In [4]:
```
import pandas as pd
```

In [5]:
```
import numpy as np
```

In [6]:
```
df=pd.read_csv(r"C:\Users\Rutu\Documents\uber - uber.csv")
```

In [7]:
```
df.head()
```

Out[7]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	drop
0	24238194	2015-05-07 19:52:06	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	
1	27835199	2009-07-17 20:04:56	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	
2	44984355	2009-08-24 21:45:00	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	
3	25894730	2009-06-26 8:22:21	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	
4	17610152	2014-08-28 17:47:00	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	

In [8]:
```
df.tail()
```

Out[8]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude
199995	42598914	2012-10-28 10:49:00	3.0	2012-10-28 10:49:00 UTC	-73.987042	40.739367	-73.986525
199996	16382965	2014-03-14 1:09:00	7.5	2014-03-14 01:09:00 UTC	-73.984722	40.736837	-74.006672
199997	27804658	2009-06-29 0:42:00	30.9	2009-06-29 00:42:00 UTC	-73.986017	40.756487	-73.858957
199998	20259894	2015-05-20 14:56:25	14.5	2015-05-20 14:56:25 UTC	-73.997124	40.725452	-73.983215
199999	11951496	2010-05-15 4:08:00	14.1	2010-05-15 04:08:00 UTC	-73.984395	40.720077	-73.985508

In [9]:
```
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 [10]:
```
df.shape
```

Out[10]:
```
(200000, 9)
```

In [11]:
```
df.describe()
```

Out[11]:

	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passe
count	2.000000e+05	200000.000000	200000.000000	200000.000000	199999.000000	199999.000000	200
mean	2.771250e+07	11.359955	-72.527638	39.935885	-72.525292	39.923890	
std	1.601382e+07	9.901776	11.437787	7.720539	13.117408	6.794829	
min	1.000000e+00	-52.000000	-1340.648410	-74.015515	-3356.666300	-881.985513	
25%	1.382535e+07	6.000000	-73.992065	40.734796	-73.991407	40.733823	
50%	2.774550e+07	8.500000	-73.981823	40.752592	-73.980093	40.753042	
75%	4.155530e+07	12.500000	-73.967153	40.767158	-73.963659	40.768001	
max	5.542357e+07	499.000000	57.418457	1644.421482	1153.572603	872.697628	

In [12]:
```
df.corr()
```

Out[12]:

	Unnamed: 0	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
Unnamed: 0	1.000000	0.000589	0.000230	-0.000341	0.000270	0.000271
fare_amount	0.000589	1.000000	0.010457	-0.008481	0.008986	-0.011014
pickup_longitude	0.000230	0.010457	1.000000	-0.816461	0.833026	-0.846324
pickup_latitude	-0.000341	-0.008481	-0.816461	1.000000	-0.774787	0.702367
dropoff_longitude	0.000270	0.008986	0.833026	-0.774787	1.000000	-0.917010
dropoff_latitude	0.000271	-0.011014	-0.846324	0.702367	-0.917010	1.000000
passenger_count	0.002257	0.010150	-0.000414	-0.001560	0.000033	-0.000659

In [13]:
```
df.dropna(inplace=True)
```

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    0
```

```
dropoff_latitude     0
passenger_count      0
dtype: int64
```

In [15]:
```
df.nunique().sort_values()
```

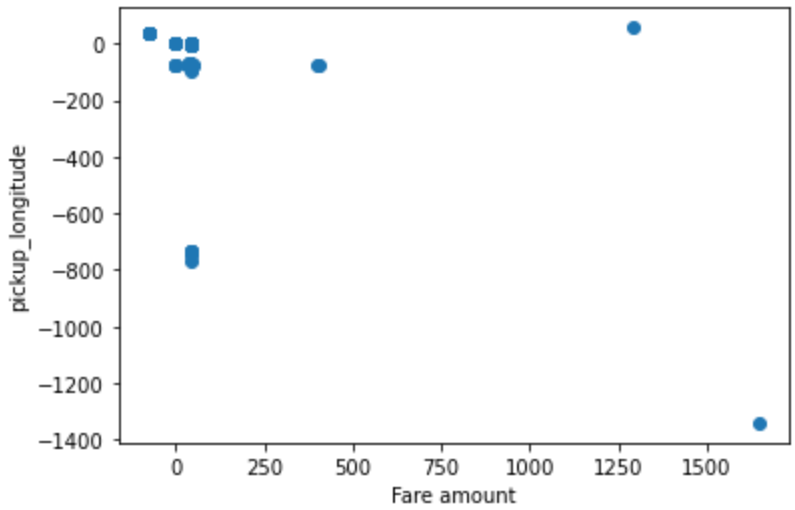
Out[15]:
```
passenger_count            8
fare_amount             1240
pickup_longitude       71013
dropoff_longitude      76836
pickup_latitude        83773
dropoff_latitude       90526
key                   196628
pickup_datetime       196628
Unnamed: 0            199999
dtype: int64
```

In [16]:
```
df[["passenger_count"]].value_counts()
```

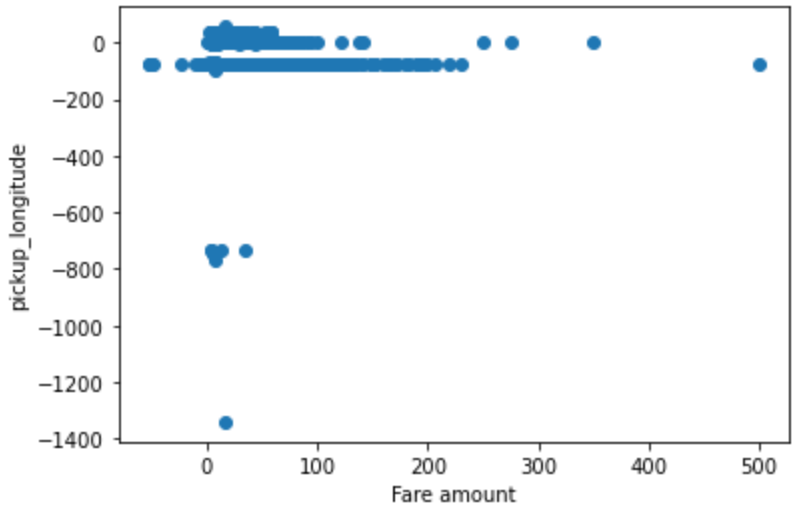
Out[16]:
```
passenger_count
1                138425
2                 29428
5                 14009
3                  8881
4                  4276
6                  4271
0                   708
208                   1
dtype: int64
```

In [17]:
```
import matplotlib.pyplot as plt
```

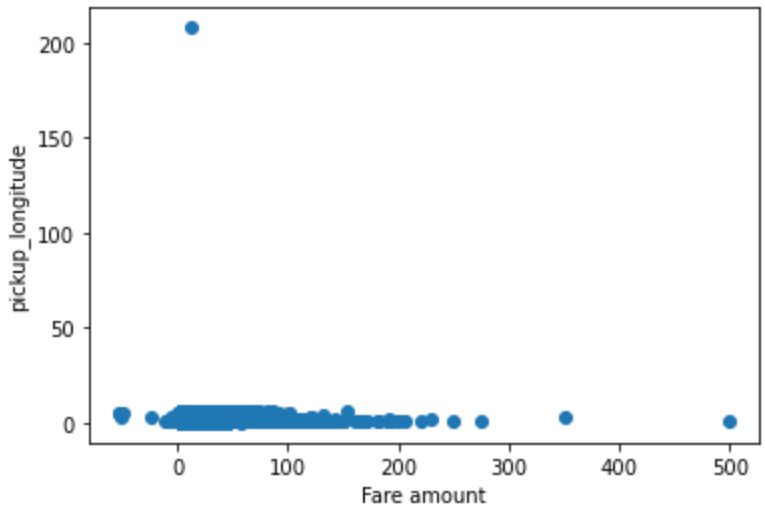
In [18]:
```
x = df['pickup_latitude']
y = df['pickup_longitude']
plt.scatter(x,y)
plt.xlabel("Fare amount")
plt.ylabel("pickup_longitude")
plt.show()
```



In [19]:
```
x = df['fare_amount']
y = df['pickup_longitude']
plt.scatter(x,y)
plt.xlabel("Fare amount")
plt.ylabel("pickup_longitude")
plt.show()
```



In [20]:
```
x = df['fare_amount']
y = df['passenger_count']
plt.scatter(x,y)
plt.xlabel("Fare amount")
plt.ylabel("pickup_longitude")
plt.show()
```



In [21]:
```
counter = 0
rs,cs = df.shape

df.drop_duplicates(inplace=True)
df.drop(['pickup_latitude','pickup_longitude',
         'dropoff_latitude','dropoff_longitude'],axis=1)
```

Out[21]:

	Unnamed: 0	key	fare_amount	pickup_datetime	passenger_count
0	24238194	2015-05-07 19:52:06	7.5	2015-05-07 19:52:06 UTC	1
1	27835199	2009-07-17 20:04:56	7.7	2009-07-17 20:04:56 UTC	1
2	44984355	2009-08-24 21:45:00	12.9	2009-08-24 21:45:00 UTC	1
3	25894730	2009-06-26 8:22:21	5.3	2009-06-26 08:22:21 UTC	3
4	17610152	2014-08-28 17:47:00	16.0	2014-08-28 17:47:00 UTC	5
...	...	...	...	...	...
199995	42598914	2012-10-28 10:49:00	3.0	2012-10-28 10:49:00 UTC	1
199996	16382965	2014-03-14 1:09:00	7.5	2014-03-14 01:09:00 UTC	1
199997	27804658	2009-06-29 0:42:00	30.9	2009-06-29 00:42:00 UTC	2

199998	20259894	2015-05-20 14:56:25	14.5	2015-05-20 14:56:25 UTC	1
199999	11951496	2010-05-15 4:08:00	14.1	2010-05-15 04:08:00 UTC	1

199999 rows × 5 columns

In [28]:
```
y=df['fare_amount']
```

In [29]:
```
y
```

Out[29]:
```
0          7.5
1          7.7
2         12.9
3          5.3
4         16.0
          ...
199995     3.0
199996     7.5
199997    30.9
199998    14.5
199999    14.1
Name: fare_amount, Length: 199999, dtype: float64
```

In [30]:
```
x=df[['pickup_longitude','pickup_latitude','dropoff_longitude','dropoff_latitude']]
```

In [31]:
```
x.shape
```

Out[31]:
```
(199999, 4)
```

In [32]:
```
x
```

Out[32]:

	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
0	-73.999817	40.738354	-73.999512	40.723217
1	-73.994355	40.728225	-73.994710	40.750325
2	-74.005043	40.740770	-73.962565	40.772647
3	-73.976124	40.790844	-73.965316	40.803349
4	-73.925023	40.744085	-73.973082	40.761247
...	...	...	...	...
199995	-73.987042	40.739367	-73.986525	40.740297
199996	-73.984722	40.736837	-74.006672	40.739620
199997	-73.986017	40.756487	-73.858957	40.692588
199998	-73.997124	40.725452	-73.983215	40.695416
199999	-73.984395	40.720077	-73.985508	40.768793

199999 rows × 4 columns

In [33]:
```
from sklearn.model_selection import train_test_split
```

In [34]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=2529)
```

In [35]:
```
x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

Out[35]:
```
((139999, 4), (60000, 4), (139999,), (60000,))
```

In [36]:
```
from sklearn.linear_model import LinearRegression
```

In [37]:
```
lr=LinearRegression()
```

In [38]:
```
lr.fit(x_train,y_train)
```

Out[38]:
```
LinearRegression()
```

In [39]:
```
y_pred=lr.predict(x_test)
```

In [40]:
```
y_pred.shape
```

Out[40]:
```
(60000,)
```

In [41]:
```
 y_pred
```

Out[41]:
```
array([11.34368065, 11.34500256, 11.3425004 , ..., 11.3441611 ,
       11.34398834, 11.34360522])
```

In [50]:
```
from sklearn.metrics import mean_squared_error,r2_score,mean_absolute_error
```

In [51]:
```
r2_score(y_test, y_pred)
```

Out[51]:
```
-8.925778634871762e-05
```

In [52]:
```
mean_absolute_error(y_test, y_pred)
```

Out[52]:
```
6.025215839696142
```

In [53]:
```
mean_squared_error(y_test, y_pred)
```

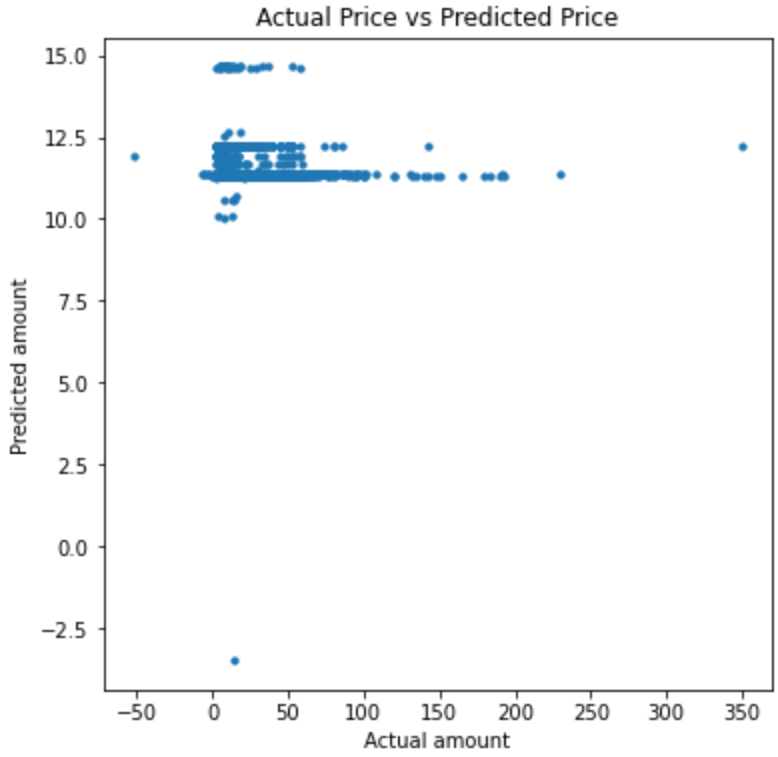
Out[53]:
```
96.85382491834356
```

In [54]:
```
mean_squared_error(y_test, y_pred)
```

Out[54]:
```
96.85382491834356
```

In [70]:
```
import matplotlib.pyplot as plt

fig,(ax) = plt.subplots(1, figsize = (6,6))
ax.scatter(y_test,y_pred,s=10)
plt.xlabel('Actual amount')
plt.ylabel('Predicted amount')
plt.title("Actual Price vs Predicted Price")
plt.show()
```


Actual Price vs Predicted Price
Predicted amount
Actual amount
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−50
0
50
100
150
200
250
300
350

In [ ]: