CRIME SCENE REPORT (TABLE 1)

In [1]:
```
import pandas as pd
import numpy as np
df=pd.read_csv("crime_scene_report.csv")
```

In [10]:
```
df
```

Out[10]:

	date	type	description	city
0	20180115	robbery	A Man Dressed as Spider-Man Is on a Robbery Spree	NYC
1	20180115	murder	Life? Dont talk to me about life.	Albany
2	20180115	murder	Mama, I killed a man, put a gun against his he...	Reno
3	20180215	murder	REDACTED REDACTED REDACTED	SQL City
4	20180215	murder	Someone killed the guard! He took an arrow to ...	SQL City
...	...	...	...	...
1223	20180430	bribery	\n	Garden Grove
1224	20180430	fraud	'Why not?' said the March Hare.\n	Houma
1225	20180430	assault	\n	Fontana
1226	20180501	assault	be NO mistake about it: it was neither more no...	Trenton
1227	20180115	murder	Security footage shows that there were 2 witne...	SQL City

1228 rows × 4 columns

In [11]:
```
df.size,df.shape
```

Out[11]:
```
(4912, (1228, 4))
```

In [12]:
```
df.dtypes
# so changes should be made to the date column
# other columns should be of string type
```

Out[12]:
```
date            int64
type           object
description    object
city           object
dtype: object
```

In [40]:
```
df.date.nunique()
# so the data is given for the 486 days
```

Out[40]:
```
486
```

In [13]:
```
df.isna().sum()
# no null values
```

Out[13]:
```
date           0
type           0
description    0
city           0
dtype: int64
```

In [14]:
```
df['description']
# some of the values in this column have incomplete values , like (3-4 tab spaces ,
```

Out[14]:
```
0       A Man Dressed as Spider-Man Is on a Robbery Spree
1                      Life? Dont talk to me about life.
2       Mama, I killed a man, put a gun against his he...
3                              REDACTED REDACTED REDACTED
4       Someone killed the guard! He took an arrow to ...
                              ...
1223                                                   \n
1224                    ‘Why not?’ said the March Hare.\n
1225                                                   \n
1226    be NO mistake about it: it was neither more no...
1227    Security footage shows that there were 2 witne...
Name: description, Length: 1228, dtype: object
```

In [21]:
```
df[df['description'] =='\n']
# see this kind of data should be replaced by "description not given"
```

Out[21]:

	date	type	description	city
30	20170107	smuggling	\n	Savannah
43	20170111	murder	\n	Springdale
45	20170112	smuggling	\n	Melbourne
48	20170113	assault	\n	Reading
50	20170114	blackmail	\n	Orange
...	...	...	...	...
1201	20180424	arson	\n	New York
1204	20180425	blackmail	\n	Chula Vista
1213	20180427	arson	\n	Leominster
1223	20180430	bribery	\n	Garden Grove
1225	20180430	assault	\n	Fontana

289 rows × 4 columns

In [8]:
```
df[df['description'] ="  *    *    *    *    *    *    *\n"]
# see this kind of data should be replaced by "description not given"
```

Out[8]:

	date	type	description	city
508	20170708	murder	*******\n	St. Louis

In [9]:
```
df[df['description'] ="    *    *    *    *    *    *\n"]
# see this kind of data should be replaced by "description not given"
```

Out[9]:

	date	type	description	city
415	20170606	assault	******\n	Elk Grove
1057	20180218	blackmail	******\n	Temecula

Drivers License (Table 2)

In [75]:
```
df=pd.read_csv("drivers_license.csv")
```

In [23]:
```
df
```

Out[23]:

	id	age	height	eye_color	hair_color	gender	plate_number	car_make	car_model
0	100280	72	57	brown	red	male	P24L4U	Acura	MDX
1	100460	63	72	brown	brown	female	XF02T6	Cadillac	SRX
2	101029	62	74	green	green	female	VKY5KR	Scion	xB
3	101198	43	54	amber	brown	female	Y5NZ08	Nissan	Rogue
4	101255	18	79	blue	grey	female	5162Z1	Lexus	GS
...	...	...	...	...	...	...	...	...	...
10002	999923	19	77	amber	black	female	5L0ZI4	GMC	Sierra 3500
10003	999940	71	61	green	green	male	1B8QN8	Mitsubishi	Eclipse
10004	999981	67	69	brown	blue	female	1684K3	Land Rover	LR2
10005	999986	49	58	green	grey	male	F8F64H	Lexus	LS
10006	999993	18	63	black	black	female	6UZO2O	Cadillac	DeVille

10007 rows × 9 columns

In [24]:
```
df.size,df.shape
```

Out[24]:
```
(90063, (10007, 9))
```

In [25]:
```
df.dtypes
# so no changes should be made to the columns schema , they are in correct data typ
# here the height column values are in inches
```

Out[25]:
```
id              int64
age             int64
height          int64
eye_color      object
hair_color     object
gender         object
plate_number   object
car_make       object
car_model      object
dtype: object
```

In [26]:
```
df.isna().sum()
# no null values
```

Out[26]:
```
id              0
age             0
height          0
eye_color       0
hair_color      0
gender          0
plate_number    0
car_make        0
car_model       0
dtype: int64
```

Facebook Event Checking(Table 3)

In [76]:
```
df=pd.read_csv("facebook_event_checkin.csv")
```

In [28]:
```
df
```

Out[28]:

	person_id	event_id	event_name	date
0	28508	5880	Nudists are people who wear one-button suits.\n	20170913
1	63713	3865	but that's because it's the best book on anyth...	20171009
2	63713	3999	If Murphy's Law can go wrong, it will.\n	20170502
3	63713	6436	Old programmers never die. They just branch t...	20170926
4	82998	4470	Help a swallow land at Capistrano.\n	20171022
...	...	...	...	...
20006	99716	1143	SQL Symphony Concert	20171206
20007	99716	1143	SQL Symphony Concert	20171212
20008	99716	1143	SQL Symphony Concert	20171229
20009	67318	4719	The Funky Grooves Tour	20180115
20010	67318	1143	SQL Symphony Concert	20171206

20011 rows × 4 columns

In [29]:
```
df.size,df.shape
```

Out[29]:
```
(80044, (20011, 4))
```

In [30]:
```
df.dtypes
# here changes should be made to the date column with a proper schema(date type)
```

Out[30]:
```
person_id      int64
event_id       int64
event_name    object
date           int64
dtype: object
```

In [77]:
```
df.isna().sum()
# no null values
```

Out[77]:
```
person_id     0
event_id      0
event_name    0
date          0
dtype: int64
```

In [39]:
```
df.date.nunique()
# so the data is given for the 486 days
```

Out[39]:
```
486
```

Get Fit Now Check In (Table 4)

In [78]:
```
df=pd.read_csv("get_fit_now_check_in.csv")
```

In [43]:
```
df
```

Out[43]:

	membership_id	check_in_date	check_in_time	check_out_time
0	NL318	20180212	329	365
1	NL318	20170811	469	920
2	NL318	20180429	506	554
3	NL318	20180128	124	759
4	NL318	20171027	418	1019
...	...	...	...	...
2698	4KB72	20170422	1016	1114
2699	4KB72	20170630	408	885
2700	48Z7A	20180109	1600	1730
2701	48Z55	20180109	1530	1700
2702	90081	20180109	1600	1700

2703 rows × 4 columns

In [44]:
```
df.size,df.shape
```

Out[44]:
```
(10812, (2703, 4))
```

In [45]:
```
df.dtypes
# here changes should be made to the date column with a proper schema(date type)
# and the check in time and check out time is given in the minutes (considering 0 a
# so getting the difference between the check out time and check in time will give
```

Out[45]:
```
membership_id     object
check_in_date      int64
check_in_time      int64
check_out_time     int64
dtype: object
```

In [79]:
```
df.isna().sum()
# no null values
```

Out[79]:
```
membership_id     0
check_in_date     0
check_in_time     0
check_out_time    0
dtype: int64
```

In [46]:
```
df.membership_id.nunique()
# 184 unique members are there
```

Out[46]:
```
184
```

Get Fit Now Member (Table 5)

In [80]:
```
df=pd.read_csv("get_fit_now_member.csv")
```

In [48]:
```
df
```

Out[48]:

	id	person_id	name	membership_start_date	membership_status
0	NL318	65076	Everette Koepke	20170926	gold
1	AOE21	39426	Noe Locascio	20171005	regular
2	2PN28	63823	Jeromy Heitschmidt	20180215	silver
3	0YJ24	80651	Waneta Wellard	20171206	gold
4	3A08L	32858	Mei Bianchin	20170401	silver
...	...	...	...	...	...
179	2V137	41693	Wendell Dulany	20171219	silver
180	4KB72	79110	Emile Hege	20170522	regular
181	48Z7A	28819	Joe Germuska	20160305	gold
182	48Z55	67318	Jeremy Bowers	20160101	gold
183	90081	16371	Annabel Miller	20160208	gold

184 rows × 5 columns

In [49]:
```
df.size,df.shape
```

Out[49]:
```
(920, (184, 5))
```

In [50]:
```
df.dtypes
# here changes should be made to the date column with a proper schema(date type)
```

Out[50]:
```
id                     object
person_id               int64
name                   object
membership_start_date   int64
membership_status      object
dtype: object
```

In [81]:
```
df.isna().sum()
# no null values
```

Out[81]:
```
id                      0
person_id               0
name                    0
membership_start_date   0
membership_status       0
dtype: int64
```

In [52]:
```
df.membership_status.unique()
```

Out[52]:
```
array(['gold', 'regular', 'silver'], dtype=object)
```

In [55]:
```
df.person_id.nunique()
```

Out[55]:
```
184
```

In [56]:
```
df.id.nunique()
# we can conclude by seeing the above cells , that each person has a unique person
```

Out[56]:
```
184
```

Income (Table 6)

In [82]:
```
df=pd.read_csv("income.csv")
```

In [59]:
```
df
```

Out[59]:

	ssn	annual_income
0	100009868	52200
1	100169584	64500
2	100300433	74400
3	100355733	35900
4	100366269	73000
...	...	...
7509	999679296	54400
7510	999762859	77000
7511	999824984	82000
7512	999910617	82600
7513	999942603	11500

7514 rows × 2 columns

In [60]:
```
df.size,df.shape
```

Out[60]:
```
(15028, (7514, 2))
```

In [61]:
```
df.dtypes
```

Out[61]:
```
ssn              int64
annual_income    int64
dtype: object
```

In [83]:
```
df.isna().sum()
# no null values
```

Out[83]:
```
ssn              0
annual_income    0
dtype: int64
```

In [63]:
```
df.annual_income.min(),df.annual_income.max()
# so the salary ranges from 10000 to 498500
```

Out[63]:
```
(10000, 498500)
```

In [64]:
```
df.ssn.nunique()
#Social Security Number is a unique nine-digit identification number issued by the
#SSN are used to verify a person's identity and employment eligibility and are cons
```

Out[64]:
```
7514
```

Interview (Table 7)

In [10]:
```
df=pd.read_csv("interview.csv")
```

In [66]:
```
df
```

Out[66]:

	person_id	transcript
0	28508	'I deny it!' said the March Hare.\n
1	63713	\n
2	86208	way, and the whole party swam to the shore.\n
3	35267	lessons in here? Why, there's hardly room for ...
4	33856	\n
...	...	...
4986	37357	Alice did not wish to offend the Dormouse agai...
4987	10206	time,' she said, 'than waste it in asking ridd...
4988	14887	I heard a gunshot and then saw a man run out. ...
4989	16371	I saw the murder happen, and I recognized the ...
4990	67318	I was hired by a woman with a lot of money. I ...

4991 rows × 2 columns

In [68]:
```
df.size,df.shape
```

Out[68]:
```
(9982, (4991, 2))
```

In [67]:
```
df.dtypes
```

Out[67]:
```
person_id      int64
transcript    object
dtype: object
```

In [85]:
```
df.isna().sum()
# no null values
```

Out[85]:
```
person_id     0
transcript    0
dtype: int64
```

In [69]:
```
df['transcript']
# some of the values in this column have incomplete values , like (3-4 tab spaces ,
```

Out[69]:
```
0                    ‘I deny it!’ said the March Hare.\n
1                                                     \n
2          way, and the whole party swam to the shore.\n
3       lessons in here? Why, there’s hardly room for ...
4                                                     \n
                              ...
4986    Alice did not wish to offend the Dormouse agai...
4987    time,’ she said, ‘than waste it in asking ridd...
4988    I heard a gunshot and then saw a man run out. ...
4989    I saw the murder happen, and I recognized the ...
4990    I was hired by a woman with a lot of money. I ...
Name: transcript, Length: 4991, dtype: object
```

In [70]:
```
df[df['transcript'] =='\n']
# see this kind of data should be transformed corrrectly
```

Out[70]:

	person_id	transcript
1	63713	\n
4	33856	\n
5	82799	\n
10	54206	\n
12	34615	\n
...	...	...
4960	22220	\n
4968	89706	\n
4972	54954	\n
4977	36345	\n
4982	41577	\n

1253 rows × 2 columns

In [11]:
```
df[df['transcript'] ="  *    *    *    *    *    *    *\n"]
# see this kind of data should be replaced by "transcription not given"
```

Out[11]:

	person_id	transcript
285	20247	* * * * * * *\n
629	84681	* * * * * * *\n
856	14373	* * * * * * *\n
945	29124	* * * * * * *\n
1236	14297	* * * * * * *\n
1472	44397	* * * * * * *\n
3152	36469	* * * * * * *\n
3197	27407	* * * * * * *\n
3434	90497	* * * * * * *\n
3805	52335	* * * * * * *\n
3851	70911	* * * * * * *\n
4407	19948	* * * * * * *\n

In [12]:
```
df[df['transcript'] ==" * * * * * *\n"]
# see this kind of data should be replaced by "transcription not given"
```

Out[12]:

	person_id	transcript
2249	24347	* * * * * *\n
2847	12103	* * * * * *\n
4444	10304	* * * * * *\n

Person (Table 8)

In [86]:
```
df=pd.read_csv("person.csv")
```

In [72]:
```
df
```

Out[72]:

	id	name	license_id	address_number	address_street_name	ssn
0	10000	Christoper Peteuil	993845	624	Bankhall Ave	747714076
1	10007	Kourtney Calderwood	861794	2791	Gustavus Blvd	477972044
2	10010	Muoi Cary	385336	741	Northwestern Dr	828638512
3	10016	Era Moselle	431897	1987	Wood Glade St	614621061
4	10025	Trena Hornby	550890	276	Daws Hill Way	223877684
...	...	...	...	...	...	...
10006	99936	Luba Benser	274427	680	Carnage Blvd	685095054
10007	99941	Roxana Mckimley	975942	1613	Gate St	512136801
10008	99965	Cherie Zeimantz	287627	3661	The Water Ave	362877324
10009	99982	Allen Cruse	251350	3126	N Jean Dr	348734531
10010	99990	Vance Hunten	830407	3056	Lancefield St	896677562

10011 rows × 6 columns

In [73]:
```
df.size,df.shape
```

Out[73]:
```
(60066, (10011, 6))
```

In [74]:
```
df.dtypes
# no correction required for the schema of datatype here
```

Out[74]:
```
id                    int64
name                 object
license_id            int64
address_number        int64
address_street_name  object
ssn                   int64
dtype: object
```

In [87]:
```
df.isna().sum()
# no null values
```

Out[87]:
```
id                   0
name                 0
license_id           0
address_number       0
address_street_name  0
ssn                  0
dtype: int64
```