In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
patients= pd.read_csv(r"C:\Users\Jai Shree Shyam\Desktop\data cleaning files\patients.csv")
treatments= pd.read_csv(r"C:\Users\Jai Shree Shyam\Desktop\data cleaning files\treatments.csv")
adverse_reaction= pd.read_csv(r"C:\Users\Jai Shree Shyam\Desktop\data cleaning files\adverse_reactions.csv")
treatments_cut= pd.read_csv(r"C:\Users\Jai Shree Shyam\Desktop\data cleaning files\treatments_cut.csv")
```

In [3]:
```
print(patients.shape)
print(treatments.shape)
print(adverse_reaction.shape)
print(treatments_cut.shape)
```

```
(503, 14)
(280, 7)
(34, 3)
(70, 7)
```

In [4]:
```
# Export data for manual assessment
with pd.ExcelWriter("clinical_trails.xlsx") as writer:
    patients.to_excel(writer, sheet_name="patients")
    treatments.to_excel(writer,sheet_name="treatments")
    treatments_cut.to_excel(writer,sheet_name= "treatments_cut")
    adverse_reaction.to_excel(writer,sheet_name="adverse_reaction")
```

Data Description

This is a dataset about 503patients of which 350 patients participated in a clinical trail. None of the patients were using Novodra (a popular injectale insulin) or auralin(the oral insulin) being researched as their primary source of insulin before. All were experiencing elevated hbaIc levels.

All 350 patients were treated with Novodra to establish a baseline. HbaIc level and insulin dose. After 4 weeks which is it enough time to capture all the change in hba1c that can be attributed by the switch to auralin or novodra.

175 patients switched to auralin to 24 weeks. 175 patients switched or remain to novodra for 24 weeks.

Data about patients feeling some adverse effect is also recorded.

Description about columns of the data.

Table:- Patients

- patient_id- patient id of all the patient is given 1 to 503 total 503 patients are there.
- assigned_sex- patient sex is given ["Male", "Female"].
- given_name- name of the patient is given.
- surname- surname of each patient is given.
- address- Address is given starting from a digits and followed by alphabets.
- city- city name is given of particular person living in that locality.
- State- State name is given of particular person belongs to that state.
- zip_code- Zip code of his or her locality is given and must be of 5 digit.
- country- country name is given, united states is given because all the data regarding patients is of united states.
- birthdate- date of birth of the patients is given, in which year they borned.
- weight- weight of the patients is given in (pounds)
- height- height of the patients is given in (inches)
- bmi- body mass index is given of every patient.

Table:- treatments and treatement_cut

- given_name- volunteer name is given who take part in clinical trials.
- surname- surname of particular volunteer is given who take part in clinical trials.
- auralin- auralin is the name of oral insulin drug which is given to volunteer consisting usage of auralin [41u - 48u] is given starting usage before 24 weeks and after 24 weeks.
- Novodra- novodra is the name of the injectable insulin drug which is given to volunteer consisting usage of novodra [41u - 48u] is given starting usage before 24 weeks and after 24 weeks.
- hba1c_start- It measures the level of sugar present in hemaglobin. hba1c_start means the rate before the volunteer has.
- hba1c_end - It measures the level of sugar present in hemaglobin. hba1c_end means the rate after the volunteer has after 24 weeks.
- hba1c_change- It measures the change in hba1c_start and hba1c_end after 24 weeks after clinical trials is completed.

Table:- adverse_reaction

- given_name:- volunteers name is given who participated in the clinical trails.
- surname:- surname is given of the volunteers who participated in the clinical trails.
- adverse_reaction:- adverse_reaction is given to certain volunteer after taking trails of the doses.

Data Assesment

Dirty Data [content issue]

Table [Patient]

- given_name column has row number 9 has dsvid must be correct as David. [Accuracy]
- state column has some name written fully and some are in abbrevations. [Consistency]
- zip_code column has must be 5 digit but have 4 digit is given. [Validity]
- some columns have nan values in address, city,state,zip_code, country and contact consitisting of 12 patients.[completeness].
- john Doe patients is duplicated 5 times in the data have to delete it.[validity]
- patient_id 211 female Camilla Zaitseva has weight around 48 pounds which is not possible.[Accuracy]
- patient_id 5 male Tim Neudorf has height of 27 inches which is not possible.[Accuracy]

Table [treatments and treatments_cut]

- given_name and surname having all small letter have to change in capitalize first word.[consistency]
- auralin and novodra having additional unit u h ave to delete it.[validity]
- hbaIc change coloumn have some Missing value.[completeness].
- patient_id 139 joseph day is repeated or duplicated in the data[validity]
- treatments and treatments_cut colummn have some misvalue return in hba1c coloumn [Accuracy.]

Table [adverse_reaction]

- column Name given, surname must be in capitalize first word. [consistency].

Messy Data [structural issue]

Table [Patient]

- Column contact has mixed phone no. and and email id have to seprate one column into two column phone no. and email id.

Table[treatments and treatments_cut]

- Column auralin and novodra having starting and ending both dosage have to split into two different columns.
- auralin and Novodra both have to in one column.

Table[adverse_reaction]

- adverse_reaction doesn't have seprate table must have to merge with treatment table.

In [5]:
```
## Making copy of all the original files.
patients_df= patients.copy()
treatments_df=treatments.copy()
treatments_cut_df=treatments_cut.copy()
adverse_reactions_df= adverse_reaction.copy()
```

Patients_df

In [6]:
```
## Patients
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       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 [7]:
```
patients_df[patients_df["address"].isna()]
# These 12 patients has missing values in all {address, city,state,zip_code, country, contact}
```

Out[7]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdate	weight	height	bmi
209	210	female	Lalita	Eldarkhanov	NaN	NaN	NaN	NaN	NaN	NaN	8/14/1950	143.4	62	26.2
219	220	male	Mỹ	Quynh	NaN	NaN	NaN	NaN	NaN	NaN	4/9/1978	237.8	69	35.1
230	231	female	Elisabeth	Knudsen	NaN	NaN	NaN	NaN	NaN	NaN	9/23/1976	165.9	63	29.4
234	235	female	Martina	Tománková	NaN	NaN	NaN	NaN	NaN	NaN	4/7/1936	199.5	65	33.2
242	243	male	John	O'Brian	NaN	NaN	NaN	NaN	NaN	NaN	2/25/1957	205.3	74	26.4
249	250	male	Benjamin	Mehler	NaN	NaN	NaN	NaN	NaN	NaN	10/30/1951	146.5	69	21.6
257	258	male	Jin	Kung	NaN	NaN	NaN	NaN	NaN	NaN	5/17/1995	231.7	69	34.2
264	265	female	Wafiyyah	Asfour	NaN	NaN	NaN	NaN	NaN	NaN	11/3/1989	158.6	63	28.1
269	270	female	Flavia	Fiorentino	NaN	NaN	NaN	NaN	NaN	NaN	10/9/1937	175.2	61	33.1
278	279	female	Generosa	Cabán	NaN	NaN	NaN	NaN	NaN	NaN	12/16/1962	124.3	69	18.4
286	287	male	Lewis	Webb	NaN	NaN	NaN	NaN	NaN	NaN	4/1/1979	155.3	68	23.6
296	297	female	Chỉ	Lâm	NaN	NaN	NaN	NaN	NaN	NaN	5/14/1990	181.1	63	32.1

In [8]:
```
# patient. duplicated
patients_df[patients_df.duplicated(subset=["given_name","surname"])]
# john Doe patient is duplicated 5 times in the data, have to delete as it becomes redundant information for us.
```

Out[8]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdate	weight	height	bmi
229	230	male	John	Doe	123 Main Street	New York	NY	12345.0	United States	johndoe@email.com1234567890	1/1/1975	180.0	72	24.4
237	238	male	John	Doe	123 Main Street	New York	NY	12345.0	United States	johndoe@email.com1234567890	1/1/1975	180.0	72	24.4
244	245	male	John	Doe	123 Main Street	New York	NY	12345.0	United States	johndoe@email.com1234567890	1/1/1975	180.0	72	24.4
251	252	male	John	Doe	123 Main Street	New York	NY	12345.0	United States	johndoe@email.com1234567890	1/1/1975	180.0	72	24.4
277	278	male	John	Doe	123 Main Street	New York	NY	12345.0	United States	johndoe@email.com1234567890	1/1/1975	180.0	72	24.4

In [9]:
```
# patient_df [describe]
patients_df.describe()
```

Out[9]:

	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 [10]:
```
# patients_df table consists weight as min of 48 pounds.

patients_df[patients_df["weight"]==48.800000]
# weight is not correctly given, should be change as it affects the distribution of the data and mean.
```

Out[10]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdate	weight	height	bmi
210	211	female	Camilla	Zaitseva	4689 Briarhill Lane	Wooster	OH	44691.0	United States	330-202-2145CamillaZaitseva@superrito.com	11/26/1938	48.8	63	19.1

In [11]:
```
# patients_df table consists height as min of 27 inches
patients_df[patients_df["height"]==27.000000]

# height is 27 inches which is not possible as comparision to weight which is 192.3 pounds
```

Out[11]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdate	weight	height	bmi
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	36303.0	United States	334-515-7487TimNeudorf@cuvox.de	2/18/1928	192.3	27	26.1

treatments_df and treatments_cut_df

In [12]:
```
print(treatments_df.head())
print(treatments_cut_df.head())
```

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

   hba1c_change
0           NaN
1          0.97
2           NaN
3          0.35
4          0.32
  given_name    surname    auralin    novodra  hba1c_start  hba1c_end  \
0      jožka  resanovič  22u - 30u          -         7.56       7.22
1  inunnguaq   heilmann  57u - 67u          -         7.85       7.45
2      alwin   svensson  36u - 39u          -         7.78       7.34
3        thể      lương          -  61u - 64u         7.64       7.22
4     amanda    ribeiro  36u - 44u          -         7.85       7.47

   hba1c_change
0          0.34
1           NaN
2           NaN
3          0.92
4          0.38
```

In [13]:
```
print(treatments_df.tail())
print(treatments_cut_df.tail())
```

```
    given_name       surname    auralin    novodra  hba1c_start  hba1c_end  \
275     albina      zetticci  45u - 51u          -         7.93       7.73
276       john   teichelmann          -  49u - 49u         7.90       7.58
277     mathea       lillebø  23u - 36u          -         9.04       8.67
278     vallie        prince  31u - 38u          -         7.64       7.28
279     samúel  guðbrandsson  53u - 56u          -         8.00       7.64

     hba1c_change
275          0.20
276           NaN
277          0.37
278          0.36
279          0.36
   given_name     surname    auralin    novodra  hba1c_start  hba1c_end  \
65     rovzan     kishiev  32u - 37u          -         7.75       7.41
66      jakob    jakobsen          -  28u - 26u         7.96       7.51
67      bernd   schneider  48u - 56u          -         7.74       7.44
68      berta  napolitani          -  42u - 44u         7.68       7.21
69     armina       sauvé  36u - 46u          -         7.86       7.40

    hba1c_change
65          0.34
66          0.95
67          0.30
68           NaN
69           NaN
```

In [14]:
```
treatments_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 280 entries, 0 to 279
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   given_name    280 non-null    object
 1   surname       280 non-null    object
 2   auralin       280 non-null    object
 3   novodra       280 non-null    object
 4   hba1c_start   280 non-null    float64
 5   hba1c_end     280 non-null    float64
 6   hba1c_change  171 non-null    float64
dtypes: float64(3), object(4)
memory usage: 15.4+ KB
```

In [15]:
```
treatments_cut_df.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 [16]:
```
treatments_df[treatments_df.duplicated()]
# joseph day volunteer is repeated one time have to delete it.
```

Out[16]:

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

In [17]:
```
treatments_df[treatments_df.duplicated(subset=["given_name","surname"])]
```

Out[17]:

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

In [18]:
```
treatments_cut_df[treatments_cut_df.duplicated()]
```

Out[18]:

given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change

In [19]:
```
treatments_cut_df[treatments_cut_df.duplicated(subset=["given_name","surname"])]
```

Out[19]:

given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change

In [20]:
```
# treatment [describe()]
treatments_df.describe()
```

Out[20]:

	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 [21]:
```
treatments_df[treatments_df["hba1c_start"]==9.950000]
```

Out[21]:

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
166	annie	allen	36u - 42u	-	9.95	9.58	0.37

In [22]:
```
treatments_df.sort_values("hba1c_start")
# it seems normal as many volunteer is having hba1c_start in 9.something range.
```

Out[22]:

	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 [23]:
```
treatments_df.sort_values("hba1c_change",na_position="first")
# there is incorrect values of hba1c_change that is difference between hba1c_start and hba1c_end
```

Out[23]:

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
0	veronika	jindrová	41u - 48u	-	7.63	7.20	NaN
2	yukitaka	takenaka	-	39u - 36u	7.68	7.25	NaN
8	saber	ménard	-	54u - 54u	8.08	7.70	NaN
9	asia	woźniak	30u - 36u	-	7.76	7.37	NaN
10	joseph	day	29u - 36u	-	7.70	7.19	NaN
...	...	...	...	...	...	...	...
49	jackson	addison	-	42u - 42u	7.99	7.51	0.98
17	gina	cain	-	36u - 36u	7.88	7.40	0.98
32	laura	ehrlichmann	-	43u - 40u	7.95	7.46	0.99
245	wu	sung	-	47u - 48u	7.61	7.12	0.99
138	giovana	rocha	-	23u - 21u	7.87	7.38	0.99

280 rows × 7 columns

In [24]:
```
# treatments_cut_df
treatments_cut_df.describe()
# Same problem with treatments_cut_df
```

Out[24]:

	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

Data Cleaning

Patients table

In [25]:
```
patients_df[patients_df["address"].isna()]
## filling nan values in some columns not possible as, we do it by our own
## we consult the regarding some missing data of which we can't do anything to team mates.
```

Out[25]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdate	weight	height	bmi
209	210	female	Lalita	Eldarkhanov	NaN	NaN	NaN	NaN	NaN	NaN	8/14/1950	143.4	62	26.2
219	220	male	Mỹ	Quynh	NaN	NaN	NaN	NaN	NaN	NaN	4/9/1978	237.8	69	35.1
230	231	female	Elisabeth	Knudsen	NaN	NaN	NaN	NaN	NaN	NaN	9/23/1976	165.9	63	29.4
234	235	female	Martina	Tománková	NaN	NaN	NaN	NaN	NaN	NaN	4/7/1936	199.5	65	33.2
242	243	male	John	O'Brian	NaN	NaN	NaN	NaN	NaN	NaN	2/25/1957	205.3	74	26.4
249	250	male	Benjamin	Mehler	NaN	NaN	NaN	NaN	NaN	NaN	10/30/1951	146.5	69	21.6

257	258	male	Jin	Kung	NaN	NaN	NaN	NaN	NaN	NaN	5/17/1995	231.7	69	34.2
264	265	female	Wafiyyah	Asfour	NaN	NaN	NaN	NaN	NaN	NaN	11/3/1989	158.6	63	28.1
269	270	female	Flavia	Fiorentino	NaN	NaN	NaN	NaN	NaN	NaN	10/9/1937	175.2	61	33.1
278	279	female	Generosa	Cabán	NaN	NaN	NaN	NaN	NaN	NaN	12/16/1962	124.3	69	18.4
286	287	male	Lewis	Webb	NaN	NaN	NaN	NaN	NaN	NaN	4/1/1979	155.3	68	23.6
296	297	female	Chỉ	Lâm	NaN	NaN	NaN	NaN	NaN	NaN	5/14/1990	181.1	63	32.1

treatments_cut_df and treatments_df

hbalc change coloumn have some Missing value.[completeness].

In [26]:
```
# let's concate treatments_cut_df and treatment_df column

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

In [27]:
```
treatments_df["hba1c_change"]=treatments_df["hba1c_start"]-treatments_df["hba1c_end"]
```

In [28]:
```
treatments_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 350 entries, 0 to 69
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   given_name    350 non-null    object 
 1   surname       350 non-null    object 
 2   auralin       350 non-null    object 
 3   novodra       350 non-null    object 
 4   hba1c_start   350 non-null    float64
 5   hba1c_end     350 non-null    float64
 6   hba1c_change  350 non-null    float64
dtypes: float64(3), object(4)
memory usage: 21.9+ KB
```

In [29]:
```
treatments_df.sort_values("hba1c_change")
```

Out[29]:

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
275	albina	zetticci	45u - 51u	-	7.93	7.73	0.20
70	tosh	jensen	-	51u - 48u	7.93	7.69	0.24
237	manouck	wubbels	55u - 62u	-	7.66	7.40	0.26
144	nicoline	østergaard	24u - 32u	-	7.84	7.57	0.27
200	nicolas	ferreira	43u - 51u	-	7.99	7.72	0.27
...	...	...	...	...	...	...	...
273	kate	wilkinson	36u - 39u	-	7.72	7.20	0.52
187	león	reynoso	-	38u - 40u	7.59	7.06	0.53
51	kristoffer	martinsen	29u - 37u	-	9.18	8.64	0.54
178	kerman	dandonneau	41u - 50u	-	7.82	7.28	0.54
80	hideki	haraguchi	-	37u - 35u	7.59	7.05	0.54

350 rows × 7 columns

Patient_df

In [30]:
```
patients_df.head()
```

Out[30]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdate	weight	height	bmi
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	92390.0	United States	951-719-9170ZoeWellish@superrito.com	7/10/1976	121.7	66	19.6
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	61812.0	United States	PamelaSHill@cuvox.de+1 (217) 569-3204	4/3/1967	118.8	66	19.2
2	3	male	Jae	Debord	1493 Poling Farm Road	York	Nebraska	68467.0	United States	402-363-6804JaeMDebord@gustr.com	2/19/1980	177.8	71	24.8
3	4	male	Liêm	Phan	2335 Webster Street	Woodbridge	NJ	7095.0	United States	PhanBaLiem@jourrapide.com+1 (732) 636-8246	7/26/1951	220.9	70	31.7
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	36303.0	United States	334-515-7487TimNeudorf@cuvox.de	2/18/1928	192.3	27	26.1

In [31]:
```
# Extracting phone and email id from contact.
```

In [32]:
```
patients_df['Phone'] = patients_df['contact'].str.extract(r'([+]?[0-9]+[\s+]?[\(]?[\-]?[0-9]+[\)]?[\s+]?[0-9]+[\s+]?[\-]?[0-9]+)')
patients_df['E_mail'] = patients_df['contact'].str.extract(r'([a-zA-Z][a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]+)')
```

In [33]:
```
patients_df.drop(["contact"],axis=1,inplace=True)
```

Column auralin and novodra having starting and ending both dosage have to split into two different columns. auralin and Novodra both have to in one column.

In [34]:
```
treatments_df=treatments_df.melt(id_vars=["given_name","surname","hba1c_start","hba1c_end","hba1c_change"],var_name="type",value_name="dosage")
```

In [35]:
```
treatments_df=treatments_df[treatments_df["dosage"]!="-"]
```

In [36]:
```
treatments_df=treatments_df[treatments_df["dosage"]!="-"]
```

In [37]:
```
treatments_df.set_index(np.arange(1,351),inplace=True)
```

In [38]:
```
treatments_df["dosage_start"]=treatments_df["dosage"].str.split("-").str.get(0)
treatments_df["dosage_ends"]=treatments_df["dosage"].str.split("-").str.get(1)
```

In [39]:
```
treatments_df["dosage_start"]=treatments_df["dosage_start"].str.replace("u","")
treatments_df["dosage_ends"]=treatments_df["dosage_ends"].str.replace("u","")
```

In [40]:
```
treatments_df
```

Out[40]:

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	dosage	dosage_start	dosage_ends
1	veronika	jindrová	7.63	7.20	0.43	auralin	41u - 48u	41	48
2	skye	gormanston	7.97	7.62	0.35	auralin	33u - 36u	33	36
3	sophia	haugen	7.65	7.27	0.38	auralin	37u - 42u	37	42
4	eddie	archer	7.89	7.55	0.34	auralin	31u - 38u	31	38
5	asia	woźniak	7.76	7.37	0.39	auralin	30u - 36u	30	36
...	...	...	...	...	...	...	...	...	...
346	christopher	woodward	7.51	7.06	0.45	novodra	55u - 51u	55	51
347	maret	sultygov	7.67	7.30	0.37	novodra	26u - 23u	26	23

348	lixue	hsueh	9.21	8.80	0.41	novodra	22u - 23u	22	23
349	jakob	jakobsen	7.96	7.51	0.45	novodra	28u - 26u	28	26
350	berta	napolitani	7.68	7.21	0.47	novodra	42u - 44u	42	44

350 rows × 9 columns

In [41]:
```
treatments_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 350 entries, 1 to 350
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   dosage        350 non-null    object
 7   dosage_start  350 non-null    object
 8   dosage_ends   350 non-null    object
dtypes: float64(3), object(6)
memory usage: 27.3+ KB
```

In [42]:
```
treatments_df["dosage_start"]=treatments_df["dosage_start"].astype("int")
treatments_df["dosage_ends"]=treatments_df["dosage_ends"].astype("int")
```

In [43]:
```
treatments_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 350 entries, 1 to 350
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   dosage        350 non-null    object
 7   dosage_start  350 non-null    int32
 8   dosage_ends   350 non-null    int32
dtypes: float64(3), int32(2), object(4)
memory usage: 24.6+ KB
```

treatments_df table is cleaned properly.

Adverse_reaction

In [44]:
```
treatments_df=treatments_df.merge(adverse_reactions_df, how="left", on=["given_name","surname"])
```

In [45]:
```
treatments_df["adverse_reaction"].fillna("No_Reaction",inplace=True)
```

In [46]:
```
treatments_df.drop(["dosage"],axis=1,inplace=True)
```

In [47]:
```
patients_df["zip_code"].fillna("Not_available",inplace=True)
```

In [48]:
```
patients_df=patients_df[patients_df["zip_code"] !="Not_available"]
```

In [49]:
```
patients_df["zip_code"]=patients_df["zip_code"].astype("int")
```

In [50]:
```
patients_df["state"].replace({"NJ":"New Jersey"},inplace=True)
```

In [51]:
```
patients_df["state"].replace({"TX":"Texas",
                              "CA":"California",
                              "NY":"New York",
                              "MA":"Massachusetts",
                              "PA":"Pennsylvania",
                              "GA":"Georgia",
                              "OH":"Ohio",
                              "OK":"Oklahoma",
                              "MI":"Michigan",
                              "LA":"Louisiana",
                              "VA":"Virginia",
                              "MS":"Mississippi",
                              "WI":"Wisconsin",
                              "IL":"Illinois",
                              "IN":"Indiana",
                              "MN":"Minnesota",
                              "FL":"Florida",
                              "AL":" Alabama",
                              "TN":"Tennessee",
                              "WA":"Washington",
                              "NC":"North Carolina",
                              "KY":"     Kentucky",
                              "MO":"Missouri ",
                              "ID":"Idaho",
                              "NV":"Nevada",
                              "KS":"Kansas",
                              "SC":"South Carolina",
                              "IA":"Iowa",
                              "CT":"Connecticut",
                              "ME":"Maine",
                              "RI":"Rhode",
                              "ND":"North Dakota",
                              "CO":"Colorado",
                              "AZ":"Arizona",
                              "AR":"Arkansas",
                              "MD":"Maryland",
                              "DE":"Delaware",
                              "WV":"West Virginia",
                              "SD":"South Dakota",
                              "OR":"Oregon",
                              "NE":"Nebraska",
                              "MT":"Montana",
                              "VT":"Vermont",
                              "DC":"Washington DC",
                              "AK":"Alaska",
                              "WY":"Wyoming",
                              "NH":"New Hampshire",
                              "NM":"New Mexico"},inplace=True)
```

In [57]:
```
patients_df
```

Out[57]:

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	birthdate	weight	height	bmi	Phone	E_mail
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	92390	United States	7/10/1976	121.7	66	19.6	951-719-9170	ZoeWellish@superrito.com
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	61812	United States	4/3/1967	118.8	66	19.2	+1 (217) 569-3204	PamelaSHill@cuvox.de
2	3	male	Jae	Debord	1493 Poling Farm Road	York	Nebraska	68467	United States	2/19/1980	177.8	71	24.8	402-363-6804	JaeMDebord@gustr.com
3	4	male	Liêm	Phan	2335 Webster Street	Woodbridge	New Jersey	7095	United States	7/26/1951	220.9	70	31.7	+1 (732) 636-8246	PhanBaLiem@jourrapide.com
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	Alabama	36303	United States	2/18/1928	192.3	27	26.1	334-515-7487	TimNeudorf@cuvox.de
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
498	499	male	Mustafa	Lindström	2530 Victoria Court	Milton Mills	Maine	3852	United States	4/10/1959	181.1	72	24.6	207-477-0579	MustafaLindstrom@jourrapide.com
499	500	male	Ruman	Bisliev	494 Clarksburg Park Road	Sedona	Arizona	86341	United States	3/26/1948	239.6	70	34.4	928-284-4492	RumanBisliev@gustr.com
500	501	female	Jinke	de Keizer	649 Nutter Street	Overland Park	Missouri	64110	United States	1/13/1971	171.2	67	26.8	816-223-6007	JinkedeKeizer@teleworm.us
501	502	female	Chidalu	Onyekaozulu	3652 Boone Crockett Lane	Seattle	Washington	98109	United States	2/13/1952	176.9	67	27.7	1 360 443 2060	ChidaluOnyekaozulu@jourrapide.com
502	503	male	Pat	Gersten	2778 North Avenue	Burr	Nebraska	68324	United States	5/3/1954	138.2	71	19.3	402-848-4923	PatrickGersten@rhyta.com

491 rows × 15 columns

In [58]:
```
treatments_df
```

Out[58]:

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	type	dosage_start	dosage_ends	adverse_reaction
0	veronika	jindrová	7.63	7.20	0.43	auralin	41	48	No_Reaction
1	skye	gormanston	7.97	7.62	0.35	auralin	33	36	No_Reaction
2	sophia	haugen	7.65	7.27	0.38	auralin	37	42	No_Reaction
3	eddie	archer	7.89	7.55	0.34	auralin	31	38	No_Reaction
4	asia	woźniak	7.76	7.37	0.39	auralin	30	36	No_Reaction
...	...	...	...	...	...	...	...	...	...
345	christopher	woodward	7.51	7.06	0.45	novodra	55	51	nausea
346	maret	sultygov	7.67	7.30	0.37	novodra	26	23	No_Reaction
347	lixue	hsueh	9.21	8.80	0.41	novodra	22	23	injection site discomfort
348	jakob	jakobsen	7.96	7.51	0.45	novodra	28	26	hypoglycemia
349	berta	napolitani	7.68	7.21	0.47	novodra	42	44	injection site discomfort

350 rows × 9 columns

In [59]:
```
# Export data for manual assessment
with pd.ExcelWriter("clinicals_trails.xlsx") as writer:
    patients_df.to_excel(writer, sheet_name="patients_df")
    treatments_df.to_excel(writer,sheet_name="treatments_df")
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: