In [9]:
```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import seaborn as sns
import matplotlib.pyplot as plt
```

In [10]:
```
data=pd.read_csv("uberr.csv")
```

In [3]:
```
data
```

Out[3]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pick
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	
...	...	...	...	...	...	...
199995	42598914	2012-10-28 10:49:00.00000053	3.0	2012-10-28 10:49:00 UTC	-73.987042	
199996	16382965	2014-03-14 01:09:00.0000008	7.5	2014-03-14 01:09:00 UTC	-73.984722	
199997	27804658	2009-06-29 00:42:00.00000078	30.9	2009-06-29 00:42:00 UTC	-73.986017	
199998	20259894	2015-05-20 14:56:25.0000004	14.5	2015-05-20 14:56:25 UTC	-73.997124	
199999	11951496	2010-05-15 04:08:00.00000076	14.1	2010-05-15 04:08:00 UTC	-73.984395	

200000 rows × 9 columns

In [4]:
```
data.shape
```

Out[4]:
```
(200000, 9)
```

In [5]:
```
data.head(5)
```

Out[5]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_lat
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.73
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.72
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.74
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.79
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.74

In [6]:
```
data.tail(5)
```

Out[6]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	picku
199995	42598914	2012-10-28 10:49:00.00000053	3.0	2012-10-28 10:49:00 UTC	-73.987042	
199996	16382965	2014-03-14 01:09:00.0000008	7.5	2014-03-14 01:09:00 UTC	-73.984722	
199997	27804658	2009-06-29 00:42:00.00000078	30.9	2009-06-29 00:42:00 UTC	-73.986017	
199998	20259894	2015-05-20 14:56:25.0000004	14.5	2015-05-20 14:56:25 UTC	-73.997124	
199999	11951496	2010-05-15 04:08:00.00000076	14.1	2010-05-15 04:08:00 UTC	-73.984395	

In [7]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 9 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   Unnamed: 0         200000 non-null  int64  
 1   key                200000 non-null  object 
 2   fare_amount        200000 non-null  float64
 3   pickup_datetime    200000 non-null  object 
 4   pickup_longitude   200000 non-null  float64
 5   pickup_latitude    200000 non-null  float64
 6   dropoff_longitude  199999 non-null  float64
 7   dropoff_latitude   199999 non-null  float64
 8   passenger_count    200000 non-null  int64  
dtypes: float64(5), int64(2), object(2)
memory usage: 13.7+ MB
```

In [38]:
```
data.to_csv('reseruber.csv')
```

In [39]:
```
data.describe()
```

Out[39]:

	Unnamed: 0	key	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitud
count	2.000000e+05	0.0	200000.000000	200000.000000	200000.000000	199999.00000
mean	2.771250e+07	NaN	11.359955	-72.527638	39.935885	-72.52529
std	1.601382e+07	NaN	9.901776	11.437787	7.720539	13.11740
min	1.000000e+00	NaN	-52.000000	-1340.648410	-74.015515	-3356.66630
25%	1.382535e+07	NaN	6.000000	-73.992065	40.734796	-73.99140
50%	2.774550e+07	NaN	8.500000	-73.981823	40.752592	-73.98009
75%	4.155530e+07	NaN	12.500000	-73.967154	40.767158	-73.96365
max	5.542357e+07	NaN	499.000000	57.418457	1644.421482	1153.57260

In [12]:
```
list(data)
```

Out[12]:
```
['Unnamed: 0',
 'key',
 'fare_amount',
 'pickup_datetime',
 'pickup_longitude',
 'pickup_latitude',
 'dropoff_longitude',
 'dropoff_latitude',
 'passenger_count']
```

In [13]:
```
data['pickup_datatime']=pd.to_datetime(data['pickup_datetime'])
```

In [14]:
```
data['year']=data['pickup_datatime'].dt.year
```

In [15]:
```
data['date']=data['pickup_datatime'].dt.date
```

In [16]:
```
data['time']=data['pickup_datatime'].dt.time
```

In [17]:
```
data['month']=data['pickup_datatime'].dt.month
```

In [18]:
```
print(data[['pickup_datetime','date','time','year','month']].head())
```

```
           pickup_datetime        date      time  year  month
0  2015-05-07 19:52:06 UTC  2015-05-07  19:52:06  2015      5
1  2009-07-17 20:04:56 UTC  2009-07-17  20:04:56  2009      7
2  2009-08-24 21:45:00 UTC  2009-08-24  21:45:00  2009      8
3  2009-06-26 08:22:21 UTC  2009-06-26  08:22:21  2009      6
4  2014-08-28 17:47:00 UTC  2014-08-28  17:47:00  2014      8
```

In [19]:
```
data
```

Out[19]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pick
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	
...	...	...	...	...	...	...
199995	42598914	2012-10-28 10:49:00.00000053	3.0	2012-10-28 10:49:00 UTC	-73.987042	
199996	16382965	2014-03-14 01:09:00.0000008	7.5	2014-03-14 01:09:00 UTC	-73.984722	
199997	27804658	2009-06-29 00:42:00.00000078	30.9	2009-06-29 00:42:00 UTC	-73.986017	
199998	20259894	2015-05-20 14:56:25.0000004	14.5	2015-05-20 14:56:25 UTC	-73.997124	
199999	11951496	2010-05-15 04:08:00.00000076	14.1	2010-05-15 04:08:00 UTC	-73.984395	

200000 rows × 14 columns

In [20]:
```
data.groupby('year')['passenger_count'].sum()
```

Out[20]:
```
year
2009    51398
2010    50849
2011    53079
2012    54156
2013    53343
2014    50923
2015    23159
Name: passenger_count, dtype: int64
```

In [21]:
```
data.groupby('month')['passenger_count'].sum()
```

Out[21]:
```
month
1     29432
2     28028
3     31032
4     31061
5     31847
6     29959
7     25693
8     24314
9     25349
10    27492
11    25944
12    26756
Name: passenger_count, dtype: int64
```

In [22]:
```
data.groupby('date')['passenger_count'].sum()
```

Out[22]:
```
date
2009-01-01    113
2009-01-02    113
2009-01-03    147
2009-01-04    132
2009-01-05    109
              ...
2015-06-26    145
2015-06-27    133
2015-06-28    123
2015-06-29     99
2015-06-30    103
Name: passenger_count, Length: 2372, dtype: int64
```

In [23]:
```
data['year']=pd.to_datetime(data['date']).dt.year
```

In [24]:
```
result=data.groupby('year')['passenger_count'].sum().reset_index()
result
```

Out[24]:

	year	passenger_count
0	2009	51398
1	2010	50849
2	2011	53079
3	2012	54156
4	2013	53343
5	2014	50923
6	2015	23159

In [25]:
```
result=data.groupby('month')['passenger_count'].sum().reset_index()
result
```

Out[25]:

	month	passenger_count
0	1	29432
1	2	28028
2	3	31032
3	4	31061
4	5	31847
5	6	29959
6	7	25693
7	8	24314
8	9	25349
9	10	27492
10	11	25944
11	12	26756

In [26]:
```
# Check the column names
print(data.columns)
```

```
Index(['Unnamed: 0', 'key', 'fare_amount', 'pickup_datetime',
       'pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'passenger_count', 'pickup_datatime', 'year',
       'date', 'time', 'month'],
      dtype='object')
```

In [27]:
```
data_numeric = data.select_dtypes(include='number')
cor_mat = data_numeric.corr()
```

In [28]:
```
data['key'] = pd.to_numeric(data['key'], errors='coerce')
```

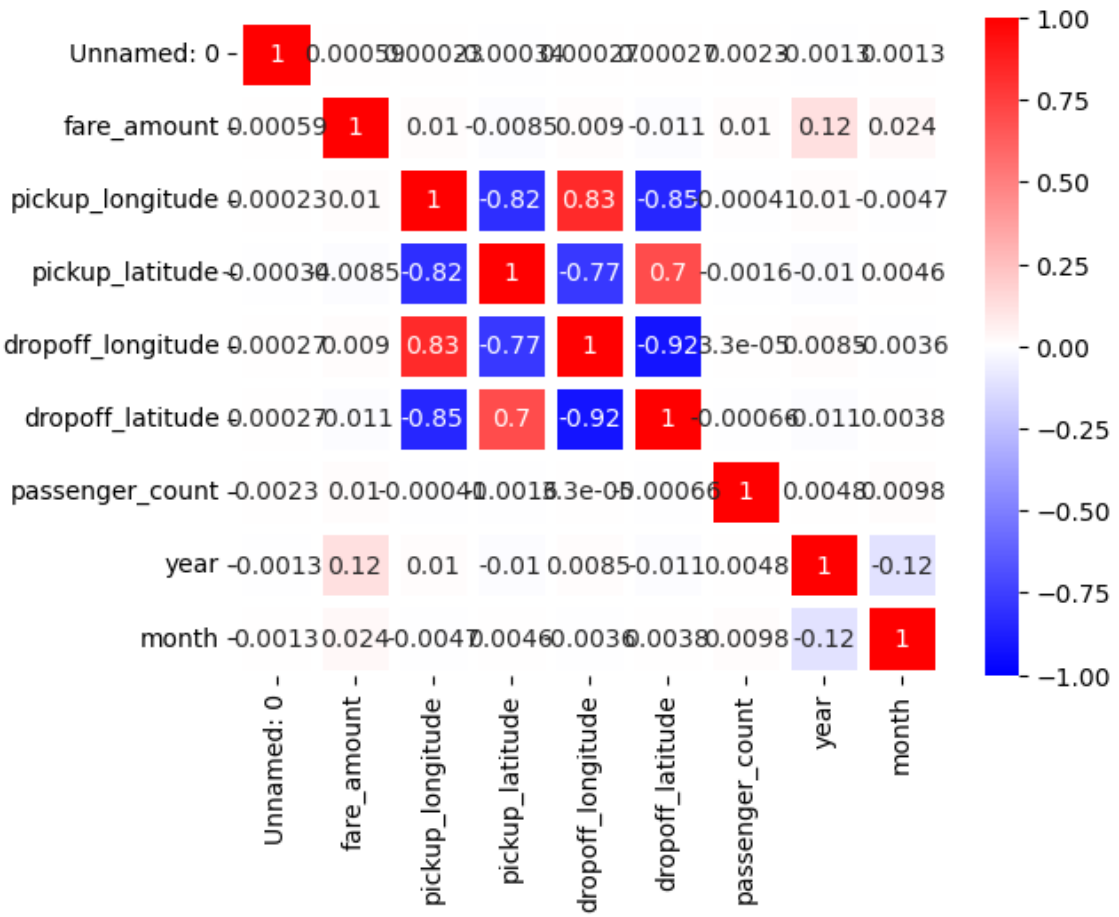
In [29]:
```
cor_mat
```

Out[29]:

	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitu
Unnamed: 0	1.000000	0.000589	0.000230	-0.000341	0.0002
fare_amount	0.000589	1.000000	0.010457	-0.008481	0.0089
pickup_longitude	0.000230	0.010457	1.000000	-0.816461	0.8330
pickup_latitude	-0.000341	-0.008481	-0.816461	1.000000	-0.7747
dropoff_longitude	0.000270	0.008986	0.833026	-0.774787	1.0000
dropoff_latitude	0.000271	-0.011014	-0.846324	0.702367	-0.9170
passenger_count	0.002257	0.010150	-0.000414	-0.001560	0.0000
year	-0.001324	0.118335	0.009966	-0.010233	0.0084
month	0.001299	0.023814	-0.004665	0.004625	-0.0036

In [30]:
```
import seaborn as sns
sns.heatmap(cor_mat,vmin=-1,annot=True,linewidth=5,cmap='bwr')
```

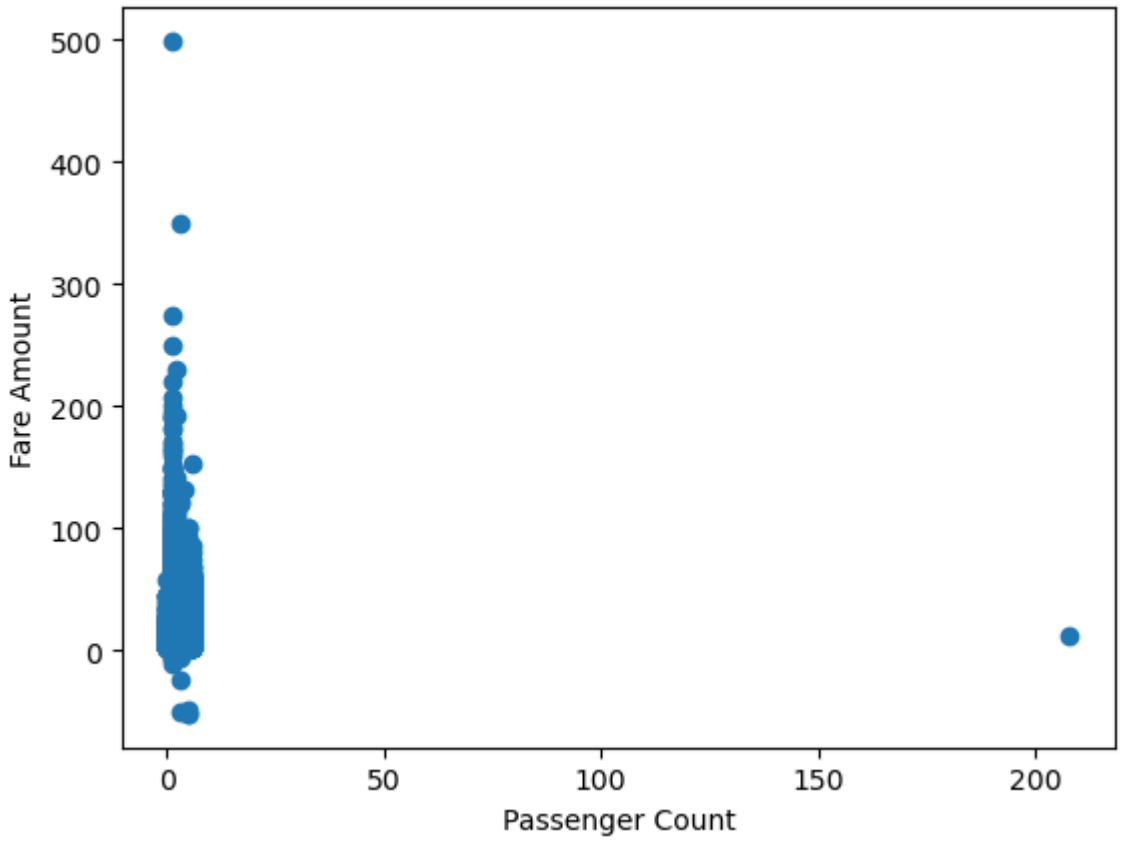
Out[30]: <Axes: >



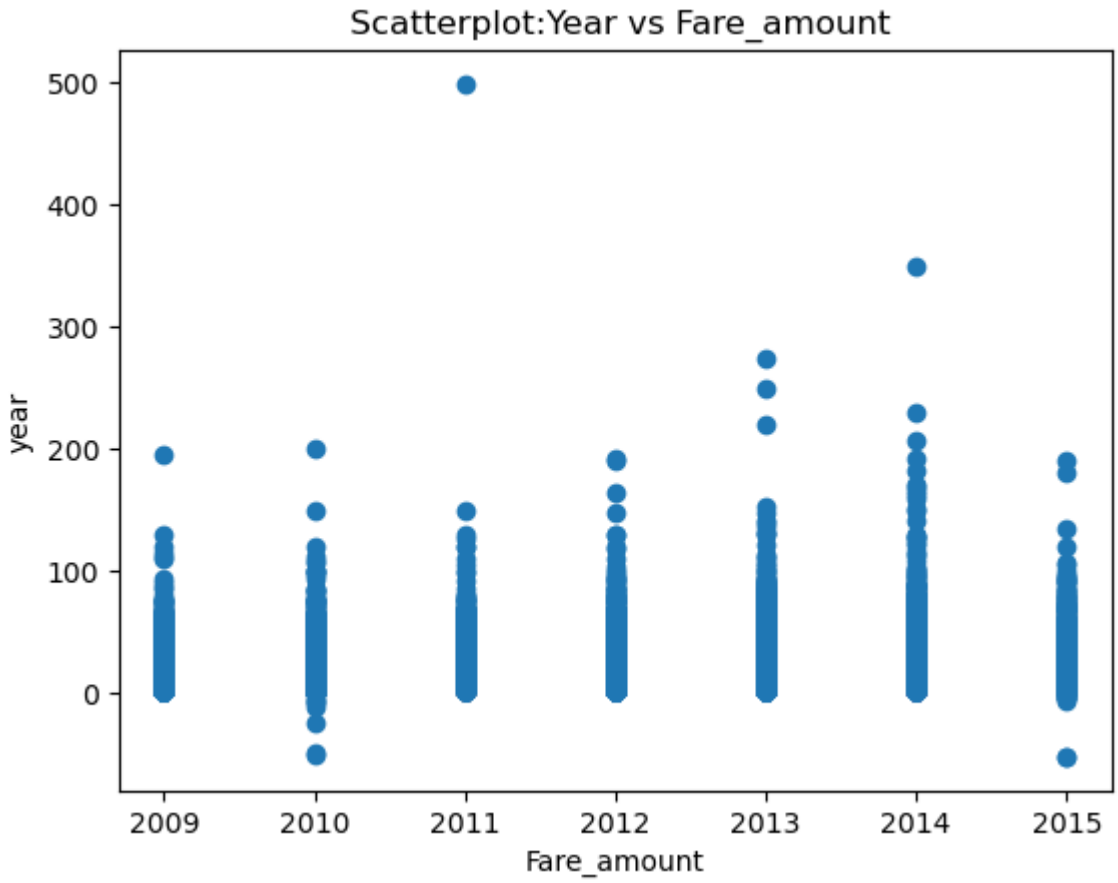
In [31]:
```
data.isnull().sum()
```

Out[31]:
```
Unnamed: 0             0
key                200000
fare_amount             0
pickup_datetime         0
pickup_longitude        0
pickup_latitude         0
dropoff_longitude       1
dropoff_latitude        1
passenger_count         0
pickup_datatime         0
year                    0
date                    0
time                    0
month                   0
dtype: int64
```

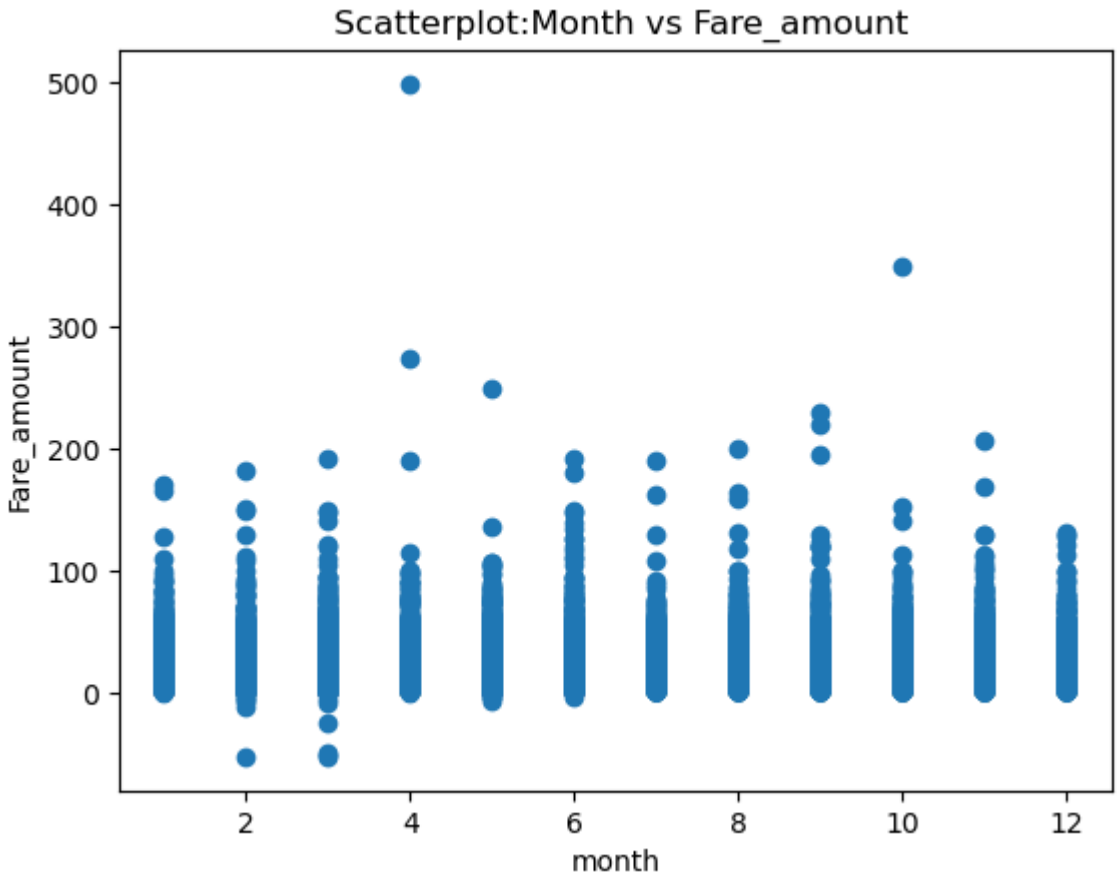
In [32]:
```
plt.scatter(data['passenger_count'], data['fare_amount'])
plt.xlabel('Passenger Count')
plt.ylabel('Fare Amount')
plt.show()
```



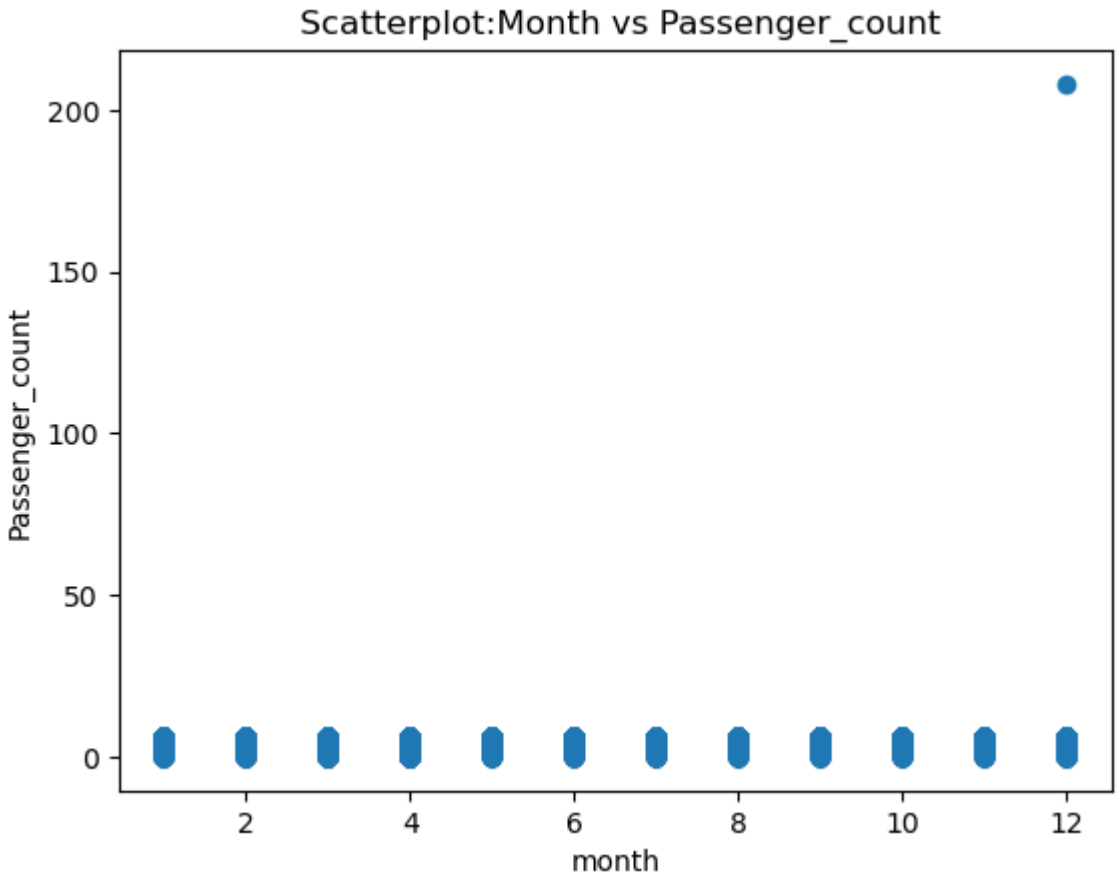
In [33]:
```
plt.scatter(data['year'],data['fare_amount'])
plt.ylabel('year')
plt.xlabel('Fare_amount')
plt.title(' Scatterplot:Year vs Fare_amount')
plt.show()
```



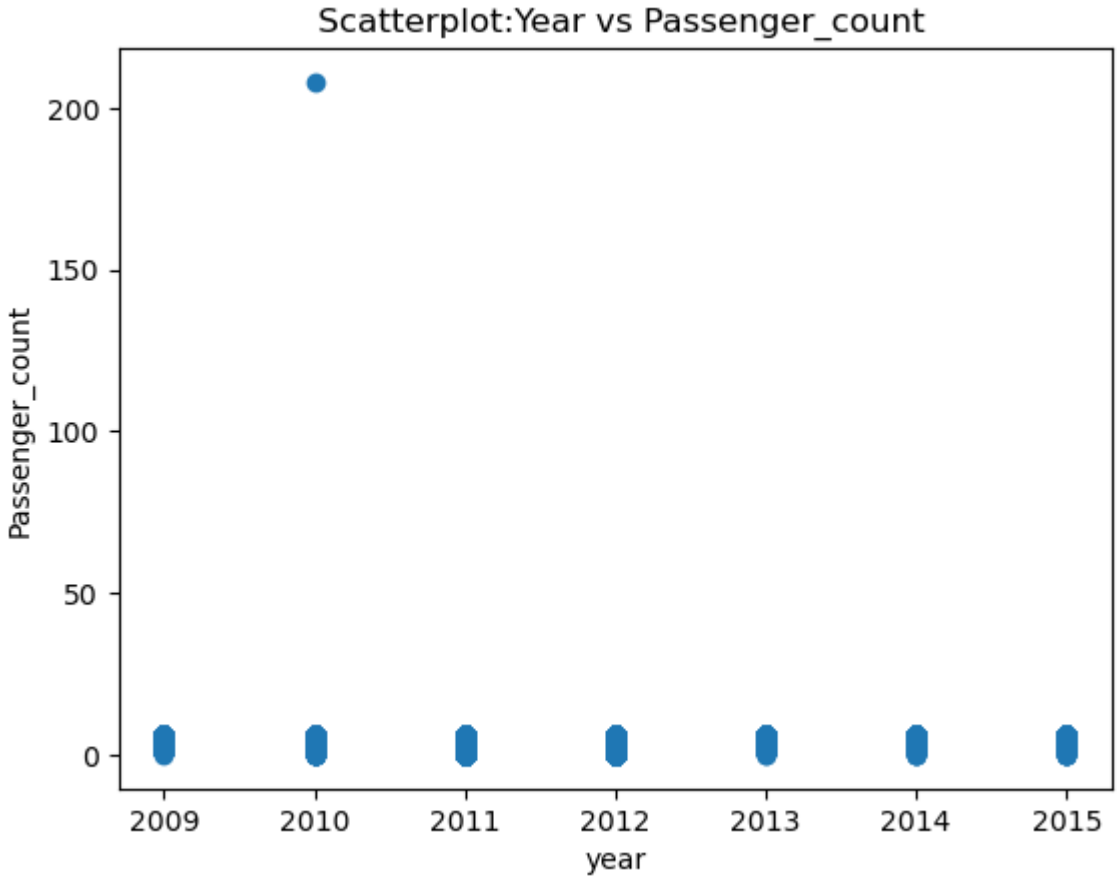
In [34]:
```
plt.scatter(data['month'],data['fare_amount'])
plt.xlabel('month')
plt.ylabel('Fare_amount')
plt.title(' Scatterplot:Month vs Fare_amount')
plt.show()
```



In [35]:
```
plt.scatter(data['month'],data['passenger_count'])
plt.xlabel('month')
plt.ylabel('Passenger_count')
plt.title(' Scatterplot:Month vs Passenger_count')
plt.show()
```



In [36]:
```
plt.scatter(data['year'],data['passenger_count'])
plt.xlabel('year')
plt.ylabel('Passenger_count')
plt.title(' Scatterplot:Year vs Passenger_count')
plt.show()
```



In [ ]:
```

```