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

In [2]:
```
df=pd.read_csv('uber.csv')
```

In [3]:
```
df
```

Out[3]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pi
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]:
```
df.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 [5]:
```
df['pickup_datetime'].value_counts()
```

Out[5]:
```
pickup_datetime
2014-04-13 18:19:00 UTC    4
2010-03-14 12:00:00 UTC    4
2009-02-12 12:46:00 UTC    4
2011-02-18 18:55:00 UTC    3
2009-03-12 17:12:00 UTC    3
                          ..
2013-03-08 07:16:00 UTC    1
2013-05-17 21:33:31 UTC    1
2009-10-24 04:05:00 UTC    1
2013-05-16 16:12:00 UTC    1
2010-05-15 04:08:00 UTC    1
Name: count, Length: 196629, dtype: int64
```

In [6]:
```
df['pickup_datetime']=pd.to_datetime(df['pickup_datetime'])
```

In [7]:
```
df['year']=df['pickup_datetime'].dt.year
df['month']=df['pickup_datetime'].dt.month
df['time']=df['pickup_datetime'].dt.time
df['date']=df['pickup_datetime'].dt.date
```

In [8]:
```
df
```

Out[8]:

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

200000 rows × 13 columns

In [10]:
```
df.groupby('year').count()
```

Out[10]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	
year							
2009	30536	30536	30536	30536	30536	30536	
2010	30194	30194	30194	30194	30194	30194	
2011	31945	31945	31945	31945	31945	31945	
2012	32396	32396	32396	32396	32396	32396	
2013	31195	31195	31195	31195	31195	31195	
2014	29968	29968	29968	29968	29968	29968	
2015	13766	13766	13766	13766	13766	13766	

In [11]:
```
df.groupby('month').count()
```

Out[11]:

month	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude
1	17668	17668	17668	17668	17668	17668
2	16695	16695	16695	16695	16695	16695
3	18763	18763	18763	18763	18763	18763
4	18606	18606	18606	18606	18606	18606
5	18859	18859	18859	18859	18859	18859
6	17787	17787	17787	17787	17787	17787
7	15095	15095	15095	15095	15095	15095
8	14221	14221	14221	14221	14221	14221
9	15266	15266	15266	15266	15266	15266
10	16212	16212	16212	16212	16212	16212
11	15312	15312	15312	15312	15312	15312
12	15516	15516	15516	15516	15516	15516

In [12]:
```
df.groupby('time').count()
```

Out[12]:

time	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude
00:00:00	79	79	79	79	79	79
00:00:02	1	1	1	1	1	1
00:00:03	3	3	3	3	3	3
00:00:07	4	4	4	4	4	4
00:00:09	2	2	2	2	2	2
...	...	...	...	...	...	...
23:59:54	4	4	4	4	4	4
23:59:55	2	2	2	2	2	2
23:59:57	1	1	1	1	1	1
23:59:58	2	2	2	2	2	2
23:59:59	4	4	4	4	4	4

59072 rows × 12 columns

In [13]:
```
df.groupby('date').count()
```

Out[13]:

date	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dr
2009-01-01	63	63	63	63	63	63	
2009-01-02	60	60	60	60	60	60	
2009-01-03	84	84	84	84	84	84	
2009-01-04	75	75	75	75	75	75	
2009-01-05	64	64	64	64	64	64	
...	...	...	...	...	...	...	
2015-06-26	81	81	81	81	81	81	
2015-06-27	75	75	75	75	75	75	
2015-06-28	65	65	65	65	65	65	
2015-06-29	63	63	63	63	63	63	
2015-06-30	66	66	66	66	66	66	

2372 rows × 12 columns

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 13 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  datetime64[ns, UTC]
 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              
 9   year               200000 non-null  int32              
 10  month              200000 non-null  int32              
 11  time               200000 non-null  object             
 12  date               200000 non-null  object             
dtypes: datetime64[ns, UTC](1), float64(5), int32(2), int64(2), object
(3)
memory usage: 18.3+ MB
```

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
Unnamed: 0           0
key                  0
fare_amount          0
pickup_datetime      0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    1
dropoff_latitude     1
passenger_count      0
year                 0
month                0
time                 0
date                 0
dtype: int64
```

In [16]:
```
del df['dropoff_longitude']
del df['dropoff_latitude']
del df['pickup_datetime']
```

In [17]:
```
del df['pickup_longitude']
del df['Unnamed: 0']
del df['pickup_latitude']
del df['key']
```

In [18]:
```
df
```

Out[18]:

	fare_amount	passenger_count	year	month	time	date
0	7.5	1	2015	5	19:52:06	2015-05-07
1	7.7	1	2009	7	20:04:56	2009-07-17
2	12.9	1	2009	8	21:45:00	2009-08-24
3	5.3	3	2009	6	08:22:21	2009-06-26
4	16.0	5	2014	8	17:47:00	2014-08-28
...	...	...	...	...	...	...
199995	3.0	1	2012	10	10:49:00	2012-10-28
199996	7.5	1	2014	3	01:09:00	2014-03-14
199997	30.9	2	2009	6	00:42:00	2009-06-29
199998	14.5	1	2015	5	14:56:25	2015-05-20
199999	14.1	1	2010	5	04:08:00	2010-05-15

200000 rows × 6 columns

In [19]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 6 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   fare_amount      200000 non-null  float64
 1   passenger_count  200000 non-null  int64  
 2   year             200000 non-null  int32  
 3   month            200000 non-null  int32  
 4   time             200000 non-null  object 
 5   date             200000 non-null  object 
dtypes: float64(1), int32(2), int64(1), object(2)
memory usage: 7.6+ MB
```

In [20]:
```
#df=pd.get_dummies('time')
#df=pd.get_dummies('date')
```

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 6 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   fare_amount      200000 non-null  float64
 1   passenger_count  200000 non-null  int64  
 2   year             200000 non-null  int32  
 3   month            200000 non-null  int32  
 4   time             200000 non-null  object 
 5   date             200000 non-null  object 
dtypes: float64(1), int32(2), int64(1), object(2)
memory usage: 7.6+ MB
```

In [22]:
```
df
```

Out[22]:

	fare_amount	passenger_count	year	month	time	date
0	7.5	1	2015	5	19:52:06	2015-05-07
1	7.7	1	2009	7	20:04:56	2009-07-17
2	12.9	1	2009	8	21:45:00	2009-08-24
3	5.3	3	2009	6	08:22:21	2009-06-26
4	16.0	5	2014	8	17:47:00	2014-08-28
...	...	...	...	...	...	...
199995	3.0	1	2012	10	10:49:00	2012-10-28
199996	7.5	1	2014	3	01:09:00	2014-03-14
199997	30.9	2	2009	6	00:42:00	2009-06-29
199998	14.5	1	2015	5	14:56:25	2015-05-20
199999	14.1	1	2010	5	04:08:00	2010-05-15

200000 rows × 6 columns

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

In [24]:
```
result=df.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=df.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

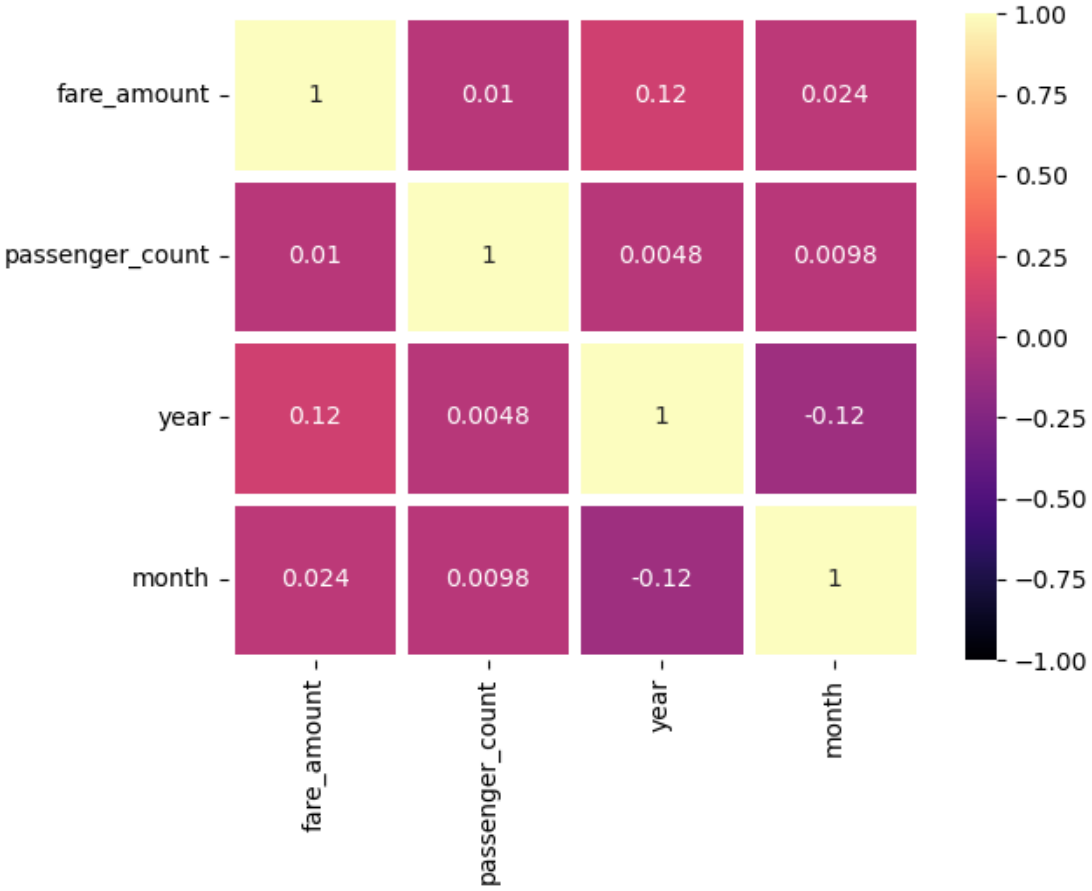
In [36]:
```
df_numeric = df.select_dtypes(include='number')
cor_mat = df_numeric.corr()
```

In [ ]:
```

```

In [37]:
```
import seaborn as sns
sns.heatmap(cor_mat,vmax=1,vmin=-1,annot=True,linewidth=5,cmap='magma')
```

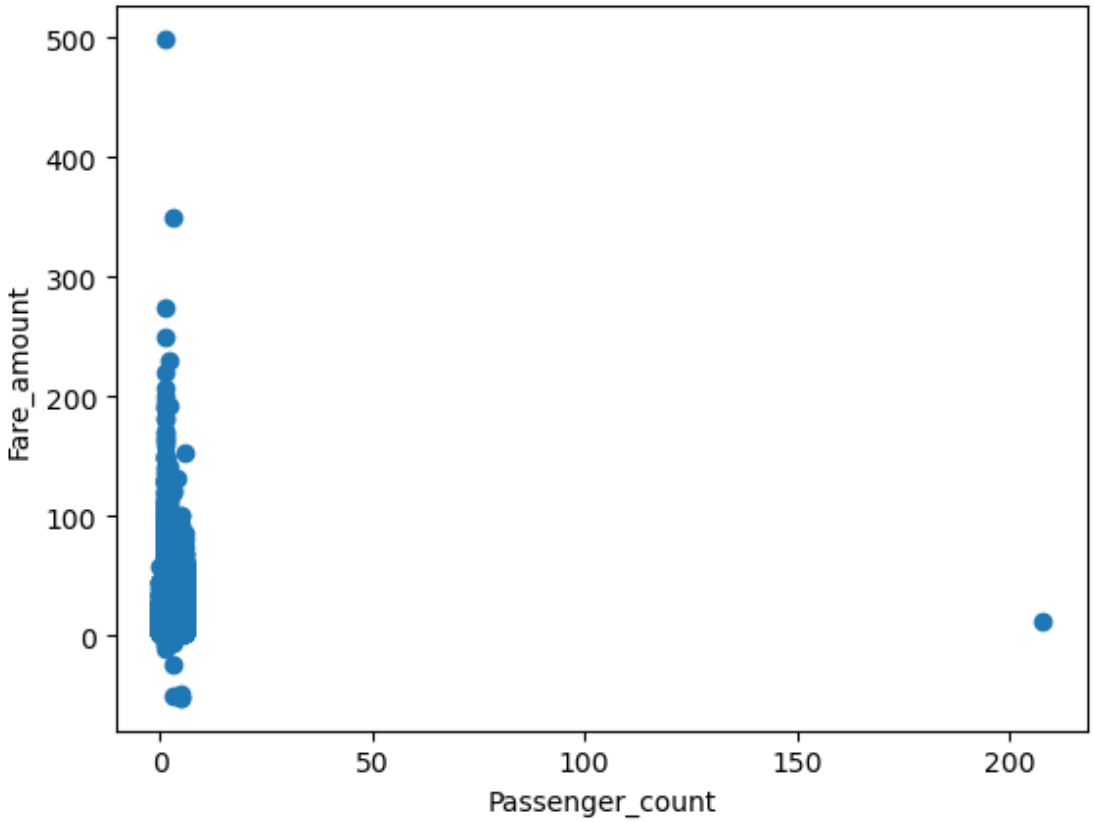
Out[37]: <Axes: >



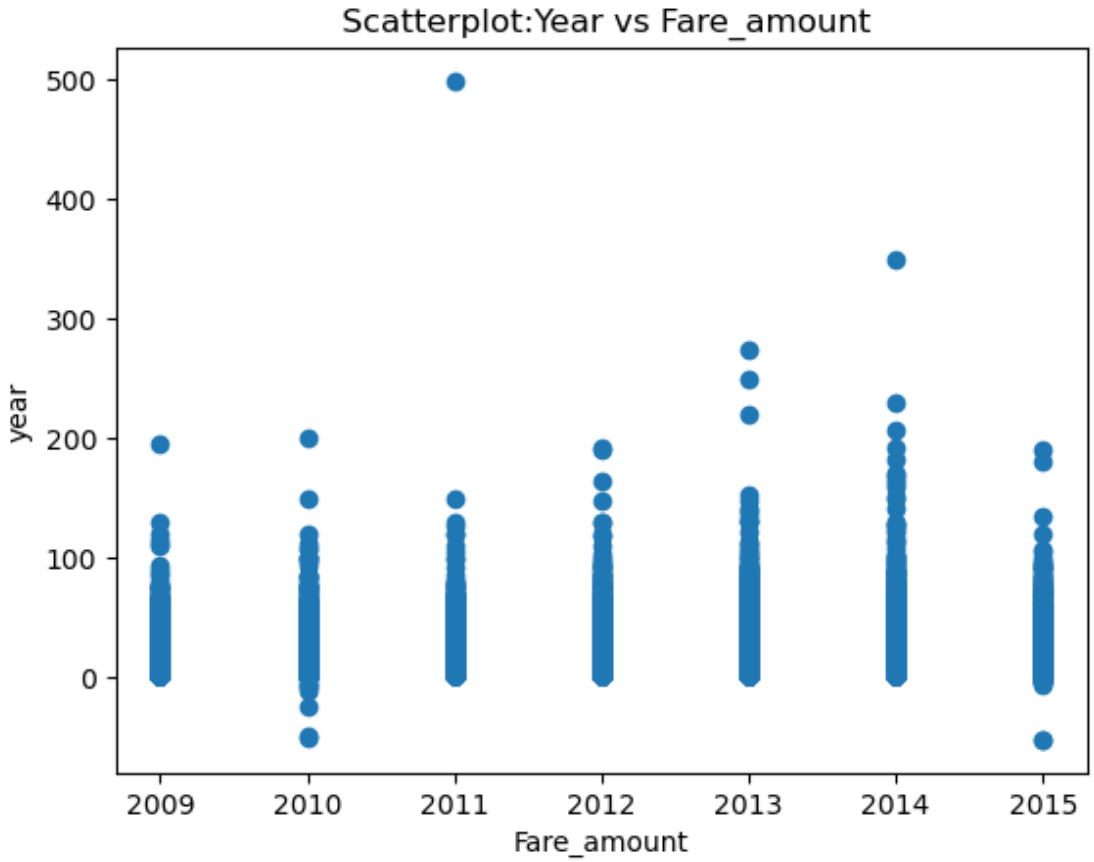
In [38]:
```
df.isnull().sum()
```

Out[38]:
```
fare_amount        0
passenger_count    0
year               0
month              0
time               0
date               0
dtype: int64
```

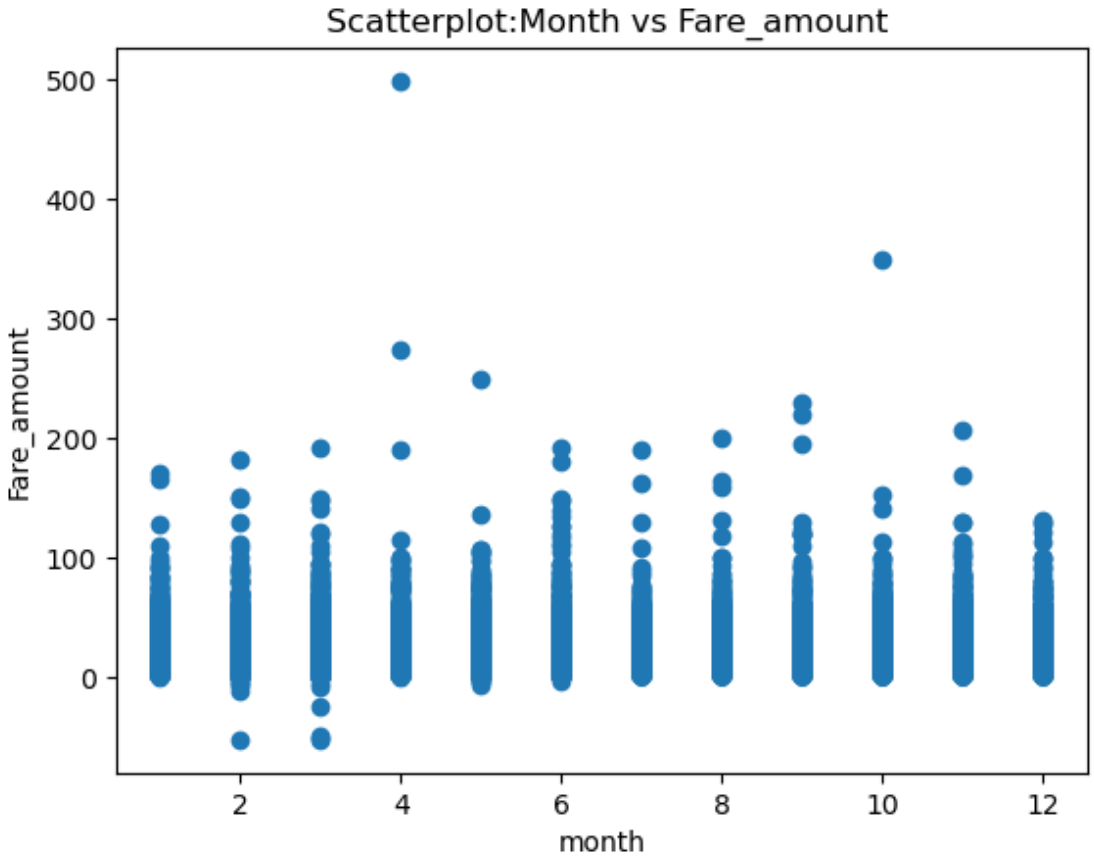
In [39]:
```
plt.scatter(df['passenger_count'],df['fare_amount'])
plt.xlabel('Passenger_count')
plt.ylabel('Fare_amount')
plt.show()
```



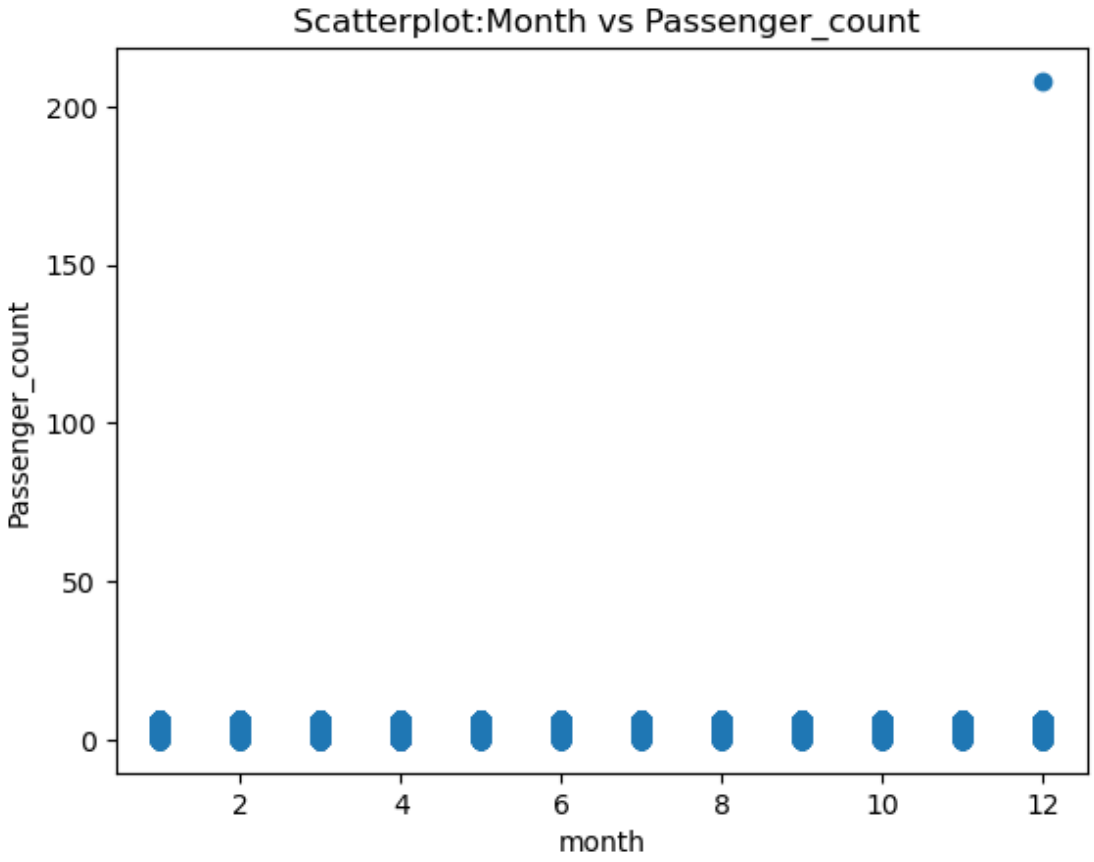
In [40]:
```
plt.scatter(df['year'],df['fare_amount'])
plt.ylabel('year')
plt.xlabel('Fare_amount')
plt.title(' Scatterplot:Year vs Fare_amount')
plt.show()
```



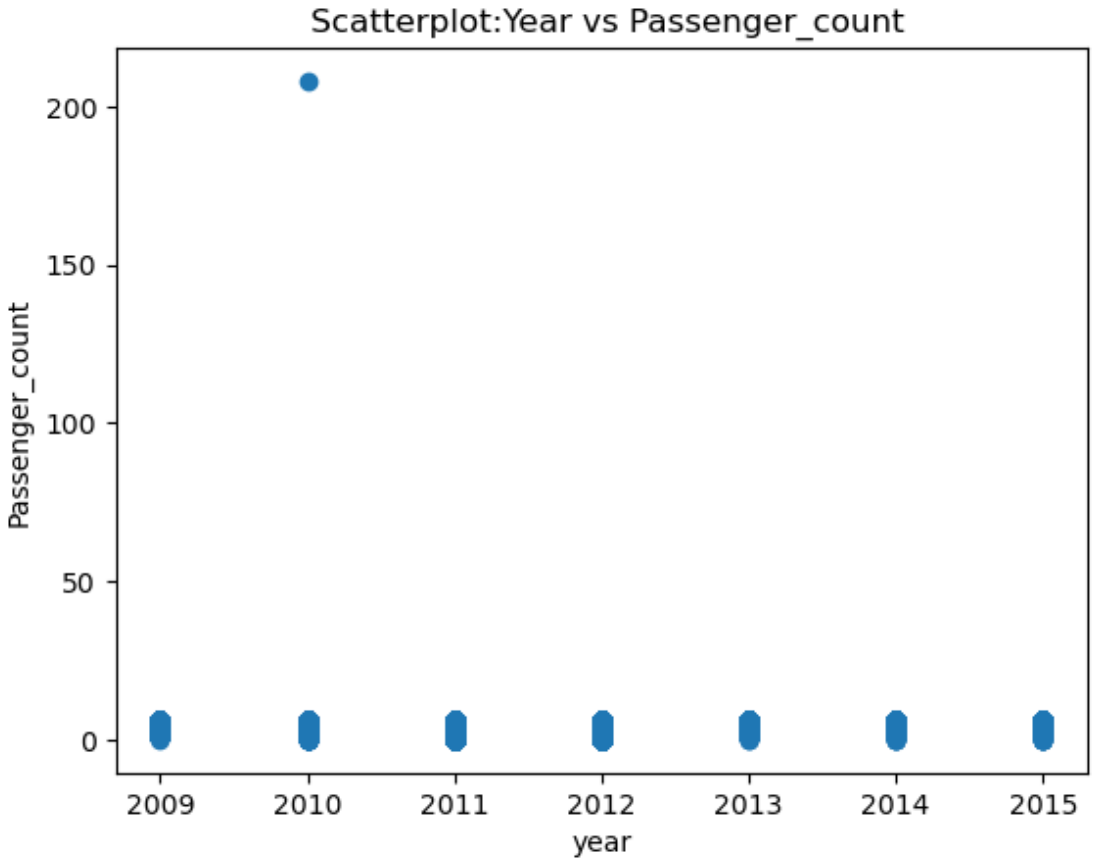
In [41]:
```
plt.scatter(df['month'],df['fare_amount'])
plt.xlabel('month')
plt.ylabel('Fare_amount')
plt.title(' Scatterplot:Month vs Fare_amount')
plt.show()
```



In [42]:
```
plt.scatter(df['month'],df['passenger_count'])
plt.xlabel('month')
plt.ylabel('Passenger_count')
plt.title(' Scatterplot:Month vs Passenger_count')
plt.show()
```



In [43]:
```
plt.scatter(df['year'],df['passenger_count'])
plt.xlabel('year')
plt.ylabel('Passenger_count')
plt.title(' Scatterplot:Year vs Passenger_count')
plt.show()
```



In [44]:
```
df.to_csv('newfile_uber.csv')
```

In [ ]:

In [ ]:

In [ ]:

In [ ]: