KMeans

In [42]:
```
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
%matplotlib inline
```

In [5]:
```
df = pd.read_csv("C:/Users/bhava/OneDrive/Pictures/Desktop/GPTC/5th SEM/EXC
df.head()
```

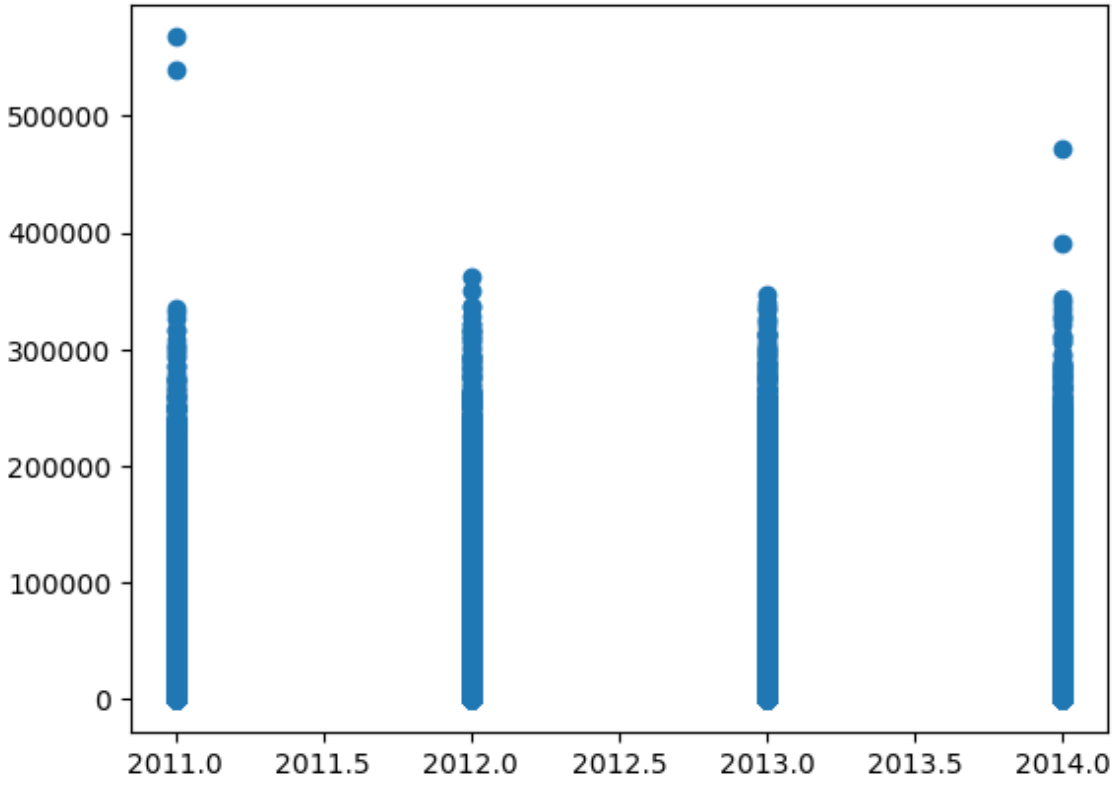
```
C:\Users\bhava\AppData\Local\Temp\ipykernel_24984\583893607.py:1: DtypeWar
ning: Columns (3,4,5,6,12) have mixed types. Specify dtype option on impor
t or set low_memory=False.
  df = pd.read_csv("C:/Users/bhava/OneDrive/Pictures/Desktop/GPTC/5th SEM/
EXCEL/Salaries.csv")
```

Out[5]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalP
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373

In [7]:
```
plt.scatter(df["Year"],df["TotalPay"])
```

Out[7]:
```
<matplotlib.collections.PathCollection at 0x27a1dd064f0>
```



In [9]:
```
km = KMeans(n_clusters=3)
km
```

Out[9]:
```
KMeans(n_clusters=3)
```

In [11]:
```
y_predicted = km.fit_predict(df[["Year","TotalPay"]])
y_predicted
```

Out[11]:
```
array([2, 2, 2, ..., 1, 1, 1])
```

In [13]:
```
df["cluster"] = y_predicted
df.head()
```

Out[13]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalP
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373

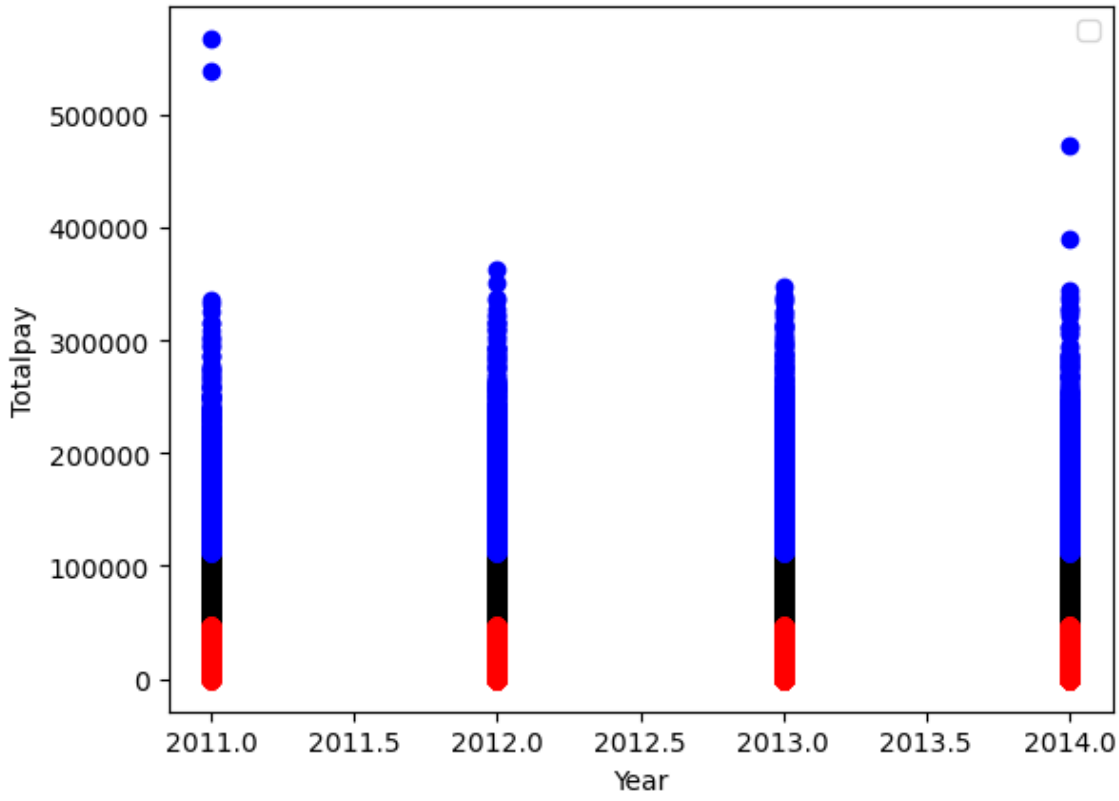
In [16]:
```
df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
df3 = df[df.cluster==2]

plt.scatter(df1.Year,df1["TotalPay"],color="black")
plt.scatter(df2.Year,df2["TotalPay"],color="red")
plt.scatter(df3.Year,df3["TotalPay"],color="blue")

plt.xlabel("Year")
plt.ylabel("Totalpay")
plt.legend()
```

```
No artists with labels found to put in legend.  Note that artists whose la
bel start with an underscore are ignored when legend() is called with no a
rgument.
```

Out[16]:
```
<matplotlib.legend.Legend at 0x27a1edbd8e0>
```



In [21]:
```
scaler = MinMaxScaler()
scaler.fit(df[["TotalPay"]])
df
```

Out[21]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefi
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	Na
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	Na
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	Na
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	Na
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.6	9737.0	182234.59	Na
...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.(
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	N Provide
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	N Provide
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	N Provide
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.(

148654 rows × 14 columns

In [24]:
```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[["Year","TotalPay"]])
y_predicted
```

Out[24]:
```
array([2, 2, 2, ..., 0, 0, 0])
```

In [25]:
```
df["cluster"]=y_predicted
df.drop("cluster",axis='columns',inplace=True)
df
```

Out[25]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefi
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	Na
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	Na
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	Na
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	Na
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.6	9737.0	182234.59	Na
...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.(
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	N Provide
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	N Provide
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	N Provide
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.(

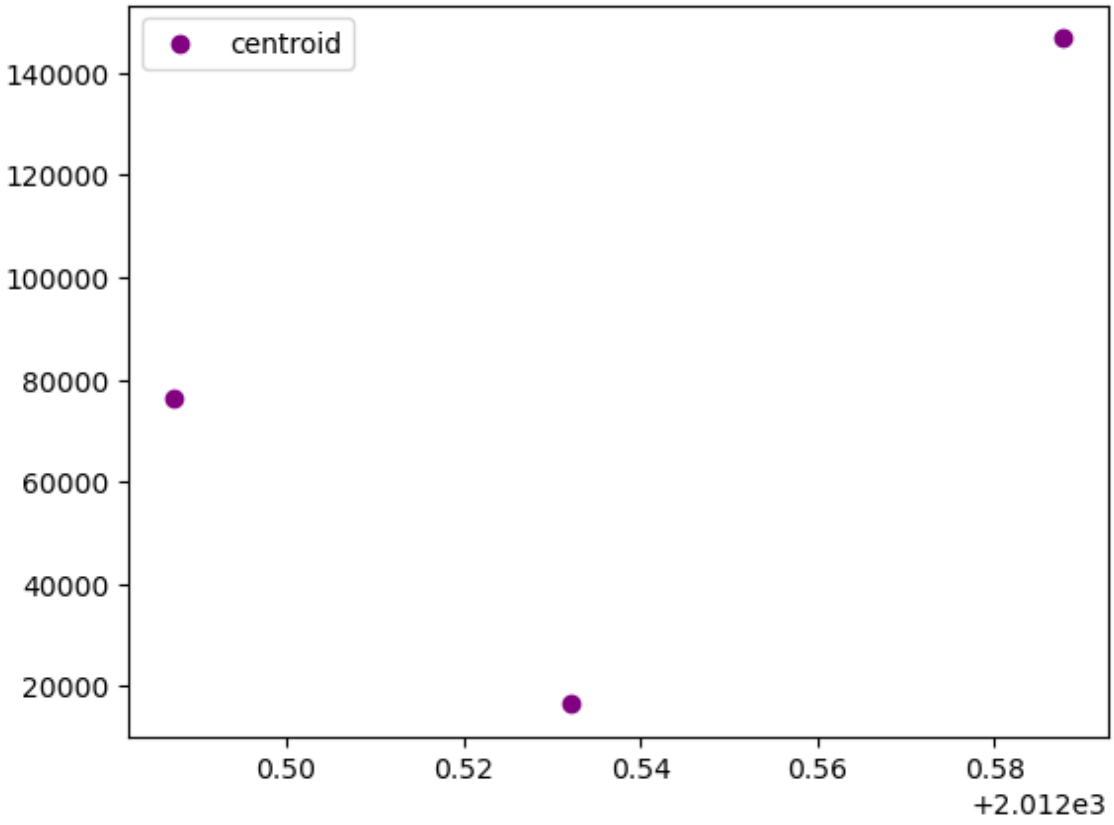
148654 rows × 13 columns

In [26]:
```
km.cluster_centers_
```

Out[26]:
```
array([[  2012.53211563,  16463.68324596],
       [  2012.48717878,  76419.92413688],
       [  2012.58783621, 146813.112225  ]])
```

In [36]:
```
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color='purple
plt.legend()
```

Out[36]:
```
<matplotlib.legend.Legend at 0x27a1ed79c40>
```



In [37]:
```
k_rng = range(1,10)
sse = []
for k in k_rng:
    km = KMeans (n_clusters=k)
    km.fit(df[['Year','TotalPay']])
    sse.append(km.inertia_)
```

In [39]:
```
sse
```

Out[39]:
```
[379357672806865.25,
 136540431582782.33,
 60741428052229.336,
 34437355815825.75,
 22801548834997.56,
 17018652555897.646,
 13032003189765.281,
 10127539539649.484,
 8096501218718.167]
```

In [40]:
```
plt.xlabel('K')
plt.ylabel('Sum of squared error')
plt.plot(k_rng,sse)
```

Out[40]:
```
[<matplotlib.lines.Line2D at 0x27a203e43d0>]
```