

▼ Data Cleaning Steps

Step 1: Address the Completeness Issue

- **Concatenate** `df_treatments.csv` and `df_treatments_cut.csv` to ensure all treatment data is included.

Step 2: Address the Structural Issues

1. Fix Email and Phone Number Formatting (`patients.csv`)

- Use **regular expressions** to separate the combined email and phone number into two distinct columns.

2. Reformat Treatment Data (`treatments.csv`)

- **Split "Novodra" and "Auralin"** into two separate columns: `treatment` and `oral_dose`.
- Use the `melt()` function to restructure the data properly.
- **Extract Dose Values:**
 - Split dose data into **low dose** and **high dose** values.
 - Remove the letter 'u' from the values.
 - Create three new columns: `dose_low`, `dose_high`, and `treatment`.

3. Integrate Adverse Reactions Data (`adverse_reactions.csv` & `treatments.csv`)

- Add a new column `df_adverse_condition` to `df_treatments.csv`.
- **Merge** both datasets based on common patient names using the `merge()` function.

Step 3: Address Quality Issues

- **Handle Improper Data**
- **Remove Duplicate Records**
- **Handle Missing Values**
- **Detect and Fix Corrupt Data**
- **Fix Incorrect Data**

#Colab link:- https://colab.research.google.com/drive/1aoM0eimDwgHJUPxT_aGnLDKbhJW8ePLj?usp=sharing

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df_treatments = pd.read_csv('/content/treatments.csv')
df_treatments_cut = pd.read_csv('/content/treatments_cut.csv')
df_patients = pd.read_csv('/content/patients.csv')
df_adverse_conditions = pd.read_csv('/content/adverse_reactions.csv')
```

```
df_treatments.head()
```

	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

Next steps: [Generate code with df_treatments](#) [View recommended plots](#) [New interactive sheet](#)

```
df_treatments.shape
```

```
(280, 7)
```

```
df_adverse_conditions.head()
```

	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	

Next steps:

[Generate code with df_adverse_conditions](#)[View recommended plots](#)[New interactive sheet](#)

df_patients.head()

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact	birthdat
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	92390.0	United States	9170ZoeWellish@superrito.com	7/10/197
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	61812.0	United States	PamelaSHill@cuvovox.de+1 (217) 569-3204	4/3/196

Next steps:

[Generate code with df_patients](#)[View recommended plots](#)[New interactive sheet](#)

df_treatments.head()

	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	

Next steps:

[Generate code with df_treatments](#)[View recommended plots](#)[New interactive sheet](#)

df_treatments_cut.head()

	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	luong	- 61u - 64u		7.64	7.22	0.92	
4	amanda	ribeiro	36u - 44u	-	7.85	7.47	0.38	

Next steps:


[Generate code with df_treatments_cut](#)[View recommended plots](#)[New interactive sheet](#)

df_patients.isna().sum()



	0
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


```
df_patients[df_patients.isna().any(axis=1)]
```




	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	Chĩ	Lâm	NaN	NaN	NaN	NaN	NaN	NaN	5/14/1990	181.1	63	32.1

```
df_patients= df_patients.dropna()
```

```
df_patients
```



	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	contact
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	92390.0	United States	951-79170ZoeWellish@superrito.c
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	61812.0	United States	PamelaSHill@cuvovx.de+1 (217) 533
2	3	male	Jae	Debord	1493 Poling Farm Road	York	Nebraska	68467.0	United States	402-363-6804JaeMDebord@gustr.c
3	4	male	Liêm	Phan	2335 Webster Street	Woodbridge	NJ	7095.0	United States	PhanBaLiem@jourrapide.com+1 (7636-8
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	36303.0	United States	334-515-7487TimNeudorf@cuvovx.de
...
498	499	male	Mustafa	Lindström	2530 Victoria Court	Milton Mills	ME	3852.0	United States	207-40579MustafaLindstrom@jourrapide.c
499	500	male	Ruman	Bisliev	494 Clarksburg Park Road	Sedona	AZ	86341.0	United States	928-284-4492RumanBisliev@gustr.c
500	501	female	Jinke	de Keizer	649 Nutter Street	Overland Park	MO	64110.0	United States	816-26007JinkedeKeizer@teleworm



Next steps: [Generate code with df_patients](#) [View recommended plots](#) [New interactive sheet](#)

```
df_patients.isna().sum()
```



	0
patient_id	0
assigned_sex	0
given_name	0
surname	0
address	0
city	0
state	0
zip_code	0
country	0
contact	0
birthdate	0
weight	0
height	0
bmi	0



```
df_adverse_conditions.isna().sum()
```



	0
given_name	0
surname	0
adverse_reaction	0



```
df_treatments.isna().sum()
```



	0
given_name	0
surname	0
auralin	0
novodra	0
hba1c_start	0
hba1c_end	0
hba1c_change	109


```
df_treatments_cut.isna().sum()
```



	0
given_name	0
surname	0
auralin	0
novodra	0
hba1c_start	0
hba1c_end	0
hba1c_change	28

```
df_treatments = pd.concat([df_treatments, df_treatments_cut], ignore_index=True)
```

```
df_treatments
```



	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change
0	veronika	jindrová	41u - 48u	-	7.63	7.20	NaN
1	elliot	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
...
345	rovzan	kishiev	32u - 37u	-	7.75	7.41	0.34
346	jakob	jakobsen	- 28u - 26u		7.96	7.51	0.95
347	bernd	schneider	48u - 56u	-	7.74	7.44	0.30
348	berta	napolitani	- 42u - 44u		7.68	7.21	NaN
349	armina	sauvé	36u - 46u	-	7.86	7.40	NaN

350 rows x 7 columns

Next steps:

Generate code with df_treatments

View recommended plots

New interactive sheet

```
df_treatments.isna().sum()
```

```

0
given_name 0
surname 0
auralin 0
novodra 0
hba1c_start 0
hba1c_end 0
hba1c_change 137

```

```
df_treatments['hba1c_change'] = df_treatments['hba1c_start'] - df_treatments['hba1c_end']
```

```
df_treatments.head()
```

```

given_name  surname  auralin  novodra  hba1c_start  hba1c_end  hba1c_change
0    veronika    jindrová  41u - 48u      -         7.63      7.20         0.43
1      elliot  richardson      -  40u - 45u         7.56      7.09         0.47
2    yukitaka   takenaka      -  39u - 36u         7.68      7.25         0.43
3        skye  gormanston  33u - 36u      -         7.97      7.62         0.35
4      alissa    montez      -  33u - 29u         7.78      7.46         0.32

```

Next steps: [Generate code with df_treatments](#) [View recommended plots](#) [New interactive sheet](#)

```
df_treatments.isna().sum()
```

```

0
given_name 0
surname 0
auralin 0
novodra 0
hba1c_start 0
hba1c_end 0
hba1c_change 0

```

```
df_patients.head()
```

```

patient_id  assigned_sex  given_name  surname  address  city  state  zip_code  country  contact  birthdat
0          1        female        Zoe   Wellish  576    Rancho  California  92390.0  United  9170ZoeWellish@superrito.com  7/10/197
              Brown  Bear  Drive
1          2        female      Pamela    Hill  2370    University  Armstrong  Illinois  61812.0  United  PamelaSHill@cuvovx.de+1 (217)  4/3/19€
              Hill  Road

```

Next steps: [Generate code with df_patients](#) [View recommended plots](#) [New interactive sheet](#)

```

df_patients['phone_number'] = df_patients['contact'].str.extract(r'(\+?[0-9 \(\)-]+)')
df_patients['email'] = df_patients['contact'].str.extract(r'([a-zA-Z._%+-]+@[a-zA-Z.-]+\.[a-zA-Z]{2,})')

```

```
df_patients = df_patients.drop(columns=['contact'])
```

df_patients

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	birthdate	weight	height	bm:
0	1	female	Zoe	Wellish	576 Brown Bear Drive	Rancho California	California	92390.0	United States	7/10/1976	121.7	66	19.6
1	2	female	Pamela	Hill	2370 University Hill Road	Armstrong	Illinois	61812.0	United States	4/3/1967	118.8	66	19.4
2	3	male	Jae	Debord	1493 Poling Farm Road	York	Nebraska	68467.0	United States	2/19/1980	177.8	71	24.8
3	4	male	Liêm	Phan	2335 Webster Street	Woodbridge	NJ	7095.0	United States	7/26/1951	220.9	70	31.1
4	5	male	Tim	Neudorf	1428 Turkey Pen Lane	Dothan	AL	36303.0	United States	2/18/1928	192.3	27	26.7
...
498	499	male	Mustafa	Lindström	2530 Victoria Court	Milton Mills	ME	3852.0	United States	4/10/1959	181.1	72	24.6
499	500	male	Ruman	Bisliev	494 Clarksburg Park Road	Sedona	AZ	86341.0	United States	3/26/1948	239.6	70	34.4
500	501	female	Jinke	de Keizer	649 Nutter Street	Overland Park	MO	64110.0	United States	1/13/1971	171.2	67	26.8

Next steps: [Generate code with df_patients](#) [View recommended plots](#) [New interactive sheet](#)

df_treatments

	given_name	surname	auralin	novodra	hba1c_start	hba1c_end	hba1c_change	
0	veronika	jindrová	41u - 48u	-	7.63	7.20	0.43	
1	elliot	richardson	- 40u - 45u		7.56	7.09	0.47	
2	yukitaka	takenaka	- 39u - 36u		7.68	7.25	0.43	
3	skye	gormanston	33u - 36u	-	7.97	7.62	0.35	
4	alissa	montez	- 33u - 29u		7.78	7.46	0.32	
...	
345	rovzan	kishiev	32u - 37u	-	7.75	7.41	0.34	
346	jakob	jakobsen	- 28u - 26u		7.96	7.51	0.45	
347	bernd	schneider	48u - 56u	-	7.74	7.44	0.30	
348	berta	napolitani	- 42u - 44u		7.68	7.21	0.47	
349	armina	sauvé	36u - 46u	-	7.86	7.40	0.46	

350 rows x 7 columns

Next steps: [Generate code with df_treatments](#) [View recommended plots](#) [New interactive sheet](#)

Double-click (or enter) to edit

Melted dataframe

```
# Melt the DataFrame to create a 'treatment' and 'dose' column
df_treatments = df_treatments.melt(id_vars=['given_name', 'surname', 'hba1c_start', 'hba1c_end', 'hba1c_change'],
                                   value_vars=['auralin', 'novodra'],
                                   var_name='treatment_type',
                                   value_name='dose')

# Filter out rows where dose is missing
```

```
df_treatments = df_treatments[df_treatments['dose'] != '-']
```

```
# Display the resulting DataFrame
print(df_treatments)
```

```
↩
```

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	\
0	veronika	jindrová	7.63	7.20	0.43	
3	skye	gormanston	7.97	7.62	0.35	
6	sophia	haugen	7.65	7.27	0.38	
7	eddie	archer	7.89	7.55	0.34	
9	asia	woźniak	7.76	7.37	0.39	
...	
688	christopher	woodward	7.51	7.06	0.45	
690	maret	sulygov	7.67	7.30	0.37	
694	lixue	hsueh	9.21	8.80	0.41	
696	jakob	jakobsen	7.96	7.51	0.45	
698	berta	napolitani	7.68	7.21	0.47	

```

treatment_type    dose
0      auralin    41u - 48u
3      auralin    33u - 36u
6      auralin    37u - 42u
7      auralin    31u - 38u
9      auralin    30u - 36u
...      ...      ...
688    novodra    55u - 51u
690    novodra    26u - 23u
694    novodra    22u - 23u
696    novodra    28u - 26u
698    novodra    42u - 44u

```

[350 rows x 7 columns]

Double-click (or enter) to edit

```
df_treatments.head()
```

```
↩
```

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	treatment_type	dose
0	veronika	jindrová	7.63	7.20	0.43	auralin	41u - 48u
3	skye	gormanston	7.97	7.62	0.35	auralin	33u - 36u
6	sophia	haugen	7.65	7.27	0.38	auralin	37u - 42u
7	eddie	archer	7.89	7.55	0.34	auralin	31u - 38u
9	asia	woźniak	7.76	7.37	0.39	auralin	30u - 36u

Next steps:

[Generate code with df_treatments](#)
[View recommended plots](#)
[New interactive sheet](#)

```
df_patients.tail()
```

```
↩
```

	patient_id	assigned_sex	given_name	surname	address	city	state	zip_code	country	birthdate	weight	height	bmi
498	499	male	Mustafa	Lindström	2530 Victoria Court	Milton Mills	ME	3852.0	United States	4/10/1959	181.1	72	24.6
499	500	male	Ruman	Bisliev	494 Clarksburg Park Road	Sedona	AZ	86341.0	United States	3/26/1948	239.6	70	34.4
500	501	female	Jinke	de Keizer	649 Nutter Street	Overland Park	MO	64110.0	United States	1/13/1971	171.2	67	26.8
501	502	female	Chidalu	Onyekaozulu	3652 Boone Crockett	Seattle	WA	98109.0	United States	2/13/1952	176.9	67	27.7

```
# Split the 'dose' column into low and high dose values
```

```
df_treatments[['low_dose', 'high_dose']] = df_treatments['dose'].str.split(' - ', expand=True)
```

```
# Remove the 'u' character and convert to integers
```

```
df_treatments['low_dose'] = df_treatments['low_dose'].str.replace('u', '').astype(int)
```

```
df_treatments['high_dose'] = df_treatments['high_dose'].str.replace('u', '').astype(int)
```

```
# Keep only relevant columns
```

```
final_df = df_treatments[['given_name', 'surname', 'treatment_type', 'low_dose', 'high_dose']]
```



```
# Display the resulting DataFrame
print(final_df)
```

```
↵
   given_name  surname treatment_type  low_dose  high_dose
0    veronika  jindrová      auralin        41         48
3      skye  gormanston      auralin        33         36
6    sophia   haugen      auralin        37         42
7     eddie   archer      auralin        31         38
9      asia   woźniak      auralin        30         36
..      ...      ...      ...      ...      ...
688 christopher  woodward      novodra        55         51
690    maret    sulygov      novodra        26         23
694    lixue    hsueh      novodra        22         23
696    jakob   jakobsen      novodra        28         26
698    berta  napolitani      novodra        42         44
```

[350 rows x 5 columns]

```
df_treatments.head()
```

```
↵
   given_name  surname  hba1c_start  hba1c_end  hba1c_change  treatment_type  dose  low_dose  high_dose
0    veronika  jindrová         7.63         7.20          0.43      auralin  41u - 48u        41         48
3      skye  gormanston         7.97         7.62          0.35      auralin  33u - 36u        33         36
6    sophia   haugen         7.65         7.27          0.38      auralin  37u - 42u        37         42
7     eddie   archer         7.89         7.55          0.34      auralin  31u - 38u        31         38
9      asia   woźniak         7.76         7.37          0.39      auralin  30u - 36u        30         36
```

Next steps:

[Generate code with df_treatments](#)
[View recommended plots](#)
[New interactive sheet](#)

```
df_treatments.tail()
```

```
↵
   given_name  surname  hba1c_start  hba1c_end  hba1c_change  treatment_type  dose  low_dose  high_dose
688 christopher  woodward         7.51         7.06          0.45      novodra  55u - 51u        55         51
690    maret    sulygov         7.67         7.30          0.37      novodra  26u - 23u        26         23
694    lixue    hsueh         9.21         8.80          0.41      novodra  22u - 23u        22         23
696    jakob   jakobsen         7.96         7.51          0.45      novodra  28u - 26u        28         26
698    berta  napolitani         7.68         7.21          0.47      novodra  42u - 44u        42         44
```

```
df_treatments.isnull().sum()
```

```
↵
      0
given_name  0
surname     0
hba1c_start 0
hba1c_end   0
hba1c_change 0
treatment_type 0
dose         0
low_dose     0
high_dose    0
```

```
df_treatments
```

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	treatment_type	dose	low_dose	high_dose
0	veronika	jindrová	7.63	7.20	0.43	auralin	41u - 48u	41	48
3	skye	gormanston	7.97	7.62	0.35	auralin	33u - 36u	33	36
6	sophia	haugen	7.65	7.27	0.38	auralin	37u - 42u	37	42
7	eddie	archer	7.89	7.55	0.34	auralin	31u - 38u	31	38
9	asia	woźniak	7.76	7.37	0.39	auralin	30u - 36u	30	36
...
688	christopher	woodward	7.51	7.06	0.45	novodra	55u - 51u	55	51
690	maret	sulygov	7.67	7.30	0.37	novodra	26u - 23u	26	23
694	lixue	hsueh	9.21	8.80	0.41	novodra	22u - 23u	22	23
696	jakob	jakobsen	7.96	7.51	0.45	novodra	28u - 26u	28	26
698	berta	napolitani	7.68	7.21	0.47	novodra	42u - 44u	42	44

350 rows × 9 columns

Next steps:

[Generate code with df_treatments](#)[View recommended plots](#)[New interactive sheet](#)

df_treatments.duplicated().sum()

↔ 1

df_treatments = df_treatments.applymap(lambda x: x.lower() if isinstance(x, str) else x)

↔ <ipython-input-38-810883b0f417>:1: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
df_treatments = df_treatments.applymap(lambda x: x.lower() if isinstance(x, str) else x)

df_treatments.duplicated().sum()

↔ 1

df_treatments

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	treatment_type	dose	low_dose	high_dose
0	veronika	jindrová	7.63	7.20	0.43	auralin	41u - 48u	41	48
3	skye	gormanston	7.97	7.62	0.35	auralin	33u - 36u	33	36
6	sophia	haugen	7.65	7.27	0.38	auralin	37u - 42u	37	42
7	eddie	archer	7.89	7.55	0.34	auralin	31u - 38u	31	38
9	asia	woźniak	7.76	7.37	0.39	auralin	30u - 36u	30	36
...
688	christopher	woodward	7.51	7.06	0.45	novodra	55u - 51u	55	51
690	maret	sulygov	7.67	7.30	0.37	novodra	26u - 23u	26	23
694	lixue	hsueh	9.21	8.80	0.41	novodra	22u - 23u	22	23
696	jakob	jakobsen	7.96	7.51	0.45	novodra	28u - 26u	28	26
698	berta	napolitani	7.68	7.21	0.47	novodra	42u - 44u	42	44

350 rows × 9 columns

Next steps:

[Generate code with df_treatments](#)[View recommended plots](#)[New interactive sheet](#)

df_treatments = df_treatments.reset_index(drop=True)

df_treatments = df_treatments.drop(columns=['dose'])

df_treatments

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

350 rows × 8 columns

Next steps:

[Generate code with df_treatments](#)[View recommended plots](#)[New interactive sheet](#)

```

# Remove the unnecessary column if present
if "Unnamed: 0" in final_df.columns:
    final_df = final_df.drop(columns=["Unnamed: 0"])

# Merge datasets on given_name and surname
df_treatments = df_treatments.merge(df_adverse_conditions, on=["given_name", "surname"], how="left")

# Fill missing values in adverse_reaction column with "No Reaction"
df_treatments["adverse_reaction"] = df_treatments["adverse_reaction"].fillna("No Reaction")

# Save the updated dataframe
df_treatments.to_csv("final_dataframe_with_reactions.csv", index=False)

# Display the first few rows
print(df_treatments.head())

```

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	treatment_type	\
0	veronika	jindrová	7.63	7.20	0.43	auralin	
1	skye	gormanston	7.97	7.62	0.35	auralin	
2	sophia	haugen	7.65	7.27	0.38	auralin	
3	eddie	archer	7.89	7.55	0.34	auralin	
4	asia	woźniak	7.76	7.37	0.39	auralin	

	low_dose	high_dose	adverse_reaction
0	41	48	No Reaction
1	33	36	No Reaction
2	37	42	No Reaction
3	31	38	No Reaction
4	30	36	No Reaction

df_treatments

	given_name	surname	hba1c_start	hba1c_end	hba1c_change	treatment_type	low_dose	high_dose	adverse_reaction	
Next	0	0	0	0	0	0	0	0	0	
	1	skye	qormanston	7.97	7.62	0.35	auralin	33	36	No Reaction

```
# Check for outliers using IQR method for numerical columns
Q1 = df_treatments[["low_dose", "high_dose"]].quantile(0.25)
Q3 = df_treatments[["low_dose", "high_dose"]].quantile(0.75)
IQR = Q3 - Q1
```

```
# Define outlier boundaries
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

```
# Filter out rows with outliers
df_treatments = df_treatments[
    ~((df_treatments["low_dose"] < lower_bound["low_dose"]) | (df_treatments["low_dose"] > upper_bound["low_dose"])) &
    ~((df_treatments["high_dose"] < lower_bound["high_dose"]) | (df_treatments["high_dose"] > upper_bound["high_dose"]))
]
```

df_treatments.info()

```
<class 'pandas.core.frame.DataFrame'>
Index: 348 entries, 0 to 349
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   given_name            348 non-null   object
1   surname               348 non-null   object
2   hba1c_start           348 non-null   float64
3   hba1c_end             348 non-null   float64
4   hba1c_change          348 non-null   float64
5   treatment_type        348 non-null   object
6   low_dose              348 non-null   int64
7   high_dose             348 non-null   int64
8   adverse_reaction      348 non-null   object
dtypes: float64(3), int64(2), object(4)
memory usage: 27.7+ KB
```