In [30]:

```
# Import packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [28]:

```
# Import dataset
baseball = pd.read_excel("AL_West.xlsx")
```

In [29]:

```
baseball.drop('Division', inplace =  True, axis = 1) # Removes `Division` since every player on the list
baseball
```

Out[29]:

	Name	Team	AVG_2019	AVG_2020	OBP_2019	OBP_2020
0	Martin Maldonado	Astros	0.213	0.215	0.293	0.350
1	Yuli Gurriel	Astros	0.298	0.232	0.343	0.274
2	Jose Altuve	Astros	0.298	0.219	0.353	0.286
3	Carlos Correa	Astros	0.279	0.264	0.358	0.326
4	Alex Bregman	Astros	0.296	0.242	0.423	0.350
...	...	...	...	...	...	...
94	Chad Pinder	Athletics	0.240	0.232	0.290	0.295
95	Austin Allen	Athletics	0.215	0.194	0.282	0.219
96	Sean Murphy	Athletics	0.245	0.233	0.333	0.364
97	Beau Taylor	Athletics	0.174	0.048	0.300	0.130
98	Jake Lamb	Athletics	0.193	0.193	0.323	0.283

99 rows × 6 columns

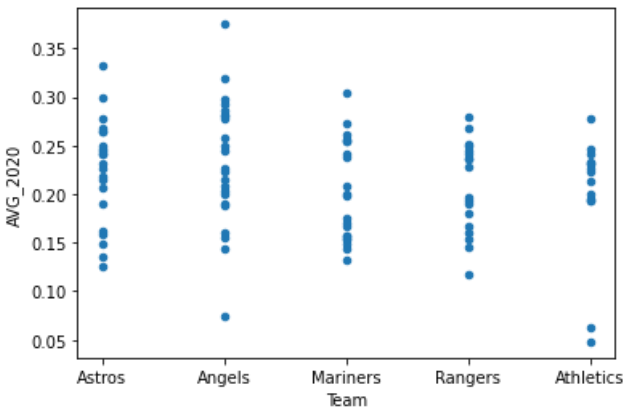
This dataset contains 99 observations of individual player statistics during the 2019 and 2020 major league baseball (MLB) seasons. Each observation includes the player's name, team, batting averages in 2019 and 2020, and on-base percentages in 2019 and 2020.

In [32]:

```
# Categoric EDA
# Use pandas to create a scatterplot
baseball.plot.scatter(x = 'Team', y = 'AVG_2020')
```

Out[32]:

```
<AxesSubplot:xlabel='Team', ylabel='AVG_2020'>
```



This scatterplot shows each player's batting average in 2020, and is organized by team to show the distribution of individual batting averages on each team in the American League West division. The Angels and Astros appear to have the most productive batting averages, meaning that they are higher than other teams on average.

In [46]:

```
# Numeric EDA
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
print("The mean batting average in the AL West in 2019 was", round(np.mean(baseball.AVG_2019), 3))
print("The standard deviation for batting average in the AL West in 2019 was", round(np.std(baseball.AVG_
print()
print("The mean batting average in the AL West in 2020 was", round(np.mean(baseball.AVG_2020), 3))
print("The standard deviation for batting average in the AL West in 2020 was", round(np.std(baseball.AVG_
print()
print("The mean on-base percentage in the AL West in 2019 was", round(np.mean(baseball.OBP_2019), 3))
```

```
print("The standard deviation for on-base percentage in the AL West in 2019 was", round(np.std(baseball.C
print()
print("The mean one-base percentage in the AL West in 2020 was", round(np.mean(baseball.OBP_2020), 3))
print("The standard deviation for on-base percentage in the AL West in 2020 was", round(np.std(baseball.C
```

```
The mean batting average in the AL West in 2019 was 0.236
The standard deviation for batting average in the AL West in 2019 was 0.045

The mean batting average in the AL West in 2020 was 0.215
The standard deviation for batting average in the AL West in 2020 was 0.057

The mean on-base percentage in the AL West in 2019 was 0.315
The standard deviation for on-base percentage in the AL West in 2019 was 0.053

The mean one-base percentage in the AL West in 2020 was 0.296
The standard deviation for on-base percentage in the AL West in 2020 was 0.066
```

In [64]:

```
# Visualization for numeric EDA
teamAVG = baseball.groupby(['Team']) \
.agg(['mean'])
teamAVG
```

Out[64]:

	AVG_2019	AVG_2020	OBP_2019	OBP_2020
	mean	mean	mean	mean
Team				
Angels	0.232130	0.232652	0.309478	0.321217
Astros	0.237909	0.225091	0.324818	0.303045
Athletics	0.239667	0.201133	0.323000	0.289000
Mariners	0.230619	0.202571	0.302952	0.284476
Rangers	0.242833	0.207944	0.319833	0.276611

In [74]:

```
data = {'Angels': .233, 'Astros': .226, 'Athletics': .201,
        'Mariners': .203, 'Rangers': .208}
teams = list(data.keys())
values = list(data.values())

plt.bar(teams, values, color = 'black',
        width = 0.5)

plt.xlabel("Teams")
plt.ylabel("Mean Batting Average")
plt.title("Mean Team Batting Averages in 2020")
plt.show()
```

Out[74]:

```
<BarContainer object of 5 artists>
```

Out[74]:

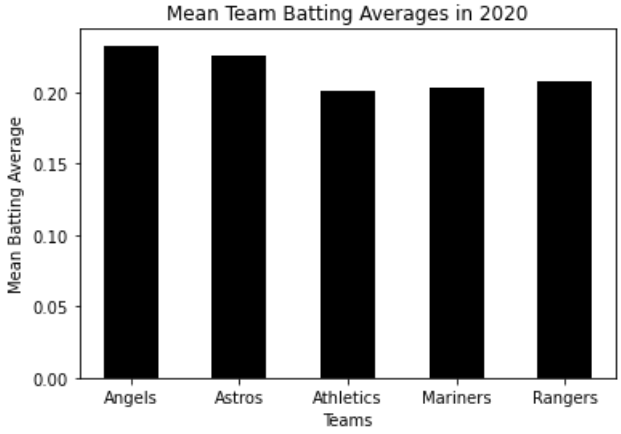
```
Text(0.5, 0, 'Teams')
```

Out[74]:

```
Text(0, 0.5, 'Mean Batting Average')
```

Out[74]:

```
Text(0.5, 1.0, 'Mean Team Batting Averages in 2020')
```



The bar plot shows that, in 2020, the Angels had the highest team batting average in the AL West, while the Athletics had the worst, which is interesting to note because the Athletics actually won the AL West that year, while the Angels didn't even make the playoffs.