task-01

September 1, 2024

0.1 DATA SCIENCE TASK 01

1 Visualisation of a categorical/continous data through BAR CHARTS / HISTOGRAMS

IMPORTING Pandas AND READING CSV FILE

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv(r'C:\Users\Windows\Desktop\projects\winemag-data_first150k.csv')
```

HAVING A QUICK VIEW OF THE DATA

[2]:
```
df.head()
```

[2]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	variety	winery
0	0	US	This tremendous 100% varietal wine hails from …	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz
1	1	Spain	Ripe aromas of fig, blackberry and cassis are …	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	NaN	Tinta de Toro	Bodega Carmen Rodríguez
2	2	US	Mac Watson honors the memory of a wine once ma…	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley
3	3	US	This spent 20 months in 30% new French oak, an…	Reserve	96	65.0	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi
4	4	France	This is the top wine from La Bégude, named aft…	La Brûlade	95	66.0	Provence	Bandol	NaN	Provence red blend	Domaine de la Bégude

```
1  Bodega Carmen Rodríguez
2                 Macauley
3                    Ponzi
4     Domaine de la Bégude
```

[3]:
```
df.shape
```

[3]:
```
(150930, 11)
```

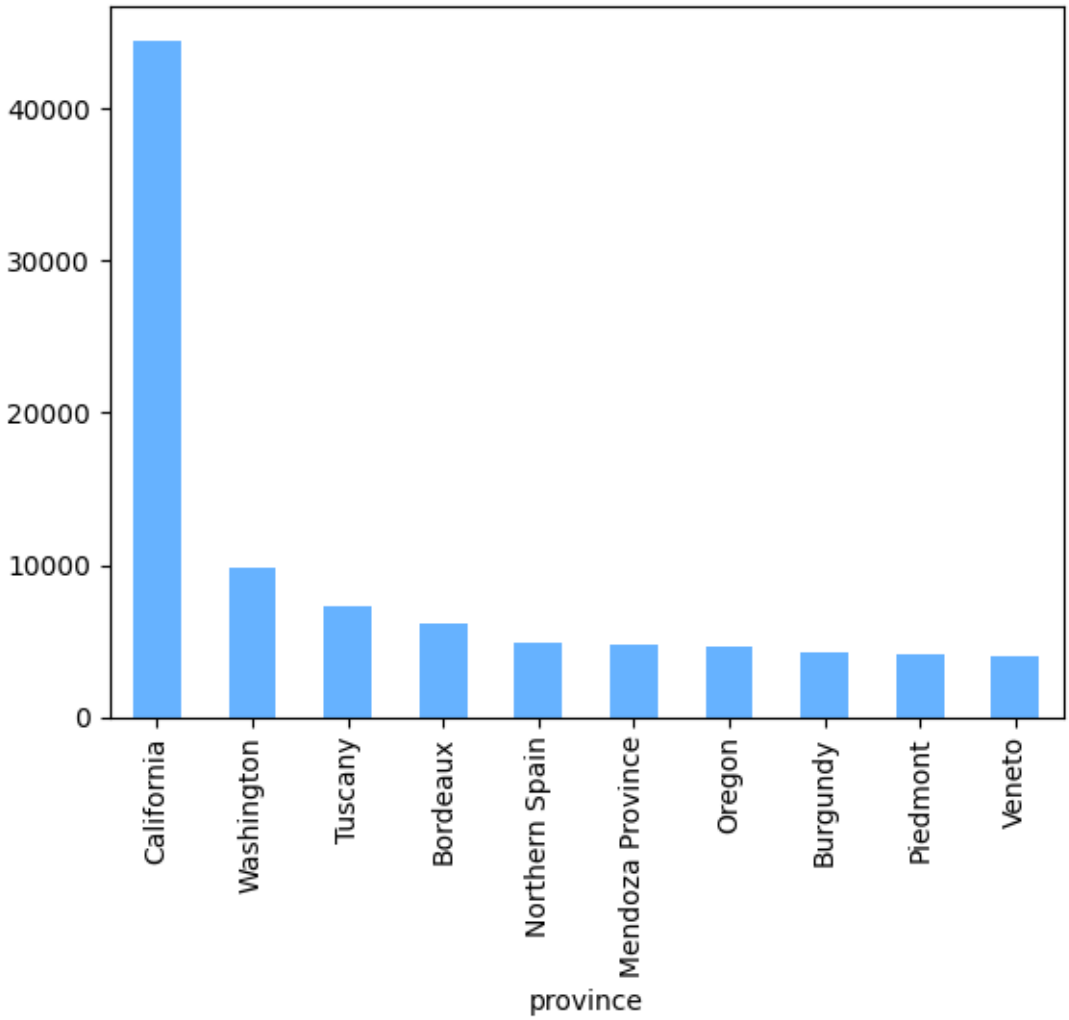
[4]:
```
df.dtypes
```

[4]:
```
Unnamed: 0       int64
country         object
description     object
designation     object
points           int64
price          float64
province        object
region_1        object
region_2        object
variety         object
winery          object
dtype: object
```

2 Bar charts and categorical data

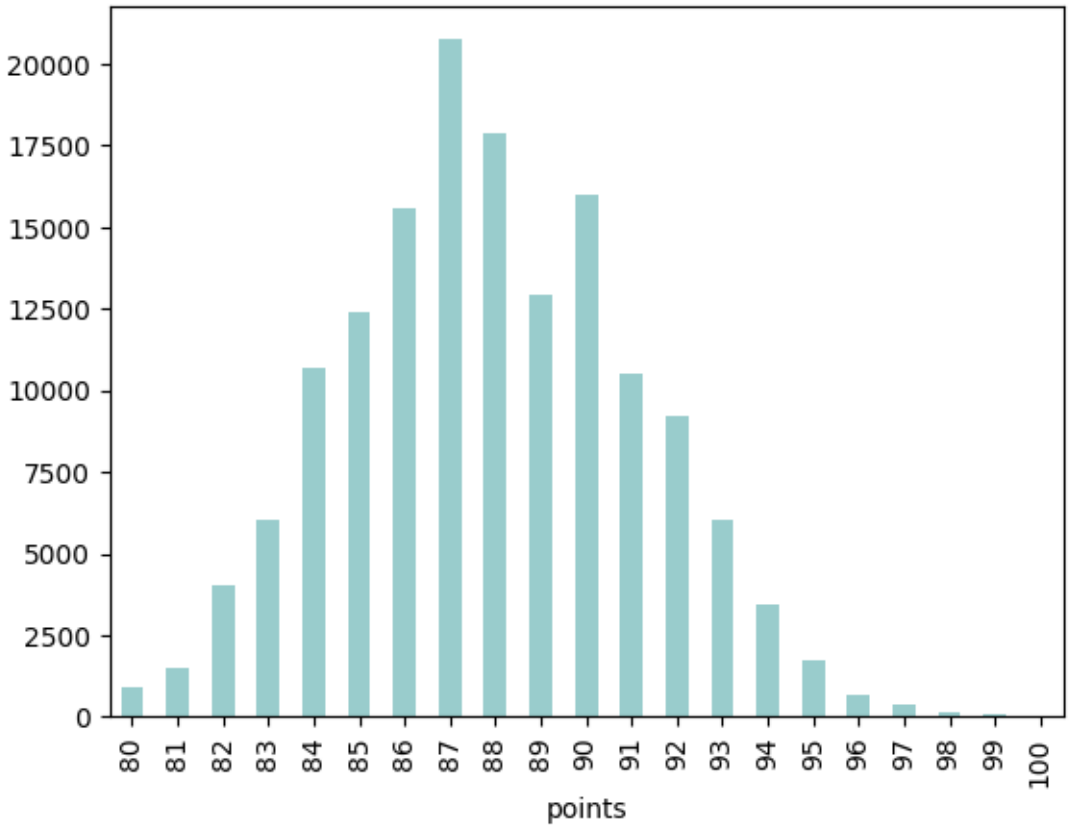
[5]:
```
df['province'].value_counts().head(10).plot(kind='bar', color='#66B2FF')
```

[5]:
```
<Axes: xlabel='province'>
```



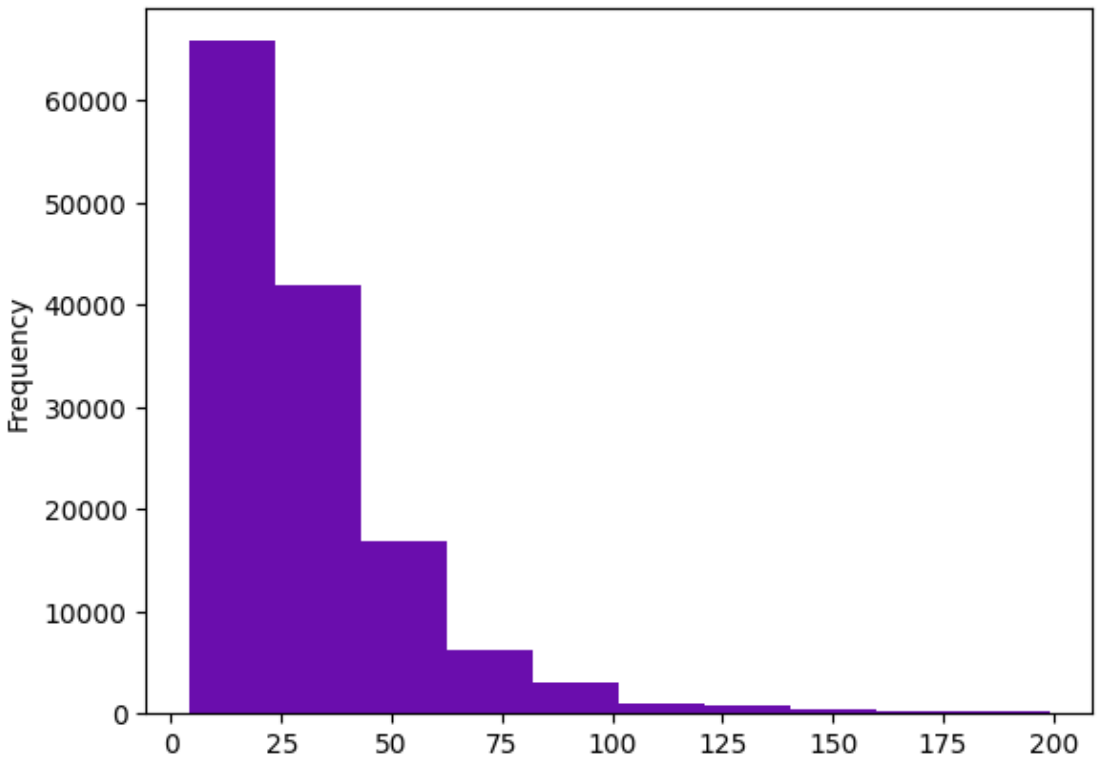
What does this plot tell us? It says California produces far more wine than any other province of the world! We might ask what percent of the total is Californian vintage? This bar chart tells us absolute numbers, but it's more useful to know relative proportions.

[6]:
```
df['points'].value_counts().sort_index().plot.bar(color='#99CCCC')
plt.show()
```



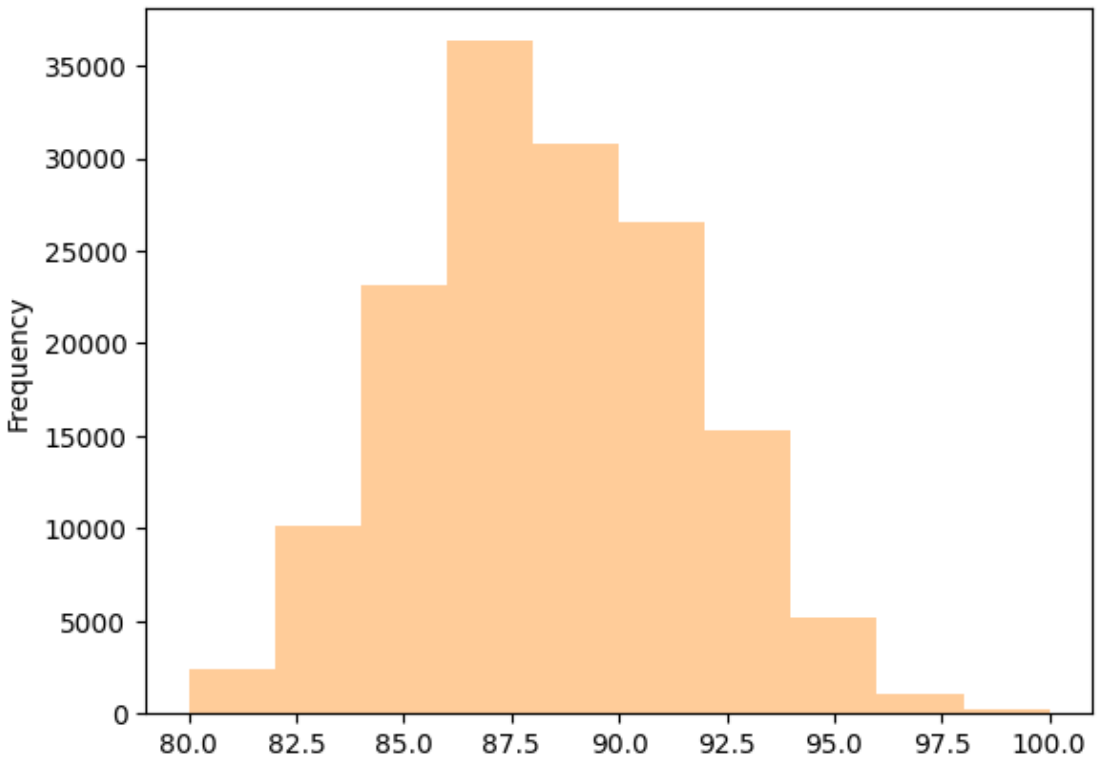
3 Histograms

[7]:
```
df[df['price'] < 200]['price'].plot.hist(color='#6A0DAD')
plt.show()
```



[8]:
```
df['points'].plot.hist(color='#FFCC99')
```

[8]:
```
<Axes: ylabel='Frequency'>
```



Histograms work best for interval variables without skew. They also work really well for ordinal categorical variables like points

4 THANK YOU !!!

[ ]: