Introductory of your dataset:

I obtained my dataset from Kaggle, and it is called 2018 NBA. There are 530 cases (which represents 530 players) and 36 variables in the dataset (36 columns and 530 rows). For each case, the varaibles being measured are field goals attempted, rebounds, steals, blocks, and many other variables. For my lab, I am interested in the variables rebounds, points, and position. I am interested to find the relationship between rebounds rebounded and position (first relationship I am analyzing) and the relationship between average amount points scored and rebounds rebounded (second relationship I am analyzing). I am interested in this dataset because I thought it would be interesting to dive deeper into these statistics to see if there is a correlation between these statsitics, overall wanting to see how one variable can influence the other. I chose this dataset specifically because it gave me a large sample size and it included the variables I wanted to analyze.`

Description of your research quesiton:

For my assignment, I am trying to answer two questions. The first question I am trying to answer is if being a center leads to more rebounds rebounded than those of non-center positions such as a Guard or Foward in the NBA. To answer this question I will be doing a hypothesis test and a bootstrapping test. This is the question I will be analyzing first.

The second question I am trying to answer is if there is a relationship between the average amount of points scored and the amount of rebounds rebounded by an NBA player. To answer this question I will be performing a regression where I will make a scatter plot with "pts" on the x-axis and "reb" on the y-axis and the yellow dots will represent being a Center and the blue dots will represent every other NBA position. Then I will graph the regression line to find the relationship between rebounds and points. I will analyze this question last.

```
from datascience import *
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
import math
from scipy import stats
import numpy as np
import pandas as pd
import warnings
import matplotlib.pyplot as plots
warnings.simplefilter(action='ignore', category=np.VisibleDeprecationWarning)
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
!unzip "/content/drive/MyDrive/archive (6).zip"
```

```
Archive:  /content/drive/MyDrive/archive (6).zip
  inflating: 2018NBA.csv
```

Exploratory Data Analysis Starts Here:

```
NBA = Table.read_table('2018NBA.csv')
NBA.show(10)
```

Unnamed: 0	games_played	player_id	season	min	fgm	fga	fg3m	fg3a	ftm	fta	oreb	dreb	reb	ast	stl	blk	turnover	pf	pts	fg_pct	fg3_pct	ft_pct	4_years	first_name	last_name	position	height_feet	height_inches	weight_pounds
0	31	1	2018	18:58	1.81	5.06	1.32	4.1	0.39	0.42	0.16	1.39	1.55	0.65	0.55	0.19	0.45	1.71	5.32	0.357	0.323	0.923	0	Alex	Abrines	G	6	6	200
1	3	14	2018	2:03	0	1	0	0	0	0	0.33	0.67	1	0.33	0	0.33	0.33	0	0	0	0	0	0	Ike	Anigbogu	C	nan	nan	nan
2	15	25	2018	10:07	0.27	1.33	0.07	0.87	0.33	0.4	0.07	0.67	0.73	1	0.4	0.07	0.33	1.2	0.93	0.2	0.077	0.833	0	Ron	Baker	G	nan	nan	nan
3	29	67	2018	13:20	2.62	5.83	0.52	1.86	0.79	1.14	0.41	1.14	1.55	0.86	0.31	0.14	0.72	1.14	6.55	0.45	0.278	0.697	0	MarShon	Brooks	G	nan	nan	nan
4	26	71	2018	8:08	0.88	2.73	0.23	1.08	0.12	0.12	0.19	1	1.19	1.08	0.46	0.19	0.62	0.85	2.12	0.324	0.214	1	0	Lorenzo	Brown	G	nan	nan	nan
5	36	90	2018	14:27	2.39	4.47	0.42	1.19	1.08	1.61	0.47	2.72	3.19	0.72	0.56	0.25	0.64	0.97	6.28	0.534	0.349	0.672	0	Omri	Casspi	F	nan	nan	nan
6	1	119	2018	0:55	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	Tyler	Davis	C	nan	nan	nan
7	47	179	2018	15:58	2.11	3.96	0	0	0.74	1.02	1.43	4.13	5.55	1.38	0.13	0.51	1.06	1.96	4.96	0.532	0	0.729	0	Marcin	Gortat	C	nan	nan	nan
8	51	241	2018	10:21	1.55	3.08	0.24	0.78	0.61	0.8	0.92	1.96	2.88	1.18	0.31	0.25	0.88	1.94	3.94	0.503	0.3	0.756	0	Amir	Johnson	C-F	6	9	240
9	42	263	2018	11:54	1.74	3.64	0	0	0.24	0.57	1.24	2.98	4.21	0.86	0.36	0.43	0.64	1.62	3.71	0.477	0	0.417	0	Kosta	Koufos	C	7	0	245

... (520 rows omitted)

Exploratory data analysis:

This is the dataset I am using for my lab consisting of 530 cases (which represent 530 NBA players) where for each player we can see their position, the average amount of rebounds they rebounded, and the average amount of points they scored. I will be using this data to evaluate the relationship between position, rebounds rebounded, and points.

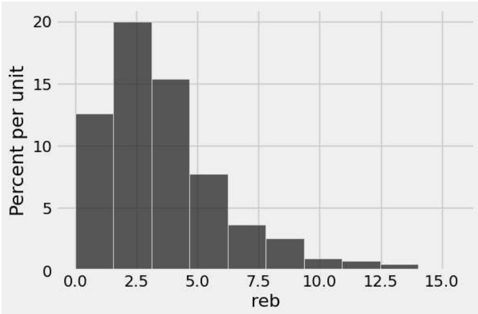
```
df.isnull().sum()
```

```
Unnamed: 0          0
games_played        0
player_id           0
season              0
min                 0
fgm                 0
fga                 0
fg3m                0
fg3a                0
ftm                 0
fta                 0
oreb                0
dreb                0
reb                 0
ast                 0
stl                 0
blk                 0
turnover            0
pf                  0
pts                 0
fg_pct              0
fg3_pct             0
ft_pct              0
4_years             0
first_name          0
last_name           0
position            0
height_feet        61
height_inches      61
weight_pounds      61
team_id             0
team_abbreviation   0
team_city           0
team_conference     0
team_division       0
team_full_name      0
team_name           0
dtype: int64
```

Exploratory data analysis:

This table categorizes each of the 530 cases into each players position. We see each players position and the amount of rebounds they averaged in a clearer table consisting of just these two variables (position and rebounds).

```
position_and_rebounds.select('reb').hist()
```




```
median_rebounds = df['reb'].median()
median_rebounds
```

```
3.0600000000000001
```

Exploratory data analysis:

This histogram gives us insight into the average rebounds rebounded for all 530 cases. It is clearly right skewed which means the mean is greater than the median: 3.61 is greater than 3.06

```
position_and_rebounds.group('position')
```

position	count
C	55
C-F	10
F	161
F-C	25
F-G	11
G	195
G-F	34
nan	39

Exploratory data analysis:

This table shows the positions of all 530 cases and how the 530 cases are distributed among these positions.

```
position_and_rebounds = position_and_rebounds.with_column(
    'is_Center', position_and_rebounds.column('position') == 'C')
position_and_rebounds.group('is_Center', np.average)
```

is_Center	position average	reb average
False		3.29135
True		6.40127

```
reb_averages = position_and_rebounds.select('is_Center', 'reb')
reb_averages
```

is_Center	reb
False	1.55
True	1
False	0.73
False	1.55
False	1.19
False	3.19
True	1
True	5.55
False	2.88
True	4.21

... (520 rows omitted)

```
means_table = position_and_rebounds.group('is_Center', np.average)
means_table
```

is_Center	position average	reb average
False		3.29135
True		6.40127

```
means_table.select('is_Center', 'reb average')
```

is_Center	reb average
False	3.29135
True	6.40127

Exploratory data analysis and code analysis:

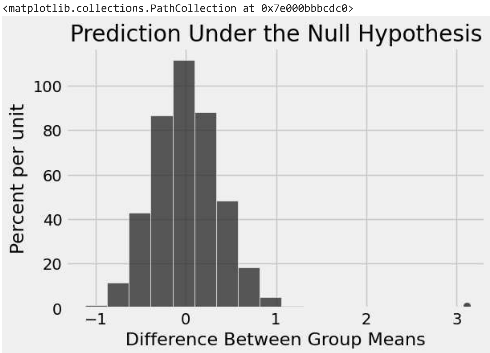
The 4 above codes I used to converted all cases that are apart of the center position to true and every other position to false. Then I calculated the average rebounds for the center position (labled true) and the average rebounds for non-center positions (labled false). We see that for centers the average amount of rebounds rebounded is 6.40 and for non-center positions the average amount of rebounds rebounded is 3.29. We see that the average amount of rebounds rebounded by centers is almost double the average amount of rebounds rebounded by the non-center positions.

```
differences = make_array()

repetitions = 5000
for i in np.arange(repetitions):
    new_difference = one_simulated_difference_of_means()
    differences = np.append(differences, new_difference)


Table().with_column('Difference Between Group Means', differences).hist()
print('Observed Difference:', observed_diff)
plots.title('Prediction Under the Null Hypothesis');
plots.scatter(3.11, .01, color = 'red')
```

```
Observed Difference: 3.10992535885
<matplotlib.collections.PathCollection at 0x7e000bbbcdc0>
```



Hypothesis Test Analysis:

This histogram displays the difference in means between the amount of rebounds rebounded for centers and non-center positions for 5000 repetitions of the dataset. The histogram is normally distributed, representing a bell-shaped curve. The red dot represents the observed difference between centers and non-center positions (observed difference = 3.110). The dot clearly shows that there are no values as extreme as the observed difference.

```
empirical_p = np.count_nonzero(differences >= observed_diff) / repetitions
empirical_p
```

```
0.0
```

Hypothesis Testing Conclusion:

The observed test statstic of 3.110 rebounds (represented by the red dot) is the most extreme value on the histogram because there is no other value that is more or as extreme as the observed test statistic. The p-value I calculated is 0.0 which supports my conclusion that there are no values as extreme as 3.110 rebounds. Based on this p-value, we can reject the null hypothesis which is there is no statistical difference between being a center versus non-center positions and the amount of rebounds they average throughout the season.

In conclusion, based on my experiment, we can say that the average rebounds for a center is more than the average rebounds for any non-center position in the NBA.

Condifdence Interval Starts Here:

```
#NBA_Center = means_table.where('is_Center', True)
#NBA_Other_Positions = means_table.where('is_Center', False)


true_values_reb_averages = reb_averages.where('is_Center', True)
true_values_reb_averages
```

is_Center	reb
True	1
True	1
True	5.55
True	4.21
True	2
True	3.64
True	3.9
True	2.17
True	0.36
True	2.2

... (45 rows omitted)

Code Analysis:

Table consists of only the players who are centers and the amount of rebounds each center averages.

```
false_values_reb_averages = reb_averages.where('is_Center', False)
false_values_reb_averages
```

is_Center	reb
False	1.55
False	0.73
False	1.55
False	1.19
False	3.19
False	2.88
False	0
False	0
False	0.59
False	0

... (465 rows omitted)

Code Analysis:

Table consists of only the non-center players and the amount of rebounds each non-center averages.

Histogram Analysis:

From our bootstrapping test we produced a histogram. This historgam shows the difference in means of rebounds between the center position and non-center positions by resampling the single dataset many times and creating simulated samples of the difference in means and plotting these on a histogram. On this histogram, there are no negative value which means for every sample there was always a positive difference between rebounds rebounded for centers and rebounds rebounded for non-center positions.

```
left = percentile(2.5, bstrap_means)
right = percentile(97.5, bstrap_means)
print(left, right)
```

```
2.15307368421 4.07587559809
```

Confidence Interval Analysis:

We are 95% confident that the true difference in mean of rebounds rebounded by Centers and anyother position is between 2.19 and 4.04.

Predicitve Method Starts Here:

```
regression = NBA.select('position','pts','reb')
regression
```

position	pts	reb
G	5.32	1.55
C	0	1
G	0.93	0.73
G	6.55	1.55
G	2.12	1.19
F	6.28	3.19
C	0	1
C	4.96	5.55
C-F	3.94	2.88
C	3.71	4.21

... (520 rows omitted)

Code Analysis:

Selected "position" , "pts" , and "reb" from the original dataset called NBA.

```
regression = regression.with_column('is_Center', regression.column('position') == 'C')
regression
```

position	pts	reb	is_Center
G	5.32	1.55	False
C	0	1	True
G	0.93	0.73	False
G	6.55	1.55	False
G	2.12	1.19	False
F	6.28	3.19	False
C	0	1	True
C	4.96	5.55	True
C-F	3.94	2.88	False
C	3.71	4.21	True

... (520 rows omitted)

Code Analysis:

Added the is_center column so we can later convert all positions under the "position" column to numerical values. Center position will turn into the value 1 and non-center positions will turn into the value 0.

```
regression = regression.drop('position')
regression
```

pts	reb	is_Center
5.32	1.55	False
0	1	True
0.93	0.73	False
6.55	1.55	False
2.12	1.19	False
6.28	3.19	False
0	1	True
4.96	5.55	True
3.94	2.88	False
3.71	4.21	True

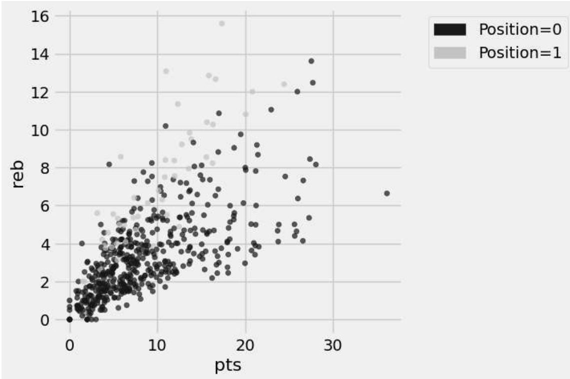
... (520 rows omitted)

Code Analysis:

Dropped the "Position" so we can then change all categorical variables in the "Position" column to values.

```
regression= regression.with_column('Position', np.multiply(1, regression.column('is_Center') == True))
regression.show(10)
```

pts	reb	is_Center	Position
5.32	1.55	False	0
0	1	True	1
0.93	0.73	False	0
6.55	1.55	False	0
2.12	1.19	False	0
6.28	3.19	False	0
0	1	True	1



Scatter-Plot Analysis:

This scatter plot shows the relationship between points and rebounds. The blue dots represent a center and the yellow dots represent non-center positions.

```
def regression_linear_mse(any_slope, any_intercept):
    x = regression.column('pts')
    y = regression.column('reb')
    fitted = any_slope*x + any_intercept
    return np.mean((y - fitted) ** 2)
```

```
minimize(regression_linear_mse)
```

```
array([ 0.27333493,  1.25911448])
```

What do these values represent:

The output of ([0.27333493, 1.25911448]) from the linear regression analysis shows the slope and the y-intercept for the regression line. The slope is 0.27 and the y-intercept is 1.26. A slope of 0.27 means that for each additional point scored results in a increase of 0.27 rebounds rebounded. The y-intercept means that when no points are scored a player is expected to have rebounded 1.26 rebounds. So, from these values we can predict the amount of rebounds rebounded from the amount of points scored for any player in the dataset.

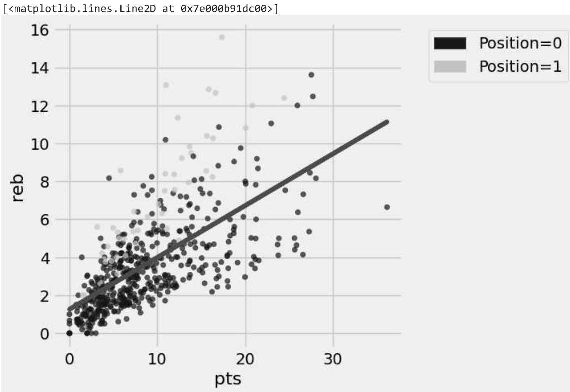
The equation of the regression line would be:

prediction of rebounds rebounded = 0.27 * pts scored + 1.26

and from this you can predict the amount of rebounds rebounded by a player by plugging in the amount of points scored for a player.

```
regression.scatter('pts', 'reb', group='Position')
x = np.array(regression.column('pts'))
y = np.array(regression.column('reb'))
slope = 0.27333493
intercept = 1.25911448
fitted = slope*x + intercept
plots.plot(x, fitted, color='red')
```

```
[<matplotlib.lines.Line2D at 0x7e000b91dc00>]
```



Scatter Plot Analysis:

This scatter plot shows the relationship between points and rebounds. The blue dots represent a center and the yellow dots represent non-center positions. The data points are mostly clusted between 0-10 pts with some points expanding out of this cluster. Of these data points, there appears to be no big outliars but there could be a potential outliar at the data point past 30 pts as it is the most far off from the cluster of data points. The red line represents the regession line, which serves as the line of best fit, which shows the trend in the data points. This line predicts the amount of rebounds rebounded from the amount of points scored by NBA players.

Conclusion:

Based on this lab, we can clearly see that there is a statistical difference in the amount of rebounds rebounded by centers and the amount of rebounds rebounded by non-center positions in the NBA.

Based on the p-value of 0.0 for the hypothesis test, we can conclude that there was no values as extreme as the observed difference of 3.110 and so we rejected the null hypothsis.

Based on the confidence interval from the bootstrapping test, we can see that since the confidence interval is completely positive, we are 95% confident that the amount of rebounds rebounded by centers will be more than the amount of rebounds rebounded by any other position.

Based on the predictive method, we can predict the amount of rebounds rebounded from the pts scored by a player using the equation: **prediction of rebounds rebounded = 0.27 * pts scored + 1.26.**

By doing this lab, I was able to gain more insight into the relationship between rebounds rebounded by a center versus non-center positions, and the relationship between pts scored and rebounds rebounded by implementing statistical tests.