In this notebook, I have picked up the *Wine Dataset* from Kaggle to explore the following-

- The top country with maximum wine production
- The province which produced the highest number of wines
- The best variety of wine in the specific province
- The price range of the wines in the same province and country

In [4]:

```
import pandas as pd
import numpy as np
```

The following are the 2 data sets, I used for analysi-

In [5]:

```
Wine_data1 = pd.read_csv('winemag-data_first150k.csv',sep=',')
Wine_data2 = pd.read_csv('winemag-data-130k-v2.csv',sep=',')
```

In the below command, I have appended the 1st dataset to the second to get the overall wine data combined in one dataset

In [6]:

```
Wine_data=Wine_data1.append(Wine_data2)
Wine_data.head()
```

Out[6]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	titl
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	NaN	NaN	Na
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	NaN	NaN	NaN	Na
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonoma	NaN	NaN	Na
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Willamette Valley	NaN	NaN	Na
4	4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol	NaN	NaN	NaN	Na

We can see some null values appearing in some fields, lets check how many columns have it.

In [7]:

```
Wine_data.isnull().any()
```

Out[7]:

```
Unnamed: 0                False
country                    True
description               False
designation                True
points                    False
price                      True
province                   True
region_1                   True
region_2                   True
taster_name                True
taster_twitter_handle      True
title                      True
variety                    True
winery                    False
dtype: bool
```

So there are null values in all columns accept 4. Let us get rid of them

In [8]:

```
Wine_data = Wine_data.dropna()
```

In [9]:

```
Wine_data.isnull().any()
```

Out[9]:

```
Unnamed: 0                False
country                   False
description               False
designation               False
points                    False
price                     False
province                  False
region_1                  False
region_2                  False
taster_name               False
taster_twitter_handle     False
title                     False
variety                   False
winery                    False
dtype: bool
```

Now we have eliminated the null values from our dataset!!!

In [10]:

```
uniq_countries=Wine_data.sort_values('country', ascending=False).drop_duplicates(['country'])
```

In [11]:

```
country_count=uniq_countries['country'].value_counts()
max_country=country_count.max()
```

In [12]:

```
print (country_count)
```

```
US    1
Name: country, dtype: int64
```

Here we see that the wine data from US is the highest.

We will use US's data and further dive deep into which region produces the more wine.

In [14]:

```
Top_provinces=Wine_data['province'].value_counts()
print (Top_provinces)
```

```
California    12900
Washington     5845
Oregon         3489
New York        153
Name: province, dtype: int64
```

We are able to find out that the top most place where Wine is produced the most is in the province of California!

In [80]:

```
Wine_data.columns
```

Out[80]:

```
Index(['Unnamed: 0', 'country', 'description', 'designation', 'points',
       'price', 'province', 'region_1', 'region_2', 'taster_name',
       'taster_twitter_handle', 'title', 'variety', 'winery'],
      dtype='object')
```

In [81]:

```
Wine_data.drop(Wine_data.columns[[0,2,7,8,9,10,11,13]], axis=1, inplace=True)
```

In [82]:

```
Wine_data.columns
```

Out[82]:

```
Index(['country', 'designation', 'points', 'price', 'province', 'variety'], dtype='object')
```

In [83]:

```
Most_variety=Wine_data['variety'].value_counts()
Most_variety.head()
```

Out[83]:

```
Pinot Noir            4788
Chardonnay            2407
Cabernet Sauvignon    2372
Red Blend             1803
Syrah                 1678
Name: variety, dtype: int64
```

The above findings show that the vast famous variety of these is the *'Pinot Noir'*

Now we will pick the top country and he province which contains the maximum data on wine and analyze the cost of the wines sold in them.

In [17]:

```
hist_province='California'
hist_country='US'

mask1=Wine_data['province'].str.contains(hist_province)
mask2=Wine_data['country'].str.contains(hist_country)
```

```
stage = Wine_data[mask1 & mask2]
```

The below shows the first 5 rows of Wine_data with the unwanted rows eliminated.

In [85]:

```
stage.head()
```

Out[85]:

	country	designation	points	price	province	variety
10	US	Mountain Cuvée	87	19.0	California	Cabernet Sauvignon
23	US	Signature Selection	87	22.0	California	Merlot
25	US	King Ridge Vineyard	87	69.0	California	Pinot Noir
60	US	Estate	86	100.0	California	Cabernet Sauvignon
64	US	Golden Horn	86	26.0	California	Sauvignon Blanc

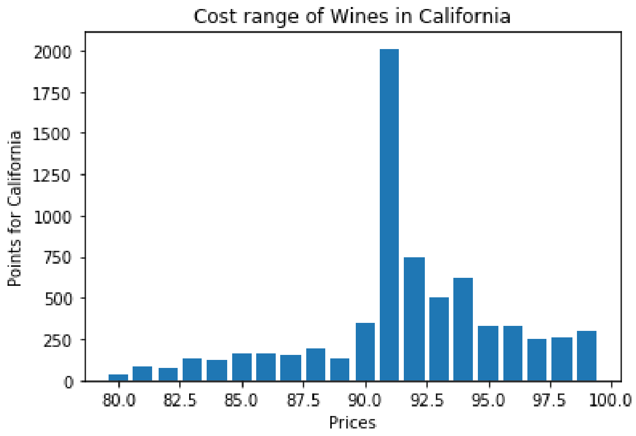
In [16]:

```
import matplotlib.pyplot as plt
```

In [19]:

```
Price_of_Wine=stage['price'].values
Points=stage['points'].values

plt.xlabel('Prices')
plt.ylabel('Points for California')
plt.title('Cost range of Wines in California')
plt.bar(Points,Price_of_Wine)
plt.show()
```



The above plot shows range of price against the points given to each designation in US, California

We see that at point 91 there is an outlier whose price is 2000 which is way beyond the 2nd costliest wine i.e. at 750

This gives us the cost range of wines produced in he California province of US!