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

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

200000 rows × 13 columns

In [9]:
```
df.groupby('year').sum()
```

Out[9]:

year	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_
2009	848710858477	305637.75	-2.232821e+06	1.229674e+06	-2.232855e+06	1.230
2010	833729967335	306002.55	-2.187446e+06	1.204660e+06	-2.187036e+06	1.204
2011	886031339250	332326.24	-2.312442e+06	1.275468e+06	-2.314091e+06	1.273
2012	900860818069	363298.45	-2.341982e+06	1.290853e+06	-2.340214e+06	1.289
2013	863791365792	396489.39	-2.257106e+06	1.238200e+06	-2.256054e+06	1.237
2014	827695471297	390094.57	-2.170868e+06	1.195865e+06	-2.170769e+06	1.195
2015	381680916250	178142.10	-1.002863e+06	5.524570e+05	-1.003968e+06	5.530

In [10]:
```
df.groupby('month').sum()
```

Out[10]:

month	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropo
1	486987125167	189499.77	-1.282882e+06	706744.720934	-1.283403e+06	7076
2	462616125495	182453.99	-1.210854e+06	667174.316142	-1.211418e+06	6673
3	518260998186	208300.37	-1.361985e+06	750790.077412	-1.362392e+06	7505
4	516426844720	210972.89	-1.349988e+06	743473.645669	-1.350249e+06	7437
5	523325048690	220246.02	-1.355818e+06	742706.361097	-1.354176e+06	7404
6	497313077455	206421.84	-1.291368e+06	711916.114951	-1.291373e+06	7111
7	419062857609	168478.59	-1.096190e+06	603516.026602	-1.095832e+06	6037
8	392180588364	159351.40	-1.028802e+06	566336.153238	-1.028017e+06	5664
9	420411601849	180011.21	-1.108592e+06	610728.102071	-1.108655e+06	6107
10	451256645918	190058.67	-1.178197e+06	648636.267774	-1.177370e+06	6487
11	425127319399	177806.02	-1.113620e+06	614098.428624	-1.115476e+06	6134
12	429532503618	178390.28	-1.127233e+06	621056.861089	-1.126626e+06	6206

In [11]:
```
df.groupby('time').sum()
```

Out[11]:

time	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropof
00:00:00	1891896212	1103.83	-5622.246162	3096.781739	-5621.801465	309
00:00:02	44076675	10.50	-74.006600	40.739723	-73.985401	4
00:00:03	144979463	34.70	-221.977395	122.206233	-221.934831	12
00:00:07	85590703	72.00	-295.941793	162.980459	-295.914283	16
00:00:09	56404478	44.90	-147.749583	81.410774	-147.926483	8
...	...	...	...	...	...	
23:59:54	100023642	40.20	-295.758908	163.043281	-295.799398	16
23:59:55	75424814	14.50	-147.926398	81.459413	-147.922748	8
23:59:57	46476086	6.00	-73.991553	40.750460	-73.986270	4
23:59:58	57946679	16.30	-147.980875	81.461953	-147.983098	8
23:59:59	131897464	37.50	-295.952207	163.045907	-295.895126	16

59072 rows × 9 columns

In [12]:
```
df.groupby('date').sum()
```

Out[12]:

date	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_la
2009-01-01	1756779842	621.20	-4587.326377	2530.820767	-4587.245918	2530.5
2009-01-02	1578947277	739.55	-4363.629330	2404.780476	-4363.733208	2404.6
2009-01-03	2355445751	935.70	-6139.758759	3382.516586	-6139.262025	3382.6
2009-01-04	2078662060	733.30	-5547.644664	3056.309580	-5548.045850	3056.4
2009-01-05	1929909985	550.95	-4586.648421	2526.734719	-4586.646650	2526.7
...	...	...	...	...	...	
2015-06-26	2150245049	1082.12	-5918.214592	3260.027855	-5991.784050	3300.4
2015-06-27	2040511693	1114.24	-5548.435600	3056.282757	-5547.671196	3056.4
2015-06-28	1671454013	905.58	-4808.014450	2648.532997	-4807.767609	2648.6
2015-06-29	1633785053	764.12	-4586.406067	2526.232811	-4586.415230	2526.3
2015-06-30	1865942856	884.66	-4807.688179	2648.659561	-4807.745293	2648.9

2372 rows × 9 columns

In [13]:
```
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  int64              
 10  month              200000 non-null  int64              
 11  time               200000 non-null  object             
 12  date               200000 non-null  object             
dtypes: datetime64[ns, UTC](1), float64(5), int64(4), object(3)
memory usage: 19.8+ MB
```

In [14]:
```
df.isnull().sum()
```

Out[14]:
```
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 [15]:
```
del df['dropoff_longitude']
del df['dropoff_latitude']
del df['pickup_datetime']
```

In [16]:
```
del df['pickup_longitude']
del df['Unnamed: 0']
del df['pickup_latitude']
del df['key']
```

In [17]:
```
df
```

Out[17]:

	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 [18]:
```
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  int64  
 3   month            200000 non-null  int64  
 4   time             200000 non-null  object 
 5   date             200000 non-null  object 
dtypes: float64(1), int64(3), object(2)
memory usage: 9.2+ MB
```

In [19]:
```
#df=pd.get_dummies('time')
#df=pd.get_dummies('date')
```

In [20]:
```
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  int64  
 3   month            200000 non-null  int64  
 4   time             200000 non-null  object 
 5   date             200000 non-null  object 
dtypes: float64(1), int64(3), object(2)
memory usage: 9.2+ MB
```

In [21]:
```
df
```

Out[21]:

	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 [22]:
```
df['year']=pd.to_datetime(df['date']).dt.year
```

In [23]:
```
result=df.groupby('year')['passenger_count'].sum().reset_index()
result
```

Out[23]:

	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 [24]:
```
result=df.groupby('month')['passenger_count'].sum().reset_index()
result
```

Out[24]:

	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 [25]:
```
cor_mat=df.corr()
```

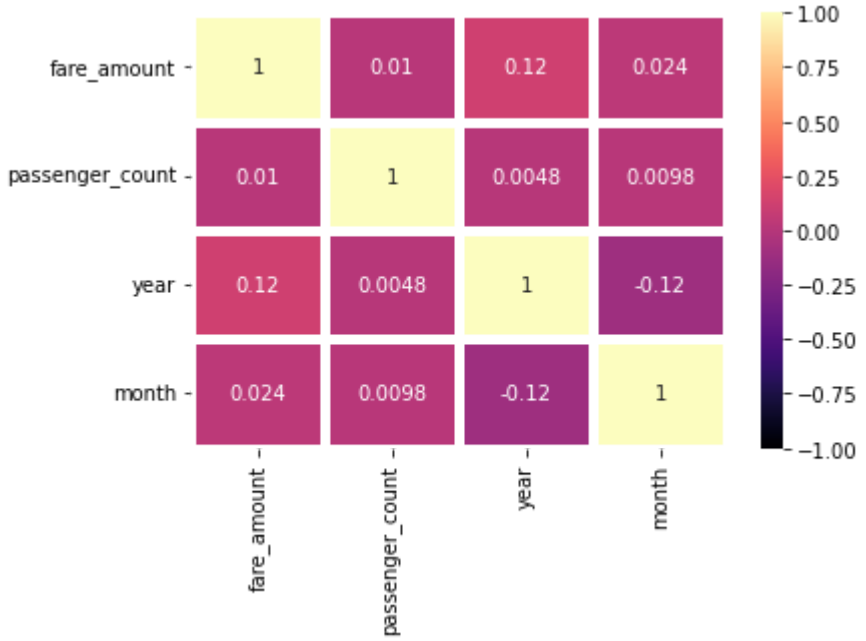
In [26]:
```
cor_mat
```

Out[26]:

	fare_amount	passenger_count	year	month
fare_amount	1.000000	0.010150	0.118335	0.023814
passenger_count	0.010150	1.000000	0.004798	0.009773
year	0.118335	0.004798	1.000000	-0.115859
month	0.023814	0.009773	-0.115859	1.000000

In [27]:
```
import seaborn as sns
sns.heatmap(cor_mat,vmax=1,vmin=-1,annot=True,linewidth=5,cmap='magma'
```

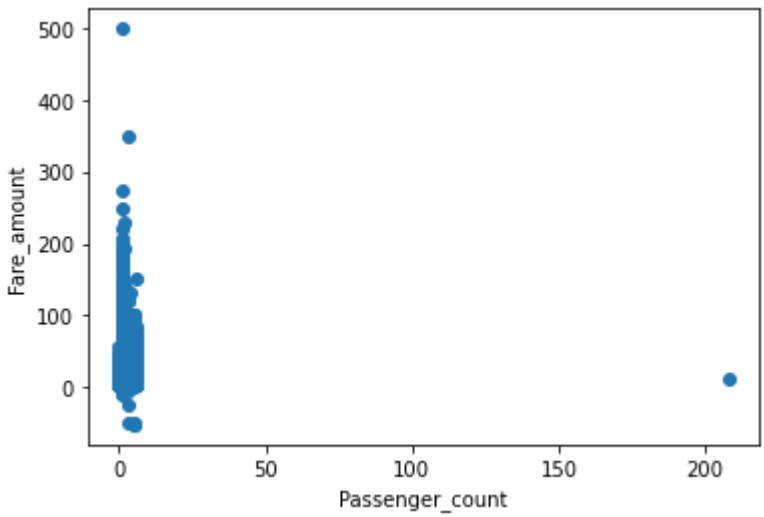
Out[27]: <AxesSubplot:>



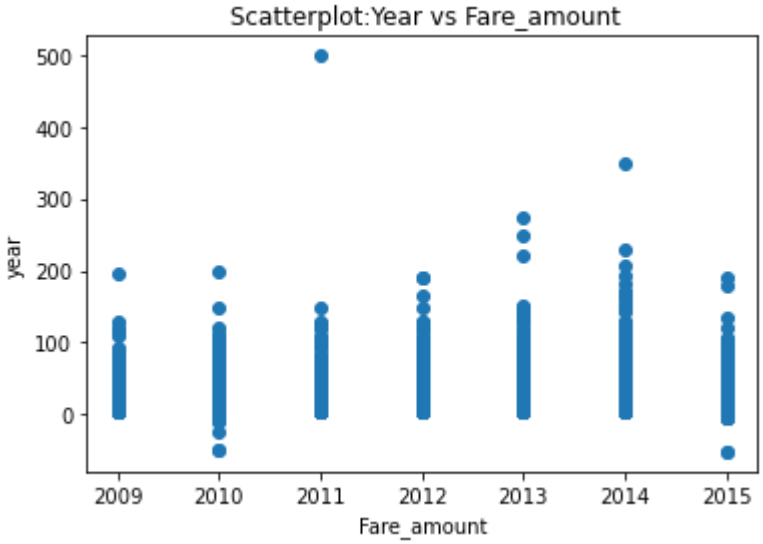
In [28]:
```
df.isnull().sum()
```

Out[28]:
```
fare_amount        0
passenger_count    0
year               0
month              0
time               0
date               0
dtype: int64
```

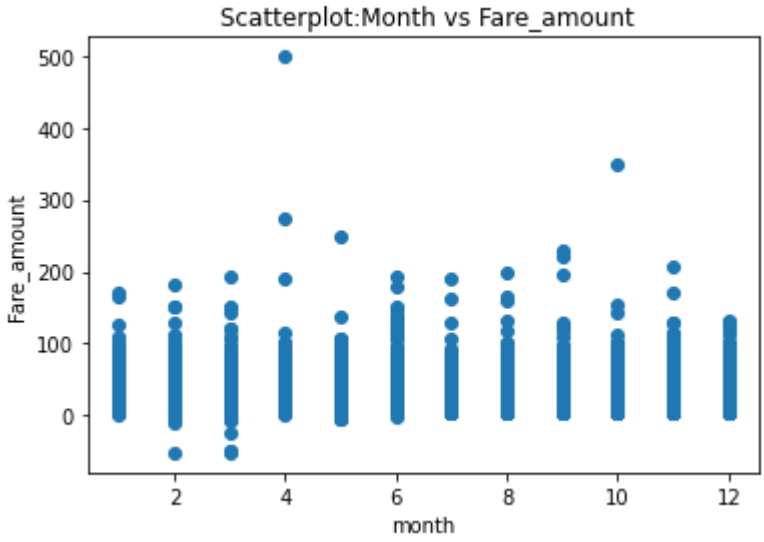
In [29]:
```
plt.scatter(df['passenger_count'],df['fare_amount'])
plt.xlabel('Passenger_count')
plt.ylabel('Fare_amount')
plt.show()
```



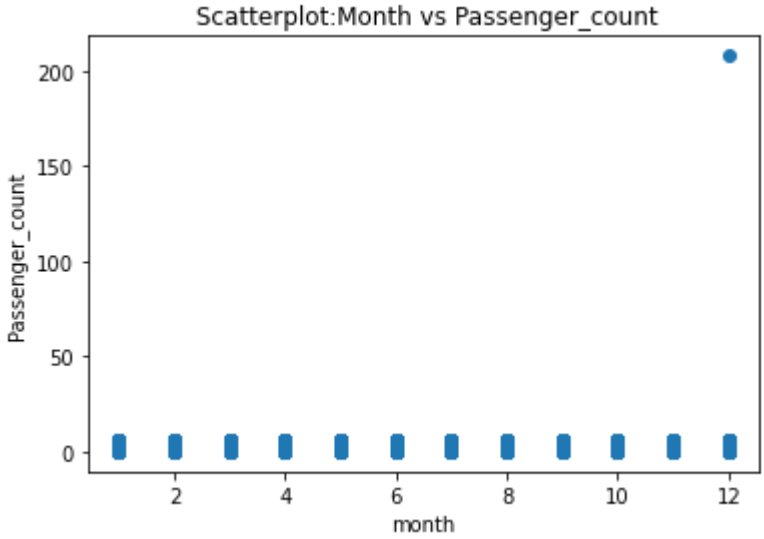
In [30]:
```
plt.scatter(df['year'],df['fare_amount'])
plt.ylabel('year')
plt.xlabel('Fare_amount')
plt.title(' Scatterplot:Year vs Fare_amount')
plt.show()
```



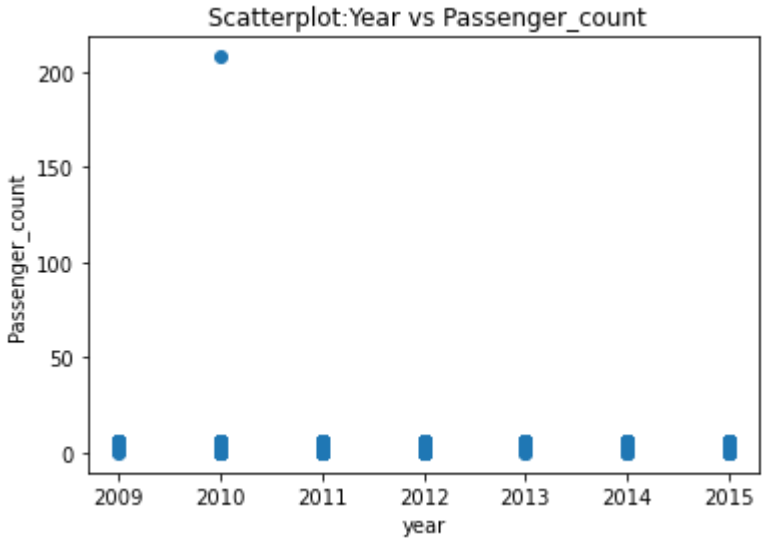
In [31]:
```
plt.scatter(df['month'],df['fare_amount'])
plt.xlabel('month')
plt.ylabel('Fare_amount')
plt.title(' Scatterplot:Month vs Fare_amount')
plt.show()
```



In [32]:
```
plt.scatter(df['month'],df['passenger_count'])
plt.xlabel('month')
plt.ylabel('Passenger_count')
plt.title(' Scatterplot:Month vs Passenger_count')
plt.show()
```



In [33]:
```
plt.scatter(df['year'],df['passenger_count'])
plt.xlabel('year')
plt.ylabel('Passenger_count')
plt.title(' Scatterplot:Year vs Passenger_count')
plt.show()
```



In [34]:
```
df.to_csv('NEW_FILE.CSV')
```

In [ ]:
```
```