'source', 'ref#', 'kind', 'url', 'email_final']
Replace empty rows in 'email_final' with new email
df['email_final'] = ...
Output the DataFrame to CSV and Excel files
df.to_csv('Reptile_final_filled_new_email.csv')
df.to_excel('Reptile_filled_new_email.xlsx')

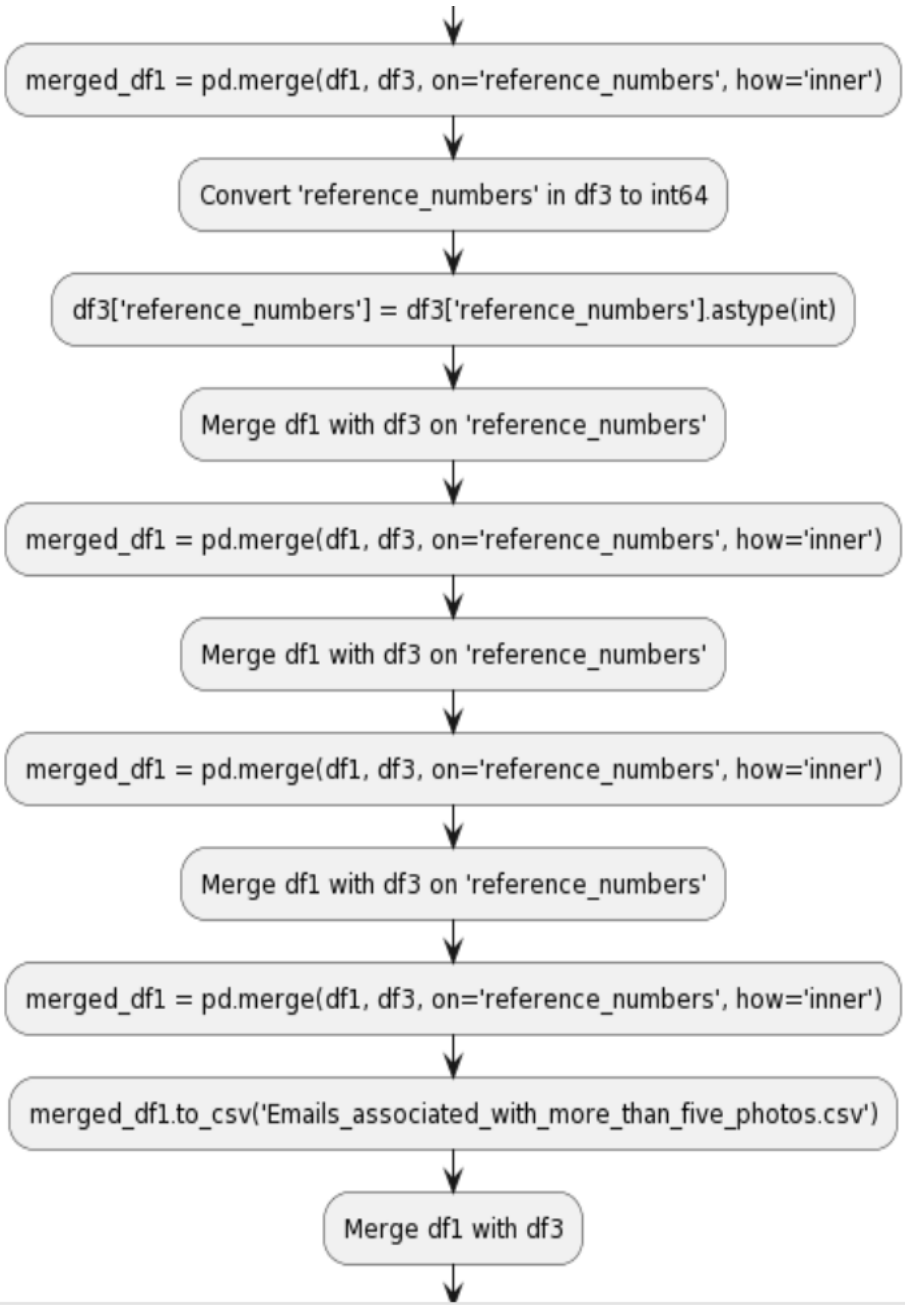

df['email'] = df['email'].astype(str)
df['author'] = df['author'].astype(str)
Define function clean_corpus(nlp)
- Lowercase the text
- Remove non-alphanumeric characters
- Remove extra spaces
Apply clean_corpus function to 'email' and 'author' columns
df['email_clean'] = df['email'].apply(clean_corpus)
df['author_clean'] = df['author'].apply(clean_corpus)
Strip whitespace, replace empty strings with NaN, and replace 'nan'
df = df.replace(r'^\s*$', np.nan, regex=True)
df['email_clean'] = ...
Rename columns to 'author' and 'email' after cleaning


Import necessary libraries
import pandas as pd
import numpy as np
import re
import csv
pd.set_option('display.max_colwidth', None)
Read CSV into Pandas DataFrame
df1 = pd.read_csv('Reptile_final_filled_new_email.csv')
Read CSV file into another DataFrame
df2 = pd.read_csv('reptiles_no_photos_csv.csv', quotechar='"', quoting=csv.QUOTE_ALL)
Read another CSV file into a third DataFrame


df3 = pd.read_csv('more_than_five_photos_in_total.csv', quotechar='"', quoting=csv.QUOTE_ALL)
Clean reference columns
df1['reference_numbers'] = df1['reference_numbers'].astype(str).str.strip()
df2['reference_numbers'] = df2['reference_numbers'].astype(str).str.strip()
Merge df1 and df2 on 'reference_numbers'
merged_df = pd.merge(df1, df2, on='reference_numbers', how='inner')
merged_df.to_csv("Trial.csv")
Replace newline characters in df2
df2['reference_numbers'] = df2['reference_numbers'].str.replace('
', ' ')
Replace newline characters in df3


Merge df1 with df3
merged_df1 = pd.merge(df1, df3, on='reference_numbers', how='inner')
merged_df.to_csv('Merged_file_with_email_and_picture.csv')
Create a new DataFrame
df = pd.read_csv("Reptiles_no_photos.csv")
Create a new DataFrame with rows where 'photos_total' > 5
filtered_df = df[df['photos_total'] > 5]
filtered_df.to_csv('Emails_more_than_five_photos_in_total.csv')


Merge df1 with df2 on 'reference_numbers'
merged_df1 = pd.merge(df1, df2, on='reference_numbers', how='left')
Convert 'reference_numbers' in df3 to numeric, setting non-numeric values to NaN
df3['reference_numbers'] = pd.to_numeric(df3['reference_numbers'], errors='coerce')
Merge df1 with df3 on 'reference_numbers'
merged_df1 = pd.merge(df1, df3, on='reference_numbers', how='inner')
Merge df1 with df3 on 'reference_numbers'
merged_df1 = pd.merge(df1, df3, on='reference_numbers', how='inner')
Merge df1 with df3 on 'reference_numbers'
merged_df1 = pd.merge(df1, df3, on='reference_numbers', how='inner')
Convert 'reference_numbers' in df3 to int64



The flow chart summarizes the workflow of data analysis from the initial data presentation to the result output level.

S.No	author	email	year	title	source	reference_	kin	url	email_final			
1	bauer, a.m.	aaron.bauer@villanova.edu	1990	Phylogenetic systematics and biogeography of the Carph	Bonner zo	368		http://bio	aaron.bauer@villanova.edu			
2	bauer a m	aaron.bauer@villanova.edu	1990	Pachydactylus mariquensis latirostris.	J. Herp. As	382		http://ww	aaron.bauer@villanova.edu			
3	bauer, aaron m. vindum, jens v.	aaron.bauer@villanova.edu	1990	A checklist and key to the herpetofauna of New Caledon	Proc. Cal.	14254	p	http://ww	aaron.bauer@villanova.edu			
4	bauer, a. m. russell, a. p. shadwick,	aaron.bauer@villanova.edu	1990	Skin mechanics and morphology of the gecko Sphaeroda	American	14255			aaron.bauer@villanova.edu			
5	bauer, aaron m. russell, anthony p.	aaron.bauer@villanova.edu	1990	Recent advances in the search for the living giant gecko	Cryptozoo	14256			aaron.bauer@villanova.edu			
6	bauer, a. m.	aaron.bauer@villanova.edu	1990	Phylogeny and biogeography of the geckos of southern A	Bonn: Zoo	19687	c		aaron.bauer@villanova.edu			
7	russell, a.p., bauer, a.m.	aaron.bauer@villanova.edu	1990	Hypertrophied phalangeal chondroepiphyses in the gekk	Journal of	19914		https://do	aaron.bauer@villanova.edu	arussell@ucalgary.ca		
8	russell, a. p., and a. m. bauer.	aaron.bauer@villanova.edu	1990	Digit I in pad-bearing gekkonine geckos: alternative desig	Memoirs c	26519			aaron.bauer@villanova.edu	arussell@ucalgary.ca		
9	russell, a. p., and a. m. bauer.	aaron.bauer@villanova.edu	1990	Oedura and Afroedura (Reptilia: Gekkonidae) revisited: s	Memoirs c	26520			aaron.bauer@villanova.edu	arussell@ucalgary.ca		
10	dubois, alain	adubois@mnhn.fr	1990	Nomenclature of parthenogenetic, gynogenetic and "hyt	Alytes 8 (3	15224			adubois@mnhn.fr			
11	hernando, a. y b. alvarez.	ahernando@infovia.com.ar	1990	CARIOTIPO DE Mabuya frenata (COPE, 1862) (SAURIA, S	Facena 8:	31192		http://exa	ahernando@infovia.com.ar			
12	disi, a.m.	ahmadmdisi@yahoo.com	1990	Venomous snakes in Jordan. In, Snakes of Medical Impo	National U	26746			ahmadmdisi@yahoo.com			
13	mu oz alonso, luis antonio	amunoz@ecosur.mx	1990	Estudio Herpetofaunistico del Parque Ecologico Estatal	BoletÃ-n d	13238		http://soc	amunoz@ecosur.mx			
14	rasmussen,a.r. andersen,m.	arr@kons.dk	1990	The sea snake Kerilia jerdoni Gray (1849): First records fr	The Snake	4492	p		arr@kons.dk			
15	cattaneo, a.	augustocattaneo@hotmail.com	1990	I serpenti delle isole greche di Kythnos e Kea (Cicladi occ	Att. Soc. it	33419	p	http://vip	augustocattaneo@hotmail.com			
16	flaschendr ger, axel	axel.flaeschendraeger@t	1990	Anolis bahorucoensis bahorucoensis (Noble & Hassler, 1	Herpetofa	13210	s	http://ww	axel.flaeschendraeger@t			
17	orlov,n.l. b.s. tuniyev.	azemiops@zin.ru	1990	Three Species in the Vipera kaznakowi Complex (Eurosib	Asiatic Her	4074	p	http://bio	azemiops@zin.ru			
18	bergna, s. y b. b. alvarez.	balvarez@exa.unne.edu.ar	1990	COMPOSICIÃ"N Y DISTRIBUCIÃ"N DE LA OFODIOFAUNA	Facena 8:	31193		http://exa	balvarez@exa.unne.edu.ar			
19	brattstrom, b. h.	bayard@hughes.net	1990	Biogeography of the Islas Revillagigedo, Mexico.	J. Biogeog	14429	rm	https://do	bayard@hughes.net			
20	sch tti, beat kramer, eugen touzet, j	beatschaetti@hotmail.com	1990	Systematic remarks on a rare crotalid snake from Ecuad	Revue Suis	9386		http://ww	beatschaetti@hotmail.com			
21	schatti, b. guillod, m.	beatschaetti@hotmail.com	1990	Bemerkungen zur Rassengliederung bei der Philippinische	Herpetofa	11459	s	http://ww	beatschaetti@hotmail.com			
22	beck, d.d.	beckd@cwu.edu	1990	Ecology and behavior of the Gila monster in southweste	Journal of	24532		http://ww	beckd@cwu.edu			
23	eidenm ller, b.	bernd.eidenmueller@t	1990	Beobachtungen bei der Haltung und Nachzucht von Vara	Salamandr	16135	o	http://sala	bernd.eidenmueller@t			
24	aguilar cortes r., camarillo r j l. and l	bezy@comcast.com	1990	Distribution, species status and reproductive mode of th	Southwest	19		http://ww	bezy@comcast.com			
25	browne cooper, r. maryan, b.	brad.maryan@museum.wa.gov.	1990	Observations of Ctenotus angusticeps.	Herpetofa	13205		https://ww	brad.maryan@museum.wa.gov.au			
26	tuniyev, b.s.	btuniyev@mail.ru	1990	On the Independence of the Colchis Center of Amphibian	Asiatic Her	27898	p	http://bio	btuniyev@mail.ru			
27	villa j d wilson l d	bufodoc@aol.com	1990	Ungaliophis Muller. Central American dwarf boas.	Catalogue	8861	p	https://re	bufodoc@aol.com			
28	wilson l d	bufodoc@aol.com	1990	Tantilla striata Dunn.	Catalogue	8875	p	https://re	bufodoc@aol.com			
29	wilson l d	bufodoc@aol.com	1990	Tantilla oaxacae Wilson and Meyer.	Catalogue	8876	p	https://re	bufodoc@aol.com			
30	wilson l d	bufodoc@aol.com	1990	Tantilla insulamontana Wilson & Mena.	Catalogue	8877	p	https://re	bufodoc@aol.com			

Table: 1, Shows the output of the dataset of the initial thirty rows with a new email column

41926	peterson cr, giltz pd		2022	Reviewing Observations for the Idaho Amphibian and Re	Biodiversit	75843		https://doi.org/10.3897/biss.6.95052
41927	bour, roger and josef f. schmidtler.		2022	Nikolaus Michael Oppelâ€™s Drawings, Watercolors, an	ISHBH, Sal	75846		http://ww bour@mnhn.fr
41928	park s m, rahman mm, ham c h, sung h c		2022	The first record of an invasive reptile species, Pelomedus	Check List	75850		https://doi.org/10.15560/18.5.989
41929	joseph ouni, mehdi, cann, john william p. mccord.		2022	THE MORPHOLOGICAL IDENTITY OF ELSEYA DENTATA (T	The Batagı	75857	p o	
41930	joseph ouni, mehdi, william p. mccord cann, john		2022	A NEW SPECIES OF ELSEYA (TESTUDINES: CHELIDAE) FR	The Batagı	75858	p o	
41931	joseph ouni, mehdi, cann, john william p. mccord.		2022	STRANGERS IN THE RIVER: FIRST DOCUMENTED SYMPA	The Batagı	75859	p o	
41932	joseph ouni, mehdi, william p. mccord cann, john		2022	AN ILLUSTRATED GUIDE TO EXTERNAL MORPHOLOGICA	The Batagı	75860	p o	
41933	deshmukh ub, mungole aj, scanferla a, zaher h.		2022	Katariana nomen novum: a replacement name for the pr	Zootaxa 5	75862		oi: 10.11646/zootaxa.5178.6.7.
41934	srikanthan, a. n., adhikari, o. d., kumar mallik, a., campbell, p. d., bhu		2022	Taxonomic revaluation of the Ahaetulla prasina (H. Boie	European	75868	p	https://doi.org/10.5852/ejt.2022.839.1937
41935	liang t, wang l and shi l		2022	Sexual and natural selection interplay in sexual head sha	Front. Eco	75874		doi: 10.3389/fevo.2022.1016885
41936	mahrdt, c. r., k. r. beaman, j. h. valdez villavicencio, and t. j. papenfus		2022	Bipes biporus.	Catalogu	75876	p	
41937	ishihara, m.a., domingos, f.m.c.b., gomides, s.c. i. a. novelli, g. r. colli		2022	Genetic structure of Enyalius capetinga (Squamata, Leios	Genetica	75886		https://doi.org/10.1007/s10709-022-00170-w
41938	bernstein, j. m., voris, h. k., stuart, b. l., phimmachak, s., seateun, s., s		2022	Undescribed Diversity in a Widespread, Common Group	Ichthyolog	75889	p	https://www.researchgate.net/profile/Justin-Bernstein/publication/364329744_Undes
41939	shan s, wang y		2022	ï»¿Complete mitochondrial genomes of Boiga kraepelini	ZooKeys 1	75891		https://doi.org/10.3897/zookeys.1124.87861
41940	cer aco, l.m.p., de lima, r.f., bell, r.c., melo, m.		2022	Biodiversity in the Gulf of Guinea Oceanic Islands: A Synt	In: CerÃ-a	75895		https://doi.org/10.1007/978-3-031-06153-0_1
41941	thill vl, moniz ha, teglas mb, wasley mkj, feldman cr.		2022	Preying dangerously: black widow spider venom resistan	R. Soc. Op	75897		https://doi.org/10.1098/rsos.221012
41942	l onard, k. a., emmanuel, a. n. g., n guessan, g., simplice, k. g., roland,		2022	First record of Seven Species of Lizards in TaÃ¯ National	Internatio	75903		https://www.zoologicaljournal.com/article/45/2-1-22-290.pdf
41943	alvarado, r., alvarado, e. v., l pez, l. i., uma a, d., mora, j. m.		2022	Predation of a Juvenile Iguana rhinolopha (Squamata: Ig	Caribbean	75905		https://doi.org/10.18475/cjos.v52i2.a7
41944	phan, x. t., van ngo, b., nguyen, h. d., hoang, q. t., ngo, c. d., bui, c. t.,		2022	Evaluating and Reconstructing the Genetic Diversity of B	Russian Jo	75908		http://www.rjh.folium.ru/index.php/rjh/article/view/1842
41945	tamar, k., moravec, j.		2022	First exact record and phylogenetic position of the geckc	Zootaxa, 5	75925		https://doi.org/10.11646/zootaxa.5200.5.3
41946	yu, xin akalili b.t. mohd zanudin, mohd uzair rusli, david t. booth and		2022	Diet reflects opportunistic feeding habit of the Asian wat	Animal Bic	75927		https://brill.com/view/journals/ab/ab-overview.xml
41947	cyriac, vivek p. kiran b. srinivasa, lohith kumar and gerard martin		2022	Should I stay or should I go: escape behaviour of Russell	Animal Bic	75928		https://bri vivek.cyriac@gmail.com
41948	zhou, xianwen hui luo, dan zeng, yazhou hu, pei wang, gang xiong anc		2022	Sex-relevant genes in the embryo stage of Chinese soft-s	Animal Bic	75929		https://brill.com/view/journals/ab/ab-overview.xml
41949	champini, barbara de godoi andr luis da cruz and wilfried klein		2022	Heart position and pulmonary vasculature in snakes with	Animal Bic	75930		https://brill.com/view/journals/ab/ab-overview.xml
41950	xiong, jianli yinlong bai, guanglu li and zhangqiang you		2022	Sexual dimorphism in the mountain dragon, Diploderma	Animal Bic	75931		https://doi.org/10.1163/15707563-bja10085
41951	brown ar, comai k, mannino d, mccullough h, donekal y, meyers hc, e		2022	A community-science approach identifies genetic variant	PLoS ONE	75934		https://doi.org/10.1371/journal. pone.0276376
41952	hidalgo licona, luis fernando mar a guadalupe guti rrez may n, c sar a		2022	Ecogeographic and Morphometric Variation in the Mexic	Ichthyolog	75938		doi:
41953	decena, s. c. p., macasait jr, d. r., ara syrus.decena@vsu.edu.ph		2023	Species Accounts, Assemblage, and Microhabitats of Am	Philippine	75785	p	Northeast syrus.decena@vsu.edu.ph
41954	hosseinian yousefkhani, s.s., cavalcanti, m.j.		2023	Species richness and areas of endemism of Lacertidae ar	Journal of	75844	p	DOI:
41955	kuhn, arianna, marcelo gehara, mamy s.m. andrianarimalala, nirhy rabibiso			Drivers of unique and asynchronous population dynamic	Journal of	75282		https://doi.org/10.1111/jbi.14315
41956	van den burg, matthijs p., hannah madden, timothy p. van wagensveld erik			Hurricane-associated population decrease in a critically	Biotropica	75321		https://doi.org/10.1111/btp.13087

Table: 2 shows the output of the last thirty rows of the table 1 dataset

	Unnamed:	author	email	year	title	source	reference_	kind	url	email_final	Species	photos_to	photo_rdb	photo_Cal	photo_Flic	photo_rep	photo_link	type_speci
0	11744	greer, a.e.	g.shea@us	2004	A new cha	Journal of	21187	p	http://ww	g.shea@usyd.edu.au	Sphenomorphus fuscolineatus							
1	13575	sadlier, r.a	aaron.bau	2006	A new gen	Rec. Austr	23936	p	http://aus	aaron.bauer@villanova.edu ross.sac	Celatiscincus similis	1						
2	15237	horner, p.	paul.horne	2007	Systematic	The Beagle	25874	p		paul.horner@nt.gov.au	Cryptoblepharus furvus							
3	15237	horner, p.	paul.horne	2007	Systematic	The Beagle	25874	p		paul.horner@nt.gov.au	Cryptoblepharus xenikos							
4	15237	horner, p.	paul.horne	2007	Systematic	The Beagle	25874	p		paul.horner@nt.gov.au	Cryptoblepharus yulensis	4		3		1		
5	15421	mcmahan,	zugg@si.e	2007	Burmese H	Proc. Cal.	26144	p		zugg@si.edu cmcmahan@fieldmuse	Hemidactylus thayene							
6	16819	lue, kuang	yang and si	2008	Two New	Herpetolo	26947	p o	http://www.jstor.org/action/showPublication?		Takydromus luyeanus	8						
7	17200	sadlier, r.a	aaron.bau	2009	A New Live	Pacîˆc Sci	26961	p	https://do	aaron.bauer@villanova.edu ross.sac	Kanakysaurus zebratus							
8	17407	k hler, j. v	gkoehler@	2009	A further n	African Jo	28383	p	http://ww	gkoehler@senckenberg.de	Paracontias kankana							
9	17477	oliver , pa	iskandar@	2009	A new spe	Zootaxa 2	27555	p	http://ww	iskandar@sith.itb.ac.id	Cyrtodactylus nuaulu	1						
10	18573	welton, l.j.	furcifer@k	2010	Phylogeny	Zootaxa 2	28756	p	http://ww	furcifer@ku.edu lwelton@byu.net	Cyrtodactylus jambangan							
11	18896	siler, came	rafe@ku.e	2010	Phylogeny	Herpetolo	29730	p	http://ww	rafe@ku.edu camsiler@ku.edu	Brachymeles tungaoi							
12	19840	siler, came	camsiler@	2011	Phylogeny	Herpetolo	31028	p	http://ww	camsiler@ku.edu	Brachymeles bicolandia							
13	19840	siler, came	camsiler@	2011	Phylogeny	Herpetolo	31028	p	http://ww	camsiler@ku.edu	Brachymeles cobos							
14	19840	siler, came	camsiler@	2011	Phylogeny	Herpetolo	31028	p	http://ww	camsiler@ku.edu	Brachymeles brevidactylus							
15	20275	oliver, pau	paul.oliver	2011	A new spe	Zootaxa 2	30149	p	https://ww	paul.oliver@anu.edu.au paul.oliver@	Cyrtodactylus boreoclivus	1					1	
16	21298	siler, came	camsiler@	2012	Phylogeny	Herpetolo	32417	p	http://ww	camsiler@ku.edu	Brachymeles samad							
17	21935	vasconcel	salvador.c	2012	An integra	Zoological	31040	p	http://onli	salvador.carranza@ibe.upf raquel.v	Tarentola fogoensis							
18	21944	hedges, s.l	sbh@temp	2012	A new skin	Zootaxa 3	31534	p		sbh@temple.edu	Spondylurus caicosae							
19	22155	zug, g. r., i	zugg@si.e	2012	Lizards of	Pacific Sci	31375	p	http://ww	zugg@si.edu	Emoia mokolahi							
20	22223	trape, j.f.	trape, s. chi	2012	LÃ©zards, IRD Orstor		31928	p	https://ho	ois.trape@ird.fr	Cophoscincopus senegalensis							
21	23103	aguilar, ce	caguilarp@	2013	Integrative	ZooKeys 3	34481	p	http://ww	caguilarp@gmail.com rocio.aguilar@	Liolaemus chavin	1						
22	23103	aguilar, ce	caguilarp@	2013	Integrative	ZooKeys 3	34481	p	http://ww	caguilarp@gmail.com rocio.aguilar@	Liolaemus pachacutec							
23	23103	aguilar, ce	caguilarp@	2013	Integrative	ZooKeys 3	34481	p	http://ww	caguilarp@gmail.com rocio.aguilar@	Liolaemus wari	1						
24	23264	glaw, franl	frank.glaw	2013	New insigh	Org Divers	34322	p	http://dx.c	frank.glaw@zsm.mwn.de	Liopholidophis oligolepis							
25	23379	goicoeche	iriva@mnc	2013	A Taxonon	American I	34316	p	http://ww	iriva@mncn.csic.es n.goicoechear@	Proctoporus carabaya							
26	23379	goicoeche	iriva@mnc	2013	A Taxonon	American I	34316	p	http://ww	iriva@mncn.csic.es n.goicoechear@	Proctoporus iridescens							
27	23379	goicoeche	iriva@mnc	2013	A Taxonon	American I	34316	p	http://ww	iriva@mncn.csic.es n.goicoechear@	Proctoporus kiziriani							
28	23777	linkem, ch	rafe@ku.e	2013	Systematic	Zootaxa 3	33777	p	http://biot	rafe@ku.edu cwlinkem@gmail.com	Parvoscincus abstrusus							
29	23843	sadlier, ros	rosss@aus	2013	A new spe	Zootaxa 3	33685	p	http://biot	rosss@austmus.gov.au ross.sadlier@	Caledoniscincus notialis							

Table: 3, Shows the output of the initial thirty rows of merge dataset with images and emails

Conclusions:

This report highlights the potential ability of Python programming in conjunction with social engineering techniques for data extraction tasks. The presented script demonstrates a practical application in the realm of reptile bibliography databases and is easier to abstract the missing information of pioneer data. Among the 41,956 bibliography dataset, only 22,100 email information was recorded. With the help of the data engineering and analysis process of Python 6386 emails were filled in the empty rows and 319 images with the respective author's emails. Furthermore, the fusion of social engineering and Python programming provides a potent combination for data extraction, retrieval of information, and analysis.

Future Study:

In future iterations of this project could explore more advanced social engineering tactics, refine data cleaning procedures, and implement machine learning algorithms for improved accuracy in author-email, and species-image association with other required information. Additionally, the script could be adapted to handle larger datasets and integrate web scraping techniques for a more comprehensive reptile photo retrieval.

References:

[1] NumPy. (2022). Fundamental package for scientific computing with Python (Version 1.21.0). Travis Oliphant. [https://numpy.org/] (https://numpy.org/)

[2] Pandas. (2022). Powerful data structures for data manipulation and analysis (Version 1.3.1). Wes McKinney. [https://pandas.pydata.org/] (https://pandas.pydata.org/)

[3] Stack Overflow: [(https://stackoverflow.com/questions/45946202/how-to-iterate-over-a-list-in-python)] (https://stackoverflow.com/questions/45946202/how-to-iterate-over-a-list-in-python)

[4] ChatGPT: Model Name: GPT-3.5, ChatGPT, Creator(s): OpenAI

[5] PlantUML web server [(https://www.plantuml.com/plantuml/uml/)]

Appendix:

Codes execution and Output

```
# Import the nesessaries libraries
import pandas as pd
import numpy as np
# Set the pandas option to display full column width
pd.set_option('display.max_colwidth', None)
```

✓ 3.1s

```
# Read an excelfile into Pands DataFrame
df_orig = pd.read_excel('C:/Users/bhojr/Desktop/Reptile_bibliography/Reptile_bibliography Nov 2022.xlsx')
# Create the copy of the original DataFrame
df = df_orig.copy()
```

✓ 4.5s

```
# Remove the duplicated column
df = df.loc[:, ~df.columns.duplicated()]
# Select only the email and authors column from the DataFrame
#df = df[['email', 'author']]
# Convert the email and author columns to strings
df['email'] = df['email'].astype(str)
df['author'] = df['author'].astype(str)
# Check the some initial data in a data frem
df.head()
```

```
# Check the column
df.columns
```

✓ 0.0s

```
Index(['author', 'email', 'year', 'title', 'source', 'ref#', 'kind', 'url',
       'email_final'],
      dtype='object')
```

```
# Define function to clean the text data
import re
def clean_corpus(nlp):
    words= nlp.lower()
    mytext=re.sub(r'[^a-zA-Z0-9\,\@\.]', ' ',words)
    mytext=re.sub(r' +', ' ', mytext)
    return mytext.strip()
```

✓ 0.0s

```
# List the authors having similar emails
correct_author = []
author = []
emailaddress = []
for index, row in df.iterrows():
    if row['Likely_Author1'] == 'Yes' and row['first_author'] != 'nan':
        if row['first_author'] not in author:
            author.append(row['first_author'])
            emailaddress.append(row['email'])
    elif row['Likely_Author2'] == 'Yes' and row['second_author'] != 'nan':
        if row['second_author'] not in author:
            author.append(row['second_author'])
            emailaddress.append(row['email'])
    elif row['Likely_Author3'] == 'Yes' and row['third_author'] != 'nan':
        if row['third_author'] not in author:
            author.append(row['third_author'])
            emailaddress.append(row['email'])
    elif row['Likely_Author4'] == 'Yes' and row['forth_author'] != 'nan':
        if row['forth_author'] not in author:
            author.append(row['forth_author'])
            emailaddress.append(row['email'])
    elif row['Likely_Author5'] == 'Yes' and row['fifth_author'] != 'nan':
        if row['fifth_author'] not in author:
            author.append(row['fifth_author'])
            emailaddress.append(row['email'])
```

```
# Dictionary for mapping the authors and email address
author_dict = dict(zip(author, emailaddress))
#author_dict
df = df.replace({"first_author": author_dict,
                 "second_author": author_dict,
                 "third_author": author_dict,
                 "forth_author": author_dict,
                 "fifth_author": author_dict,
                 "sndlast_author": author_dict,
                 "last_author": author_dict})
def find_email(text):
    email = re.findall(r'[\w\.-]+@[\w\.-]+',str(text))
    return ",".join(email)

columns = ['email', 'first_author', 'second_author', 'third_author',
       'forth_author', 'fifth_author', 'last_author', 'sndlast_author']
for column in columns:
    df[column] = df[column].apply(lambda x: find_email(x))
df.columns
# df = df[['author', 'email', 'first_author', 'second_author', 'third_author',
# 'forth_author', 'fifth_author', 'last_author', 'sndlast_author']]
df["email_final"] = df["email"].map(str) + ' ' + \
                    df["first_author"].map(str) + ' ' + \
                    df["second_author"].map(str) + ' ' +\
                    df["third_author"].map(str) + ' ' +\
                    df["forth_author"].map(str) + ' ' +\
                    df["fifth_author"].map(str) + ' ' +\
```

```
# Total length of Data
len(df)
```

✓ 0.0s

```
41956
```

```
# Replace the empty rows with new email
df ['email_final']= df[['email_final']].apply(lambda x: x.str.strip()).replace('', np.nan).replace('nan',None)
df ['email']= df[['email']].apply(lambda x: x.str.strip()).replace('', np.nan).replace('nan',None)
```

✓ 0.0s

```
# List the five samples in a data
df.sample(5)
```

author	email	year	title	source	ref#	kind	url
casale, paolo freggi, daniela basso, roberto argano,	paolo.casale@uniroma1.it	2005.0	Size at Male Maturity, Sexing Methods and Adult Sex Ratio in Loggerhead Turtles (Caretta caretta) from Italian Waters Investigated Through	The Herpetological Journal 15: 145-148	61515	NaN	https://www.ingentaconnect.com/content/bhs/thj

```
        elif row['Likely_Author5'] == 'Yes' and row['fifth_author'] != 'nan':
            if row['fifth_author'] not in author:
                author.append(row['fifth_author'])
                emailaddress.append(row['email'])
        elif row['Likely_Author6'] == 'Yes' and row['sndlast_author'] != 'nan':
            if row['sndlast_author'] not in author:
                author.append(row['sndlast_author'])
                emailaddress.append(row['email'])
        if row['Likely_Author7'] == 'Yes' and row['last_author'] != 'nan':
            if row['last_author'] not in author:
                author.append(row['last_author'])
                emailaddress.append(row['email'])
    df['author'] = df['author'].astype(str)
```

✓ 5.5s

```
                        df["sndlast_author"].map(str) + ' ' +\
                        df["last_author"].map(str)

    df["email_final"] = df["email_final"].apply(lambda x: ' '.join(pd.unique(x.split())))
    df['email_final'] = df['email_final'].str.split(',').str[0]
    df.columns

    #df = df[['author', 'email', 'email_final']]
    df.sample(2)
    df = df[['author', 'email', 'year', 'title', 'source', 'ref#', 'kind', 'url',
             'email_final']]
```

✓ 35.8s

```
# Total email in the data
sum(pd.notnull(df['email']))
```

```
19856
```

```
#Total null email after the new email refilled in a data
sum(pd.isnull(df['email_final']))
```

```
15714
```

```
# Total emails after filled in the data
sum(pd.notnull(df['email_final']))
```

```
26242
```

```
# Create the out put file with filled emails into csv format
df.to_csv('Reptile_final_filled_new_email.csv')

# Create the output file with filled emails into xlsx format
df.to_excel('Reptile_filled_new_email.xlsx')
```

```
# Apply the 'clean_corpus' function to the email and author columns

df['email_clean'] =  df.email.apply(clean_corpus)
df['author_clean'] =  df.author.apply(clean_corpus)

# Strip whitespace, replace empty string with Nan and replace 'nan'
df = df.replace(r'^\s*$', np.nan, regex=True)
df['email_clean'] = df[['email_clean']].apply(lambda x: x.str.strip()).replace('', np.nan).replace('nan', None)
df['author_clean'] = df[['author_clean']].apply(lambda x: x.str.strip()).replace('', np.nan).replace('nan', None)
# Rename columns to author and email after the cleaning
df  = df[['author_clean', 'email_clean', 'year', 'title', 'source', 'ref#', 'kind', 'url']]
df.rename({'author_clean':'author', 'email_clean': 'email'}, axis =1 , inplace = True)
```

✓ 0.5s

```
#Split the data by author name
df['first_author'] = df['author'].str.split(',').str[0]
df['second_author'] = df['author'].str.split(',').str[1]
df['third_author'] = df['author'].str.split(',').str[2]
df['forth_author'] = df['author'].str.split(',').str[3]
df['fifth_author'] = df['author'].str.split(',').str[4]
df['last_author'] = df['author'].str.split(',').str[-1]
df['sndlast_author'] = df['author'].str.split(',').str[-2]
```

✓ 0.5s

```
# Split the data by email address components
df['username_email'] = df['email'].str.split('@').str[0]
df['first_author'] = df['first_author'].astype(str)
df['second_author'] = df['second_author'].astype(str)
df['third_author'] = df['third_author'].astype(str)
df['forth_author'] = df['forth_author'].astype(str)
df['fifth_author'] = df['fifth_author'].astype(str)
df['sndlast_author'] = df['sndlast_author'].astype(str)
df['last_author'] = df['last_author'].astype(str)
df['username_email'] = df['username_email'].astype(str)
df['Likely_Author1'] = df.apply(lambda x: 'Yes' if x['first_author'] in x['username_email'] else 'No',axis=1)
df['Likely_Author2'] = df.apply(lambda x: 'Yes' if x['second_author'] in x['username_email'] else 'No',axis=1)
df['Likely_Author3'] = df.apply(lambda x: 'Yes' if x['third_author'] in x['username_email'] else 'No',axis=1)
df['Likely_Author4'] = df.apply(lambda x: 'Yes' if x['forth_author'] in x['username_email'] else 'No',axis=1)
df['Likely_Author5'] = df.apply(lambda x: 'Yes' if x['fifth_author'] in x['username_email'] else 'No',axis=1)
df['Likely_Author6'] = df.apply(lambda x: 'Yes' if x['sndlast_author'] in x['username_email'] else 'No',axis=1)
df['Likely_Author7'] = df.apply(lambda x: 'Yes' if x['last_author'] in x['username_email'] else 'No',axis=1)
df['email_duplicate'] = df.loc[:, 'email']
```