Codecademy NBA Trends Project

Analyze National Basketball Association (NBA) data to look at associations between teams, win-rates, playoff appearances, and more.

In this project, you'll analyze data from the NBA (National Basketball Association) and explore possible associations.

This data was originally sourced from 538's Analysis of the Complete History Of The NBA and contains the original, unmodified data from Basketball Reference as well as several additional variables 538 added to perform their own analysis.

You can read more about the data and how it's being used by 538 here. For this project we've limited the data to just 5 teams and 10 columns (plus one constructed column, `point_diff`, the difference between `pts` and `opp_pts`).

You will create several charts and tables in this project, so you'll need to use `plt.clf()` between plots in your code so that the plots don't layer on top of one another.

In [1]:
```
import pandas as pd
import numpy as np
from scipy.stats import pearsonr, chi2_contingency
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
#to make the output look nicer
np.set_printoptions(suppress=True, precision = 2)
```

In [3]:
```
nba = pd.read_csv('nba_games.csv')
nba.head()
```

Out[3]:

	game_id	year_id	fran_id	opp_fran	game_location	is_playoffs	pts	opp_
0	194611010TRH	1947	Knicks	Huskies	A	0	68	
1	194611020CHS	1947	Knicks	Stags	A	0	47	
2	194611020PRO	1947	Celtics	Steamrollers	A	0	53	
3	194611050BOS	1947	Celtics	Stags	H	0	55	
4	194611070STB	1947	Knicks	Bombers	A	0	68	

In [4]:
```
# Subset Data to 2010 Season, 2014 Season
nba_2010 = nba[nba.year_id == 2010]
nba_2014 = nba[nba.year_id == 2014]
```

Task 1

The data has been subset for you into two smaller datasets: games from 2010 (named nba_2010) and games from 2014 (named nba_2014). To start, let's focus on the 2010 data.

Suppose you want to compare the knicks to the nets with respect to points earned per game. Using the pts column from the nba_2010 DataFrame, create two series named knicks_pts (fran_id = "Knicks") and nets_pts(fran_id = "Nets") that represent the points each team has scored in their games.

In [5]:
```
knicks_pts_2010 = nba_2010.pts[nba_2010.fran_id=="Knicks"]
nets_pts_2010 = nba_2010.pts[nba_2010.fran_id=="Nets"]
```

Task 2

Calculate the difference between the two teams' average points scored and save the result as diff_means_2010. Based on this value, do you think fran_id and pts are associated? Why or why not?

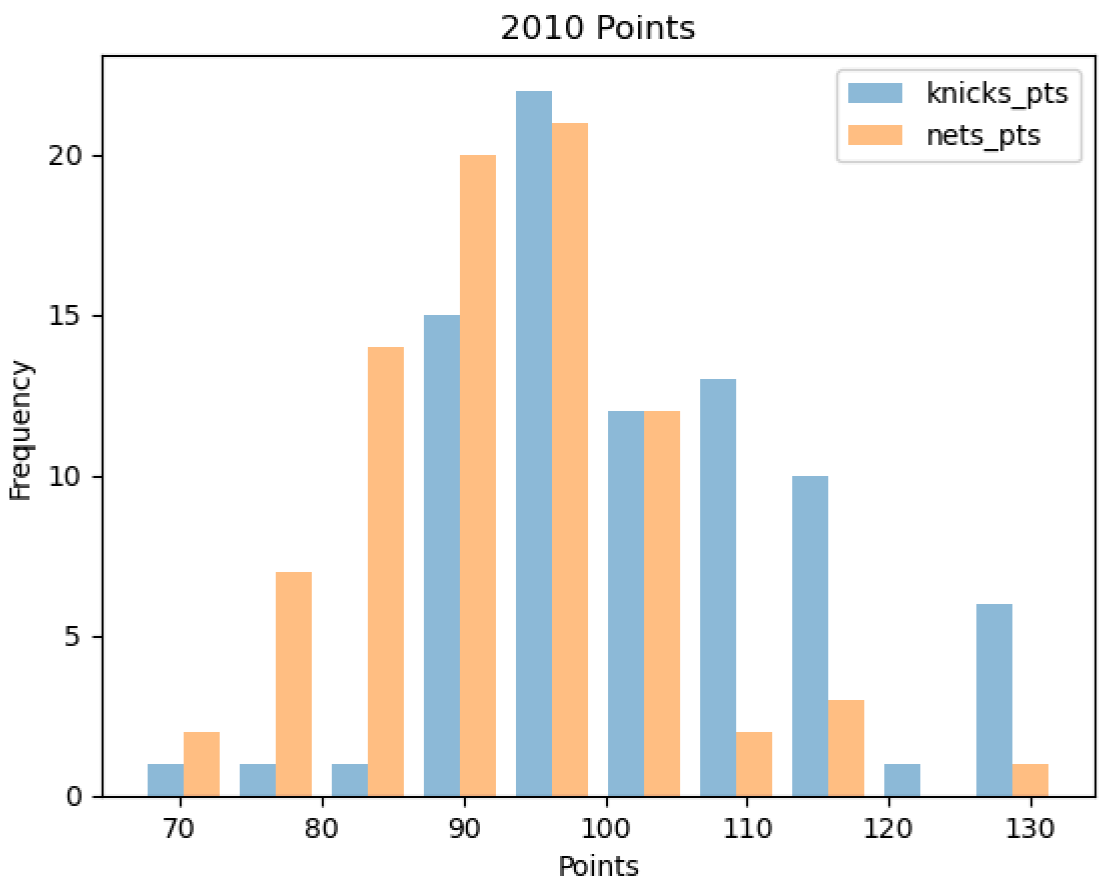
In [7]:
```
knicks_pts_average_2010=knicks_pts_2010.mean()
nets_pets_average_2010=nets_pts_2010.mean()

print(knicks_pts_average_2010)
print(nets_pets_average_2010)
diff_means_2010 = knicks_pts_average_2010-nets_pets_average_2010
print(abs(diff_means_2010))
```

```
102.10975609756098
92.3780487804878
9.731707317073173
```

Task 3

Rather than comparing means, it's useful look at the full distribution of values to understand whether a difference in means is meaningful. Create a set of overlapping histograms that can be used to compare the points scored for the Knicks compared to the Nets. Use the series you created in the previous step (1) and the code below to create the plot. Do the distributions appear to be the same?

In [8]:
```
# Plot overlapping histograms
plt.hist([knicks_pts_2010, nets_pts_2010], histtype='bar', stacked=False, la
plt.xlabel('Points')
plt.ylabel('Frequency')
plt.title('2010 Points')
plt.legend()
plt.show()
```



Task 4

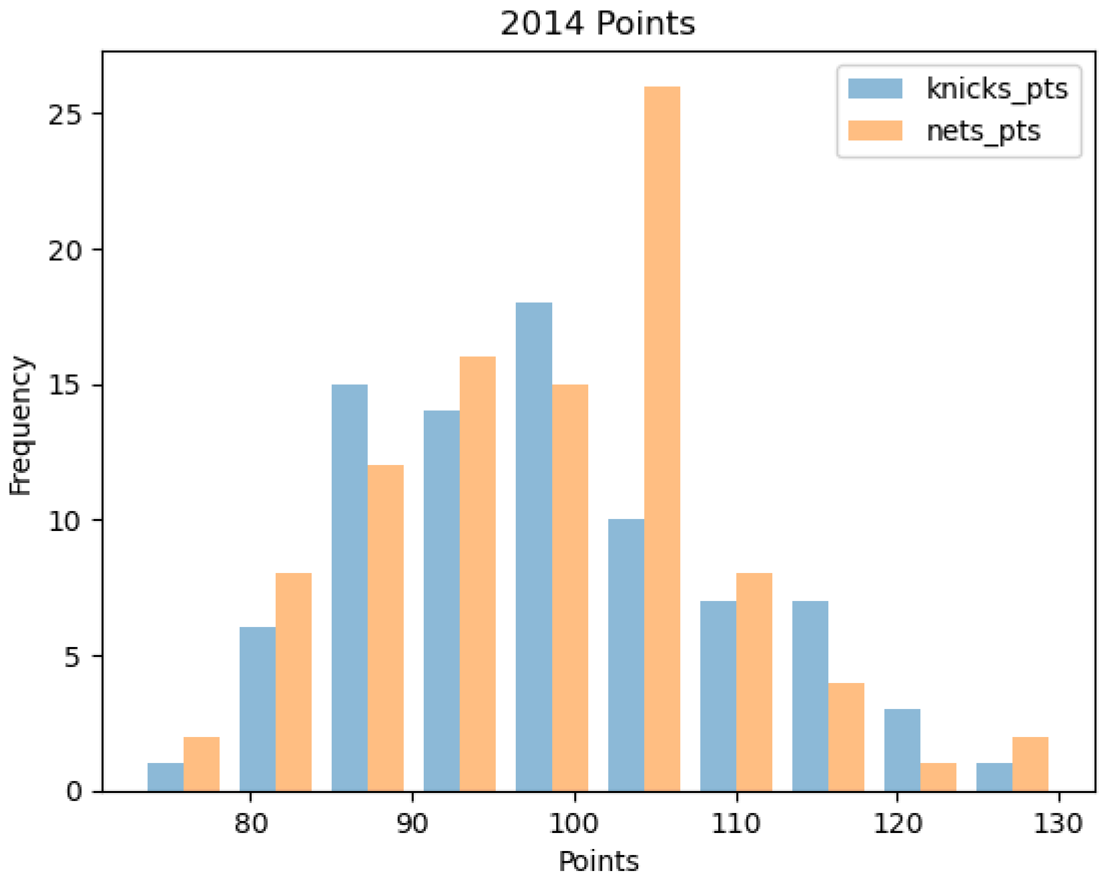
Now, let's compare the 2010 games to 2014. Replicate the steps from Tasks 2 and 3 using `nba_2014`. First, calculate the mean difference between the two teams points scored. Save and print the value as `diff_means_2014`. Did the difference in points get larger or smaller in 2014? Then, plot the overlapping histograms. Does the mean difference you calculated make sense?

In [12]:

```
knicks_pts_2014 = nba_2014.pts[nba_2014.fran_id=="Knicks"]
nets_pts_2014 = nba_2014.pts[nba_2014.fran_id=="Nets"]

knicks_pts_average_2014=knicks_pts_2014.mean()
nets_pets_average_2014=nets_pts_2014.mean()

print(knicks_pts_average_2014)
print(nets_pets_average_2014)
diff_means_2014 = knicks_pts_average_2014-nets_pets_average_2014
print(abs(diff_means_2014))

plt.hist([knicks_pts_2014, nets_pts_2014], histtype='bar', stacked=False, la
plt.xlabel('Points')
plt.ylabel('Frequency')
plt.title('2014 Points')
plt.legend()
```

```
plt.show()
plt.close()
```

```
98.58536585365853
98.13829787234043
0.44706798131809933
```

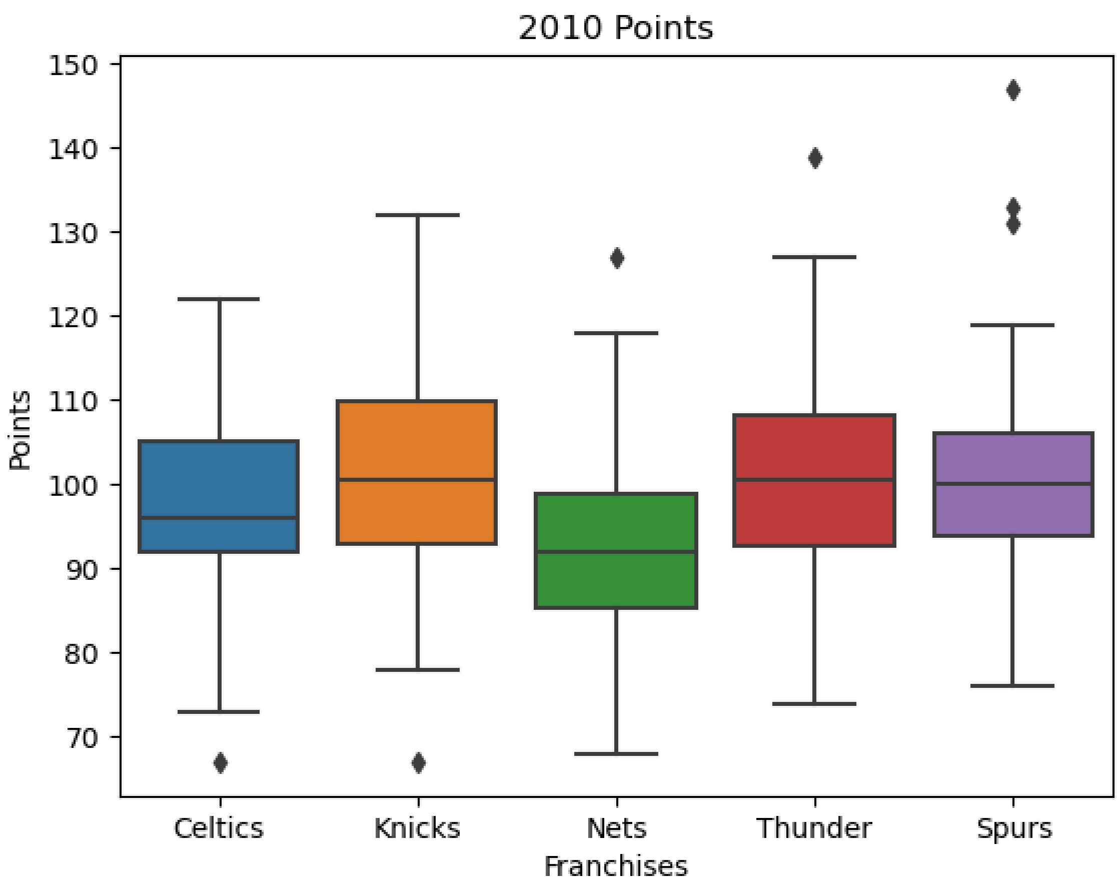


Task 5

For the remainder of this project, we'll focus on data from 2010. Let's now include all teams in the dataset and investigate the relationship between franchise and points scored per game.

Using nba_2010, generate side-by-side boxplots with points scored (pts) on the y-axis and team (fran_id) on the x-axis. Is there any overlap between the boxes? Does this chart suggest that fran_id and pts are associated? Which pairs of teams, if any, earn different average scores per game?

In [20]:
```
sns.boxplot(data = nba_2010, x = 'fran_id', y = 'pts')
plt.xlabel('Franchises')
plt.ylabel('Points')
plt.title('2010 Points')

plt.show()
print(nba_2010[nba_2010.fran_id == 'Nets'].pts.max())
```




```
127
```

Task 6

We'd like to know if teams tend to win more games at home compared to away.

The variable, `game_result`, indicates whether a team won a particular game ('W' stands for “win” and 'L' stands for “loss”). The variable, `game_location`, indicates whether a team was playing at home or away ('H' stands for “home” and 'A' stands for “away”).

Data scientists will often calculate a contingency table of frequencies to help them determine if categorical variables are associated. Calculate a table of frequencies that shows the counts of game_result and game_location.

Save your result as `location_result_freq` and print your result. Based on this table, do you think the variables are associated?`

In [22]:
```
location_result_freq = pd.crosstab(nba_2010.game_result, nba_2010.game_locat
print(location_result_freq)
```

```
game_location    A    H
game_result
L              133  105
W               92  120
```

Task 7

Convert this table of frequencies to a table of proportions and save the result as `location_result_proportions`.

In [23]:
```
location_result_proportions = location_result_freq/len(nba_2010.game_result)
print(location_result_proportions)
```

```
game_location         A         H
game_result
L              0.295556  0.233333
W              0.204444  0.266667
```

Task 8

Using the contingency table created above (Task 6), calculate the expected contingency table (if there were no association) and the Chi-Square statistic.

Does the actual contingency table look similar to the expected table — or different? Based on this output, do you think there is an association between these variables?

In [28]:
```
chi2, pval, dof, expected = chi2_contingency(location_result_freq)
print(np.round(expected))

'''Comparing the observed frequency table:

game_location    A    H
game_result
L              133  105
W               92  120

to the expected frequency table:

[[119. 119.]
 [106. 106.]]

There is a slight difference and the actual difference can be calculated by
and when we calculate the chi squared statistic it is larger than 4 which ir
'''

print(location_result_freq-np.round(expected))

print(chi2)
```

```
[[119. 119.]
 [106. 106.]]
game_location     A     H
game_result
L              14.0 -14.0
W             -14.0  14.0
6.501704455367053
```

For a 2x2 table, Chi-squared greater than about 4 indicates an association. We're not there

Task 9

For each game, 538 has calculated the probability that each team will win the game. We want to know if teams with a higher probability of winning (according to 538) also tend to win games by more points.

In the data, 538's prediction is saved as `forecast`. The `point_diff` column gives the margin of victory/defeat for each team (positive values mean that the team won; negative values mean that they lost).

Using `nba_2010`, calculate the covariance between `forecast` (538's projected win probability) and `point_diff` (the margin of victory/defeat) in the dataset. Save and print your result. Looking at the matrix, what is the covariance between these two variables?

In [44]:
```
forecast_pointDiff_cov = np.cov(nba_2010.forecast,nba_2010.point_diff)
print(forecast_pointDiff_cov)

'''The covariance here is 1.37
The positive value (1.37) indicates a positive covariation between forecast
This means that as the forecasted points increase,
the actual point differential (difference between points scored by two teams
```

```
[[  0.05    1.37]
 [  1.37 186.56]]
```

Out[44]:
```
'The positive value (1.37) indicates a positive covariation between forecas
t and point_diff.This means that as the forecasted points increase, the act
ual point differential (difference between points scored by two teams) also
tends to increase.'
```

Task 10

Because 538’s forecast variable is reported as a probability (not a binary), we can calculate the strength of the correlation.

Using nba_2010, calculate the correlation between `forecast` and `point_diff`. Call this `point_diff_forecast_corr`. Save and print your result. Does this value suggest an association between the two variables?

In [46]:
```
point_diff_forecast_corr,p =  pearsonr(nba_2010.forecast, nba_2010.point_dif

print(point_diff_forecast_corr)

'''
```

```
The correlation here is normalized so it can only be between -1 and 1.
Since this is positive and a value of 0.4, it indicates a moderately strong
```

```
0.440208870846808 54
```

Out[46]:

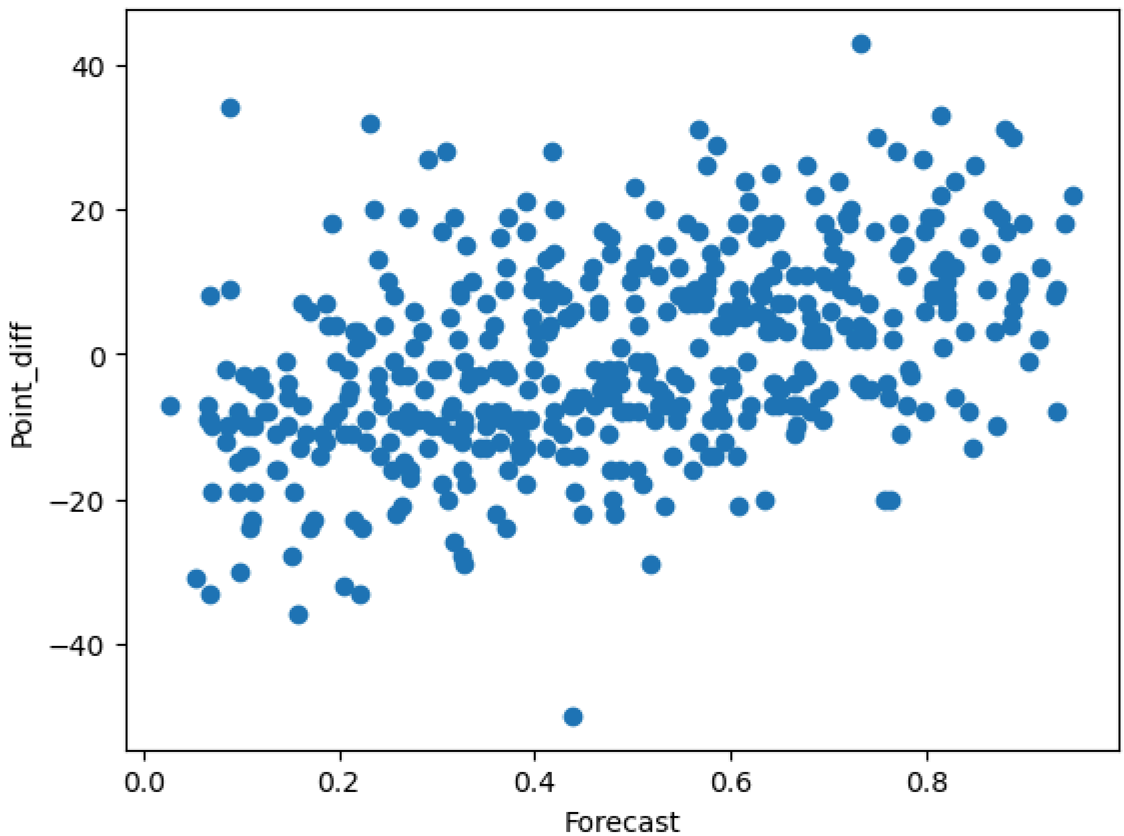
```
'\nThe correlation here is normalized so it can only be between -1 and 1. S
ince this is positive and a value of 0.4, it indicates a moderately strong
positive relationship between these variables'
```

Task 11

Generate a scatter plot of `forecast` (on the x-axis) and `point_diff` (on the y-axis). Does the correlation value make sense?

In [47]:

```
plt.scatter(x = nba_2010.forecast, y = nba_2010.point_diff)
plt.xlabel('Forecast')
plt.ylabel('Point_diff')
plt.show()
```



As the scatterplot shows, there is a somewhat positive linear relationship between forecast and point differential.