In [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

In [ ]:

```

```

In [3]:

```
df = pd.read_csv("Mall_Customers.csv")
```

In [4]:

```
df.head()
```

Out[4]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

In [5]:

```
df.shape
```

Out[5]:

```
(200, 5)
```

In [6]:

```
encode = pd.get_dummies(df['Gender'],drop_first=True,prefix='Gender',dtype='int8')
```

In [7]:

```
df = pd.concat([df,encode],axis=1)
```

In [8]:

```
df
```

Out[8]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Gender_Male
0	1	Male	19	15	39	1
1	2	Male	21	15	81	1
2	3	Female	20	16	6	0
3	4	Female	23	16	77	0
4	5	Female	31	17	40	0
...	...	...	...	...	...	...
195	196	Female	35	120	79	0
196	197	Female	45	126	28	0
197	198	Male	32	126	74	1
198	199	Male	32	137	18	1
199	200	Male	30	137	83	1

200 rows × 6 columns

In [9]:

```
df.drop(['Gender'],axis=1,inplace=True)
```

In [10]:

```
df.rename(columns={'Gender_Male':'Gender'},inplace=True)
```

In [ ]:

In [11]:

```
df.head()
```

Out[11]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)	Gender
0	1	19	15	39	1
1	2	21	15	81	1
2	3	20	16	6	0
3	4	23	16	77	0
4	5	31	17	40	0

In [12]:

```
df.dtypes
```

Out[12]:

```
CustomerID                 int64
Age                        int64
Annual Income (k$)         int64
Spending Score (1-100)     int64
Gender                      int8
dtype: object
```

In [13]:

```
df.isnull().sum()
```

Out[13]:

```
CustomerID                 0
Age                        0
Annual Income (k$)         0
Spending Score (1-100)     0
Gender                     0
dtype: int64
```

In [14]:

```
df.drop(["CustomerID"],axis=1, inplace=True)
```

In [ ]:

```
df
```

In [ ]:

```

```

In [ ]:

```

```

#################### VISUALIZATION #########################

In [ ]:

```

```

In [15]:

```
data= "Mall_Customers.csv"
```

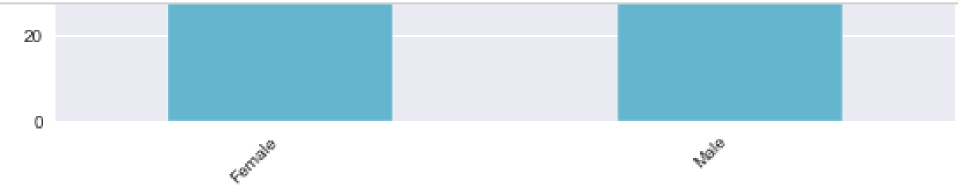
In [16]:

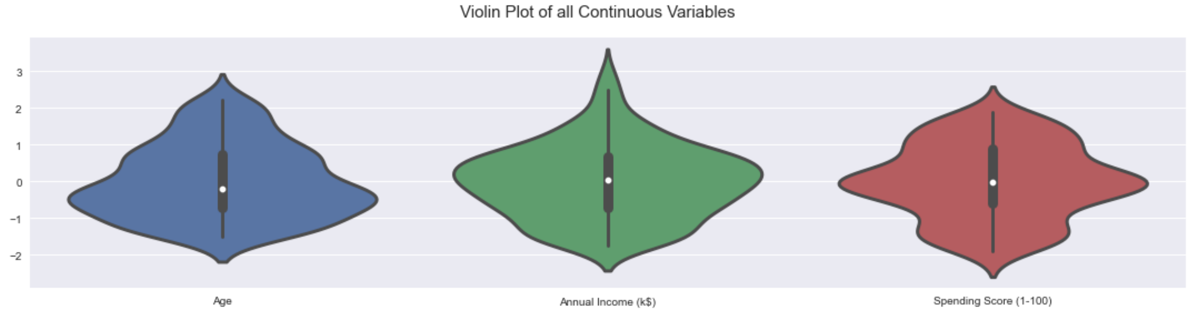
```
from autoviz.AutoViz_Class import AutoViz_Class
av = AutoViz_Class()
```

```
Imported AutoViz_Class version: 0.0.81. Call using:
    from autoviz.AutoViz_Class import AutoViz_Class
    AV = AutoViz_Class()
    AV.AutoViz(filename, sep=',', depVar='', dfte=None, header=0, verb
ose=0,
                            lowess=False,chart_format='svg',max_rows_a
nalyzed=150000,max_cols_analyzed=30)
Note: verbose=0 or 1 generates charts and displays them in your local
Jupyter notebook.
      verbose=2 saves plots in your local machine under AutoViz_Plots
directory and does not display charts.
```

In [17]:

```
av.AutoViz(data,
           sep=",",
          depVar="",
          dfte=None,
          header=0,
          verbose=0,
          lowess=False,
          chart_format="svg",
          max_rows_analyzed=10000,
          max_cols_analyzed=10,)
```





In [ ]:

```
wcss
```

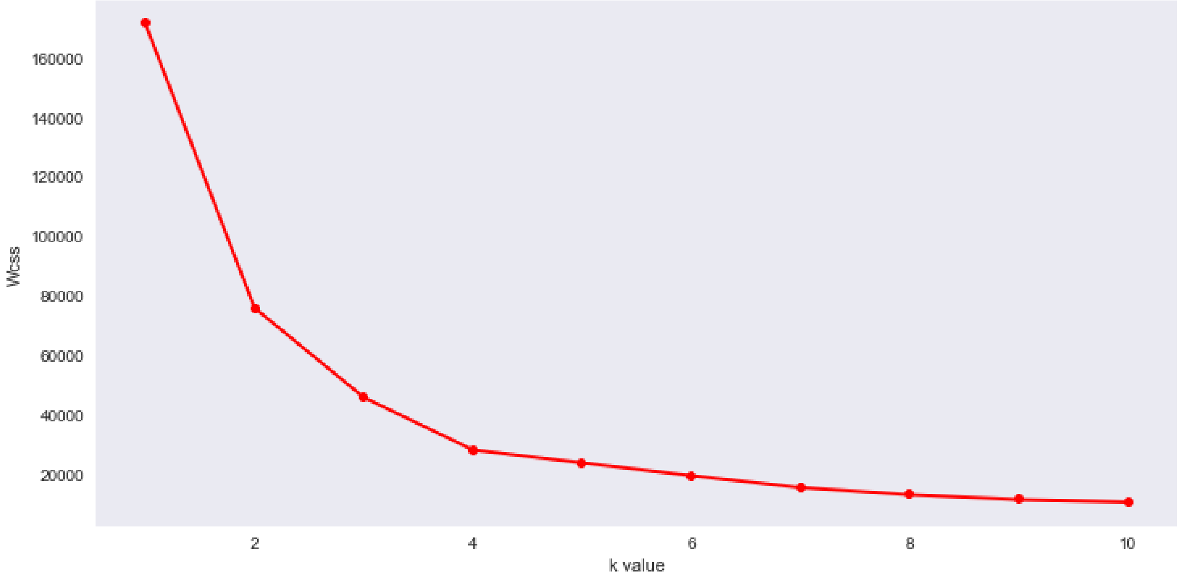
In [21]:

```
x1 = df.loc[:,['Age','Spending Score (1-100)']].values

wcss=[]

for k in range(1,11):
    kmeans = KMeans(n_clusters=k,init='k-means++')
    kmeans.fit(x1)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color = 'red' , marker = '8')
plt.xlabel('k value')
plt.ylabel('Wcss')
plt.show()
```



In [42]:

```
kmeans = KMeans(n_clusters=4)

label = kmeans.fit_predict(x1)

#print(label)
```

In [43]:

```
print(kmeans.cluster_centers_)
```

```
[[27.61702128 49.14893617]
 [30.1754386  82.35087719]
 [55.70833333 48.22916667]
 [43.29166667 15.02083333]]
```

In [44]:

```
kmeans.cluster_centers_[:,1]
```

Out[44]:

```
array([49.14893617, 82.35087719, 48.22916667, 15.02083333])
```

In [ ]:

```

```

In [45]:

```
kmeans.cluster_centers_[:,0]
```

Out[45]:

```
array([27.61702128, 30.1754386 , 55.70833333, 43.29166667])
```

In [46]:

```
kmeans.cluster_centers_[:,1]
```

Out[46]:

```
array([49.14893617, 82.35087719, 48.22916667, 15.02083333])
```

In [47]:

```
kmeans.cluster_centers_
```

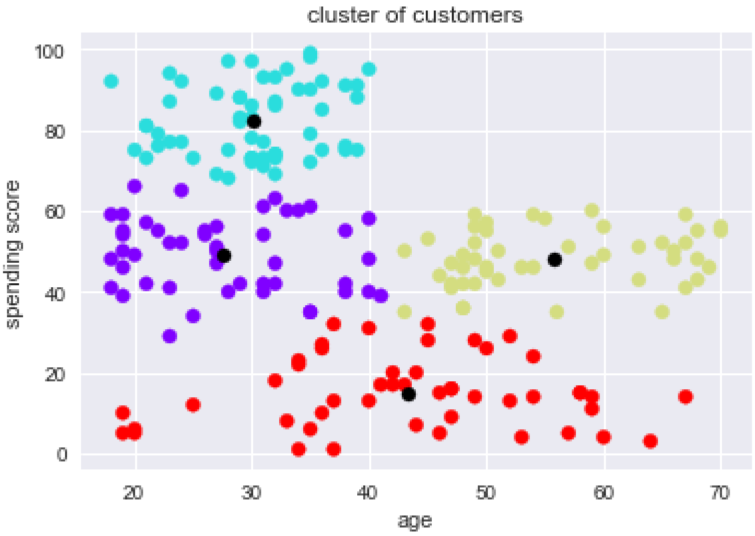
Out[47]:

```
array([[27.61702128, 49.14893617],
       [30.1754386 , 82.35087719],
       [55.70833333, 48.22916667],
       [43.29166667, 15.02083333]])
```

In [48]:

```
plt.scatter(x1[:,0],x1[:,1],c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='black')

plt.title("cluster of customers")
plt.xlabel("age")
plt.ylabel("spending score")
plt.show()
```



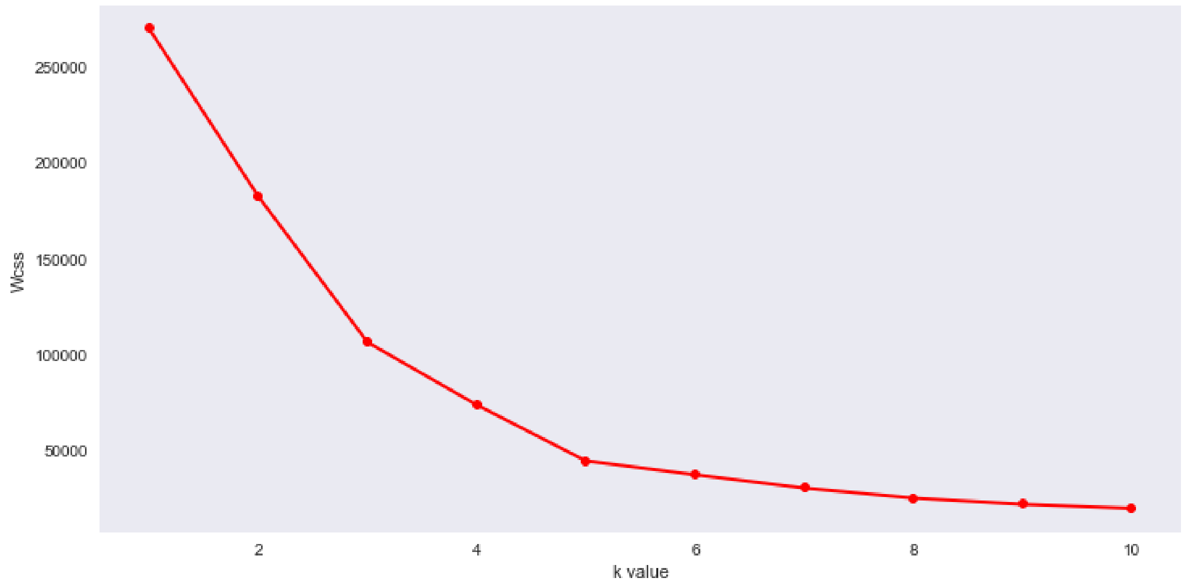
In [49]:

```
x2 = df.loc[:,['Annual Income (k$)','Spending Score (1-100)']].values

wcss=[]

for k in range(1,11):
    kmeans = KMeans(n_clusters=k,init='k-means++')
    kmeans.fit(x2)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color = 'red' , marker = '8')
plt.xlabel('k value')
plt.ylabel('Wcss')
plt.show()
```



In [55]:

```
(wcss)
```

Out[55]:

```
[269981.28000000014,
 182440.30762987016,
 106348.37306211119,
 73679.78903948837,
 44448.45544793369,
 37233.81451071002,
 30273.394312070028,
 25061.304119069322,
 21826.936303231643,
 19701.35225128174]
```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```

In [51]:

```
kmeans = KMeans(n_clusters=5)

label = kmeans.fit_predict(x2)

#print(label)
```

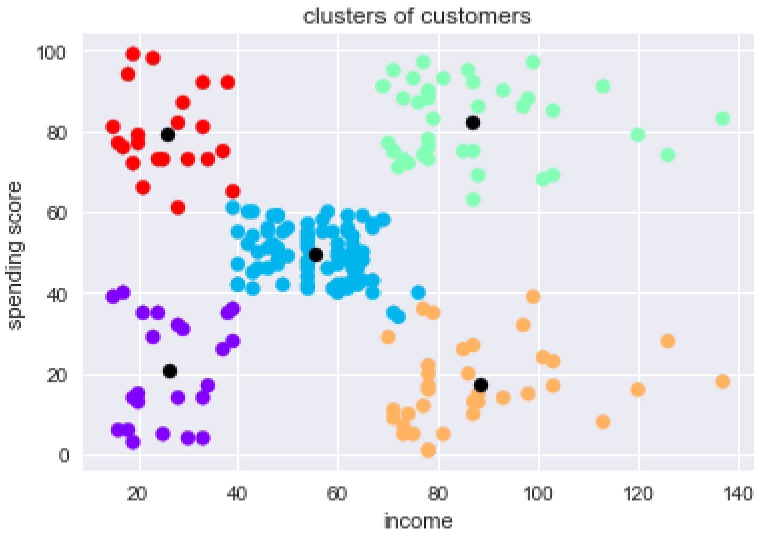
In [53]:

```
#label
```

In [54]:

```
plt.scatter(x2[:,0],x2[:,1],c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='black')

plt.title("clusters of customers")
plt.xlabel("income")
plt.ylabel("spending score")
plt.show()
```



In [57]:

```
df.head(3)
```

Out[57]:

	Age	Annual Income (k$)	Spending Score (1-100)	Gender
0	19	15	39	1
1	21	15	81	1
2	20	16	6	0

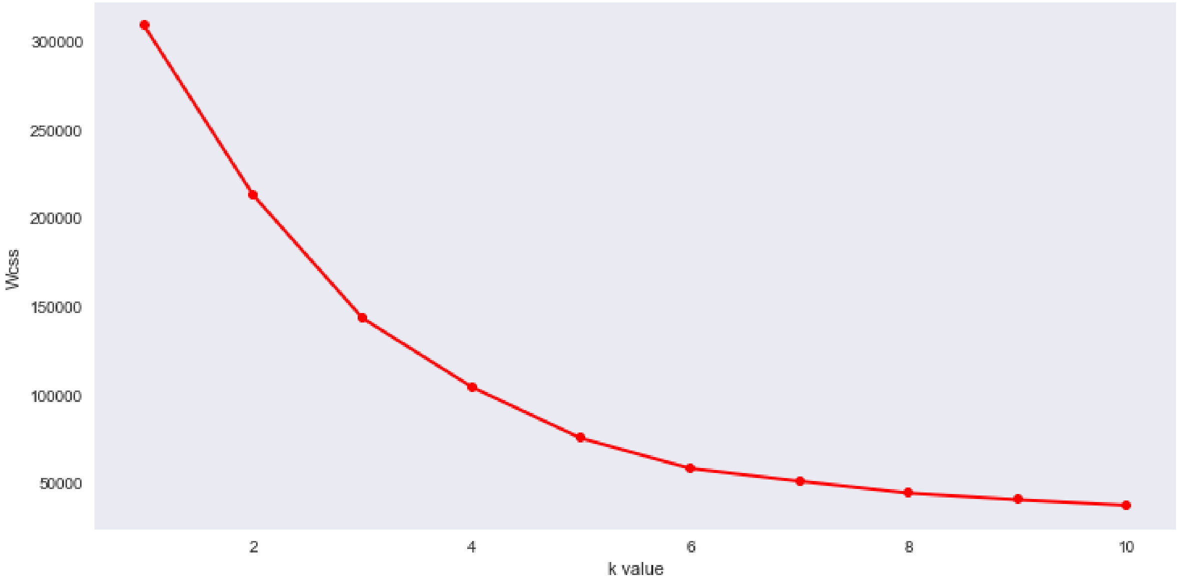
In [76]:

```
x3 = df

wcss=[]

for k in range(1,11):
    kmeans = KMeans(n_clusters=k,init='k-means++')
    kmeans.fit(x3)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color = 'red' , marker = '8')
plt.xlabel('k value')
plt.ylabel('Wcss')
plt.show()
```



In [77]:

```
x3
```

Out[77]:

	Age	Annual Income (k$)	Spending Score (1-100)	Gender
0	19	15	39	1
1	21	15	81	1
2	20	16	6	0
3	23	16	77	0
4	31	17	40	0
...	...	...	...	...
195	35	120	79	0
196	45	126	28	0
197	32	126	74	1
198	32	137	18	1
199	30	137	83	1

200 rows × 4 columns

In [79]:

```
kmeans = KMeans(n_clusters=6)

label = kmeans.fit_predict(x3)
```

In [80]:

```
#score(x3,label)


no_of_clusters = [2, 3, 4, 5, 6,7,8,9,10]

for n_clusters in no_of_clusters:

    cluster = KMeans(n_clusters = n_clusters)
    cluster_labels = cluster.fit_predict(x3)

    # The silhouette_score gives the
    # average value for all the samples.
    silhouette_avg = silhouette_score(x3, cluster_labels)

    print(n_clusters, silhouette_avg)
```

```
2 0.29307334005502633
3 0.383798873822341
4 0.40553486600451777
5 0.4440669204743008
6 0.45205475380756527
7 0.44096462877395787
8 0.4259878450877001
9 0.3884448555855653
10 0.38162205767837293
```

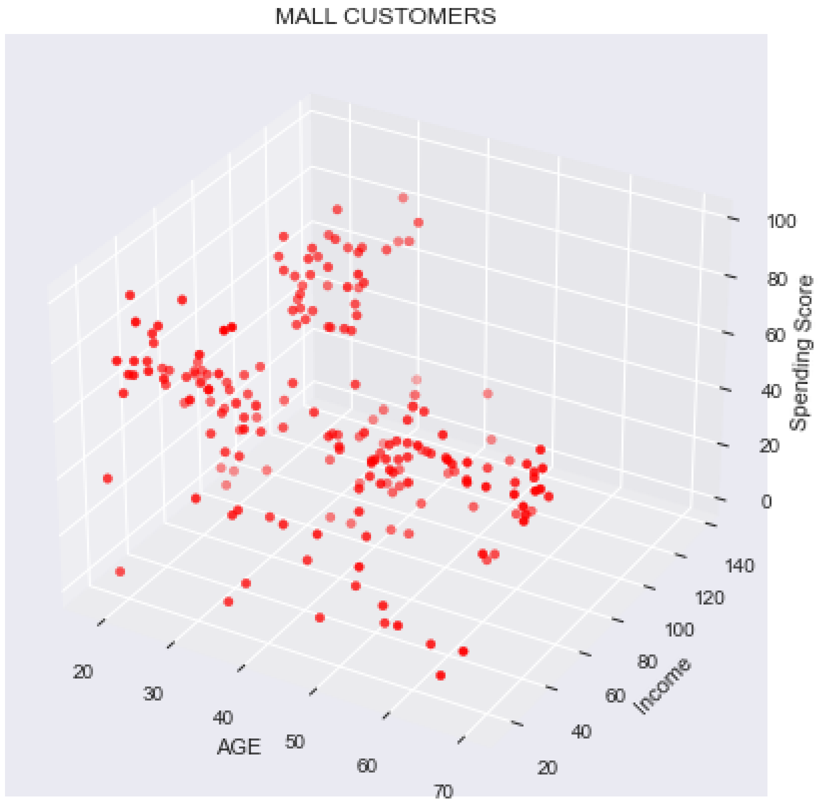
In [81]:

```
from mpl_toolkits import mplot3d


# Creating figure
fig = plt.figure(figsize = (10, 7))
ax = plt.axes(projection ="3d")

# Creating plot
ax.scatter3D(df['Age'], df['Annual Income (k$)'], df['Spending Score (1-100)'], colc
plt.title("MALL CUSTOMERS")

# show plot
plt.xlabel("AGE")
plt.ylabel("Income")
ax.set_zlabel("Spending Score")
plt.show()
```



In [ ]:

```
df
```

In [ ]:

```
x = df.columns
x
```

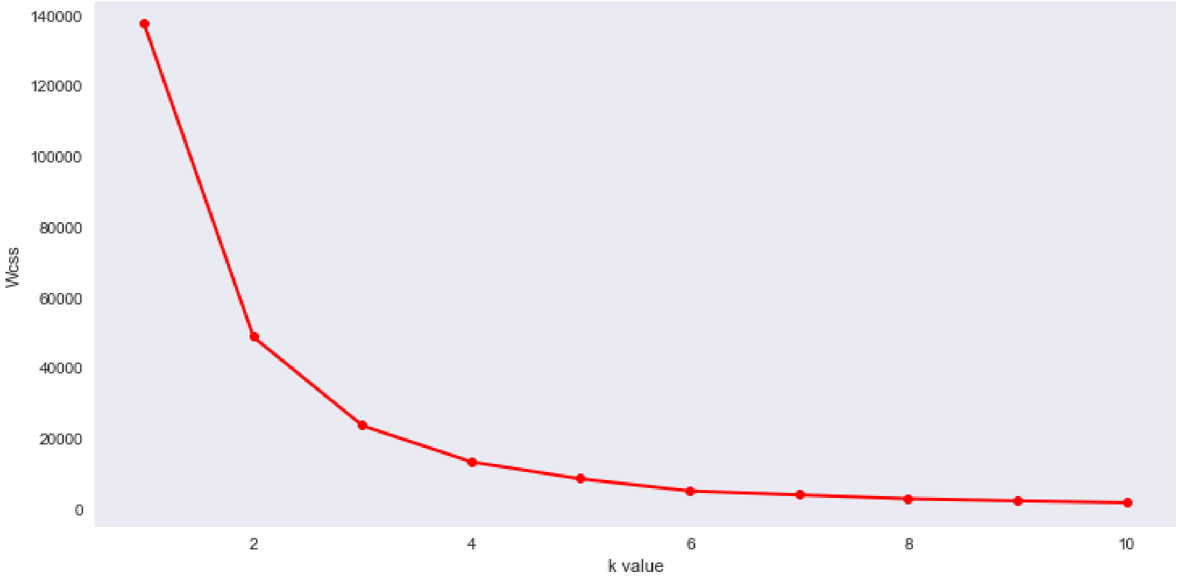
In [28]:

```
x1 = df.loc[:,['Gender','Annual Income (k$)']].values

wcss=[]

for k in range(1,11):
    kmeans = KMeans(n_clusters=k,init='k-means++')
    kmeans.fit(x1)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color = 'red' , marker = '8')
plt.xlabel('k value')
plt.ylabel('Wcss')
plt.show()
```



In [ ]:

```

```