Decoding ml task 1

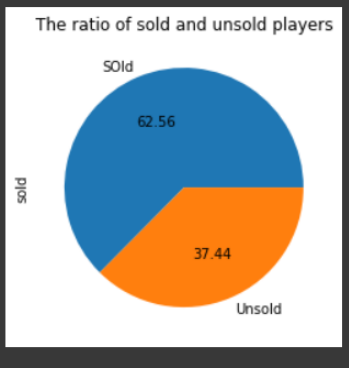
Name: Pasumarthy Sri Vishnu Vardhan

EDA

Dataset: IPL auction dataset

1) Importing necessary packages numpy, pandas, seaborn, matplotlibpyplot, warnings
2) Reading the dataset using pandas
3) Understanding the dataset
   dataset.head() : it gives you first 5 rows in the dataset
   dataset.tail() : it gives you last 5 rows from the dataset
   dataset.columns: it returns you column names
   dataset.info(): it returns you null count and datatype of each column
   I observed there are null values In Team column and other columns
4) By observing null values in dataset I split the dataset into sold players dataset and unsold players dataset

```
[56] dataset["sold"].value_counts().plot(kind="pie" ,labels=["SOld", "Unsold"], autopct="%.2f")
     plt.title("The ratio of sold and unsold players")
     plt.show()
```

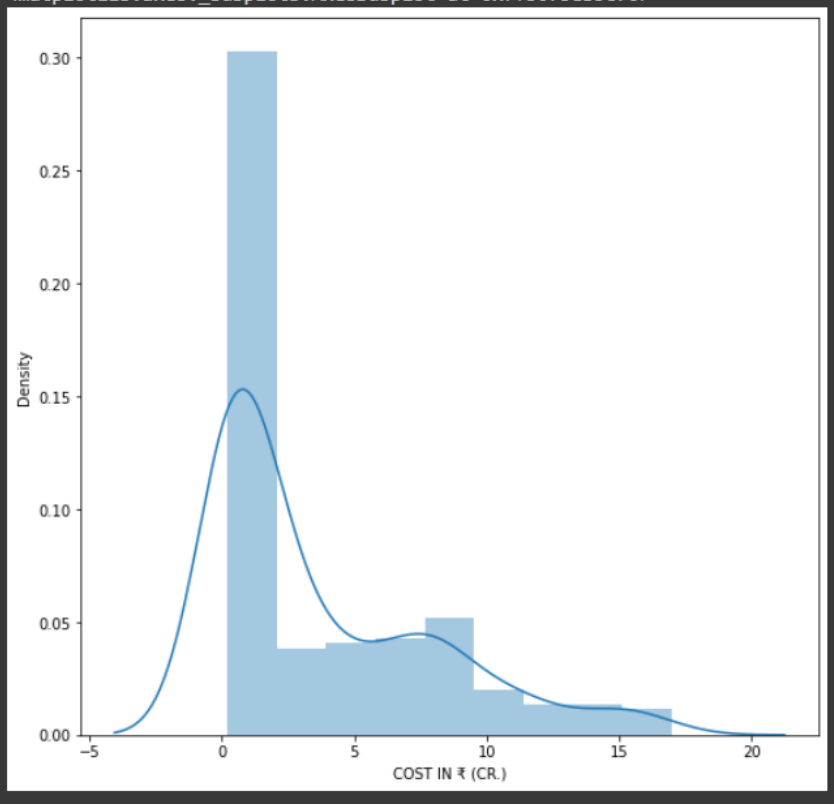


The above pie chart describes the percentage of sold and unsold players

5) Now I started exploratory data analysis on sold players
   a. Let the dataset name for sold players be dataset_sold
   b. Then I observed the price of each player starting from 0.2cr to 15 Crs
   c. So I decided to plot a dist plot

```
plt.figure(figsize=(9,9))
sns.distplot(dataset_sold["COST IN ₹ (CR.)"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3073eb5c70>
```

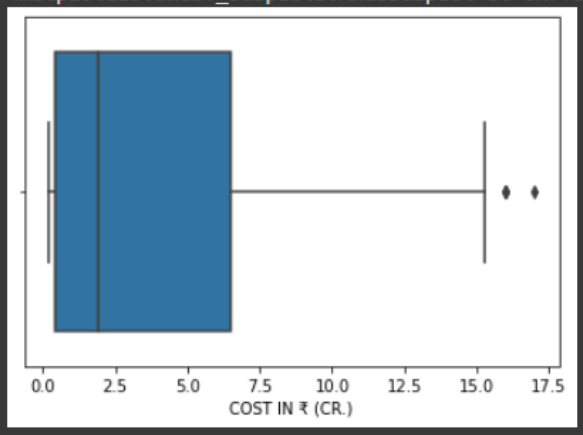


So the most of the players price at 0.2-3 cr

d. Then I tried to calculate mean or average of each price of player, before going to checked for outliers

```
[64] sns.boxplot(dataset_sold["COST IN ₹ (CR.)"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3073e07850>
```



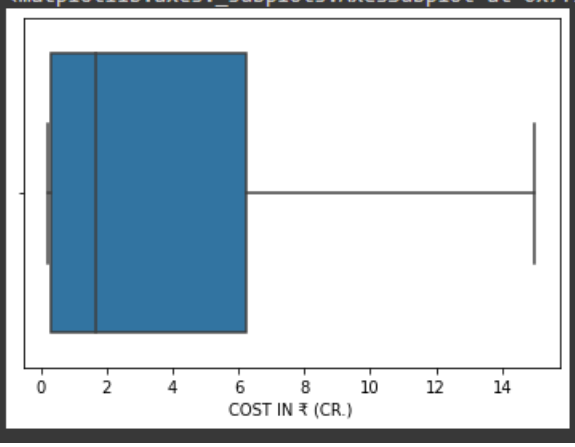
so I found an outliers and I removed them and calculated mean

So the average price of each player is 3.42 crs

```
[65] new_dataset=dataset_sold[dataset_sold["COST IN ₹ (CR.)"]<=15]

[66] sns.boxplot(new_dataset["COST IN ₹ (CR.)"])

<matplotlib.axes._subplots.AxesSubplot at 0x7f3073d8be80>
```



e. The above chart is out come After removing outlier

f. Dataset[‘COST IN $(CR.)”].describe() : it returns a five tuple summary of the dataset or row

```
[67] new_dataset["COST IN ₹ (CR.)"].describe()

count    232.000000
mean       3.429526
std        3.868938
min        0.200000
25%        0.300000
50%        1.650000
75%        6.250000
max       15.000000
Name: COST IN ₹ (CR.), dtype: float64
```

From the above plot we can say that

mean of the player price is 3.42cr

minimum price of player is 0.2cr

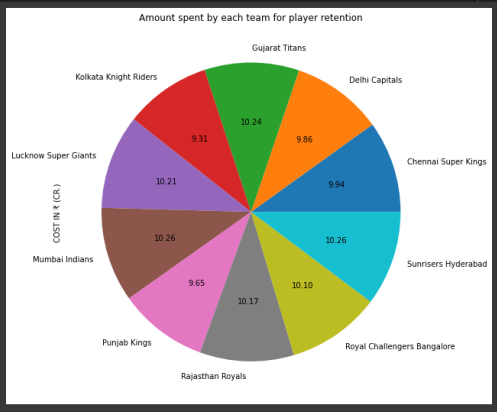
25% percentile of the player's price is 0.3cr

50% percentile of the player's price is 1.65cr

75% percentile of the player's price is 6.25cr

g. Amount spent by each team for player retention

```
# data_budget= data_team_budget.groupby(data_team_budget["Team"])
piechart=data_team_budget.groupby("Team").sum()["COST IN ₹ (CR.)"].plot( kind="pie",autopct="%.2f", figsize=(9,9), title="Amount spent by each team for player retention")
plt.show()
```



Inference from the above pie chart

Mumbai Indians and sunrisers Hyderabad has highest budget : 10.26 cr

Kolkata knight rider has least budget : 9.31cr

to get this above pie chart : First I grouped by teams column

h. Next, I tried to figure out number of players in each type in a team

Type of players

```
[71] bar_plot=pd.crosstab(dataset_sold["Team"],dataset_sold["TYPE"])
```

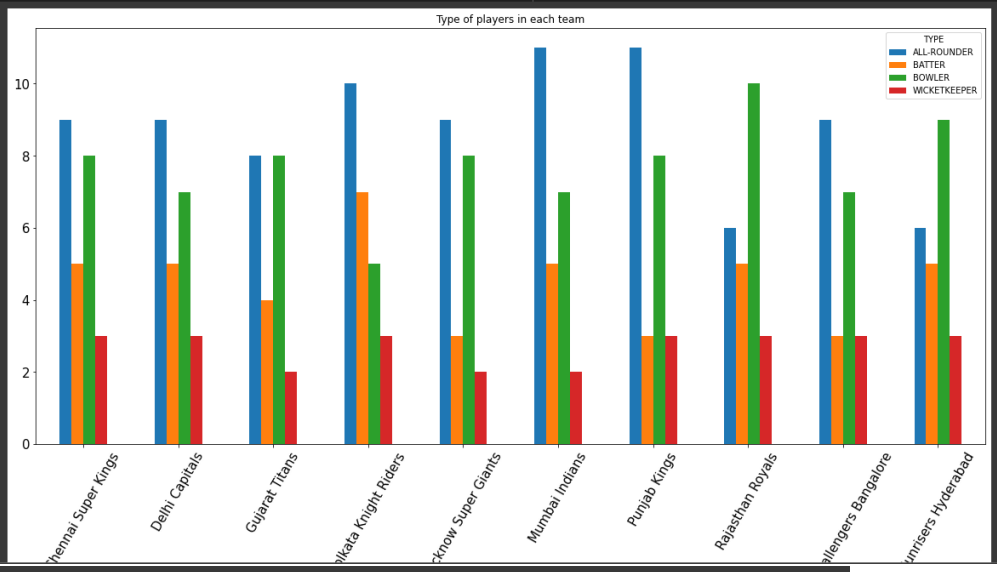
```
[72] bar_plot.head()
```

TYPE	ALL-ROUNDER	BATTER	BOWLER	WICKETKEEPER
Team				
Chennai Super Kings	9	5	8	3
Delhi Capitals	9	5	7	3
Gujarat Titans	8	4	8	2
Kolkata Knight Riders	10	7	5	3
Lucknow Super Giants	9	3	8	2

crosstab(“column1”,”column2”) : A crosstab is a table showing the relationship between two or more variables. Where the table only shows the relationship between two categorical variables, a crosstab is also known as a contingency table.

```
bar_plot.plot(kind="bar", figsize=(20,9), fontsize=15,rot=60,title="Type of players in each team")

plt.show()
```



From the above chart we observe that

1)Mumbai indians and punjab kings have more all rounder players

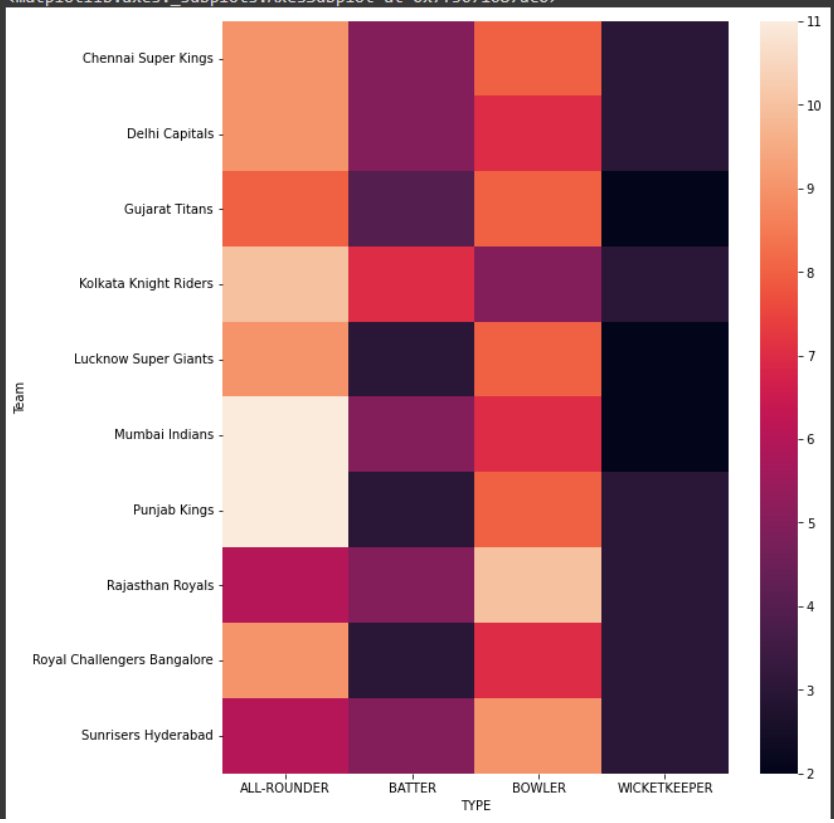
2) Rajasthan Royals has more number of bowlers

3) kolkata knight riders has more number of batter

2) using heatmap

```
[74] plt.figure(figsize=(9,11))
     sns.heatmap(pd.crosstab(dataset_sold["Team"],dataset_sold["TYPE"]))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3071087dc0>
```



6) EDA on unsold data

a)

```
[75] dataset_unsold=dataset[dataset["Team"]=="Unsold"]
```

```
[76] dataset_unsold.head()
```

	Player	Base Price	TYPE	COST IN ₹ (CR.)	Cost IN $ (000)	2021 Squad	Team	sold
237	Suresh Raina	2 Cr	BATTER	NaN	NaN	CSK	Unsold	0
238	Steve Smith	2 Cr	BATTER	NaN	NaN	DC	Unsold	0
239	Shakib Al Hasan	2 Cr	ALL-ROUNDER	NaN	NaN	KKR	Unsold	0
240	Amit Mishra	1.5 Cr	BOWLER	NaN	NaN	DC	Unsold	0
241	Adil Rashid	2 Cr	BOWLER	NaN	NaN	PBKS	Unsold	0

b)

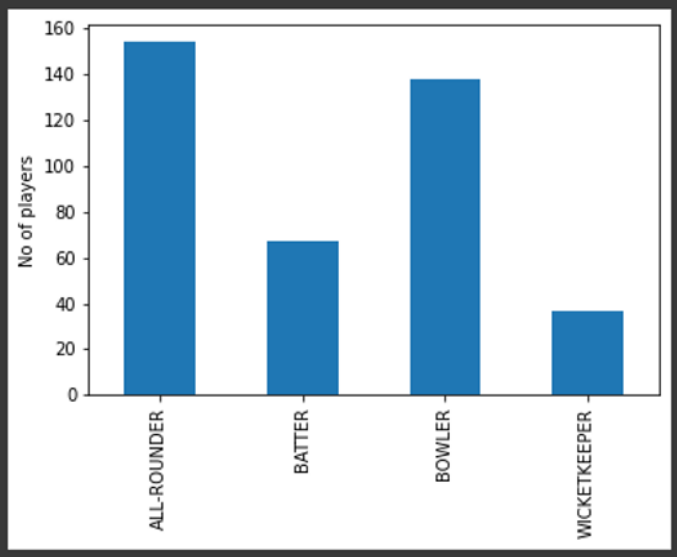
```
[77] dataset_unsold.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 396 entries, 237 to 632
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Player           396 non-null    object
 1   Base Price       396 non-null    object
 2   TYPE             396 non-null    object
 3   COST IN ₹ (CR.)  0 non-null      float64
 4   Cost IN $ (000)  0 non-null      float64
 5   2021 Squad       40 non-null     object
 6   Team             396 non-null    object
 7   sold             396 non-null    int64
dtypes: float64(2), int64(1), object(5)
memory usage: 27.8+ KB
```

c)

```
[78] bar1=pd.crosstab(dataset_unsold["TYPE"],dataset_unsold["sold"])

bar1.plot(kind="bar", legend=False)
plt.ylabel("No of players")
plt.xlabel("")
plt.show()
```



The above graph represents the types in no of unsold players

From the above bar chart
observations:

1. No of All-rounder unsold :150
2. No of Batters unsold: 60
3. No of bowlers unsold: 130
4. No of wicketkeeper unsold: 40

The End