Untitled

July 28, 2025

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

[2]:
```
data = pd.read_csv('data/trip.csv')
```

[3]:
```
data.head()
```

[3]:
```
     passenger_name    tpep_pickup_datetime   tpep_dropoff_datetime  \
0     Pamela Duffy   03/25/2017 8:55:43 AM   03/25/2017 9:09:47 AM
1  Michelle Foster   04/11/2017 2:53:28 PM   04/11/2017 3:19:58 PM
2       Tina Combs   12/15/2017 7:26:56 AM   12/15/2017 7:34:08 AM
3      Anthony Ray   05/07/2017 1:17:59 PM   05/07/2017 1:48:14 PM
4  Brianna Johnson  04/15/2017 11:32:20 PM  04/15/2017 11:49:03 PM

  payment_method  passenger_count  trip_distance  fare_amount  tip_amount  \
0     Debit Card                6           3.34         13.0        2.76
1     Debit Card                1           1.80         16.0        4.00
2     Debit Card                1           1.00          6.5        1.45
3           Cash                1           3.70         20.5        6.39
4     Debit Card                1           4.37         16.5        0.00

   tolls_amount
0           0.0
1           0.0
2           0.0
3           0.0
4           0.0
```

[4]:
```
# Q. info()                    .

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22701 entries, 0 to 22700
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype
```

```
---  ------                  --------------  -----
 0   passenger_name          22701 non-null  object
 1   tpep_pickup_datetime    22701 non-null  object
 2   tpep_dropoff_datetime   22701 non-null  object
 3   payment_method          22701 non-null  object
 4   passenger_count         22701 non-null  int64
 5   trip_distance           22701 non-null  float64
 6   fare_amount             22698 non-null  float64
 7   tip_amount              22701 non-null  float64
 8   tolls_amount            22701 non-null  float64
dtypes: float64(4), int64(1), object(4)
memory usage: 1.6+ MB
```

[5]:
```
# Q. describe()                    .

data.describe()
```

[5]:
```
       passenger_count  trip_distance   fare_amount    tip_amount  \
count     22701.000000   22701.000000  22698.000000  22701.000000
mean          1.643584       2.913400     13.024009      1.835745
std           1.304942       3.653023     13.240074      2.800537
min           0.000000       0.000000   -120.000000      0.000000
25%           1.000000       0.990000      6.500000      0.000000
50%           1.000000       1.610000      9.500000      1.350000
75%           2.000000       3.060000     14.500000      2.450000
max          36.000000      33.960000    999.990000    200.000000

       tolls_amount
count  22701.000000
mean       0.312514
std        1.399153
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max       19.100000
```

[6]:
```
# Q.          .

data.duplicated()

dp_name  = data.duplicated( keep=False)
print(data[dp_name])
```

```
    passenger_name   tpep_pickup_datetime  tpep_dropoff_datetime  \
16     Sarah Gross  08/15/2017 7:48:08 PM  08/15/2017 8:00:37 PM
17     Sarah Gross  08/15/2017 7:48:08 PM  08/15/2017 8:00:37 PM
203   Lisa Bullock  02/13/2017 4:25:41 PM  02/13/2017 4:55:35 PM
```

```
204   Lisa Bullock  02/13/2017 4:25:41 PM  02/13/2017 4:55:35 PM

    payment_method  passenger_count  trip_distance  fare_amount  tip_amount  \
16            Cash                1            3.6         12.5        2.85
17            Cash                1            3.6         12.5        2.85
203           Cash                1            4.2         21.0        0.00
204           Cash                1            4.2         21.0        0.00

     tolls_amount
16            0.0
17            0.0
203           0.0
204           0.0
```

[7]:
```
print(data[data['passenger_name'] == 'Sarah Gross'])
print(data[data['passenger_name'] == 'Lisa Bullock'])
```

```
   passenger_name   tpep_pickup_datetime  tpep_dropoff_datetime  \
16    Sarah Gross  08/15/2017 7:48:08 PM  08/15/2017 8:00:37 PM
17    Sarah Gross  08/15/2017 7:48:08 PM  08/15/2017 8:00:37 PM

   payment_method  passenger_count  trip_distance  fare_amount  tip_amount  \
16            Cash                1            3.6         12.5        2.85
17            Cash                1            3.6         12.5        2.85

    tolls_amount
16            0.0
17            0.0
    passenger_name   tpep_pickup_datetime  tpep_dropoff_datetime  \
203   Lisa Bullock  02/13/2017 4:25:41 PM  02/13/2017 4:55:35 PM
204   Lisa Bullock  02/13/2017 4:25:41 PM  02/13/2017 4:55:35 PM

    payment_method  passenger_count  trip_distance  fare_amount  tip_amount  \
203           Cash                1            4.2         21.0         0.0
204           Cash                1            4.2         21.0         0.0

     tolls_amount
203           0.0
204           0.0
```

[8]:
```
# Q.          .

data = data.drop_duplicates( )
```

[9]:
```
data
```

[9]:
```
          passenger_name    tpep_pickup_datetime   tpep_dropoff_datetime  \
0           Pamela Duffy   03/25/2017 8:55:43 AM   03/25/2017 9:09:47 AM
1         Michelle Foster   04/11/2017 2:53:28 PM   04/11/2017 3:19:58 PM
2              Tina Combs   12/15/2017 7:26:56 AM   12/15/2017 7:34:08 AM
3             Anthony Ray   05/07/2017 1:17:59 PM   05/07/2017 1:48:14 PM
4         Brianna Johnson  04/15/2017 11:32:20 PM  04/15/2017 11:49:03 PM
…                    …                   …                   …
22696      Austin Johnson   02/24/2017 5:37:23 PM   02/24/2017 5:40:39 PM
22697    Monique Williams   08/06/2017 4:43:59 PM   08/06/2017 5:24:47 PM
22698         Drew Graves   09/04/2017 2:54:14 PM   09/04/2017 2:58:22 PM
22699  Jonathan Copeland  07/15/2017 12:56:30 PM   07/15/2017 1:08:26 PM
22700     Benjamin Miller   03/02/2017 1:02:49 PM   03/02/2017 1:16:09 PM

      payment_method  passenger_count  trip_distance  fare_amount  tip_amount  \
0        Debit Card                6           3.34         13.0        2.76
1        Debit Card                1           1.80         16.0        4.00
2        Debit Card                1           1.00          6.5        1.45
3              Cash                1           3.70         20.5        6.39
4        Debit Card                1           4.37         16.5        0.00
…              …                 …              …            …           …
22696          Cash                3           0.61          4.0        0.00
22697          Cash                1          16.71         52.0       14.64
22698    Debit Card                1           0.42          4.5        0.00
22699    Debit Card                1           2.36         10.5        1.70
22700          Cash                1           2.10         11.0        2.35

       tolls_amount
0              0.00
1              0.00
2              0.00
3              0.00
4              0.00
…              …
22696          0.00
22697          5.76
22698          0.00
22699          0.00
22700          0.00

[22699 rows x 9 columns]
```

[10]:
```
data.isna().sum()
```

[10]:
```
passenger_name           0
tpep_pickup_datetime     0
tpep_dropoff_datetime    0
payment_method           0
```

```
passenger_count          0
trip_distance            0
fare_amount              3
tip_amount               0
tolls_amount             0
dtype: int64
```

[11]:
```
# Q.                .

data.isna( ).mean( )
```

[11]:
```
passenger_name           0.000000
tpep_pickup_datetime     0.000000
tpep_dropoff_datetime    0.000000
payment_method           0.000000
passenger_count          0.000000
trip_distance            0.000000
fare_amount              0.000132
tip_amount               0.000000
tolls_amount             0.000000
dtype: float64
```

[12]:
```
# Q.        .

data = data.dropna()
```

[13]:
```
data.isna().mean()
```

[13]:
```
passenger_name           0.0
tpep_pickup_datetime     0.0
tpep_dropoff_datetime    0.0
payment_method           0.0
passenger_count          0.0
trip_distance            0.0
fare_amount              0.0
tip_amount               0.0
tolls_amount             0.0
dtype: float64
```

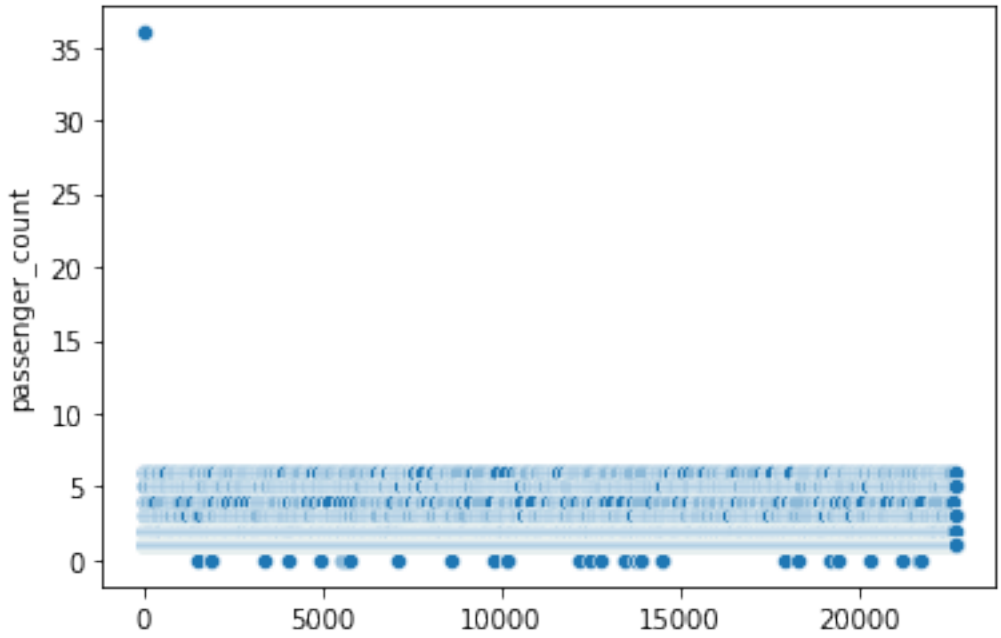
[14]:
```
# passenger_count            .

data['passenger_count'].sort_values()
```

[14]:
```
12804     0
19458     0
5565      0
5670      0
```

```
13718     0
          ..
416       6
4322      6
14500     6
0         6
64       36
Name: passenger_count, Length: 22696, dtype: int64
```

[15]:
```
# passenger_count   scatter plot    .

sns.scatterplot(x = data.index, y = data['passenger_count'])
```

[15]:
```
<AxesSubplot:ylabel='passenger_count'>
```



[16]:
```
# passenger_count          .
# (passenger_count  6      )

data = data[data['passenger_count'] <= 6]
```

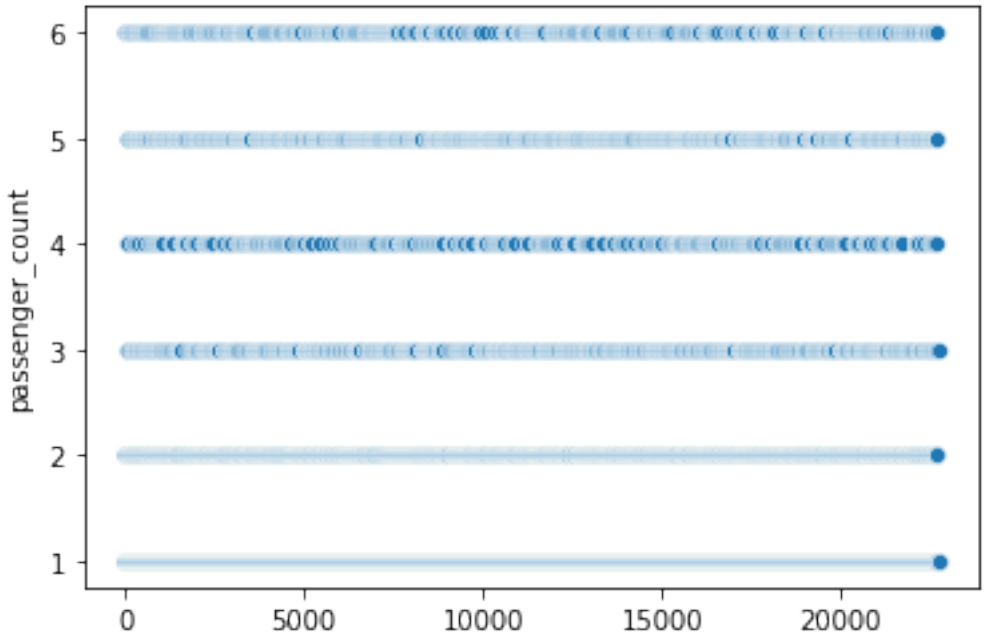
[17]:
```
# passenger_count          .
# (passenger_count  0   )

len(data[data['passenger_count'] == 0])
```

[17]:
```
33
```

[18]:
```
# passenger_count          .

data = data[data['passenger_count'] != 0]
```

[19]:
```
# passenger_count  scatter plot       .

sns.scatterplot(x = data.index, y = data['passenger_count'])
```

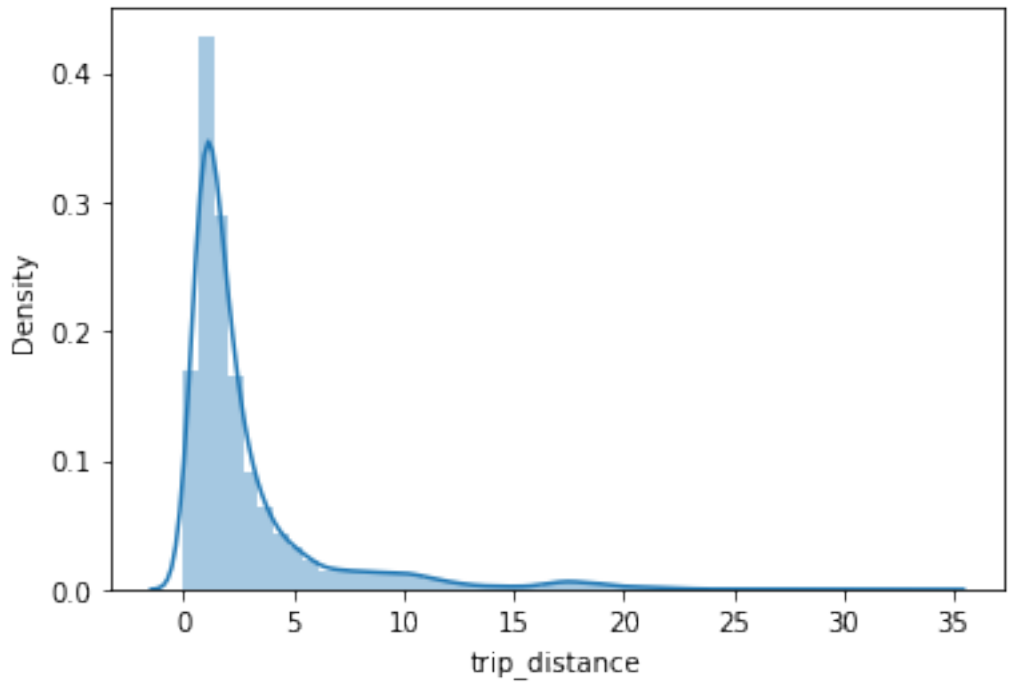
[19]: <AxesSubplot:ylabel='passenger_count'>



[20]:
```
# Q. trip_distance          .

sns.distplot(data['trip_distance'])
```

```
/opt/conda/lib/python3.9/site-packages/seaborn/distributions.py:2619:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```

[20]: <AxesSubplot:xlabel='trip_distance', ylabel='Density'>

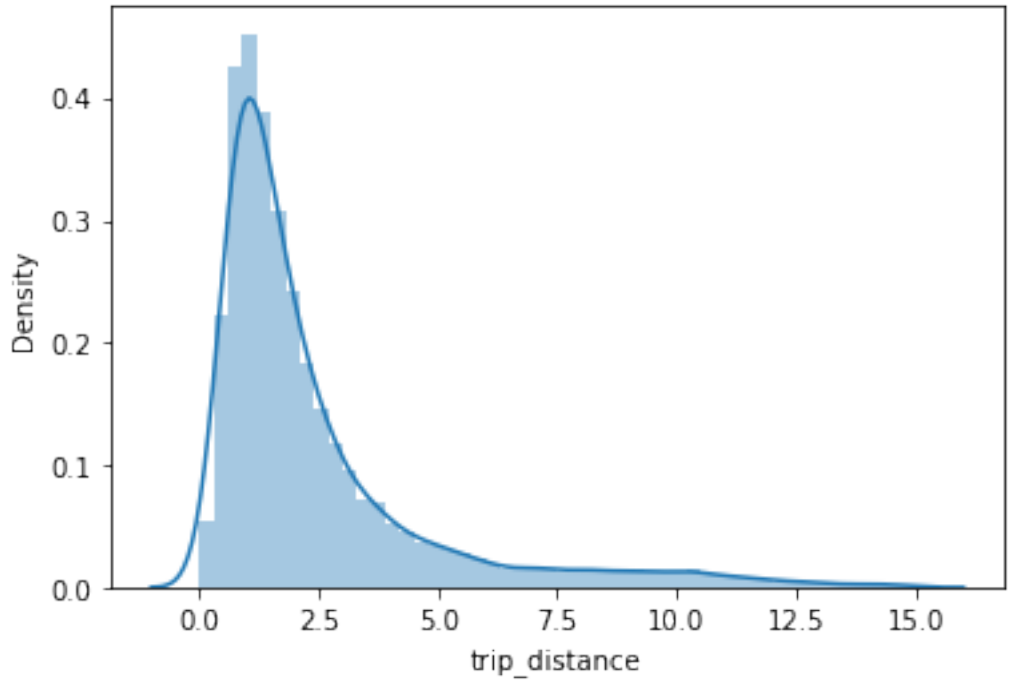


[21]:
```
data = data[data['trip_distance'] <= 15]
```

[22]:
```
# Q. trip_distance        .

sns.distplot(data['trip_distance'])
```

```
/opt/conda/lib/python3.9/site-packages/seaborn/distributions.py:2619:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```

[22]: <AxesSubplot:xlabel='trip_distance', ylabel='Density'>



[27]:
```
data.describe()
```

[27]:

	passenger_count	trip_distance	tip_amount	tolls_amount
count	22041.000000	22041.000000	22041.000000	22041.000000
mean	1.642212	2.465928	1.674179	0.196810
std	1.283578	2.495176	2.426917	1.075136
min	1.000000	0.000000	0.000000	0.000000
25%	1.000000	0.970000	0.000000	0.000000
50%	1.000000	1.600000	1.350000	0.000000
75%	2.000000	2.830000	2.350000	0.000000
max	6.000000	15.000000	200.000000	18.000000

[23]:
```
# Q. fare_amount             .
# (fare_amount  0     )
len(data[data['fare_amount'] < 0])
```

[23]: 14

[24]:
```
# Q. fare_amount        .

data = data[data['fare_amount'] > 0]
```

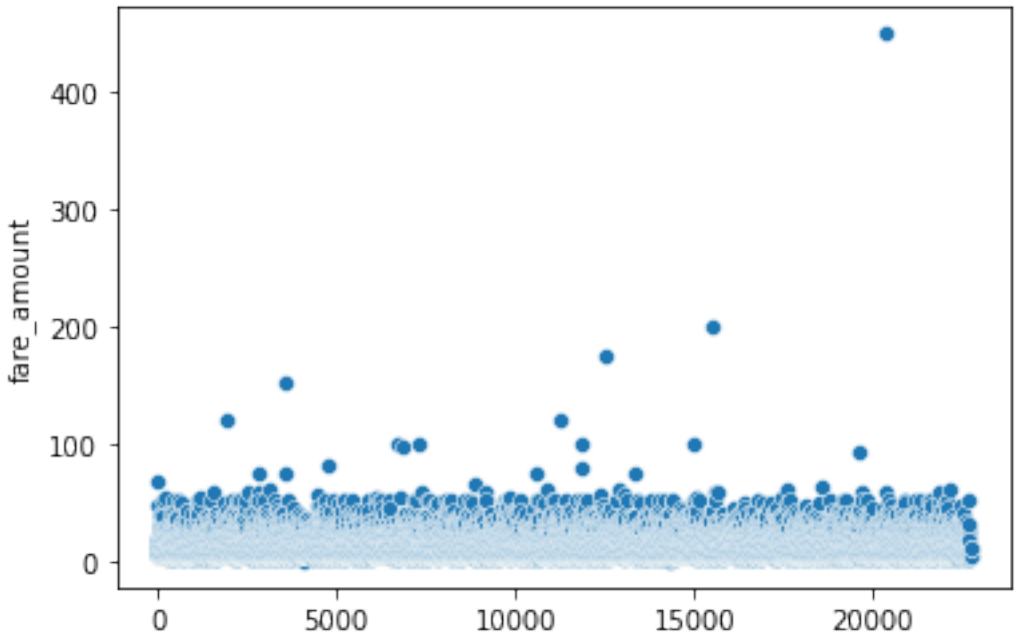
[25]:
```
data.sort_values('fare_amount')
```

[25]:
```
          passenger_name    tpep_pickup_datetime   tpep_dropoff_datetime  \
14285         Mark Reed   05/03/2017 7:44:28 PM   05/03/2017 7:44:38 PM
4063   Phillip Gonzalez   08/12/2017 8:49:29 PM   08/12/2017 9:18:50 PM
13972     Matthew Blake   02/23/2017 9:21:25 AM   02/23/2017 9:21:57 AM
9190    Valerie Vasquez   03/31/2017 5:29:19 AM   03/31/2017 5:29:32 AM
21595   Timothy Ramirez   04/14/2017 9:18:30 PM   04/14/2017 9:18:32 PM
…                  …                   …                   …
3584     Matthew Chavez  01/01/2017 11:53:01 PM  01/01/2017 11:53:42 PM
12513  Mr. Wesley Reyes   12/17/2017 6:24:24 PM   12/17/2017 6:24:42 PM
15476     James Dyer MD   06/06/2017 8:55:01 PM   06/06/2017 8:55:06 PM
20314   Nicholas Thomas   12/19/2017 9:40:46 AM   12/19/2017 9:40:55 AM
8478      Alexis Hanson   02/06/2017 5:50:10 AM   02/06/2017 5:51:08 AM

      payment_method  passenger_count  trip_distance  fare_amount  tip_amount  \
14285     Debit Card                1           0.00         0.01        0.00
4063            Cash                4           4.50         0.01        0.00
13972           Cash                1           0.00         1.00        0.00
9190            Cash                1           0.01         2.50        0.00
21595    Credit Card                1           1.20         2.50        0.00
…                …                …              …            …           …
3584     Credit Card                1           7.30       152.00        0.00
12513           Cash                1           0.00       175.00       46.69
15476     Debit Card                1           0.00       200.00       11.00
20314           Cash                2           0.00       450.00        0.00
8478     Credit Card                1           2.60       999.99      200.00

       tolls_amount
14285          0.00
4063          10.50
13972          0.00
9190           0.00
21595          0.00
…              …
3584           0.00
12513         11.75
15476          0.00
20314          0.00
8478           0.00

[22021 rows x 9 columns]
```

[26]:
```
# Q. fare_amount        .
data = data[data['fare_amount'] < 900]
```

[27]:
```
# Q. fare_amount  scatter plot    .

sns.scatterplot(x = data.index, y = data['fare_amount'])
```

[27]: <AxesSubplot:ylabel='fare_amount'>



[28]:
```
# fare_amount  150     150     .

def fare_func(x):
    if x > 150:
        return 150
    else:
        return x
```

[29]:
```
data['fare_amount'].apply(fare_func)
```

[29]:
```
0        13.0
1        16.0
2         6.5
3        20.5
4        16.5
         …
22695     7.5
22696     4.0
22698     4.5
22699    10.5
22700    11.0
Name: fare_amount, Length: 22020, dtype: float64
```

[30]:
```
data['fare_amount'] = data['fare_amount'].apply(lambda x: 150 if x > 150 else x)
```

[31]:
```
data.sort_values('fare_amount')
```

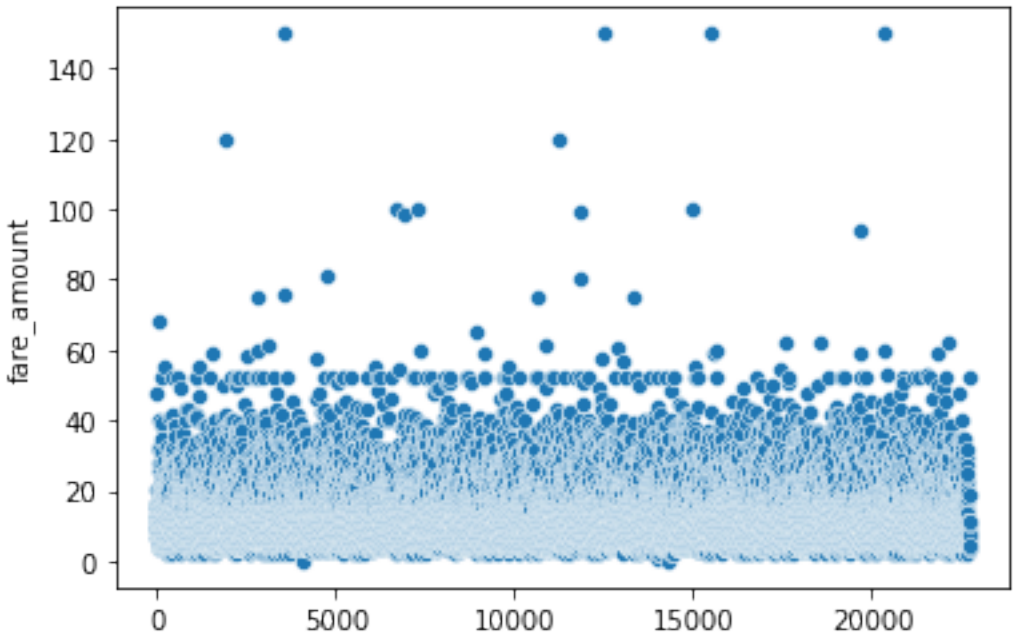
[31]:

	passenger_name	tpep_pickup_datetime	tpep_dropoff_datetime	payment_method	passenger_count	trip_distance	fare_amount	tip_amount	tolls_amount
14285	Mark Reed	05/03/2017 7:44:28 PM	05/03/2017 7:44:38 PM	Debit Card	1	0.0	0.01	0.00	0.00
4063	Phillip Gonzalez	08/12/2017 8:49:29 PM	08/12/2017 9:18:50 PM	Cash	4	4.5	0.01	0.00	10.50
13972	Matthew Blake	02/23/2017 9:21:25 AM	02/23/2017 9:21:57 AM	Cash	1	0.0	1.00	0.00	0.00
248	Erik Perez	09/18/2017 8:50:53 PM	09/18/2017 8:51:03 PM	Cash	1	0.0	2.50	0.00	0.00
18699	Donna Silva	12/01/2017 12:03:08 PM	12/01/2017 12:03:13 PM	Cash	1	0.0	2.50	0.00	0.00
…	…	…	…	…	…	…	…	…	…
1930	Cameron Long	06/16/2017 6:30:08 PM	06/16/2017 7:18:50 PM	Debit Card	2	12.5	120.00	5.00	12.50
20314	Nicholas Thomas	12/19/2017 9:40:46 AM	12/19/2017 9:40:55 AM	Cash	2	0.0	150.00	0.00	0.00
15476	James Dyer MD	06/06/2017 8:55:01 PM	06/06/2017 8:55:06 PM	Debit Card	1	0.0	150.00	11.00	0.00
3584	Matthew Chavez	01/01/2017 11:53:01 PM	01/01/2017 11:53:42 PM	Credit Card	1	7.3	150.00	0.00	0.00
12513	Mr. Wesley Reyes	12/17/2017 6:24:24 PM	12/17/2017 6:24:42 PM	Cash	1	0.0	150.00	46.69	11.75

[22020 rows x 9 columns]

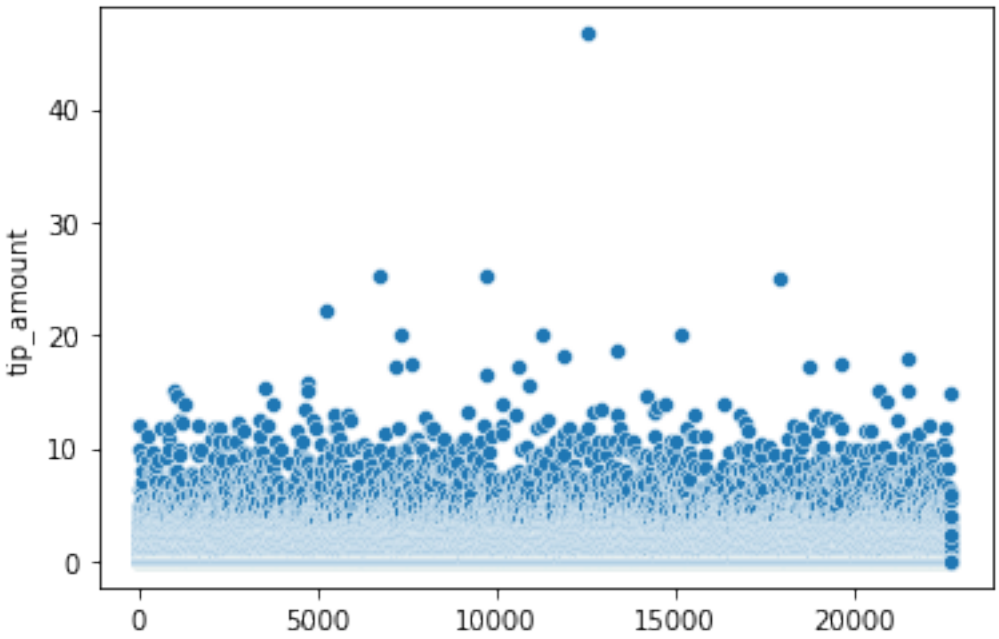
[32]:
```
# Q. tip_amount  scatter plot    .

sns.scatterplot(x = data.index, y = data['fare_amount'])
```

[32]: <AxesSubplot:ylabel='fare_amount'>



[33]:
```
# Q. tip_amount        .

sns.scatterplot(x = data.index, y = data['tip_amount'])
```

[33]: <AxesSubplot:ylabel='tip_amount'>



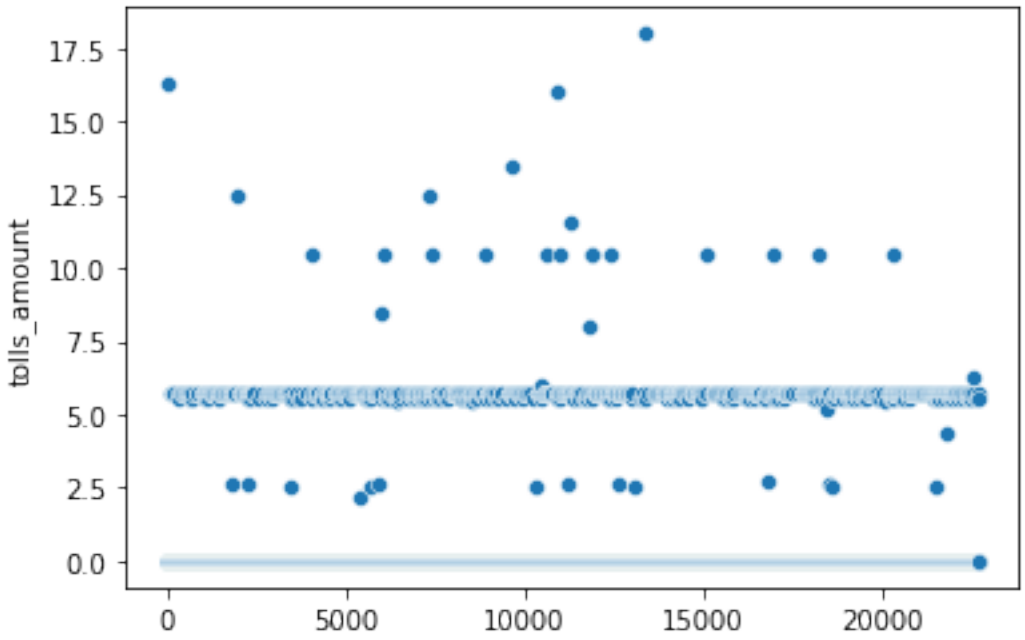
[34]:
```
# Q. tip_amount        .
data = data[data['tip_amount'] < 40]
```

[35]:
```
len(data)
```

[35]:
```
22019
```

[36]:
```
# Q. tolls_amount  scatter plot    .

sns.scatterplot(x = data.index, y = data['tolls_amount'])
```

[36]:
```
<AxesSubplot:ylabel='tolls_amount'>
```



[37]:
```
data.head(30)
```

[37]:

	passenger_name	tpep_pickup_datetime	tpep_dropoff_datetime	\
0	Pamela Duffy	03/25/2017 8:55:43 AM	03/25/2017 9:09:47 AM	
1	Michelle Foster	04/11/2017 2:53:28 PM	04/11/2017 3:19:58 PM	
2	Tina Combs	12/15/2017 7:26:56 AM	12/15/2017 7:34:08 AM	
3	Anthony Ray	05/07/2017 1:17:59 PM	05/07/2017 1:48:14 PM	
4	Brianna Johnson	04/15/2017 11:32:20 PM	04/15/2017 11:49:03 PM	
5	Justin Smith	03/25/2017 8:34:11 PM	03/25/2017 8:42:11 PM	
6	Tonya Moreno	05/03/2017 7:04:09 PM	05/03/2017 8:03:47 PM	

7	Hannah Foley	08/15/2017 5:41:06 PM	08/15/2017 6:03:05 PM
8	Katie Whitney	02/04/2017 4:17:07 PM	02/04/2017 4:29:14 PM
9	Amanda Jones	11/10/2017 3:20:29 PM	11/10/2017 3:40:55 PM
10	Cory Jensen	03/04/2017 11:58:00 AM	03/04/2017 12:13:12 PM
12	Ryan Reyes	06/09/2017 7:00:26 PM	06/09/2017 7:20:11 PM
13	Jessica Mooney	11/06/2017 11:35:05 PM	11/06/2017 11:42:57 PM
14	Heidi May	02/22/2017 3:18:31 PM	02/22/2017 3:42:50 PM
15	Anthony Richard	06/02/2017 6:41:39 AM	06/02/2017 6:57:47 AM
16	Sarah Gross	08/15/2017 7:48:08 PM	08/15/2017 8:00:37 PM
18	Susan Robinson	07/10/2017 1:36:31 PM	07/10/2017 1:48:43 PM
19	Cynthia Mendoza	04/10/2017 6:12:58 PM	04/10/2017 6:17:39 PM
20	Zachary James	03/05/2017 4:01:07 AM	03/05/2017 4:14:11 AM
21	Marissa Scott	12/30/2017 11:52:44 PM	12/30/2017 11:58:57 PM
22	Jacqueline Mclean DVM	10/11/2017 12:34:49 PM	10/11/2017 1:22:38 PM
23	Krista Stewart	01/06/2017 8:12:07 PM	01/06/2017 8:18:37 PM
24	Mike Taylor	06/27/2017 12:08:22 AM	06/27/2017 12:13:45 AM
25	Heather Johnson	02/13/2017 10:29:33 AM	02/13/2017 10:34:11 AM
26	Tiffany Ramirez	01/14/2017 7:58:42 PM	01/14/2017 8:05:59 PM
27	James Taylor	11/04/2017 1:27:59 AM	11/04/2017 1:44:05 AM
28	Gabriela Bryan	11/24/2017 10:48:13 AM	11/24/2017 10:52:57 AM
29	Janet Hogan MD	11/22/2017 10:24:17 AM	11/22/2017 10:38:52 AM
31	Jessica Cohen	08/09/2017 9:01:50 PM	08/09/2017 9:14:28 PM
32	Katherine Martin	04/12/2017 11:07:56 AM	04/12/2017 11:19:29 AM

	payment_method	passenger_count	trip_distance	fare_amount	tip_amount	\
0	Debit Card	6	3.34	13.00	2.76	
1	Debit Card	1	1.80	16.00	4.00	
2	Debit Card	1	1.00	6.50	1.45	
3	Cash	1	3.70	20.50	6.39	
4	Debit Card	1	4.37	16.50	0.00	
5	Debit Card	6	2.30	9.00	2.06	
6	Cash	1	12.83	47.50	9.86	
7	Debit Card	1	2.98	16.00	1.78	
8	Cash	1	1.20	9.00	0.00	
9	Cash	1	1.60	13.00	2.75	
10	Cash	1	1.77	11.50	2.46	
12	Debit Card	1	3.00	15.00	3.35	
13	Credit Card	1	2.39	9.50	2.16	
14	Cash	1	3.30	17.50	4.55	
15	Credit Card	1	5.93	19.00	3.00	
16	Cash	1	3.60	12.50	2.85	
18	Cash	2	1.71	9.50	0.00	
19	Cash	2	0.63	5.00	0.00	
20	Credit Card	2	2.77	11.50	3.20	
21	Debit Card	1	1.10	6.50	0.00	
22	Debit Card	1	12.30	68.25	12.00	
23	Cash	1	0.52	5.50	1.00	

```
24    Credit Card                1          1.70          7.00         2.05
25           Cash                1          0.90          5.50         1.25
26     Debit Card                1          1.72          8.00         2.79
27    Credit Card                1          2.70         13.00         2.85
28           Cash                1          0.85          5.50         0.00
29           Cash                1          2.30         11.00         2.35
31           Cash                1          2.30         10.50         1.77
32           Cash                4          1.50          9.00         0.00

    tolls_amount
0           0.00
1           0.00
2           0.00
3           0.00
4           0.00
5           0.00
6           0.00
7           0.00
8           0.00
9           0.00
10          0.00
12          0.00
13          0.00
14          0.00
15          0.00
16          0.00
18          0.00
19          0.00
20          0.00
21          0.00
22         16.26
23          0.00
24          0.00
25          0.00
26          0.00
27          0.00
28          0.00
29          0.00
31          0.00
32          0.00
```

[38]:
```
# payment_method                .

data['payment_method'].unique()
```

[38]:
```
array(['Debit Card', 'Cash', 'Credit Card'], dtype=object)
```

[39]:
```
data['payment_method'].nunique()
```

[39]:
```
3
```

[40]:
```
data['payment_method'].value_counts()
```

[40]:
```
Cash           10862
Debit Card      5601
Credit Card     5556
Name: payment_method, dtype: int64
```

[41]:
```
# Q. 'Debit Card'  'Credit Card'    'Card'    .
# ( : replace()       .)

data = data.replace({'Debit Card': 'Card', 'Credit Card': 'Card'})
```

[42]:
```
data['payment_method'].value_counts()
```

[42]:
```
Card    11157
Cash    10862
Name: payment_method, dtype: int64
```

[43]:
```
example = 'Susan Robinson'
```

[44]:
```
example.split()
```

[44]:
```
['Susan', 'Robinson']
```

[45]:
```
# Q. passenger_name              passenger_first_name       .

data['passenger_first_name'] = data['passenger_name'].str.split(' ').str[-1]
```

[46]:
```
data.head()
```

[46]:

	passenger_name	tpep_pickup_datetime	tpep_dropoff_datetime	payment_method	passenger_count	trip_distance	fare_amount	tip_amount
0	Pamela Duffy	03/25/2017 8:55:43 AM	03/25/2017 9:09:47 AM	Card	6	3.34	13.0	2.76
1	Michelle Foster	04/11/2017 2:53:28 PM	04/11/2017 3:19:58 PM	Card	1	1.80	16.0	4.00
2	Tina Combs	12/15/2017 7:26:56 AM	12/15/2017 7:34:08 AM	Card	1	1.00	6.5	1.45
3	Anthony Ray	05/07/2017 1:17:59 PM	05/07/2017 1:48:14 PM	Cash	1	3.70	20.5	6.39
4	Brianna Johnson	04/15/2017 11:32:20 PM	04/15/2017 11:49:03 PM	Card	1	4.37	16.5	0.00

```
   tolls_amount passenger_first_name
0           0.0                Duffy
1           0.0               Foster
2           0.0                Combs
3           0.0                  Ray
4           0.0              Johnson
```

[47]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22019 entries, 0 to 22700
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   passenger_name        22019 non-null  object
 1   tpep_pickup_datetime  22019 non-null  object
 2   tpep_dropoff_datetime 22019 non-null  object
 3   payment_method        22019 non-null  object
 4   passenger_count       22019 non-null  int64
 5   trip_distance         22019 non-null  float64
 6   fare_amount           22019 non-null  float64
 7   tip_amount            22019 non-null  float64
 8   tolls_amount          22019 non-null  float64
 9   passenger_first_name  22019 non-null  object
dtypes: float64(4), int64(1), object(5)
memory usage: 1.8+ MB
```

[48]:
```
# Q. tpep_pickup_datetime    object    datetime     .

data['tpep_pickup_datetime'] = pd.to_datetime(data['tpep_pickup_datetime'])
```

[49]:
```
# Q. tpep_dropoff_datetime    object    datetime     .

data['tpep_dropoff_datetime'] = pd.to_datetime(data['tpep_dropoff_datetime'])
```

[50]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22019 entries, 0 to 22700
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   passenger_name        22019 non-null  object
 1   tpep_pickup_datetime  22019 non-null  datetime64[ns]
 2   tpep_dropoff_datetime 22019 non-null  datetime64[ns]
 3   payment_method        22019 non-null  object
```

```
 4   passenger_count        22019 non-null  int64
 5   trip_distance          22019 non-null  float64
 6   fare_amount            22019 non-null  float64
 7   tip_amount             22019 non-null  float64
 8   tolls_amount           22019 non-null  float64
 9   passenger_first_name   22019 non-null  object
dtypes: datetime64[ns](2), float64(4), int64(1), object(3)
memory usage: 1.8+ MB
```

[51]:
```
# Q.              travel_time       .

data['travel_time'] = data['tpep_dropoff_datetime'] -␣
 ↪data['tpep_pickup_datetime']
```

[52]:
```
data.head()
```

[52]:
```
    passenger_name tpep_pickup_datetime tpep_dropoff_datetime payment_method  \
0     Pamela Duffy  2017-03-25 08:55:43   2017-03-25 09:09:47           Card
1  Michelle Foster  2017-04-11 14:53:28   2017-04-11 15:19:58           Card
2       Tina Combs  2017-12-15 07:26:56   2017-12-15 07:34:08           Card
3      Anthony Ray  2017-05-07 13:17:59   2017-05-07 13:48:14           Cash
4  Brianna Johnson  2017-04-15 23:32:20   2017-04-15 23:49:03           Card

   passenger_count  trip_distance  fare_amount  tip_amount  tolls_amount  \
0                6           3.34         13.0        2.76           0.0
1                1           1.80         16.0        4.00           0.0
2                1           1.00          6.5        1.45           0.0
3                1           3.70         20.5        6.39           0.0
4                1           4.37         16.5        0.00           0.0

  passenger_first_name     travel_time
0                Duffy 0 days 00:14:04
1               Foster 0 days 00:26:30
2                Combs 0 days 00:07:12
3                  Ray 0 days 00:30:15
4              Johnson 0 days 00:16:43
```

[53]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22019 entries, 0 to 22700
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   passenger_name         22019 non-null  object
 1   tpep_pickup_datetime   22019 non-null  datetime64[ns]
 2   tpep_dropoff_datetime  22019 non-null  datetime64[ns]
 3   payment_method         22019 non-null  object
```

```
 4   passenger_count        22019 non-null  int64
 5   trip_distance          22019 non-null  float64
 6   fare_amount            22019 non-null  float64
 7   tip_amount             22019 non-null  float64
 8   tolls_amount           22019 non-null  float64
 9   passenger_first_name   22019 non-null  object
 10  travel_time            22019 non-null  timedelta64[ns]
dtypes: datetime64[ns](2), float64(4), int64(1), object(3), timedelta64[ns](1)
memory usage: 2.0+ MB
```

[54]:
```
# Q. travel_time              .
data['travel_time'] = pd.to_timedelta(data['travel_time'], unit='second')

data['travel_time'] = data['travel_time'].dt.total_seconds()
```

[55]:
```
data.head()
```

[55]:

	passenger_name	tpep_pickup_datetime	tpep_dropoff_datetime	payment_method	\
0	Pamela Duffy	2017-03-25 08:55:43	2017-03-25 09:09:47	Card	
1	Michelle Foster	2017-04-11 14:53:28	2017-04-11 15:19:58	Card	
2	Tina Combs	2017-12-15 07:26:56	2017-12-15 07:34:08	Card	
3	Anthony Ray	2017-05-07 13:17:59	2017-05-07 13:48:14	Cash	
4	Brianna Johnson	2017-04-15 23:32:20	2017-04-15 23:49:03	Card	

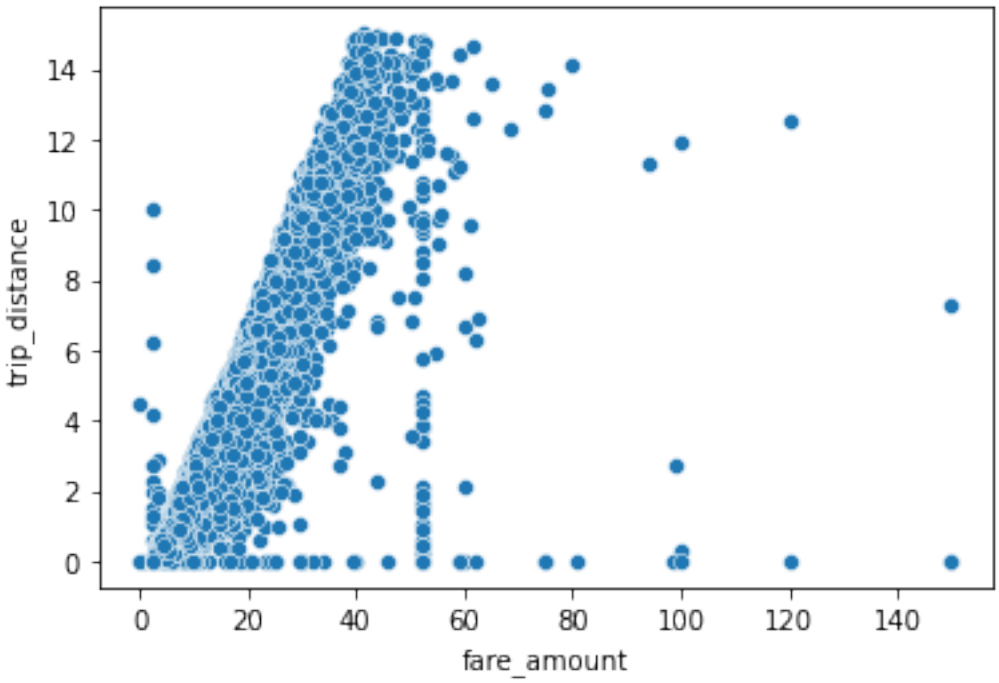
	passenger_count	trip_distance	fare_amount	tip_amount	tolls_amount	\
0	6	3.34	13.0	2.76	0.0	
1	1	1.80	16.0	4.00	0.0	
2	1	1.00	6.5	1.45	0.0	
3	1	3.70	20.5	6.39	0.0	
4	1	4.37	16.5	0.00	0.0	

	passenger_first_name	travel_time
0	Duffy	844.0
1	Foster	1590.0
2	Combs	432.0
3	Ray	1815.0
4	Johnson	1003.0

[56]:
```
# Q.          total_amount       .

data['total_amount'] = data['fare_amount'] +␣
 ↪data['tip_amount']+data['tolls_amount']
```

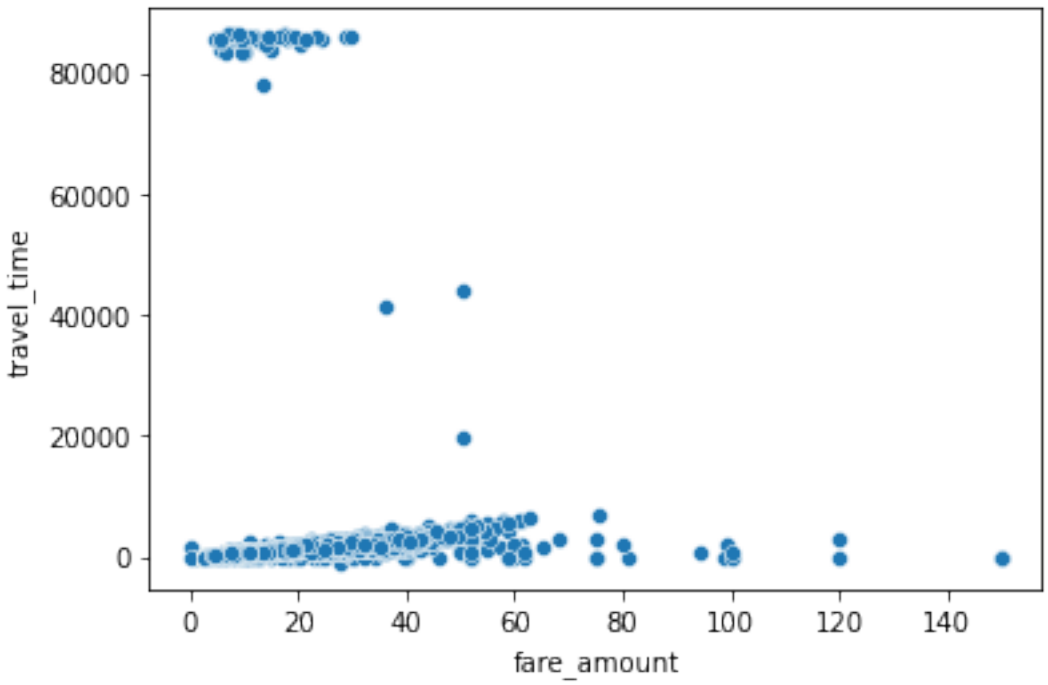
[57]:
```
# Q. fare_amount  trip_distance      scatter plot     .

sns.scatterplot(x = data['fare_amount'], y = data['trip_distance'])
```

[57]: <AxesSubplot:xlabel='fare_amount', ylabel='trip_distance'>



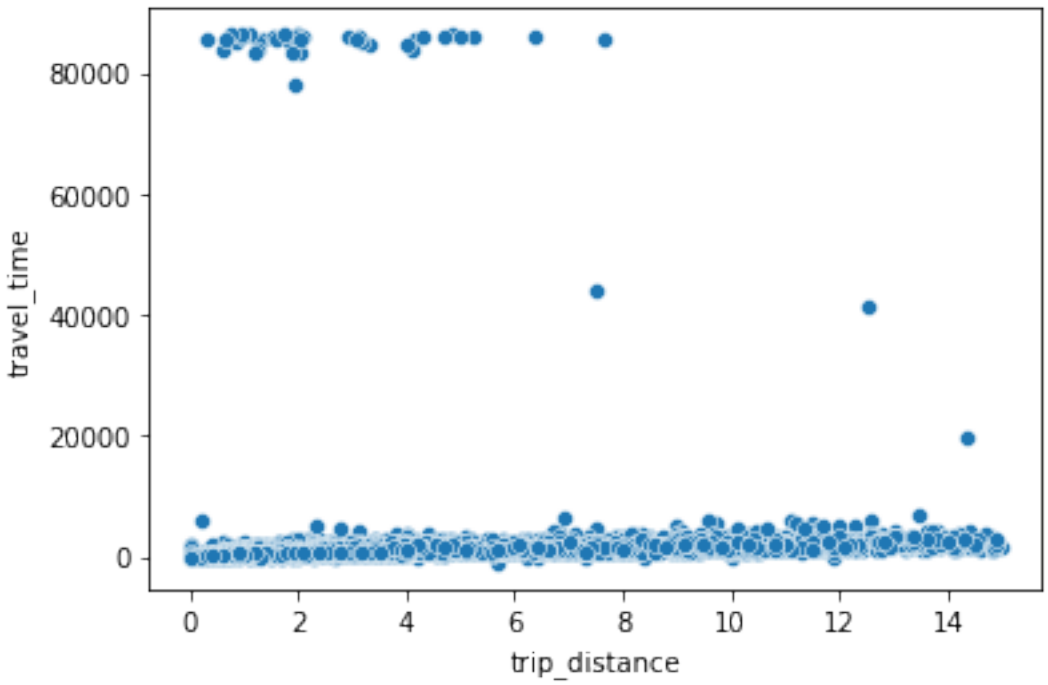
[60]:
```
# Q. fare_amount  travel_time      scatter plot     .

sns.scatterplot(x = data['fare_amount'], y = data['travel_time'])
```

[60]:
```
0         844.0
1        1590.0
2         432.0
3        1815.0
4        1003.0
          ...
22695     567.0
22696     196.0
22698     248.0
22699     716.0
22700     800.0
Name: travel_time, Length: 22019, dtype: float64
```



[61]:
```
# Q. trip_distance  travel_time      scatter plot     .

sns.scatterplot(x = data['trip_distance'], y = data['travel_time'])
```

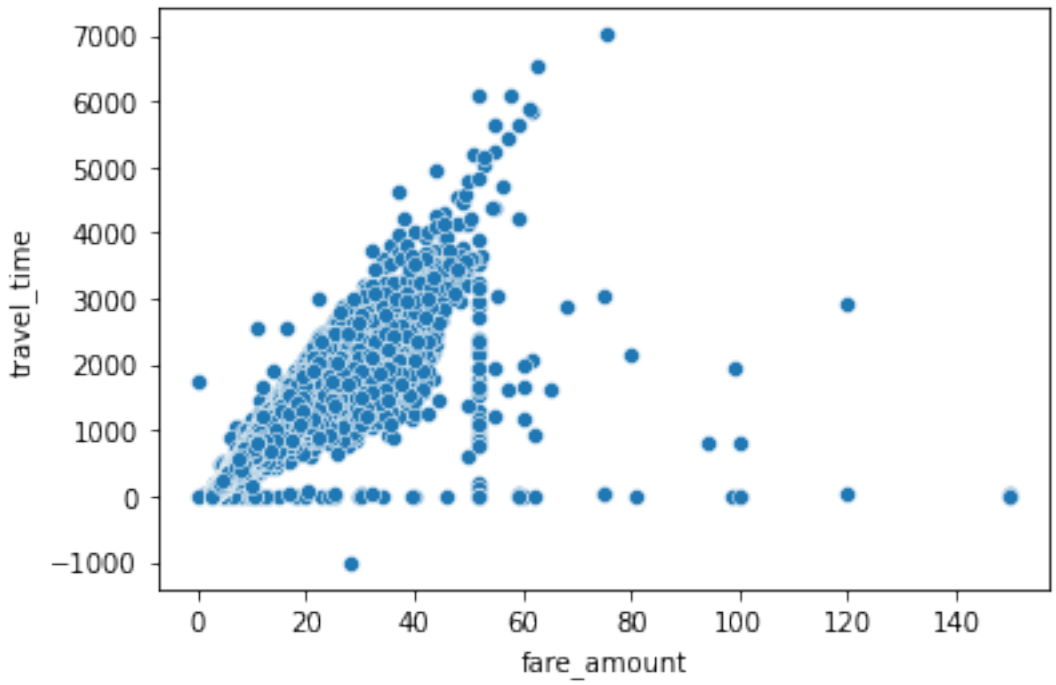
[61]: <AxesSubplot:xlabel='trip_distance', ylabel='travel_time'>



[65]:
```
# Q. scatter plot     travel_time        .
data = data[data['travel_time'] < 15000]
```

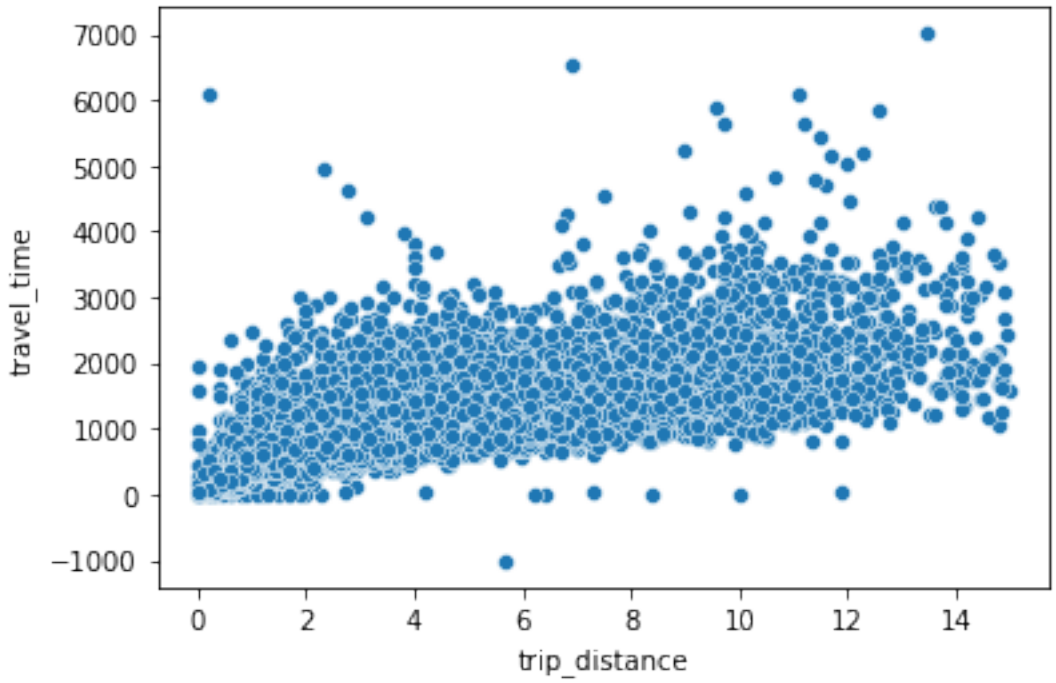
[66]:
```
sns.scatterplot(x = data['fare_amount'], y = data['travel_time'])
```

[66]: <AxesSubplot:xlabel='fare_amount', ylabel='travel_time'>



[67]:
```
sns.scatterplot(x = data['trip_distance'], y = data['travel_time'])
```

[67]: <AxesSubplot:xlabel='trip_distance', ylabel='travel_time'>



[ ]: