SNEHA K TASK 1: DATA CLEANING AND PREPROCESSING

In [37]:
```
import pandas as pd
import numpy as np
```

NETFLIX DATA

In [39]:
```
A=pd.read_csv("netflix_titles.csv") #LOAD DATASET
A
```

Out[39]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descri
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	A father the e hi fil
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	cro path pa Cape
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To p his f fi pov drug
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	F flirta and ta a
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a coa ce kno tra
...	...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A po carto a rep ar
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While alon sp to your
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Look surviv world ov
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dra from c f superh
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A sc but boy w his wa

8807 rows × 12 columns

In [41]:
```
A.head()# TAKING FIRST FEW COLUMNS FROM THE DATASETS
```

Out[41]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

In [43]:
```
A.isnull() # CHECKING WHTHER THE DATASET HAS NULL VALUES
```

Out[43]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	False	False	False	False	True	False	False	False	False	False	False	False
1	False	False	False	True	False	False	False	False	False	False	False	False
2	False	False	False	False	False	True	False	False	False	False	False	False
3	False	False	False	True	True	True	False	False	False	False	False	False
4	False	False	False	True	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	False	False	False	False	False	False	False	False	False	False	False	False
8803	False	False	False	True	True	True	False	False	False	False	False	False
8804	False	False	False	False	False	False	False	False	False	False	False	False
8805	False	False	False	False	False	False	False	False	False	False	False	False
8806	False	False	False	False	False	False	False	False	False	False	False	False

8807 rows × 12 columns

In [45]:
```
B=A.bfill()# BACK FILLING THE VALUES THAT HAS NULL VALUES
B
```

Out[45]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descri
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	A father the e h fil
1	s2	TV Show	Blood & Water	Julien Leclercq	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	cro path pa Cape
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	India	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To p his f pov drug
3	s4	TV Show	Jailbirds New Orleans	Mike Flanagan	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	F flirt and t a
4	s5	TV Show	Kota Factory	Mike Flanagan	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a coa ce kno tra
...	...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A po carto a re ar
8803	s8804	TV Show	Zombie Dumb	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While alon s| to your
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Look surviv world ov
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dra from c f superh
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A sc bu boy v his wa

8807 rows × 12 columns

In [47]:
```
B.drop_duplicates()# REMOVING DUPLICATES
```

Out[47]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descri
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	A father the e h fil
1	s2	TV Show	Blood & Water	Julien Leclercq	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	cro path pa Cape
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	India	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To p his f pov drug
3	s4	TV Show	Jailbirds New Orleans	Mike Flanagan	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	F flirt and t a
4	s5	TV Show	Kota Factory	Mike Flanagan	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a coa ce kno tra
...	...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A po carto a re ar
8803	s8804	TV Show	Zombie Dumb	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While alon s| to your
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Look surviv world ov
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dra from c f superh
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A sc bu boy v his wa

8807 rows × 12 columns

In [49]:
```
B['country']=B['country'].str.upper() # STANDARDIZING THE TEXT
B
```

Out[49]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descr
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	UNITED STATES	September 25, 2021	2020	PG-13	90 min	Documentaries	father the h fi
1	s2	TV Show	Blood & Water	Julien Leclercq	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	SOUTH AFRICA	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	cr patt p Cape
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	INDIA	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To p his po dru
3	s4	TV Show	Jailbirds New Orleans	Mike Flanagan	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	INDIA	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	R flirt an t
4	s5	TV Show	Kota Factory	Mike Flanagan	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	INDIA	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a co c kn tr
...	...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	UNITED STATES	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A p cart a re a
8803	s8804	TV Show	Zombie Dumb	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	UNITED STATES	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While alor s t you
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	UNITED STATES	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Lool surviv world o
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	UNITED STATES	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dr from t super
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	INDIA	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A s bu boy his wa

8807 rows × 12 columns

In [51]:

```
#STANDARIZING TEXT OF SPECIFIC COLUMNS IN TEXT_COLS
text_cols = ['show_id', 'type', 'title', 'director', 'cast', 'release_year', 'rating', 'duration']
def clean_text(series):
    return (
        series.astype(str)
        .str.strip()
        .str.lower()
        .str.replace(r'\s+', ' ', regex=True)
    )
```

```
B[text_cols] = B[text_cols].apply(clean_text)
B[text_cols].head()
B
```

Out[51]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descrip
0	s1	movie	dick johnson is dead	kirsten johnson	ama qamata, khosi ngema, gail mabalane, thaban...	UNITED STATES	September 25, 2021	2020	pg-13	90 min	Documentaries	A father the e his fil
1	s2	tv show	blood & water	julien leclercq	ama qamata, khosi ngema, gail mabalane, thaban...	SOUTH AFRICA	September 24, 2021	2021	tv-ma	2 seasons	International TV Shows, TV Dramas, TV Mysteries	cros paths pa Cape T
2	s3	tv show	ganglands	julien leclercq	sami bouajila, tracy gotoas, samuel jouy, nabi...	INDIA	September 24, 2021	2021	tv-ma	1 season	Crime TV Shows, International TV Shows, TV Act...	To pr his fa fr pow drug
3	s4	tv show	jailbirds new orleans	mike flanagan	mayur more, jitendra kumar, ranjan raj, alam k...	INDIA	September 24, 2021	2021	tv-ma	1 season	Docuseries, Reality TV	Fe flirta and ta a
4	s5	tv show	kota factory	mike flanagan	mayur more, jitendra kumar, ranjan raj, alam k...	INDIA	September 24, 2021	2021	tv-ma	2 seasons	International TV Shows, Romantic TV Shows, TV ...	In a c coac ce kno tra
...	...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	movie	zodiac	david fincher	mark ruffalo, jake gyllenhaal, robert downey j...	UNITED STATES	November 20, 2019	2007	r	158 min	Cult Movies, Dramas, Thrillers	A pol cartoo a c rep an
8803	s8804	tv show	zombie dumb	ruben fleischer	jesse eisenberg, woody harrelson, emma stone, ...	UNITED STATES	July 1, 2019	2018	tv-y7	2 seasons	Kids' TV, Korean TV Shows, TV Comedies	While alone sp tow young
8804	s8805	movie	zombieland	ruben fleischer	jesse eisenberg, woody harrelson, emma stone, ...	UNITED STATES	November 1, 2019	2009	r	88 min	Comedies, Horror Movies	Looki survive world t ov
8805	s8806	movie	zoom	peter hewitt	tim allen, courteney cox, chevy chase, kate ma...	UNITED STATES	January 11, 2020	2006	pg	88 min	Children & Family Movies, Comedies	Drag from ci l fo superhe
8806	s8807	movie	zubaan	mozez singh	vicky kaushal, sarah-jane dias, raaghav chanan...	INDIA	March 2, 2019	2015	tv-14	111 min	Dramas, International Movies, Music & Musicals	A scr but boy w his way a

8807 rows × 12 columns

In [53]:
```
# CHANGING DATE TO CONSISTENT TYPE DD-MM-YYYY
B['date_added'] = pd.to_datetime(B['date_added'], dayfirst=True, errors='coerce')
B['date_added'] = B['date_added'].dt.strftime('%d-%m-%Y')
B
```

Out[53]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descrip
0	s1	movie	dick johnson is dead	kirsten johnson	ama qamata, khosi ngema, gail mabalane, thaban...	UNITED STATES	25-09-2021	2020	pg-13	90 min	Documentaries	As father r the e his filn
1	s2	tv show	blood & water	julien leclercq	ama qamata, khosi ngema, gail mabalane, thaban...	SOUTH AFRICA	24-09-2021	2021	tv-ma	2 seasons	International TV Shows, TV Dramas, TV Mysteries	cros paths pa Cape T
2	s3	tv show	ganglands	julien leclercq	sami bouajila, tracy gotoas, samuel jouy, nabi...	INDIA	24-09-2021	2021	tv-ma	1 season	Crime TV Shows, International TV Shows, TV Act...	To pr his fa fr pow drug
3	s4	tv show	jailbirds new orleans	mike flanagan	mayur more, jitendra kumar, ranjan raj, alam k...	INDIA	24-09-2021	2021	tv-ma	1 season	Docuseries, Reality TV	Fe flirta and ta a
4	s5	tv show	kota factory	mike flanagan	mayur more, jitendra kumar, ranjan raj, alam k...	INDIA	24-09-2021	2021	tv-ma	2 seasons	International TV Shows, Romantic TV Shows, TV ...	In a c coac ce know tra
...	...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	movie	zodiac	david fincher	mark ruffalo, jake gyllenhaal, robert downey j...	UNITED STATES	20-11-2019	2007	r	158 min	Cult Movies, Dramas, Thrillers	A pol cartoo a c rep an
8803	s8804	tv show	zombie dumb	ruben fleischer	jesse eisenberg, woody harrelson, emma stone, ...	UNITED STATES	01-07-2019	2018	tv-y7	2 seasons	Kids' TV, Korean TV Shows, TV Comedies	While alone sp tov young
8804	s8805	movie	zombieland	ruben fleischer	jesse eisenberg, woody harrelson, emma stone, ...	UNITED STATES	01-11-2019	2009	r	88 min	Comedies, Horror Movies	Looki survive world t ov
8805	s8806	movie	zoom	peter hewitt	tim allen, courteney cox, chevy chase, kate ma...	UNITED STATES	11-01-2020	2006	pg	88 min	Children & Family Movies, Comedies	Drag from civ l fo superhe
8806	s8807	movie	zubaan	mozez singh	vicky kaushal, sarah-jane dias, raaghav chanan...	INDIA	02-03-2019	2015	tv-14	111 min	Dramas, International Movies, Music & Musicals	A scr but boy w his way a

8807 rows × 12 columns

In [55]:
```
#RENAMING COLUMNS HEADERS TO BE CLEAN AND CONSISTENT
B.columns = B.columns.str.upper()
```

In [57]:
```
B
```

Out[57]:

	SHOW_ID	TYPE	TITLE	DIRECTOR	CAST	COUNTRY	DATE_ADDED	RELEASE_YEAR	RATING	DURATION	LI
0	s1	movie	dick johnson is dead	kirsten johnson	ama qamata, khosi ngema, gail mabalane, thaban...	UNITED STATES	25-09-2021	2020	pg-13	90 min	Docu
1	s2	tv show	blood & water	julien leclercq	ama qamata, khosi ngema, gail mabalane, thaban...	SOUTH AFRICA	24-09-2021	2021	tv-ma	2 seasons	Int TV S Dr
2	s3	tv show	ganglands	julien leclercq	sami bouajila, tracy gotoas, samuel jouy, nabi...	INDIA	24-09-2021	2021	tv-ma	1 season	Int TV S
3	s4	tv show	jailbirds new orleans	mike flanagan	mayur more, jitendra kumar, ranjan raj, alam k...	INDIA	24-09-2021	2021	tv-ma	1 season	D
4	s5	tv show	kota factory	mike flanagan	mayur more, jitendra kumar, ranjan raj, alam k...	INDIA	24-09-2021	2021	tv-ma	2 seasons	Int T Ro Sho
...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	movie	zodiac	david fincher	mark ruffalo, jake gyllenhaal, robert downey j...	UNITED STATES	20-11-2019	2007	r	158 min	Cu
8803	s8804	tv show	zombie dumb	ruben fleischer	jesse eisenberg, woody harrelson, emma stone, ...	UNITED STATES	01-07-2019	2018	tv-y7	2 seasons	K S
8804	s8805	movie	zombieland	ruben fleischer	jesse eisenberg, woody harrelson, emma stone, ...	UNITED STATES	01-11-2019	2009	r	88 min	C Horr
8805	s8806	movie	zoom	peter hewitt	tim allen, courteney cox, chevy chase, kate ma...	UNITED STATES	11-01-2020	2006	pg	88 min	C Fami
8806	s8807	movie	zubaan	mozez singh	vicky kaushal, sarah-jane dias, raaghav chanan...	INDIA	02-03-2019	2015	tv-14	111 min	Int Mov &

8807 rows × 12 columns

In [59]:

```
#CHECKING AND FIXING THE DATA TYPES
B['DATE_ADDED'] = pd.to_datetime(B['DATE_ADDED'], errors='coerce', dayfirst=True)
B.dtypes
```

Out[59]:
```
SHOW_ID                 object
TYPE                    object
TITLE                   object
DIRECTOR                object
CAST                    object
COUNTRY                 object
DATE_ADDED      datetime64[ns]
RELEASE_YEAR            object
RATING                  object
DURATION                object
LISTED_IN               object
DESCRIPTION             object
dtype: object
```

In [ ]:

SALES DATA

In [ ]:

In [61]:
```
DF=pd.read_csv("car_prices.csv")# LOAD DATASET
DF
```

Out[61]:

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior
0	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black
1	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige
2	2014	BMW	3 Series	328i SULEV	Sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black
3	2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black
4	2014	BMW	6 Series Gran Coupe	650i	Sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black
...	...	...	...	...	...	...	...	...	...	...	...	...
558832	2015	Kia	K900	Luxury	Sedan	NaN	knalw4d4xf6019304	in	45.0	18255.0	silver	black
558833	2012	Ram	2500	Power Wagon	Crew Cab	automatic	3c6td5et6cg112407	wa	5.0	54393.0	white	black
558834	2012	BMW	X5	xDrive35d	SUV	automatic	5uxzw0c58cl668465	ca	48.0	50561.0	black	black
558835	2015	Nissan	Altima	2.5 S	sedan	automatic	1n4al3ap0fc216050	ga	38.0	16658.0	white	black
558836	2014	Ford	F-150	XLT	SuperCrew	automatic	1ftfw1et2eke87277	ca	34.0	15008.0	gray	gray

558837 rows × 16 columns

In [106…

```
DF1=DF.head()
DF1
```

Out[106…

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller
0	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black	kia motors america inc
1	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige	kia motors america inc
2	2014	BMW	3 Series	328i SULEV	Sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black	financial services remarketing (lease)
3	2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black	volvo na rep/world omni
4	2014	BMW	6 Series Gran Coupe	650i	Sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black	financial services remarketing (lease)

In [110…

```
DF.isnull() # CHECKING FOR NULL VALUES
```

Out[110…

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingpri
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
558832	False	False	False	False	False	True	False	False	False	False	False	False	False	False	Fal
558833	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
558834	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
558835	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal
558836	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fal

558837 rows × 16 columns

In [114…

```
DF.drop_duplicates()# REMOVING DUPLICATES
```

Out[114…

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior
0	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black
1	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige
2	2014	BMW	3 Series	328i SULEV	Sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black
3	2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black
4	2014	BMW	6 Series Gran Coupe	650i	Sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black
...	...	...	...	...	...	...	...	...	...	...	...	...
558832	2015	Kia	K900	Luxury	Sedan	NaN	knalw4d4xf6019304	in	45.0	18255.0	silver	black
558833	2012	Ram	2500	Power Wagon	Crew Cab	automatic	3c6td5et6cg112407	wa	5.0	54393.0	white	black
558834	2012	BMW	X5	xDrive35d	SUV	automatic	5uxzw0c58cl668465	ca	48.0	50561.0	black	black
558835	2015	Nissan	Altima	2.5 S	sedan	automatic	1n4al3ap0fc216050	ga	38.0	16658.0	white	black
558836	2014	Ford	F-150	XLT	SuperCrew	automatic	1ftfw1et2eke87277	ca	34.0	15008.0	gray	gray

558837 rows × 16 columns

In [122…

```
#STANDARDIZING TEXT
text_cols = ['make', 'model', 'trim', 'body', 'transmission', 'color', 'interior', 'seller', 'state']
def clean_text(series):
    return (
        series.astype(str)
        .str.strip()
        .str.lower()
        .str.replace(r'\s+', ' ', regex=True)
    )


DF[text_cols] = DF[text_cols].apply(clean_text)
DF[text_cols].head()
DF
```

Out[122…

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior
0	2015	kia	sorento	lx	suv	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black
1	2015	kia	sorento	lx	suv	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige
2	2014	bmw	3 series	328i sulev	sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black
3	2015	volvo	s60	t5	sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black
4	2014	bmw	6 series gran coupe	650i	sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black
...	...	...	...	...	...	...	...	...	...	...	...	...
558832	2015	kia	k900	luxury	sedan	nan	knalw4d4xf6019304	in	45.0	18255.0	silver	black
558833	2012	ram	2500	power wagon	crew cab	automatic	3c6td5et6cg112407	wa	5.0	54393.0	white	black
558834	2012	bmw	x5	xdrive35d	suv	automatic	5uxzw0c58cl668465	ca	48.0	50561.0	black	black
558835	2015	nissan	altima	2.5 s	sedan	automatic	1n4al3ap0fc216050	ga	38.0	16658.0	white	black
558836	2014	ford	f-150	xlt	supercrew	automatic	1ftfw1et2eke87277	ca	34.0	15008.0	gray	gray

558837 rows × 16 columns

In [130…

```
#CONVERTING DATE FORMAT TO CONSISTENT TYPE
DF['saledate'] = pd.to_datetime(DF['saledate'], errors='coerce', dayfirst=True)
DF['saledate'] = DF['saledate'].dt.strftime('%d-%m-%Y %H:%M:%S')
DF['saledate']
DF
```

Out[130…

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior
0	2015	kia	sorento	lx	suv	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black
1	2015	kia	sorento	lx	suv	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige
2	2014	bmw	3 series	328i sulev	sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black
3	2015	volvo	s60	t5	sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black
4	2014	bmw	6 series gran coupe	650i	sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black
...	...	...	...	...	...	...	...	...	...	...	...	...
558832	2015	kia	k900	luxury	sedan	nan	knalw4d4xf6019304	in	45.0	18255.0	silver	black
558833	2012	ram	2500	power wagon	crew cab	automatic	3c6td5et6cg112407	wa	5.0	54393.0	white	black
558834	2012	bmw	x5	xdrive35d	suv	automatic	5uxzw0c58cl668465	ca	48.0	50561.0	black	black
558835	2015	nissan	altima	2.5 s	sedan	automatic	1n4al3ap0fc216050	ga	38.0	16658.0	white	black
558836	2014	ford	f-150	xlt	supercrew	automatic	1ftfw1et2eke87277	ca	34.0	15008.0	gray	gray

558837 rows × 16 columns

In [85]:

```
#RENAMING COLUMNS HEADERS TO BE CLEAN AND UNIFORM
DF.columns = (
    DF.columns
    .str.strip()
    .str.lower()
    .str.replace(' ', '_')
)


DF.columns.tolist()
DF
```

Out[85]:

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior
0	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black
1	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige
2	2014	BMW	3 Series	328i SULEV	Sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black
3	2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black
4	2014	BMW	6 Series Gran Coupe	650i	Sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black
...	...	...	...	...	...	...	...	...	...	...	...	...
558832	2015	Kia	K900	Luxury	Sedan	NaN	knalw4d4xf6019304	in	45.0	18255.0	silver	black
558833	2012	Ram	2500	Power Wagon	Crew Cab	automatic	3c6td5et6cg112407	wa	5.0	54393.0	white	black
558834	2012	BMW	X5	xDrive35d	SUV	automatic	5uxzw0c58cl668465	ca	48.0	50561.0	black	black
558835	2015	Nissan	Altima	2.5 S	sedan	automatic	1n4al3ap0fc216050	ga	38.0	16658.0	white	black
558836	2014	Ford	F-150	XLT	SuperCrew	automatic	1ftfw1et2eke87277	ca	34.0	15008.0	gray	gray

558837 rows × 16 columns

In [136…

```
DF.dtypes
```

Out[136…
```
year              int64
make             object
model            object
trim             object
body             object
transmission     object
vin              object
state            object
condition       float64
odometer        float64
color            object
interior         object
seller           object
mmr             float64
sellingprice    float64
saledate         object
dtype: object
```

In [111…
```
#CHECKING AND FIXING DATATYPES
DF = DF.astype({
    'year': 'int',
    'make': 'category',
    'model':'category'

})
DF['saledate'] = pd.to_datetime(DF['saledate'], errors='coerce', dayfirst=True)
```

In [113…
```
DF.dtypes
```

Out[113…
```
year                                       int32
make                                    category
model                                   category
trim                                      object
body                                      object
transmission                              object
vin                                       object
state                                     object
condition                                float64
odometer                                 float64
color                                     object
interior                                  object
seller                                    object
mmr                                      float64
sellingprice                             float64
saledate        datetime64[ns, tzoffset('PST', 28800)]
dtype: object
```

In [ ]:
```

```

MEDICAL APPOINTMENT DATASET

In [67]:
```
df=pd.read_csv("Medical Appointment No Shows.csv")#LOAD DATASET
df
```

Out[67]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	D
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	
...	...	...	...	...	...	...	...	...	...	
110522	2.572134e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0	0	
110523	3.596266e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0	0	
110524	1.557663e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0	0	
110525	9.213493e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0	0	
110526	3.775115e+14	5629448	F	2016-04-27T13:30:56Z	2016-06-07T00:00:00Z	54	MARIA ORTIZ	0	0	

110527 rows × 14 columns

In [175…

```
#CHECKING FOR NULL VALUES
df.isnull()
```

Out[175…

	customerid	gender	age	annual_income_(k$)	spending_score_(1-100)
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
...	...	...	...	...	...
195	False	False	False	False	False
196	False	False	False	False	False
197	False	False	False	False	False
198	False	False	False	False	False
199	False	False	False	False	False

200 rows × 5 columns

In [71]:

```
#REMOVING DUPLICATES
df.drop_duplicates()
```

Out[71]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	D
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	
...	...	...	...	...	...	...	...	...	...	
110522	2.572134e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0	0	
110523	3.596266e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0	0	
110524	1.557663e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0	0	
110525	9.213493e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0	0	
110526	3.775115e+14	5629448	F	2016-04-27T13:30:56Z	2016-06-07T00:00:00Z	54	MARIA ORTIZ	0	0	

110527 rows × 14 columns

In [75]:

```
#STANDARDIZING TEXTS
text_cols = ['PatientId','AppointmentID','Gender','ScheduledDay','Age','Neighbourhood','Scholarship','Hipertens
def clean_text(series):
    return (
        series.astype(str)
        .str.strip()
        .str.lower()
        .str.replace(r'\s+', ' ', regex=True)
    )


df[text_cols] = df[text_cols].apply(clean_text)
df[text_cols].head()
df
```

Out[75]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensi
0	29872499824296.0	5642903	f	2016-04-29t18:38:08z	2016-04-29T00:00:00Z	62	jardim da penha	0	
1	558997776694438.0	5642503	m	2016-04-29t16:08:27z	2016-04-29T00:00:00Z	56	jardim da penha	0	
2	4262962299951.0	5642549	f	2016-04-29t16:19:04z	2016-04-29T00:00:00Z	62	mata da praia	0	
3	867951213174.0	5642828	f	2016-04-29t17:29:31z	2016-04-29T00:00:00Z	8	pontal de camburi	0	
4	8841186448183.0	5642494	f	2016-04-29t16:07:23z	2016-04-29T00:00:00Z	56	jardim da penha	0	
...	...	...	...	...	...	...	...	...	
110522	2572134369293.0	5651768	f	2016-05-03t09:15:35z	2016-06-07T00:00:00Z	56	maria ortiz	0	
110523	3596266328735.0	5650093	f	2016-05-03t07:27:33z	2016-06-07T00:00:00Z	51	maria ortiz	0	
110524	15576631729893.0	5630692	f	2016-04-27t16:03:52z	2016-06-07T00:00:00Z	21	maria ortiz	0	
110525	92134931435557.0	5630323	f	2016-04-27t15:09:23z	2016-06-07T00:00:00Z	38	maria ortiz	0	
110526	377511518121127.0	5629448	f	2016-04-27t13:30:56z	2016-06-07T00:00:00Z	54	maria ortiz	0	

110527 rows × 14 columns

In [79]:
```
#CONVERTING DATE FORMATS TO CONSISTENT TYPE
df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'], errors='coerce', dayfirst=True)
df['ScheduledDay'] = df['ScheduledDay'].dt.strftime('%d-%m-%Y %H:%M:%S')
df['ScheduledDay']
df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'], errors='coerce', dayfirst=True)
df['AppointmentDay'] = df['AppointmentDay'].dt.strftime('%d-%m-%Y %H:%M:%S')
df['AppointmentDay']
df
```

```
C:\Users\ssneh\AppData\Local\Temp\ipykernel_5852\2595799617.py:4: UserWarning: Parsing dates in %Y-%m-%dT%H:%M:%S%z format when dayfirst=True was specified. Pass `dayfirst=False` or specify a format to silence this warning.
  df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'], errors='coerce', dayfirst=True)
```

Out[79]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensi
0	29872499824296.0	5642903	f	29-04-2016 18:38:08	29-04-2016 00:00:00	62	jardim da penha	0	
1	558997776694438.0	5642503	m	29-04-2016 16:08:27	29-04-2016 00:00:00	56	jardim da penha	0	
2	4262962299951.0	5642549	f	29-04-2016 16:19:04	29-04-2016 00:00:00	62	mata da praia	0	
3	867951213174.0	5642828	f	29-04-2016 17:29:31	29-04-2016 00:00:00	8	pontal de camburi	0	
4	8841186448183.0	5642494	f	29-04-2016 16:07:23	29-04-2016 00:00:00	56	jardim da penha	0	
...	...	...	...	...	...	...	...	...	
110522	2572134369293.0	5651768	f	03-05-2016 09:15:35	07-06-2016 00:00:00	56	maria ortiz	0	
110523	3596266328735.0	5650093	f	03-05-2016 07:27:33	07-06-2016 00:00:00	51	maria ortiz	0	
110524	15576631729893.0	5630692	f	27-04-2016 16:03:52	07-06-2016 00:00:00	21	maria ortiz	0	
110525	92134931435557.0	5630323	f	27-04-2016 15:09:23	07-06-2016 00:00:00	38	maria ortiz	0	
110526	377511518121127.0	5629448	f	27-04-2016 13:30:56	07-06-2016 00:00:00	54	maria ortiz	0	

110527 rows × 14 columns

In [83]:
```
#RENAMING COLUMN HEADRERS TO BE CLEAN AND UNIFORM
df.columns = (
    df.columns
    .str.strip()
    .str.lower()
    .str.replace(' ', '_')

)


df.columns.tolist()
df
```

Out[83]:

	patientid	appointmentid	gender	scheduledday	appointmentday	age	neighbourhood	scholarship	hipertension
0	29872499824296.0	5642903	f	29-04-2016 18:38:08	29-04-2016 00:00:00	62	jardim da penha	0	1
1	558997776694438.0	5642503	m	29-04-2016 16:08:27	29-04-2016 00:00:00	56	jardim da penha	0	0
2	4262962299951.0	5642549	f	29-04-2016 16:19:04	29-04-2016 00:00:00	62	mata da praia	0	0
3	867951213174.0	5642828	f	29-04-2016 17:29:31	29-04-2016 00:00:00	8	pontal de camburi	0	0
4	8841186448183.0	5642494	f	29-04-2016 16:07:23	29-04-2016 00:00:00	56	jardim da penha	0	1
...	...	...	...	...	...	...	...	...	...
110522	2572134369293.0	5651768	f	03-05-2016 09:15:35	07-06-2016 00:00:00	56	maria ortiz	0	0
110523	3596266328735.0	5650093	f	03-05-2016 07:27:33	07-06-2016 00:00:00	51	maria ortiz	0	0
110524	15576631729893.0	5630692	f	27-04-2016 16:03:52	07-06-2016 00:00:00	21	maria ortiz	0	0
110525	92134931435557.0	5630323	f	27-04-2016 15:09:23	07-06-2016 00:00:00	38	maria ortiz	0	0
110526	377511518121127.0	5629448	f	27-04-2016 13:30:56	07-06-2016 00:00:00	54	maria ortiz	0	0

110527 rows × 14 columns

In [87]:
```
#CHECKING AND FIXING DATATYPES
df.dtypes
```

Out[87]:
```
patientid         object
appointmentid     object
gender            object
scheduledday      object
appointmentday    object
age               object
neighbourhood     object
scholarship       object
hipertension      object
diabetes          object
alcoholism        object
handcap           object
sms_received       int64
noshow            object
dtype: object
```

In [103…
```
df['scheduledday'] = pd.to_datetime(df['scheduledday'], errors='coerce', dayfirst=True)
df['appointmentday']=pd.to_datetime(df['appointmentday'],errors='coerce',dayfirst=True)
df = df.astype({
    'age': 'int',
    'gender': 'category',
    'appointmentid':'int',



})
```

In [105…
```
df.dtypes
```

Out[105…
```
patientid                 object
appointmentid              int32
gender                  category
scheduledday      datetime64[ns]
appointmentday    datetime64[ns]
age                        int32
neighbourhood             object
scholarship               object
hipertension              object
diabetes                  object
alcoholism                object
handcap                   object
sms_received               int64
noshow                    object
dtype: object
```

In [ ]:
```

```

MALL CUSTOMERS

In [178…
```
s=pd.read_csv("Mall_Customers.csv")#LOAD DATASETS
s
```

Out[178…

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

In [118…
```
s.isnull()#CHECKING FOR NULL VALUES
```

Out[118…

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
...	...	...	...	...	...
195	False	False	False	False	False
196	False	False	False	False	False
197	False	False	False	False	False
198	False	False	False	False	False
199	False	False	False	False	False

200 rows × 5 columns

In [120…
```
s.drop_duplicates()#REMOVING DUPLICATES
```

Out[120…

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

In [124…
```
#STANDARDIZING TEXT
text_cols=['CustomerID','Gender','Age','Annual Income (k$)','Spending Score (1-100)']
def clean_text(series):
    return(
```

```
        series.astype(str)
        .str.strip()
        .str.lower()
        .str.replace(r'\s+','',regex=True)
    )
s[text_cols]=s[text_cols].apply(clean_text)
s[text_cols].head()
s
```

Out[124…

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	male	19	15	39
1	2	male	21	15	81
2	3	female	20	16	6
3	4	female	23	16	77
4	5	female	31	17	40
...	...	...	...	...	...
195	196	female	35	120	79
196	197	female	45	126	28
197	198	male	32	126	74
198	199	male	32	137	18
199	200	male	30	137	83

200 rows × 5 columns

In [126…

```
#RENAMING COLUMN HEADERS TO NE CLEAN AND UNIFORM
s.columns = (
    s.columns
    .str.strip()
    .str.lower()
    .str.replace(' ', '_')

)


s.columns.tolist()
s
```

Out[126…

	customerid	gender	age	annual_income_(k$)	spending_score_(1-100)
0	1	male	19	15	39
1	2	male	21	15	81
2	3	female	20	16	6
3	4	female	23	16	77
4	5	female	31	17	40
...	...	...	...	...	...
195	196	female	35	120	79
196	197	female	45	126	28
197	198	male	32	126	74
198	199	male	32	137	18
199	200	male	30	137	83

200 rows × 5 columns

In [128…

```
#CHECKING AND FIXING DATATYPES
s.dtypes
```

Out[128…

```
customerid               object
gender                   object
age                      object
annual_income_(k$)       object
spending_score_(1-100)   object
dtype: object
```

In [168…

```
s = s.astype({
    'age': 'int',
    'gender': 'category',
    'annual_income_(k$)':'int'
```

```
})
s['spending_score_(1-100)'] = pd.to_numeric(s['spending_score_(1-100)'], errors='coerce')
```

In [170…
```
s.dtypes
```

Out[170…
```
customerid                 object
gender                   category
age                         int32
annual_income_(k$)          int32
spending_score_(1-100)      int64
dtype: object
```

In [ ]: