```
import pandas as pd

df = pd.read_csv('https://raw.githubusercontent.com/mGalarnyk/datasciencecc

df.head()
```

	Provider Number	Hospital Name	Address 1	Address 2	Address 3	City	State	ZIP Code
0	010001	SOUTHEAST ALABAMA MEDICAL CENTER	1108 ROSS CLARK CIRCLE	NaN	NaN	DOTHAN	AL	36301
1	010005	MARSHALL MEDICAL CENTER SOUTH	2505 U S HIGHWAY 431 NORTH	NaN	NaN	BOAZ	AL	35957
2	010006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	NaN	NaN	FLORENCE	AL	35631
3	010007	MIZELL MEMORIAL HOSPITAL	702 N MAIN ST	NaN	NaN	OPP	AL	36467
4	010008	CRENSHAW COMMUNITY HOSPITAL	101 HOSPITAL CIRCLE	NaN	NaN	LUVERNE	AL	36049

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.describe()
```

	Address 2	Address 3	ZIP Code	Phone Number
count	0.0	0.0	4826.000000	4.826000e+03
mean	NaN	NaN	53079.289888	5.829414e+09
std	NaN	NaN	27002.162551	2.333778e+09
min	NaN	NaN	603.000000	9.369338e+08
25%	NaN	NaN	32207.500000	3.616118e+09
50%	NaN	NaN	54153.500000	6.033435e+09
75%	NaN	NaN	75225.500000	7.877054e+09
max	NaN	NaN	99929.000000	9.898943e+09

```
df.columns
```

```
Index(['Provider Number', 'Hospital Name', 'Address 1', 'Address 2',
       'Address 3', 'City', 'State', 'ZIP Code', 'County', 'Phone Number',
       'Hospital Type', 'Hospital Ownership', 'Emergency Services'],
      dtype='object')
```

```
df.isnull().sum()
```

	0
Provider Number	0
Hospital Name	0
Address 1	0
Address 2	4826
Address 3	4826
City	0
State	0
ZIP Code	0
County	21
Phone Number	0
Hospital Type	0
Hospital Ownership	0
Emergency Services	0

dtype: int64

```
df.duplicated().sum()
```

```
np.int64(0)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4826 entries, 0 to 4825
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Provider Number     4826 non-null   object 
 1   Hospital Name       4826 non-null   object 
 2   Address 1           4826 non-null   object 
 3   Address 2           0 non-null      float64
 4   Address 3           0 non-null      float64
 5   City                4826 non-null   object 
 6   State               4826 non-null   object 
 7   ZIP Code            4826 non-null   int64  
 8   County              4805 non-null   object 
 9   Phone Number        4826 non-null   int64  
 10  Hospital Type       4826 non-null   object 
 11  Hospital Ownership  4826 non-null   object 
 12  Emergency Services  4826 non-null   object 
dtypes: float64(2), int64(2), object(9)
memory usage: 490.3+ KB
```

```
df.drop(labels= ['Address 2', 'Address 3'], axis=1, inplace= True)
```

```
df.head()
```

	Provider Number	Hospital Name	Address 1	City	State	ZIP Code	County	
0	010001	SOUTHEAST ALABAMA MEDICAL CENTER	1108 ROSS CLARK CIRCLE	DOTHAN	AL	36301	HOUSTON	3347
1	010005	MARSHALL MEDICAL CENTER SOUTH	2505 U S HIGHWAY 431 NORTH	BOAZ	AL	35957	MARSHALL	2565
2	010006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	FLORENCE	AL	35631	LAUDERDALE	2567
3	010007	MIZELL MEMORIAL HOSPITAL	702 N MAIN ST	OPP	AL	36467	COVINGTON	3344
4	010008	CRENSHAW COMMUNITY HOSPITAL	101 HOSPITAL CIRCLE	LUVERNE	AL	36049	CRENSHAW	3343

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df['County'].fillna(df['County'].mode()[0], inplace=True)
```

```
/tmp/ipython-input-732875338.py:1: FutureWarning: A value is trying to be s
The behavior will change in pandas 3.0. This inplace method will never work

For example, when doing 'df[col].method(value, inplace=True)', try using 'd

  df['County'].fillna(df['County'].mode()[0], inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4826 entries, 0 to 4825
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Provider Number     4826 non-null   object
 1   Hospital Name       4826 non-null   object
 2   Address 1           4826 non-null   object
 3   City                4826 non-null   object
 4   State               4826 non-null   object
 5   ZIP Code            4826 non-null   int64 
 6   County              4826 non-null   object
 7   Phone Number        4826 non-null   int64 
 8   Hospital Type       4826 non-null   object
 9   Hospital Ownership  4826 non-null   object
 10  Emergency Services  4826 non-null   object
dtypes: int64(2), object(9)
memory usage: 414.9+ KB
```

```
dff = pd.read_csv('dirty_reviews.csv')
```

```
dff.head()
```

	user_id	product_id	rating	review_text	review_date
0	1	101	5.0	Great product! Works perfectly.	1/15/2025
1	2	102	NaN	NaN	1/16/2025
2	3	101	4.0	good value for money	1/17/2025
3	4	103	2.0	Not as described	1/18/2025
4	5	104	5.0	EXCELLENT!!!	1/19/2025

Next steps: Generate code with dff | View recommended plots | New interactive sheet

```
dff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15 entries, 0 to 14
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   user_id      15 non-null     int64  
 1   product_id   15 non-null     int64  
 2   rating       12 non-null     float64
 3   review_text  13 non-null     object 
 4   review_date  15 non-null     object 
dtypes: float64(1), int64(2), object(2)
memory usage: 732.0+ bytes
```

```
dff.describe()
```

	user_id	product_id	rating
count	15.000000	15.000000	12.000000
mean	8.000000	105.400000	3.333333
std	4.472136	3.924283	1.497473
min	1.000000	101.000000	1.000000
25%	4.500000	101.500000	2.000000
50%	8.000000	105.000000	4.000000
75%	11.500000	108.500000	4.250000
max	15.000000	112.000000	5.000000

```
(dff.isnull().sum()/len(dff))*100
```

	0
user_id	0.000000
product_id	0.000000
rating	20.000000
review_text	13.333333
review_date	0.000000

dtype: float64

```
dff.duplicated().sum()
```

```
np.int64(0)
```

```
dff.shape
```

```
(15, 5)
```

```
dff['rating'].fillna(dff['rating'].mean(), inplace=True)
```

```
/tmp/ipython-input-2413861733.py:1: FutureWarning: A value is trying to be
The behavior will change in pandas 3.0. This inplace method will never work

For example, when doing 'df[col].method(value, inplace=True)', try using 'd

  dff['rating'].fillna(dff['rating'].mean(), inplace=True)
```

```
dff.isnull().sum()
```

	0
user_id	0
product_id	0
rating	0
review_text	2
review_date	0

dtype: int64

```
dff = dff[~dff['review_text'].isnull()]
```

```
1 dff.isnull().sum()
```

	0
user_id	0
product_id	0
rating	0
review_text	0
review_date	0

dtype: int64

Start coding or generate with AI.