

```
In [91]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [188... import warnings
warnings.filterwarnings("ignore")
```

```
In [92]: data1=pd.read_csv(r"D:\REV DATA SCIENCE\DATA cleaning\adverse_reactions.csv")
```

```
In [93]: patients=pd.read_csv(r"D:\REV DATA SCIENCE\DATA cleaning\patients.csv")
```

```
In [94]: treatments=pd.read_csv(r"D:\REV DATA SCIENCE\DATA cleaning\treatments.csv")
```

```
In [95]: treatments_cut=pd.read_csv(r"D:\REV DATA SCIENCE\DATA cleaning\treatments_cut.csv")
```

```
In [96]: data1.head(5)
```

```
Out[96]:
```

	given_name	surname	adverse_reaction
0	berta	napolitani	injection site discomfort
1	lena	baer	hypoglycemia
2	joseph	day	hypoglycemia
3	flavia	fiorentino	cough
4	manouck	wubbels	throat irritation

```
In [97]: patients.head(5)
```

```
Out[97]:
```

	patient_id	assigned_sex	given_name	surname	address	city	state	zip
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	92084
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	60009
2	3	male	Jae	Debord	1493 Poling Farm Road	York	Nebraska	68465
3	4	male	Liêm	Phan	2335 Webster Street	Woodbridge	NJ	07095
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	36023

```
In [98]: treatments.head(5)
```

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
0	veronika	jindrová	41u - 48u	-	7.63	7.20	NaN
1	elliott	richardson	-	40u - 45u	7.56	7.09	0.97
2	yukitaka	takenaka	-	39u - 36u	7.68	7.25	NaN
3	skye	gormanston	33u - 36u	-	7.97	7.62	0.35
4	alissa	montez	-	33u - 29u	7.78	7.46	0.32

```
In [99]: treatments_cut.head(5)
```

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
0	jožka	resanovič	22u - 30u	-	7.56	7.22	0.34
1	inunnguaq	heilmann	57u - 67u	-	7.85	7.45	NaN
2	alwin	svensson	36u - 39u	-	7.78	7.34	NaN
3	thế	lương	-	61u - 64u	7.64	7.22	0.92
4	amanda	ribeiro	36u - 44u	-	7.85	7.47	0.38

```
In [100]: patients.sample(3)
```

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_cod
138	139	male	Jose	Combs	718 Eden Drive	Richmond	VA	23220
495	496	male	Hajime	Tsukada	4111 Thunder Road	San Mateo	CA	94403
332	333	male	Abel	Efrem	2333 Hidden Pond Road	Old Hickory	TN	37138

```
In [101]: patients.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 503 entries, 0 to 502
Data columns (total 14 columns):
#   Column          Non-Null Count  Dtype
---  -
0   patient_id      503 non-null    int64
1   assigned_sex    503 non-null    object
2   given_name      503 non-null    object
3   surname         503 non-null    object
4   address         491 non-null    object
5   city            491 non-null    object
6   state           491 non-null    object
7   zip_code        491 non-null    float64
8   country         491 non-null    object
9   contact         491 non-null    object
10  birthdate       503 non-null    object
11  weight          503 non-null    float64
12  height          503 non-null    int64
13  bmi             503 non-null    float64
dtypes: float64(3), int64(2), object(9)
memory usage: 55.1+ KB

```

```
In [102... patients.isnull().sum()
```

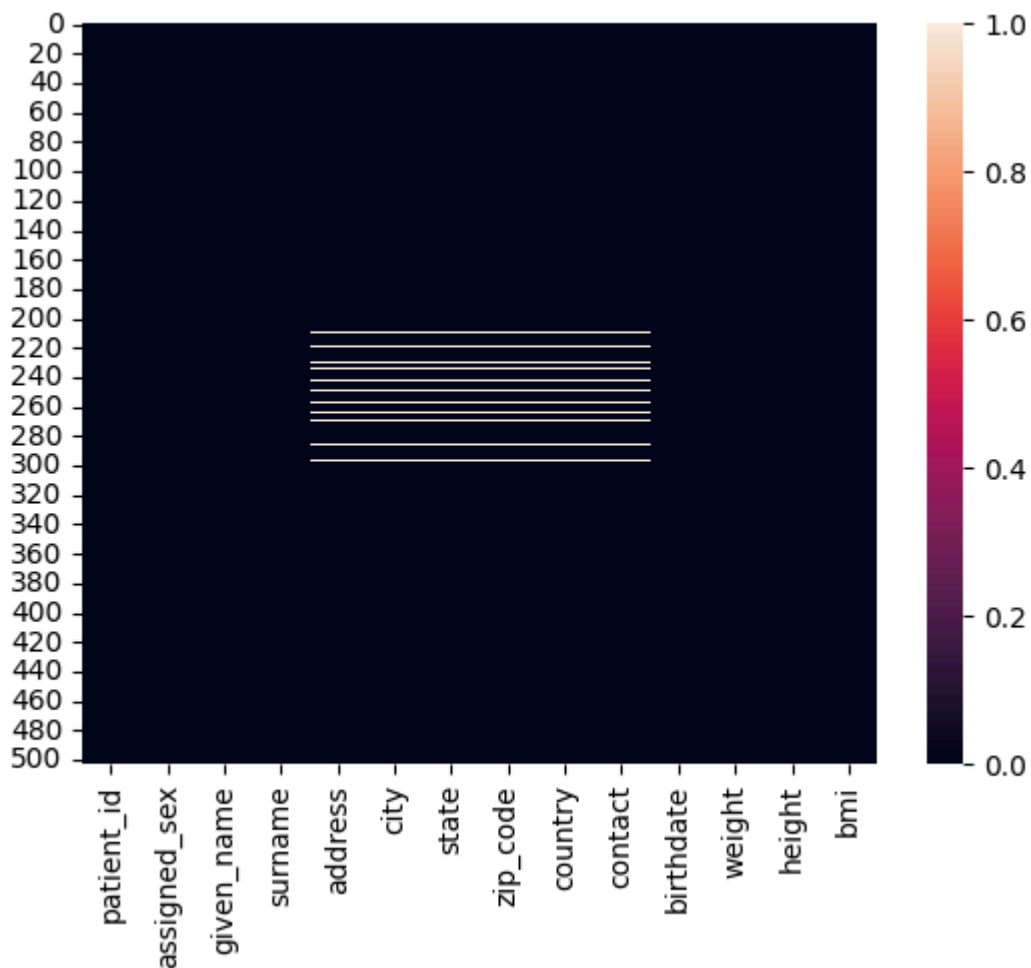
```

Out[102... patient_id      0
assigned_sex  0
given_name   0
surname      0
address      12
city         12
state        12
zip_code     12
country      12
contact      12
birthdate    0
weight       0
height       0
bmi          0
dtype: int64

```

```
In [103... sns.heatmap(patients.isnull())
```

```
Out[103... <Axes: >
```



```
In [104... patients[patients['address'].isnull()]
```

```
Out[104...
```

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code
209	210	female	Lalita	Eldarkhanov	NaN	NaN	NaN	NaN
219	220	male	Mỹ	Quynh	NaN	NaN	NaN	NaN
230	231	female	Elisabeth	Knudsen	NaN	NaN	NaN	NaN
234	235	female	Martina	Tománková	NaN	NaN	NaN	NaN
242	243	male	John	O'Brian	NaN	NaN	NaN	NaN
249	250	male	Benjamin	Mehler	NaN	NaN	NaN	NaN
257	258	male	Jin	Kung	NaN	NaN	NaN	NaN
264	265	female	Wafiyyah	Asfour	NaN	NaN	NaN	NaN
269	270	female	Flavia	Fiorentino	NaN	NaN	NaN	NaN
278	279	female	Generosa	Cabán	NaN	NaN	NaN	NaN
286	287	male	Lewis	Webb	NaN	NaN	NaN	NaN
296	297	female	Chi	Lâm	NaN	NaN	NaN	NaN

```
In [105... treatments.sample(3)
```

Out[105...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
189	mee	chung	42u - 56u	-	7.56	7.20	0.36
277	mathea	lillebø	23u - 36u	-	9.04	8.67	0.37
13	gregor	bole	-	47u - 45u	7.61	7.16	0.95

In [106...

```
treatments.isna().sum()
```

Out[106...

```
given_name      0
surname         0
auralin         0
novodra         0
hba1c_start     0
hba1c_end       0
hba1c_change    109
dtype: int64
```

In [107...

```
treatments.isna().sum()
```

Out[107...

```
given_name      0
surname         0
auralin         0
novodra         0
hba1c_start     0
hba1c_end       0
hba1c_change    109
dtype: int64
```

In [108...

```
treatments_cut.sample(3)
```

Out[108...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
1	inunnguaq	heilmann	57u - 67u	-	7.85	7.45	NaN
56	blanka	jurković	31u - 43u	-	7.77	7.33	NaN
26	firenze	fodor	-	30u - 35u	7.89	7.55	0.34

In [109...

```
treatments_cut.sample(3)
```

Out[109...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
31	galit	casárez	27u - 37u	-	7.91	7.56	0.35
24	lóa	hrafnsdóttir	22u - 31u	-	7.60	7.09	NaN
50	timothy	cotton	-	26u - 25u	7.92	7.52	0.90

```
In [110... treatments_cut.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 70 entries, 0 to 69  
Data columns (total 7 columns):  
#   Column          Non-Null Count  Dtype  
---  ---  
0   given_name      70 non-null    object  
1   surname         70 non-null    object  
2   auralin         70 non-null    object  
3   novodra         70 non-null    object  
4   hba1c_start     70 non-null    float64  
5   hba1c_end       70 non-null    float64  
6   hba1c_change    42 non-null    float64  
dtypes: float64(3), object(4)  
memory usage: 4.0+ KB
```

```
In [111... patients.duplicated().sum()
```

```
Out[111... 0
```

```
In [112... patients.duplicated(subset=['given_name','surname']).sum()
```

```
Out[112... 5
```

```
In [113... patients[patients.duplicated(subset=['given_name','surname'])]
```

```
Out[113...      patient_id  assigned_sex  given_name  surname  address  city  state  zip_code  cc
```

229	230	male	John	Doe	123 Main Street	New York	NY	12345.0	
237	238	male	John	Doe	123 Main Street	New York	NY	12345.0	
244	245	male	John	Doe	123 Main Street	New York	NY	12345.0	
251	252	male	John	Doe	123 Main Street	New York	NY	12345.0	
277	278	male	John	Doe	123 Main Street	New York	NY	12345.0	

◀ ◻ ▶

```
In [114... treatments.duplicated().sum()
```

```
Out[114... 1
```

```
In [115... treatments[treatments.duplicated()]
```

Out[115...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
136	joseph	day	29u - 36u	-	7.7	7.19	NaN

In [116... `treatments[treatments.duplicated(subset=['given_name','surname'])]`

Out[116...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
136	joseph	day	29u - 36u	-	7.7	7.19	NaN

In [117... `treatments_cut.duplicated().sum()`

Out[117... 0

In [118... `treatments_cut[treatments_cut.duplicated(subset=['given_name','surname'])]`

Out[118...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
--	------------	---------	---------	---------	-------------	-----------	--------------

In [119... `data1.sample(4)`

Out[119...

	given_name	surname	adverse_reaction
3	flavia	fiorentino	cough
33	krisztina	magyar	hypoglycemia
9	sofia	hermansen	injection site discomfort
0	berta	napolitani	injection site discomfort

In [120... `data1.duplicated().sum()`

Out[120... 0

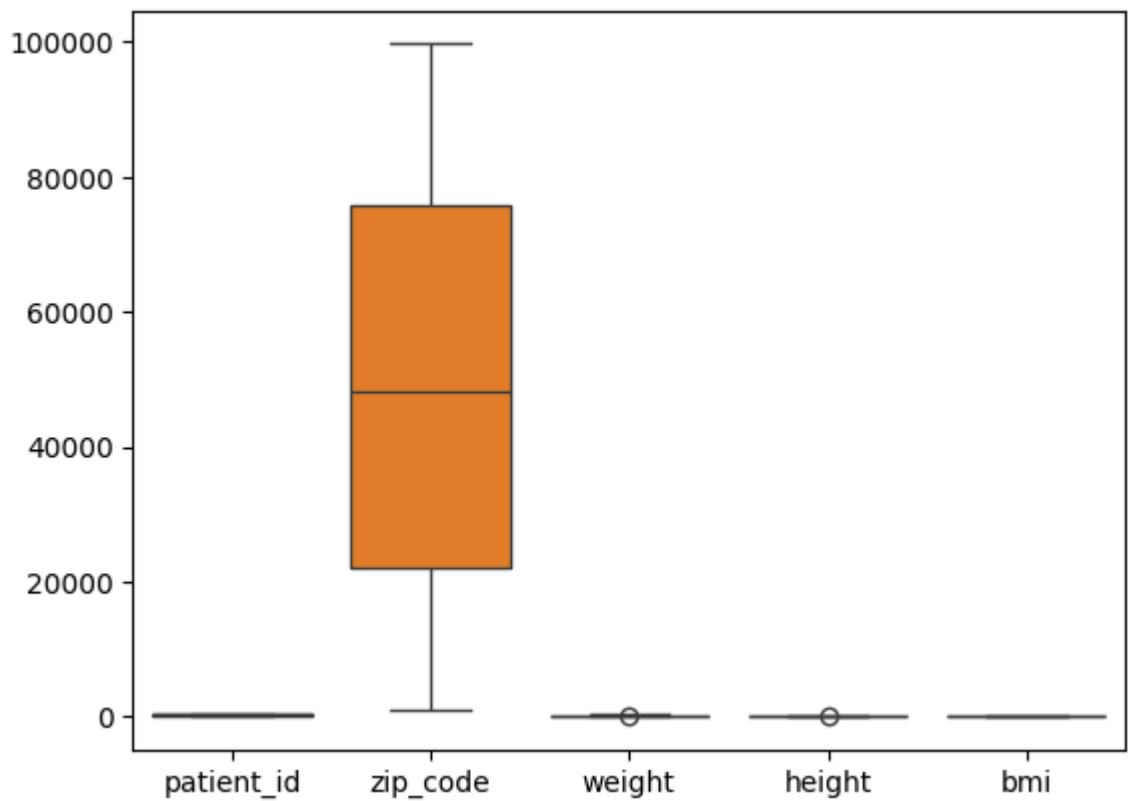
In [121... `patients.describe()`

Out[121...

	patient_id	zip_code	weight	height	bmi
count	503.000000	491.000000	503.000000	503.000000	503.000000
mean	252.000000	49084.118126	173.434990	66.634195	27.483897
std	145.347859	30265.807442	33.916741	4.411297	5.276438
min	1.000000	1002.000000	48.800000	27.000000	17.100000
25%	126.500000	21920.500000	149.300000	63.000000	23.300000
50%	252.000000	48057.000000	175.300000	67.000000	27.200000
75%	377.500000	75679.000000	199.500000	70.000000	31.750000
max	503.000000	99701.000000	255.900000	79.000000	37.700000

```
In [122... sns.boxplot(patients)
```

```
Out[122... <Axes: >
```



```
In [123... patients[patients['weight']==48.800000]
```

```
Out[123... patient_id assigned_sex given_name surname address city state zip_code
```

210	211	female	Camilla	Zaitseva	1428 Turkey Pen Lane	Dothan	AL	36303.0
-----	-----	--------	---------	----------	----------------------	--------	----	---------



```
In [124... patients[patients['height']==27.000000]
```

```
Out[124... patient_id assigned_sex given_name surname address city state zip_code
```

4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	36303.0
---	---	------	-----	---------	----------------------	--------	----	---------



```
In [125... treatments.describe()
```


Out[125...

	hba1c_start	hba1c_end	hba1c_change
count	280.000000	280.000000	171.000000
mean	7.985929	7.589286	0.546023
std	0.568638	0.569672	0.279555
min	7.500000	7.010000	0.200000
25%	7.660000	7.270000	0.340000
50%	7.800000	7.420000	0.380000
75%	7.970000	7.570000	0.920000
max	9.950000	9.580000	0.990000

In [126...

```
treatments.sort_values('hba1c_start')
```

Out[126...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
270	mika	martinsson	34u - 43u	-	7.50	7.17	0.33
113	kari	laatikainen	39u - 43u	-	7.50	7.11	NaN
126	jowita	wiśniewska	-	22u - 23u	7.50	7.08	0.92
53	nasser	mansour	-	33u - 31u	7.51	7.06	0.95
105	finlay	sheppard	-	31u - 30u	7.51	7.17	0.34
...
25	benoît	bonami	-	44u - 43u	9.82	9.40	0.92
171	justyna	kowalczyk	24u - 34u	-	9.84	9.44	NaN
81	robert	wagner	43u - 49u	-	9.84	9.52	0.32
75	mackenzie	mckay	-	44u - 43u	9.87	9.48	0.39
166	annie	allen	36u - 42u	-	9.95	9.58	0.37

280 rows × 7 columns



In [127...

```
treatments_cut.describe()
```

Out[127...

	hba1c_start	hba1c_end	hba1c_change
count	70.000000	70.000000	42.000000
mean	7.838000	7.443143	0.518810
std	0.423007	0.418706	0.270719
min	7.510000	7.020000	0.280000
25%	7.640000	7.232500	0.340000
50%	7.730000	7.345000	0.370000
75%	7.860000	7.467500	0.907500
max	9.910000	9.460000	0.970000

In [128...

```
patients_df=patients.copy()
```

In [129...

```
treatments_df=treatments.copy()
```

In [130...

```
treatments_cut_df=treatments_cut.copy()
```

In [131...

```
patients_df.isnull().sum()
```

Out[131...

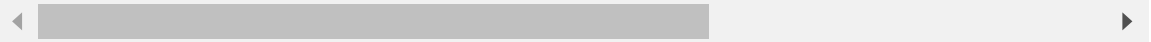
```
patient_id      0
assigned_sex    0
given_name      0
surname         0
address        12
city           12
state          12
zip_code       12
country        12
contact        12
birthdate       0
weight          0
height          0
bmi            0
dtype: int64
```

In [132...

```
patients_df[patients_df['address'].isnull()]
```

Out[132...

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code
209	210	female	Lalita	Eldarkhanov	NaN	NaN	NaN	NaN
219	220	male	Mỹ	Quynh	NaN	NaN	NaN	NaN
230	231	female	Elisabeth	Knudsen	NaN	NaN	NaN	NaN
234	235	female	Martina	Tománková	NaN	NaN	NaN	NaN
242	243	male	John	O'Brian	NaN	NaN	NaN	NaN
249	250	male	Benjamin	Mehler	NaN	NaN	NaN	NaN
257	258	male	Jin	Kung	NaN	NaN	NaN	NaN
264	265	female	Wafiyyah	Asfour	NaN	NaN	NaN	NaN
269	270	female	Flavia	Fiorentino	NaN	NaN	NaN	NaN
278	279	female	Generosa	Cabán	NaN	NaN	NaN	NaN
286	287	male	Lewis	Webb	NaN	NaN	NaN	NaN
296	297	female	Chi	Lâm	NaN	NaN	NaN	NaN



In [199...

```
patients_df.sample(10)
```

Out[199...

	patient_id	assigned_sex	given_name	surname	address	city	state
242	243	male	John	O'Brian	no-data	no-data	no-data
315	316	male	Brancaleone	Russo	2074 Parrish Avenue	San Antonio	TX
92	93	male	Michael	Smith	666 Whiteman Street	Mount Holly	NJ
378	379	female	Furuta	Osman	3300 Woodridge Lane	Memphis	TN
232	233	female	Kyouko	Ono	435 Pike Street	San Diego	CA
363	364	female	Elisabetta	Lorenzo	1014 Locust Court	Los Angeles	California
199	200	male	Zdeněk	Synek	3818 Strother Street	Dora	AL
482	483	male	Diogo	Souza	4033 White Avenue	Corpus Christi	TX
58	59	female	Yasmin	Araujo	3682 Stiles Street	Pittsburgh	PA
352	353	male	Marek	Dvořák	633 Better Street	Savannah	GA

In [201...

```
patients_df[patients_df['address']=='no-data']
```

Out[201...

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code
209	210	female	Lalita	Eldarkhanov	no-data	no-data	no-data	no-data
219	220	male	Mỹ	Quynh	no-data	no-data	no-data	no-data
230	231	female	Elisabeth	Knudsen	no-data	no-data	no-data	no-data
234	235	female	Martina	Tománková	no-data	no-data	no-data	no-data
242	243	male	John	O'Brian	no-data	no-data	no-data	no-data
249	250	male	Benjamin	Mehler	no-data	no-data	no-data	no-data
257	258	male	Jin	Kung	no-data	no-data	no-data	no-data
264	265	female	Wafiyyah	Asfour	no-data	no-data	no-data	no-data
269	270	female	Flavia	Fiorentino	no-data	no-data	no-data	no-data
278	279	female	Generosa	Cabán	no-data	no-data	no-data	no-data
286	287	male	Lewis	Webb	no-data	no-data	no-data	no-data
296	297	female	Chi	Lâm	no-data	no-data	no-data	no-data

In [134...

patients_df.info()

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 503 entries, 0 to 502
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   patient_id            503 non-null    int64
1   assigned_sex          503 non-null    object
2   given_name            503 non-null    object
3   surname               503 non-null    object
4   address               503 non-null    object
5   city                 503 non-null    object
6   state                503 non-null    object
7   zip_code             503 non-null    object
8   country              503 non-null    object
9   contact              503 non-null    object
10  birthdate            503 non-null    object
11  weight               503 non-null    float64
12  height               503 non-null    int64
13  bmi                 503 non-null    float64
dtypes: float64(2), int64(2), object(10)
memory usage: 55.1+ KB

```

In [135... treatments_df.sample(5)

Out[135...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
250	chen	yao	-	56u - 57u	7.90	7.51	0.39
90	karen	jakobsen	34u - 42u	-	7.59	7.25	0.34
157	asuna	morita	-	35u - 39u	7.58	7.25	0.33
58	gabryś	tomaszewski	29u - 37u	-	7.87	7.47	NaN
273	kate	wilkinson	36u - 39u	-	7.72	7.20	NaN

◀ ▶

In [136... treatments_df.isnull().sum()

Out[136...

```

given_name      0
surname         0
auralin         0
novodra         0
hba1c_start     0
hba1c_end       0
hba1c_change    109
dtype: int64

```

In [137... treatments_cut_df.isnull().sum()

```
Out[137... given_name      0
            surname    0
            auralin    0
            novodra    0
            hba1c_start 0
            hba1c_end   0
            hba1c_change 28
            dtype: int64
```

```
In [138... treatments_df['hba1c_change']=treatments_df['hba1c_start']+treatments_df['hba1c_
```

```
In [139... treatments_cut_df['hba1c_change']=treatments_df['hba1c_start']+treatments_df['hb
```

```
In [140... treatments_cut_df.isnull().sum()
```

```
Out[140... given_name      0
            surname    0
            auralin    0
            novodra    0
            hba1c_start 0
            hba1c_end   0
            hba1c_change 0
            dtype: int64
```

```
In [141... patients_df.sample(3)
```

```
Out[141...      patient_id  assigned_sex  given_name  surname  address  city  state  zi
```

100	101	male	Isac	Berg	1497 Hidden Meadow Drive	Binford	ND	!
349	350	male	Kristoffer	Martinsen	1865 Honeysuckle Lane	Packwood	WA	!
118	119	male	Adib	Ghanem	3457 Bridge Avenue	Delcambre	LA	!

◀ ◻ ▶

```
In [142... patients_df[['phone_number', 'email']] = patients_df['contact'].str.extract(r'(\
```

```
In [143... patients_df.head()
```

Out[143...

	patient_id	assigned_sex	given_name	surname	address	city	state	zip
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	9
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	6
2	3	male	Jae	Debord	1493 Poling Farm Road	York	Nebraska	6
3	4	male	Liêm	Phan	2335 Webster Street	Woodbridge	NJ	
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	3

In [144...

```
treatments_df=pd.concat([treatments_df,treatments_cut_df])
```

In [145...

```
treatments_df
```


Out[145...

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
0	veronika	jindrová	41u - 48u	-	7.63	7.20	14.83
1	elliott	richardson	-	40u - 45u	7.56	7.09	14.65
2	yukitaka	takenaka	-	39u - 36u	7.68	7.25	14.93
3	skye	gormanston	33u - 36u	-	7.97	7.62	15.59
4	alissa	montez	-	33u - 29u	7.78	7.46	15.24
...
65	rovzan	kishiev	32u - 37u	-	7.75	7.41	14.88
66	jakob	jakobsen	-	28u - 26u	7.96	7.51	14.95
67	bernd	schneider	48u - 56u	-	7.74	7.44	18.24
68	berta	napolitani	-	42u - 44u	7.68	7.21	15.44
69	armina	sauvé	36u - 46u	-	7.86	7.40	17.85

350 rows × 7 columns

In [146...

```
treat.isnull().sum()
```

Out[146...

```
given_name      0
surname         0
auralin         0
novodra         0
hba1c_start     0
hba1c_end       0
hba1c_change    0
dtype: int64
```

In [147...

```
treat.iloc[200]
```

Out[147...

```
given_name      nicolas
surname         ferreira
auralin         43u - 51u
novodra         -
hba1c_start     7.99
hba1c_end       7.72
hba1c_change    15.71
Name: 200, dtype: object
```

In [148...

```
treatments_df=treatments_df.melt(id_vars=['given_name' , 'surname' , 'hba1c_
```

In [149...

treatments_df

Out[149...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	dosage
0	veronika	jindrová	7.63	7.20	14.83	auralin	410
1	elliott	richardson	7.56	7.09	14.65	auralin	
2	yukitaka	takenaka	7.68	7.25	14.93	auralin	
3	skye	gormanston	7.97	7.62	15.59	auralin	330
4	alissa	montez	7.78	7.46	15.24	auralin	
...
695	rovzan	kishiev	7.75	7.41	14.88	novodra	
696	jakob	jakobsen	7.96	7.51	14.95	novodra	280
697	bernd	schneider	7.74	7.44	18.24	novodra	
698	berta	napolitani	7.68	7.21	15.44	novodra	420
699	armina	sauvé	7.86	7.40	17.85	novodra	

700 rows × 7 columns



In [150...

treatments_df=treatments_df[treatments_df['dosage_range']!='-']

In [151...

treatments_df

Out[151...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	dosage
0	veronika	jindrová	7.63	7.20	14.83	auralin	410
3	skye	gormanston	7.97	7.62	15.59	auralin	330
6	sophia	haugen	7.65	7.27	14.92	auralin	370
7	eddie	archer	7.89	7.55	15.44	auralin	310
9	asia	woźniak	7.76	7.37	15.13	auralin	300
...
688	christopher	woodward	7.51	7.06	15.34	novodra	550
690	marek	sulygov	7.67	7.30	14.70	novodra	260
694	lixue	hsueh	9.21	8.80	15.04	novodra	220
696	jakob	jakobsen	7.96	7.51	14.95	novodra	280
698	berta	napolitani	7.68	7.21	15.44	novodra	420

350 rows × 7 columns



```
In [ ]: treatments_df['start'] = treatments_df['dosage_range'].str.split(' - ').str.get(0)
treatments_df['end'] = treatments_df['dosage_range'].str.split(' - ').str.get(1)
```

```
In [153... treatments_df
```

```
Out[153...      given_name  surname  hba1c_start  hba1c_end  hba1c_change  type  dosage
0      veronika  jindrová      7.63      7.20      14.83  auralin  41i
3         skye  gormanston      7.97      7.62      15.59  auralin  33i
6        sophia    haugen      7.65      7.27      14.92  auralin  37i
7         eddie    archer      7.89      7.55      15.44  auralin  31i
9          asia   woźniak      7.76      7.37      15.13  auralin  30i
...         ...      ...      ...      ...      ...      ...
688  christopher  woodward      7.51      7.06      15.34  novodra  55i
690        maret   sulygov      7.67      7.30      14.70  novodra  26i
694        lixue    hsueh      9.21      8.80      15.04  novodra  22i
696        jakob   jakobsen      7.96      7.51      14.95  novodra  28i
698        berta  napolitani      7.68      7.21      15.44  novodra  42i
```

350 rows × 9 columns



```
In [157... treatments_df=treatments_df.drop('dosage_range',axis=1)
```

```
In [158... treatments_df
```

Out[158...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	start	end
0	veronika	jindrová	7.63	7.20	14.83	auralin	41u	4
3	skye	gormanston	7.97	7.62	15.59	auralin	33u	3
6	sophia	haugen	7.65	7.27	14.92	auralin	37u	4
7	eddie	archer	7.89	7.55	15.44	auralin	31u	3
9	asia	woźniak	7.76	7.37	15.13	auralin	30u	3
...
688	christopher	woodward	7.51	7.06	15.34	novodra	55u	!
690	maret	sulygov	7.67	7.30	14.70	novodra	26u	!
694	lixue	hsueh	9.21	8.80	15.04	novodra	22u	!
696	jakob	jakobsen	7.96	7.51	14.95	novodra	28u	!
698	berta	napolitani	7.68	7.21	15.44	novodra	42u	!

350 rows × 8 columns



In [167...

```
treatments_df['start']=treatments_df['start'].str.replace('u','')
```

In [168...

```
treatments_df
```

Out[168...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	start	end
0	veronika	jindrová	7.63	7.20	14.83	auralin	41	4
3	skye	gormanston	7.97	7.62	15.59	auralin	33	3
6	sophia	haugen	7.65	7.27	14.92	auralin	37	4
7	eddie	archer	7.89	7.55	15.44	auralin	31	3
9	asia	woźniak	7.76	7.37	15.13	auralin	30	3
...
688	christopher	woodward	7.51	7.06	15.34	novodra	55	!
690	maret	sulygov	7.67	7.30	14.70	novodra	26	!
694	lixue	hsueh	9.21	8.80	15.04	novodra	22	!
696	jakob	jakobsen	7.96	7.51	14.95	novodra	28	!
698	berta	napolitani	7.68	7.21	15.44	novodra	42	!

350 rows × 8 columns



In [169...

```
treatments_df['end']=treatments_df['end'].str.replace('u','')
```

In [170...

```
treatments_df
```

Out[170...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	start	end
0	veronika	jindrová	7.63	7.20	14.83	auralin	41	48
3	skye	gormanston	7.97	7.62	15.59	auralin	33	40
6	sophia	haugen	7.65	7.27	14.92	auralin	37	44
7	eddie	archer	7.89	7.55	15.44	auralin	31	38
9	asia	woźniak	7.76	7.37	15.13	auralin	30	37
...
688	christopher	woodward	7.51	7.06	15.34	novodra	55	62
690	marek	sulygov	7.67	7.30	14.70	novodra	26	33
694	lixue	hsueh	9.21	8.80	15.04	novodra	22	29
696	jakob	jakobsen	7.96	7.51	14.95	novodra	28	35
698	berta	napolitani	7.68	7.21	15.44	novodra	42	49

350 rows × 8 columns



In [187...

```
treatments_df[treatments_df['given_name']=='veronika']
```

Out[187...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	start	end
0	veronika	jindrová	7.63	7.2	14.83	auralin	41	48



In [173...

```
data1
```

Out[173...

	given_name	surname	adverse_reaction
0	berta	napolitani	injection site discomfort
1	lena	baer	hypoglycemia
2	joseph	day	hypoglycemia
3	flavia	fiorentino	cough
4	manouck	wubbels	throat irritation
5	jasmine	sykes	hypoglycemia
6	louise	johnson	hypoglycemia
7	albinca	komavec	hypoglycemia
8	noe	aranda	hypoglycemia
9	sofia	hermansen	injection site discomfort
10	tegan	johnson	headache
11	abel	yonatan	cough
12	abdul-nur	isa	hypoglycemia
13	leon	scholz	injection site discomfort
14	gabriele	saenger	hypoglycemia
15	jia li	teng	nausea
16	jakob	jakobsen	hypoglycemia
17	christopher	woodward	nausea
18	ole	petersen	hypoglycemia
19	finley	chandler	headache
20	anenechi	chidi	hypoglycemia
21	miłosław	wiśniewski	injection site discomfort
22	lixue	hsueh	injection site discomfort
23	merci	leroux	hypoglycemia
24	kang	mai	injection site discomfort
25	elliott	richardson	hypoglycemia
26	clinton	miller	throat irritation
27	idalia	moore	hypoglycemia
28	xiuxiu	chang	hypoglycemia
29	alex	crawford	hypoglycemia
30	monika	lončar	hypoglycemia
31	steven	roy	headache
32	cecilie	nilsen	hypoglycemia

	given_name	surname	adverse_reaction
33	krisztina	magyar	hypoglycemia

In [175... treatments_df=treatments_df.merge(data1, how='left', on=['given_name', 'surname'])

In [176... treatments_df

Out[176...

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	start	end
0	veronika	jindrová	7.63	7.20	14.83	auralin	41	
1	skye	gormanston	7.97	7.62	15.59	auralin	33	
2	sophia	haugen	7.65	7.27	14.92	auralin	37	
3	eddie	archer	7.89	7.55	15.44	auralin	31	
4	asia	woźniak	7.76	7.37	15.13	auralin	30	
...
345	christopher	woodward	7.51	7.06	15.34	novodra	55	
346	marek	sulygov	7.67	7.30	14.70	novodra	26	
347	lixue	hsueh	9.21	8.80	15.04	novodra	22	
348	jakob	jakobsen	7.96	7.51	14.95	novodra	28	
349	berta	napolitani	7.68	7.21	15.44	novodra	42	

350 rows × 9 columns



In [183... treatments_df['adverse_reaction'].notna().sum()

Out[183... 35

In [190... treatments_df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 350 entries, 0 to 349
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   given_name            350 non-null    object
1   surname               350 non-null    object
2   hba1c_start           350 non-null    float64
3   hba1c_end             350 non-null    float64
4   hba1c_change          350 non-null    float64
5   type                  350 non-null    object
6   start                 350 non-null    object
7   end                   350 non-null    object
8   adverse_reaction      35 non-null     object
dtypes: float64(3), object(6)
memory usage: 24.7+ KB
```

```
In [195... treatments_df['start']=treatments_df['start'].astype('int64')
```

```
In [194... treatments_df['end']=treatments_df['end'].astype('int64')
```

```
In [196... treatments_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 350 entries, 0 to 349
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   given_name            350 non-null    object
1   surname               350 non-null    object
2   hba1c_start           350 non-null    float64
3   hba1c_end             350 non-null    float64
4   hba1c_change          350 non-null    float64
5   type                  350 non-null    object
6   start                 350 non-null    int64
7   end                   350 non-null    int64
8   adverse_reaction      35 non-null     object
dtypes: float64(3), int64(2), object(4)
memory usage: 24.7+ KB
```

```
In [197... treatments_df
```

```
Out[197...      given_name  surname  hba1c_start  hba1c_end  hba1c_change  type  start  €
0      veronika  jindrová      7.63      7.20      14.83  auralin    41
1         skye  gormanston      7.97      7.62      15.59  auralin    33
2        sophia    haugen      7.65      7.27      14.92  auralin    37
3         eddie    archer      7.89      7.55      15.44  auralin    31
4         asia    woźniak      7.76      7.37      15.13  auralin    30
...         ...         ...         ...         ...         ...         ...
345  christopher  woodward      7.51      7.06      15.34  novodra    55
346        maret  sulygov      7.67      7.30      14.70  novodra    26
347        lixue    hsueh      9.21      8.80      15.04  novodra    22
348        jakob  jakobsen      7.96      7.51      14.95  novodra    28
349        berta  napolitani      7.68      7.21      15.44  novodra    42
```

350 rows × 9 columns



```
In [ ]:
```