[Student Id]
Member #1: Van Pham, s3788106
Member #2: Sean Tan, s3806690
Member #3: Sunny Thai, s3657606

Predicting Points From Seasonal Performance

Overview

Data source

The data we are using is from:
https://www.kaggle.com/drgilermo/nba-players-stats-20142015/download (https://www.kaggle.com/drgilermo/nba-players-stats-20142015/download)
The dataset is of the stats of players in the NBA from 2014 - 2015 and has 490 observations and consists of 1 target feature and 33 descriptive features.

Project Objective

We attempt to predict the scores of an NBA player based on the 2014-2015 end of season player statistics through regression analysis

Target Feature

Our target feature is 'PTS', which is a regression problem as it's a continuous numerical feature

Descriptive Features

Name: ID like
Games Played: Continuous
MIN: Continuous
FGM: Continuous
FGA: Continuous
FG%: Percentage
3PM: Continuous
3PA: Continuous
3P%: Percentage
FTM: Continuous
FTA: Continuous
FT%: Percentage
OREB: Continuous
DREB: Continuous
REB: Continuous
AST: Continuous
STL: Continuous
BLK: Continuous
TOV: Continuous
PF: Continuous
EFF: Continuous
AST/TOV: Continuous
STL/TOV: Continuous
Age: Continuous
Birth_Place: us, do, ua, ru, fr, it, au, ca, nz, gb, cd, ba, br, de, ch, hr, lt, tr, ng, il, gr, si, sn, ve, pr, se, jm, mx, es, gf, cm, ar, ss, pl, me, mk, ht, cg, vi, be, ge.
Birthdate: Dates
Collage: University of Connecticut, University of Oregon, University of Arizona, Michigan State University, University of Florida, University of Colorado, University of New Mexico, University of Maryland, Wake Forest University, University of California, University of Alabama, Duke University, University of Utah, St. Bonaventure University, University of Nevada, Las Vegas, University of Kentucky, Georgia Institute of Technology, Syracuse University, Washington State University, University of California, Los Angeles, Gonzaga University, University of Texas at Austin, University of Kansas, Florida State University, University of Oklahoma, Louisiana State University, Brigham Young University, University of North Carolina, University of Dayton, Stanford University, Providence College, University of Tennessee, Purdue University, Kansas State University, Blinn College, University of Memphis, Central Michigan University, Lehigh University, Wichita State University, Baylor University, Western Kentucky University, Weber State University, Villanova University, University of Michigan, Xavier University, Texas A&M University, University of Pittsburgh, University of Southern California, University of Missouri, University of Wisconsin, University of Iowa, Creighton University, Marquette University, University of Louisville, University of Louisiana at Lafayette, Indiana University, Virginia Polytechnic Institute and State University, Vanderbilt University, Towson University, Indiana University-Purdue University Indianapolis, Butler University, Georgetown University, California State University, Fresno, Belmont University, Murray State University, University of Washington, Northeastern University, University of Notre Dame, San Diego State University, Saint Joseph's University, California State University, Long Beach, Arizona State University, University of Miami, University of Arkansas, Oregon State University, Boston College, Ohio State University, University of Cincinnati, University of Nevada, Reno, Harvard University, University of Tulsa, North Carolina State University, University of Virginia, Oklahoma State University, Morehead State University, Old Dominion University, University of Georgia, Western Carolina University, Clemson University, University of Minnesota, Norfolk State University, Temple University, University of

Tennessee at Martin, Saint Mary's College of California, St. John's University, University of Illinois at Urbana-Champaign, Bucknell University, Cleveland State University, Louisiana Tech University, La Salle University, University of Detroit Mercy, Tennessee State University, Eastern Washington University, Utah Valley State College, Seton Hall University, Western Michigan University,New Mexico State University, Davidson College, Pennsylvania State University,Virginia Commonwealth University, University of Montana, DePaul University
Experience: 5, 6, R, 7, 10 , 3, 1, 2, 4, 9, 15, 8, 12, 11, 16, 14, 13, 19, 18, 17
Height: Continuous
Pos: PG, PF, C, SG, SF
Team: PHO, CHI, ORL, ATL, CHA, NJN, UTA, CLE, NYK, NOH, DAL, POR, TOR, MIA, DET, GSW, WAS, OKC, MIN, SAS, BOS, SAC, LAC, PHI, IND, LAL, HOU, MEM, DEN, MIL
Weight: Continuous
BMI: Continuous

Some descriptive features are self-explanatory, except ones with 3-4 letter descriptions which unless you are a basketball fan.

MIN = Minutes Played
(M)=made, (A)=attempted, (%)=percentage
FG = Field Goal
3P = 3 Point Field Goal
FT = Free Throw
REB = Rebound, (O)=Offensive, (D)=Defensive
AST = Assists, STL = Steal, BLK = Block, TOV = Turnover, PF = Personal Fouls

Data Preparation and Cleaning

In [1]:

```
# Importing modules
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
import patsy
import warnings
###
warnings.filterwarnings('ignore')
###
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
plt.style.use("ggplot")
```

In [2]:

```
dp = pd.read_csv("players_stats.csv")
```

In [3]:

```
dp.shape
```

Out[3]:

```
(490, 34)
```

In [4]:

```
dp.isnull().sum()
```

Out[4]:

```
Name              0
Games Played      0
MIN               0
PTS               0
FGM               0
FGA               0
FG%               0
3PM               0
3PA               0
3P%               0
FTM               0
FTA               0
FT%               0
OREB              0
DREB              0
REB               0
AST               0
STL               0
BLK               0
TOV               0
PF                0
EFF               0
AST/TOV           0
STL/TOV           0
Age              68
Birth_Place      68
Birthdate        68
Collage         140
Experience       68
Height           68
Pos              68
Team             68
Weight           68
BMI              68
dtype: int64
```

In [5]:

```
dp.describe(include='int64')
```

Out[5]:

	Games Played	MIN	PTS	FGM	FGA	3PM	3PA
count	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000	490.00000
mean	53.014286	1214.714286	502.108163	188.338776	419.526531	39.387755	112.5244
std	24.175437	820.570132	422.084232	156.265752	337.367125	47.880909	127.3857
min	1.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.0000
25%	33.000000	492.250000	145.250000	55.500000	139.000000	1.000000	6.0000
50%	61.000000	1193.000000	423.000000	156.000000	357.500000	18.000000	58.0000
75%	74.000000	1905.750000	774.000000	286.000000	642.750000	66.000000	192.0000
max	83.000000	2981.000000	2217.000000	659.000000	1471.000000	286.000000	646.0000

In [6]:

```
dp.describe(include='object')
```

Out[6]:

	Name	Birth_Place	Birthdate	Collage	Experience	Pos	Team
count	490	422	422	350	422	422	422
unique	490	41	408	112	20	5	30
top	Jerami Grant	us	September 2, 1989	Duke University	R	SG	NYK
freq	1	330	2	18	68	100	16

Justifications for dropping columns:

"Name" column as it is represented by a unique id and is fundamentally wrong for multiple regression analysis.

"Birthdate" for the same reason, as it has no predictive power towards the regressional analysis.

"Experience"

'FGM', '3PM', 'FTM', are able to calculate PTS which we are trying to predict.

In [7]:

```
data = dp.drop(columns=['Name', 'Birthdate', 'Experience', 'FGM', '3PM', 'FTM'])
```

In [8]:

```
data.describe(include='object')
```

Out[8]:

	Birth_Place	Collage	Pos	Team
count	422	350	422	422
unique	41	112	5	30
top	us	Duke University	SG	NYK
freq	330	18	100	16

In [9]:

```
data.describe(include = 'int64')
```

Out[9]:

	Games Played	MIN	PTS	FGA	3PA	FTA	OREE
count	490.000000	490.000000	490.000000	490.000000	490.00000	490.000000	490.00000(
mean	53.014286	1214.714286	502.108163	419.526531	112.52449	114.689796	54.65510:
std	24.175437	820.570132	422.084232	337.367125	127.38575	115.139240	61.06603(
min	1.000000	3.000000	0.000000	0.000000	0.00000	0.000000	0.00000(
25%	33.000000	492.250000	145.250000	139.000000	6.00000	26.250000	13.00000(
50%	61.000000	1193.000000	423.000000	357.500000	58.00000	80.000000	31.50000(
75%	74.000000	1905.750000	774.000000	642.750000	192.00000	166.750000	75.75000(
max	83.000000	2981.000000	2217.000000	1471.000000	646.00000	824.000000	437.00000(

In [10]:

```
datafresh = data
```

In [11]:

```
print("Since most NBA players were born in the US, we will categorise the Birth_Place c
olumn for a player either \n" +
        "was born in the US or Other")
datafresh.loc[data['Birth_Place'] != 'us', 'Birth_Place'] = 'Other'
datafresh['Birth_Place'].value_counts()
```

```
Since most NBA players were born in the US, we will categorise the Birth_P
lace column for a player either
was born in the US or Other
```

Out[11]:

```
us       330
Other    160
Name: Birth_Place, dtype: int64
```

In [12]:

```
print("For the same reason as Birth_Place, as Collage has too many unbalanced and spora
dic, we will cut the column + \n")
print('For the same reason, we will also drop the Team column as we are looking at indi
vidual skill as being in a + \n'
        + "certain team with certain players may influence a player's statistics")
datafresh.drop(columns = ['Collage', 'Team'])
```

```
For the same reason as Birth_Place, as Collage has too many unbalanced and
sporadic, we will cut the column +

For the same reason, we will also drop the Team column as we are looking a
t individual skill as being in a +
certain team with certain players may influence a player's statistics
```

Out[12]:

	Games Played	MIN	PTS	FGA	FG%	3PA	3P%	FTA	FT%	OREB	...	PF	EFF	AST/T
0	26	324	133	137	37.2	57	26.3	24	66.7	6	...	15	110	3
1	82	1885	954	817	42.1	313	38.7	174	83.3	32	...	189	791	1
2	47	797	243	208	44.7	48	27.1	61	72.1	46	...	83	318	0
3	32	740	213	220	41.4	9	11.1	46	65.2	48	...	88	244	0
4	76	2318	1156	965	53.8	36	30.6	141	75.9	131	...	121	1530	2
5	65	1992	1082	1010	48.1	5	40.0	165	65.5	99	...	139	1225	1
6	74	1744	545	440	44.3	210	34.8	101	81.2	31	...	148	569	1
7	27	899	374	300	40.3	68	38.2	129	82.2	19	...	64	338	1
8	5	14	4	4	25.0	0	0.0	2	100.0	1	...	1	3	0
9	69	1518	432	353	50.7	3	33.3	104	70.2	142	...	213	778	0
10	42	767	434	335	39.7	139	33.8	150	80.7	23	...	48	396	2
11	68	957	443	329	55.0	0	0.0	99	81.8	104	...	151	642	0
12	74	1366	412	357	41.2	124	27.4	118	71.2	114	...	137	646	1
13	51	683	168	153	41.2	85	35.3	16	75.0	7	...	74	205	2
14	54	661	241	186	46.8	35	37.1	77	70.1	37	...	58	247	0
15	59	1244	680	490	55.7	4	0.0	186	72.0	110	...	161	774	0
16	75	1979	694	519	57.4	46	41.3	129	61.2	159	...	225	989	1
17	26	636	255	200	55.5	2	0.0	45	73.3	57	...	58	359	1
18	4	22	3	6	16.7	6	16.7	0	0.0	0	...	3	0	1
19	82	2502	1130	961	51.4	2	0.0	365	38.9	437	...	285	1705	0
20	77	2069	604	496	46.6	212	34.9	114	59.6	44	...	100	804	2
21	81	1253	355	290	50.0	34	20.6	77	75.3	37	...	103	562	2
22	67	1286	228	201	45.8	85	24.7	48	47.9	60	...	141	455	1
23	29	785	430	361	45.4	41	36.6	107	81.3	32	...	53	373	1
24	7	36	3	5	0.0	0	0.0	4	75.0	2	...	1	6	1
25	67	1583	422	355	56.3	0	0.0	42	52.4	141	...	188	1016	1
26	40	492	194	190	43.7	41	31.7	25	60.0	17	...	51	175	0
27	82	2969	1387	1137	43.7	126	31.0	466	76.0	134	...	190	1138	0
28	57	894	298	297	42.1	23	30.4	64	64.1	52	...	87	374	1
29	68	2455	1656	1199	53.5	12	8.3	461	80.5	173	...	141	2059	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
460	79	1564	567	464	48.7	84	34.5	148	58.1	139	...	144	743	0
461	76	2288	973	1005	36.8	390	31.8	145	75.2	31	...	119	790	2
462	82	2194	693	488	54.7	0	0.0	248	64.1	274	...	189	1091	0
463	47	397	176	165	37.0	118	36.4	13	84.6	8	...	40	112	1
464	75	2665	1143	926	43.6	205	34.1	363	73.0	42	...	128	1393	3

	Games Played	MIN	PTS	FGA	FG%	3PA	3P%	FTA	FT%	OREB	...	PF	EFF	AST/T(
465	33	411	121	144	36.1	43	27.9	7	71.4	7	...	36	116	1
466	74	1058	270	165	52.1	7	14.3	139	69.8	106	...	141	460	1
467	32	603	190	167	41.9	48	37.5	47	68.1	14	...	42	213	1
468	82	1731	833	619	54.9	0	0.0	186	82.3	146	...	205	1093	1
469	79	2690	1313	1165	44.7	227	30.4	291	69.4	82	...	198	1408	2
470	2	7	4	1	100.0	0	0.0	2	100.0	0	...	0	5	0
471	75	2286	771	440	66.6	0	0.0	257	72.0	294	...	169	1528	0
472	62	995	261	239	44.8	10	20.0	64	70.3	71	...	113	407	1
473	2	74	22	30	30.0	10	20.0	5	40.0	2	...	4	12	1
474	10	76	24	20	45.0	11	54.5	0	0.0	3	...	12	32	0
475	72	2573	1292	1086	43.6	248	33.9	320	81.9	51	...	187	1153	1
476	66	1091	384	406	33.3	232	29.7	57	78.9	27	...	102	327	1
477	65	1675	650	617	41.2	243	37.0	64	81.3	31	...	77	575	2
478	76	2245	753	691	41.4	259	35.1	112	80.4	67	...	162	786	1
479	60	2024	956	752	44.8	445	38.9	145	75.2	38	...	132	872	1
480	58	984	397	353	42.5	85	27.1	94	78.7	26	...	70	423	1
481	7	67	22	31	32.3	9	0.0	4	50.0	2	...	9	17	2
482	8	69	15	19	26.3	9	22.2	6	50.0	1	...	10	14	2
483	52	951	306	306	38.6	121	34.7	34	82.4	14	...	81	242	1
484	78	2471	1085	975	42.9	406	34.2	142	77.5	96	...	231	1082	1
485	9	86	20	13	23.1	0	0.0	24	58.3	2	...	6	7	1
486	77	1902	778	677	42.2	167	34.1	177	84.2	27	...	158	720	1
487	71	2304	1143	932	48.7	20	35.0	298	76.5	225	...	175	1422	0
488	73	1730	606	529	45.4	3	0.0	160	78.8	197	...	170	929	1
489	16	75	28	30	36.7	14	21.4	5	60.0	5	...	6	17	1

490 rows × 26 columns

In [13]:

```
del datafresh['Collage']
del datafresh['Team']
```

In [14]:

```
dataclean = datafresh.dropna()
```

In [15]:

```
dataclean.describe()
```

Out[15]:

	Games Played	MIN	PTS	FGA	FG%	3PA	3P
count	422.000000	422.000000	422.000000	422.000000	422.000000	422.000000	422.00000(
mean	53.748815	1246.649289	515.890995	430.000000	43.098104	117.433649	25.91658
std	24.033596	822.555115	426.260325	339.072964	9.120924	131.089094	15.3236:
min	1.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.00000(
25%	36.000000	504.250000	154.250000	146.500000	39.725000	7.000000	16.7000(
50%	62.000000	1239.000000	432.000000	370.000000	43.250000	60.000000	31.5500(
75%	74.000000	1947.000000	788.750000	656.000000	47.600000	193.000000	36.4000(
max	82.000000	2981.000000	2217.000000	1470.000000	85.700000	646.000000	66.7000(

8 rows × 24 columns

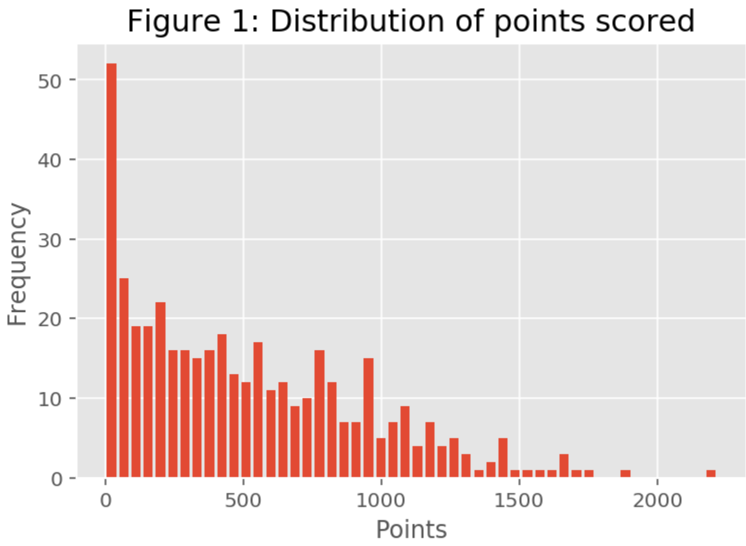
Data Exploration

As we have cleaned our data as required for our analysis, we will now continue to analysis and visual the data

Univariate Visualisation

In [16]:

```
PTS_Dist = dataclean['PTS']

plt.hist(PTS_Dist, bins = 50, rwidth=0.75)
plt.xlabel("Points")
plt.ylabel("Frequency")
plt.title('Figure 1: Distribution of points scored', fontsize = 15)

plt.show();
```



In [17]:

```
print("The median points scored in a season is" ,dataclean['PTS'].median(), "points")
print("The mean points scored in a season is", dataclean['PTS'].mean(), "points")
```

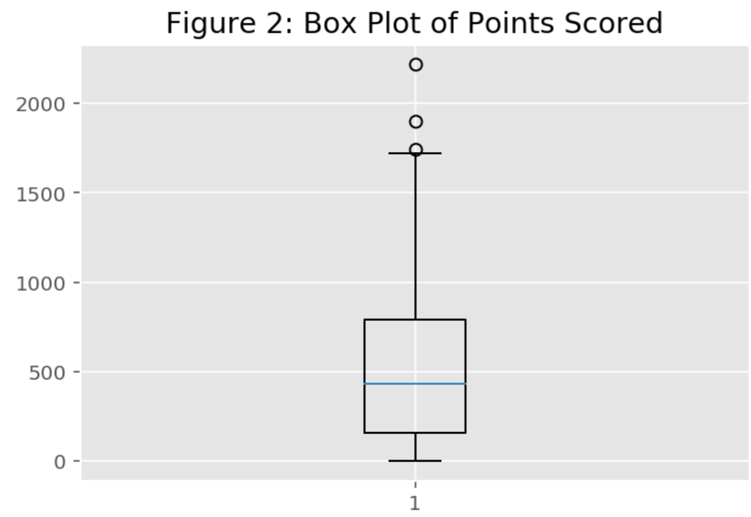
```
The median points scored in a season is 432.0 points
The mean points scored in a season is 515.8909952606635 points
```

As we can see, the distribution of points that players score during a season shows a skewness to the right, indicating that the average points that the mean is greater than the median. This suggests that the average player, i.e the player who scores 50% higher than other players does not score the average points in a season. This finding means that there are some players who score a large number of points for the season, creating a higher mean value. In the figure, this is suggested by players who scored more points than the mean of 502 points.

Next, to analyse the distribution of points in more depth, we will look at the box plot. Because of the distribution of data points, i.e. the skewness of the histrogram, a box plot will give us an idea of the dispersion of points in this set of data. Because of this robust nature of the box plot, being able to deal with the skewness of data whilst extracting meaningful information, it should provide us with more insight as to how the points scored is spread.

In [18]:

```
plt.boxplot(x = PTS_Dist)
plt.title('Figure 2: Box Plot of Points Scored')
plt.show()

print('The critical values of the box plot includes the following points of interest:')
print("the minimum points scored is",PTS_Dist.min(), "points")
print('the 25th percentile of points scored is',np.quantile(PTS_Dist, 0.25),'points')
print('the 50th percentile of points scored is',np.quantile(PTS_Dist, 0.5),'points')
print('the 75th percentile of points scored is',np.quantile(PTS_Dist, 0.75),'points')
print("the maximum points scored is",PTS_Dist.max(), "points\n")

IQR = np.quantile(PTS_Dist, 0.75)-np.quantile(PTS_Dist, 0.25)
print("Now, we can determine the interquartile range", IQR)
```




```
The critical values of the box plot includes the following points of inter
est:
the minimum points scored is 0 points
the 25th percentile of points scored is 154.25 points
the 50th percentile of points scored is 432.0 points
the 75th percentile of points scored is 788.75 points
the maximum points scored is 2217 points

Now, we can determine the interquartile range 634.5
```

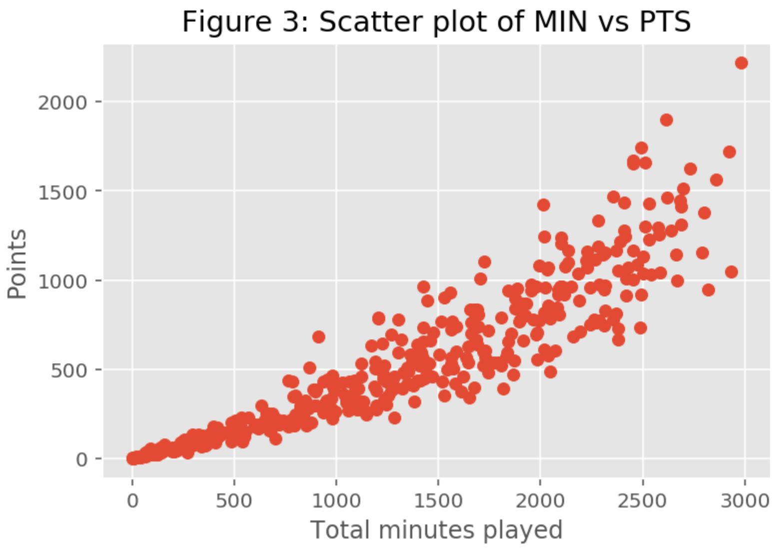
The interquanritle range of 634.5 shows us the spread of points scored by players during the season. Hence, this suggests that the spread of points from the median is 634.5 on both sides.

Bivariate Visualisation

The first bivariate plot visualisation is the of total minutes played against points. It will be a scatter plot to analyse and look at the trend of a player's points based on the number of minutes they have played in NBA.

In [19]:

```
plt.xlabel("Total minutes played")
plt.ylabel("Points")
plt.title("Figure 3: Scatter plot of MIN vs PTS")
scatter = plt.scatter(dataclean['MIN'],dataclean['PTS']);
```



Based the scatter plot above, majority of the players follow a linear pattern as the total minutes played increases. From this data we can form a hypothesis indicating that a majority of player's future result in one game will not be much different from their previous results. The outliers on this scatter plot also follow a slightly linear pattern just above or below the linear pattern of the average player. This also means that it is possible to predict the future score of outlier players.

Next we try to analyse a player's age against their total points. We suspect a peak of points at the prime age of players.

In [20]:

```
plt.xlabel("Age")
plt.ylabel("Points scored")
plt.title("Figure 4: Scatter plot of AGE vs PTS")
scatter = plt.scatter(dataclean['Age'],dataclean['PTS']);
```

The data points are scattered fairly evenly throughout the above scatter plot. There also seems to be a concentration of players in the middle age ranges. A player's age does not provide any useful measures for the points that they are expected to score.

In [21]:

```
print('Youngest player: ',dataclean['Age'].min())
print('Oldest player: ',dataclean['Age'].max())
```

```
Youngest player:  20.0
Oldest player:  39.0
```

In [22]:

```
age20 = dataclean[dataclean['Age'] == 20]
age21 = dataclean[dataclean['Age'] == 21]
age22 = dataclean[dataclean['Age'] == 22]
age23 = dataclean[dataclean['Age'] == 23]
age24 = dataclean[dataclean['Age'] == 24]
age25 = dataclean[dataclean['Age'] == 25]
age26 = dataclean[dataclean['Age'] == 26]
age27 = dataclean[dataclean['Age'] == 27]
age28 = dataclean[dataclean['Age'] == 28]
age29 = dataclean[dataclean['Age'] == 29]
age30 = dataclean[dataclean['Age'] == 30]
age31 = dataclean[dataclean['Age'] == 31]
age32 = dataclean[dataclean['Age'] == 32]
age33 = dataclean[dataclean['Age'] == 33]
age34 = dataclean[dataclean['Age'] == 34]
age35 = dataclean[dataclean['Age'] == 35]
age36 = dataclean[dataclean['Age'] == 36]
age37 = dataclean[dataclean['Age'] == 37]
age38 = dataclean[dataclean['Age'] == 38]
age39 = dataclean[dataclean['Age'] == 39]
```

In [23]:

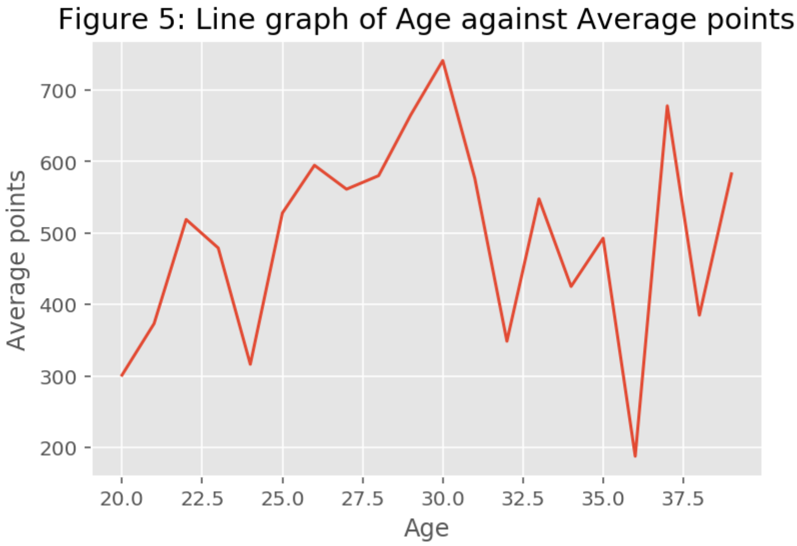
```
average20 = age20['PTS'].mean()
average21 = age21['PTS'].mean()
average22 = age22['PTS'].mean()
average23 = age23['PTS'].mean()
average24 = age24['PTS'].mean()
average25 = age25['PTS'].mean()
average26 = age26['PTS'].mean()
average27 = age27['PTS'].mean()
average28 = age28['PTS'].mean()
average29 = age29['PTS'].mean()
average30 = age30['PTS'].mean()
average31 = age31['PTS'].mean()
average32 = age32['PTS'].mean()
average33 = age33['PTS'].mean()
average34 = age34['PTS'].mean()
average35 = age35['PTS'].mean()
average36 = age36['PTS'].mean()
average37 = age37['PTS'].mean()
average38 = age38['PTS'].mean()
average39 = age39['PTS'].mean()
```

In [24]:

```
xline = [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39
]
yline = [average20, average21, average22, average23, average24, average25, average26, a
verage27, average28, average29, average30, average31, average32, average33, average34,
average35, average36, average37, average38, average39]
```

In [25]:

```
plt.plot(xline, yline);
plt.title('Figure 5: Line graph of Age against Average points')
plt.xlabel('Age')
plt.ylabel('Average points')
plt.show()
```

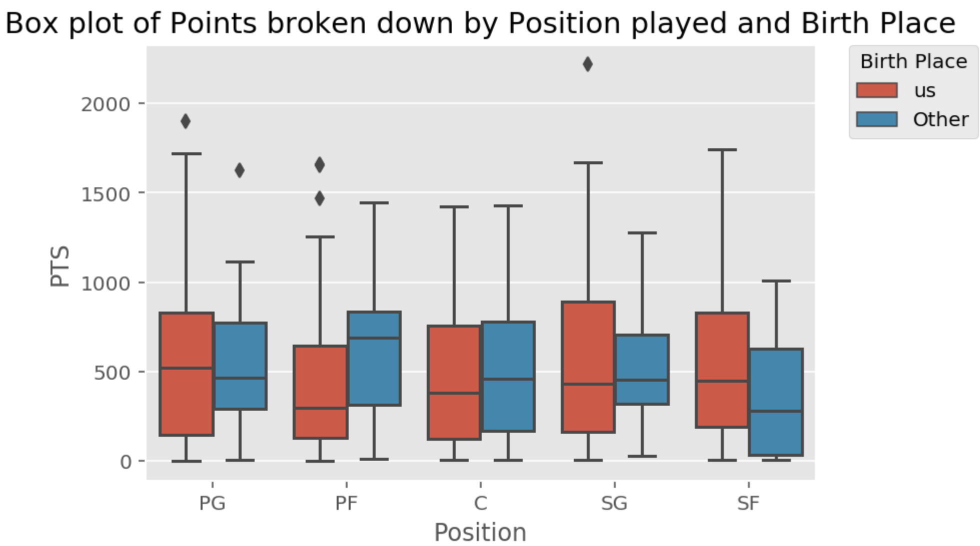


There are multiple peaks and troughs corresponding to the average points per age range. However, there does not seem to be any indicative evidence that shows that the points a player scores on average is correlated to their maturity.

Multivariate Visualisation

In [26]:

```
bp=sns.boxplot(x="Pos",
               y="PTS",
               hue="Birth_Place",
               data=dataclean);
plt.title("Box plot of Points broken down by Position played and Birth Place")

# position the legend outside the chart
bp.legend(bbox_to_anchor=(1.05, 1),loc=2, borderaxespad=0.,
          title="Birth Place");

# set the x-axis title
bp.set_xlabel("Position");
```



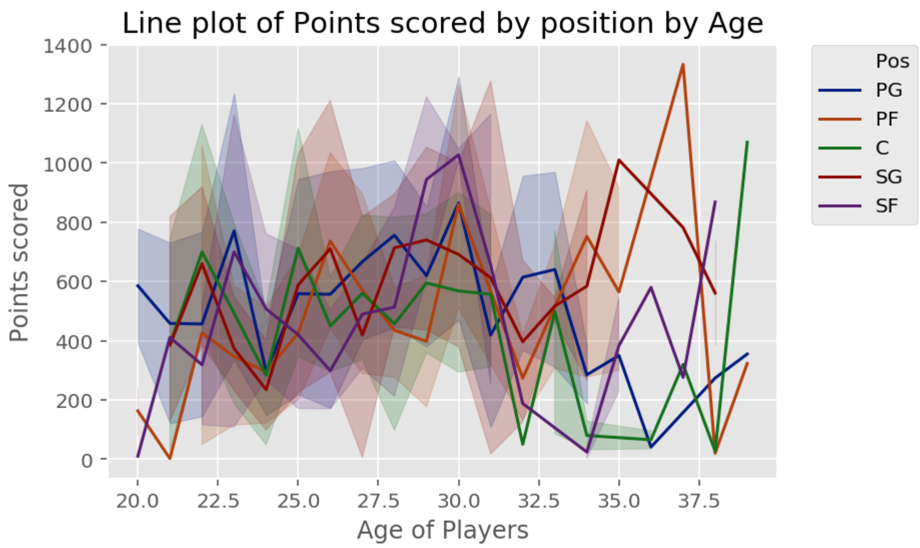
For NBA players playing the position of PG, C and SG, they seem quite uniform in that players born outside of the US can score just as many points as players playing those positions that were born in the US.

There are two box plots of interest: The first being the position of PF where players born outside the US score a bit more when juxtaposed to players born in the US. Conversely, players who were born in the US and played the SF position scored more than their counterparts born outside of the States.

Despite these differences, the main feature which is points a player scores in fairly uniform between all of these positions which suggests that regardless of a position that a player plays, they still have opportunities to score the ball.

In [27]:

```
lp = sns.lineplot(x="Age",
                  y="PTS",
                  hue="Pos",
                  palette=sns.color_palette('dark', n_colors=5),
                  data=dataclean);

# position the legend outside the chart
lp.legend(bbox_to_anchor=(1.05, 1),
          loc=2,
          borderaxespad=0.0);

# add a title and axes labels
lp.set(title='Line plot of Points scored by position by Age',
       xlabel='Age of Players',
       ylabel='Points scored');
```



All the line plots show a similar pattern where a player who reaches their mid career, say somewhere near 30 years old, will generally score more points than the average in terms of the position that they play.

We also observe that for 3 postions: PF, SG and C, they score the most points when they are near the end of their career compared to other players playing the same position.

Conversely, for players playing the positions of SF and PG, the points they score are generally lower than the other age ranges in their position.

There are several possible explanations for these trends. One is that the positions of SF and PG are more physically demanding than the other roles, where younger players are more quick and agile. Another possible explanation could be due to injuries which is associated with aging.

In [28]:

```
data.columns = [colname.replace(' ', '_') for colname in list(data.columns)]
data.columns = [colname.replace('/', '_') for colname in list(data.columns)]
data.columns = [colname.replace('3', 'Three') for colname in list(data.columns)]
data.columns = [colname.replace('%', 'P') for colname in list(data.columns)]
data.head()
```

Out[28]:

	Games_Played	MIN	PTS	FGA	FGP	ThreePA	ThreePP	FTA	FTP	OREB	...	PF	E
0	26	324	133	137	37.2	57	26.3	24	66.7	6	...	15	1
1	82	1885	954	817	42.1	313	38.7	174	83.3	32	...	189	7
2	47	797	243	208	44.7	48	27.1	61	72.1	46	...	83	3
3	32	740	213	220	41.4	9	11.1	46	65.2	48	...	88	2
4	76	2318	1156	965	53.8	36	30.6	141	75.9	131	...	121	15

5 rows × 26 columns

In [29]:

```
regression = data
regression.head
```

Out[29]:

```
<bound method NDFrame.head of      Games_Played   MIN   PTS   FGA    FGP  
ThreePA  ThreePP  FTA   FTP  \
0              26   324   133   137   37.2      57    26.3   24   66.7
1              82  1885   954   817   42.1     313    38.7  174   83.3
2              47   797   243   208   44.7      48    27.1   61   72.1
3              32   740   213   220   41.4       9    11.1   46   65.2
4              76  2318  1156   965   53.8      36    30.6  141   75.9
5              65  1992  1082  1010   48.1       5    40.0  165   65.5
6              74  1744   545   440   44.3     210    34.8  101   81.2
7              27   899   374   300   40.3      68    38.2  129   82.2
8               5    14     4     4   25.0       0     0.0    2  100.0
9              69  1518   432   353   50.7       3    33.3  104   70.2
10             42   767   434   335   39.7     139    33.8  150   80.7
11             68   957   443   329   55.0       0     0.0   99   81.8
12             74  1366   412   357   41.2     124    27.4  118   71.2
13             51   683   168   153   41.2      85    35.3   16   75.0
14             54   661   241   186   46.8      35    37.1   77   70.1
15             59  1244   680   490   55.7       4     0.0  186   72.0
16             75  1979   694   519   57.4      46    41.3  129   61.2
17             26   636   255   200   55.5       2     0.0   45   73.3
18              4    22     3     6   16.7       6    16.7    0    0.0
19             82  2502  1130   961   51.4       2     0.0  365   38.9
20             77  2069   604   496   46.6     212    34.9  114   59.6
21             81  1253   355   290   50.0      34    20.6   77   75.3
22             67  1286   228   201   45.8      85    24.7   48   47.9
23             29   785   430   361   45.4      41    36.6  107   81.3
24              7    36     3     5    0.0       0     0.0    4   75.0
25             67  1583   422   355   56.3       0     0.0   42   52.4
26             40   492   194   190   43.7      41    31.7   25   60.0
27             82  2969  1387  1137   43.7     126    31.0  466   76.0
28             57   894   298   297   42.1      23    30.4   64   64.1
29             68  2455  1656  1199   53.5      12     8.3  461   80.5
..            ...   ...   ...   ...    ...     ...     ...  ...    ...
460            79  1564   567   464   48.7      84    34.5  148   58.1
461            76  2288   973  1005   36.8     390    31.8  145   75.2
462            82  2194   693   488   54.7       0     0.0  248   64.1
463            47   397   176   165   37.0     118    36.4   13   84.6
464            75  2665  1143   926   43.6     205    34.1  363   73.0
465            33   411   121   144   36.1      43    27.9    7   71.4
466            74  1058   270   165   52.1       7    14.3  139   69.8
467            32   603   190   167   41.9      48    37.5   47   68.1
468            82  1731   833   619   54.9       0     0.0  186   82.3
469            79  2690  1313  1165   44.7     227    30.4  291   69.4
470             2     7     4     1  100.0       0     0.0    2  100.0
471            75  2286   771   440   66.6       0     0.0  257   72.0
472            62   995   261   239   44.8      10    20.0   64   70.3
473             2    74    22    30   30.0      10    20.0    5   40.0
474            10    76    24    20   45.0      11    54.5    0    0.0
475            72  2573  1292  1086   43.6     248    33.9  320   81.9
476            66  1091   384   406   33.3     232    29.7   57   78.9
477            65  1675   650   617   41.2     243    37.0   64   81.3
478            76  2245   753   691   41.4     259    35.1  112   80.4
479            60  2024   956   752   44.8     445    38.9  145   75.2
480            58   984   397   353   42.5      85    27.1   94   78.7
481             7    67    22    31   32.3       9     0.0    4   50.0
482             8    69    15    19   26.3       9    22.2    6   50.0
483            52   951   306   306   38.6     121    34.7   34   82.4
484            78  2471  1085   975   42.9     406    34.2  142   77.5
485             9    86    20    13   23.1       0     0.0   24   58.3
```

```
486           77  1902   778   677   42.2      167     34.1  177   84.2
487           71  2304  1143   932   48.7       20     35.0  298   76.5
488           73  1730   606   529   45.4        3      0.0  160   78.8
489           16    75    28    30   36.7       14     21.4    5   60.0

     OREB  ...   PF   EFF  AST_TOV  STL_TOV   Age  Birth_Place  Height  Po
s  \
0       6  ...   15   110     3.29     0.50  29.0           us   185.0   P
G
1      32  ...  189   791     1.66     0.34  30.0           us   180.0   P
G
2      46  ...   83   318     0.87     0.55  20.0           us   202.5   P
F
3      48  ...   88   244     0.68     0.43  24.0           us   205.0   P
F
4     131  ...  121  1530     2.44     0.68  29.0        Other   205.0
C
5      99  ...  139  1225     1.66     0.69  30.0           us   205.0
C
6      31  ...  148   569     1.38     0.93  33.0           us   195.0   S
G
7      19  ...   64   338     1.58     0.33  24.0           us   195.0   S
G
8       1  ...    1     3     0.00     0.00  24.0           us   210.0
C
9     142  ...  213   778     0.43     0.46  22.0        Other   212.5
C
10     23  ...   48   396     2.21     0.63  27.0        Other   195.0   S
G
11    104  ...  151   642     0.68     0.30  27.0        Other   215.0
C
12    114  ...  137   646     1.07     1.27  25.0           us   202.5   S
F
13      7  ...   74   205     2.60     1.27  23.0           us   195.0   S
G
14     37  ...   58   247     0.76     1.06  28.0           us   195.0   S
F
15    110  ...  161   774     0.58     0.37   NaN        Other     NaN  Na
N
16    159  ...  225   989     1.06     0.40  28.0           us   202.5   P
F
17     57  ...   58   359     1.00     0.54   NaN        Other     NaN  Na
N
18      0  ...    3     0     1.00     0.00  24.0           us   192.5   S
G
19    437  ...  285  1705     0.46     0.61  22.0           us   207.5
C
20     44  ...  100   804     2.59     1.01  31.0           us   195.0   S
F
21     37  ...  103   562     2.73     0.31  39.0           us   187.5   P
G
22     60  ...  141   455     1.52     1.15  24.0           us   197.5   S
G
23     32  ...   53   373     1.10     0.05  30.0        Other   210.0
C
24      2  ...    1     6     1.00     1.00  34.0        Other   202.5   S
F
25    141  ...  188  1016     1.70     0.37  31.0        Other   210.0
C
26     17  ...   51   175     0.92     0.25  26.0        Other   202.5   P
F
```

```
27    134  ...  190  1138     0.96     0.49   NaN       Other     NaN  Na
N
28     52  ...   87   374     1.33     0.75  22.0       Other   200.0   P
F
29    173  ...  141  2059     1.57     1.05  22.0          us   205.0   P
F
..    ...  ...  ...   ...      ...      ...   ...         ...     ...
...
460   139  ...  144   743     0.94     0.49  28.0          us   200.0   P
F
461    31  ...  119   790     2.65     0.52  23.0          us   182.5   P
G
462   274  ...  189  1091     0.47     0.41   NaN       Other     NaN  Na
N
463     8  ...   40   112     1.10     0.30  24.0          us   190.0   S
G
464    42  ...  128  1393     3.89     0.50  28.0          us   177.5   P
G
465     7  ...   36   116     1.74     0.43  21.0       Other   187.5   P
G
466   106  ...  141   460     1.05     1.50  30.0          us   202.5   P
F
467    14  ...   42   213     1.40     1.10  23.0          us   190.0   S
G
468   146  ...  205  1093     1.49     0.24   NaN       Other     NaN  Na
N
469    82  ...  198  1408     2.12     0.41  26.0          us   195.0   S
G
470     0  ...    0     5     0.00     0.00   NaN       Other     NaN  Na
N
471   294  ...  169  1528     0.80     0.40  33.0          us   212.5
C
472    71  ...  113   407     1.12     0.51  35.0          us   200.0   P
F
473     2  ...    4    12     1.33     0.50  23.0          us   190.0   S
G
474     3  ...   12    32     0.25     0.25  27.0       Other   202.5   S
F
475    51  ...  187  1153     1.45     0.59  23.0          us   190.0   S
G
476    27  ...  102   327     1.84     1.00  38.0          us   195.0   S
G
477    31  ...   77   575     2.10     0.66  28.0          us   190.0   S
G
478    67  ...  162   786     1.43     0.68  28.0          us   197.5   S
F
479    38  ...  132   872     1.72     0.95  29.0          us   192.5   S
G
480    26  ...   70   423     1.33     0.79  24.0          us   195.0   S
G
481     2  ...    9    17     2.25     0.13  32.0          us   180.0   S
G
482     1  ...   10    14     2.00     1.50  24.0          us   180.0   P
G
483    14  ...   81   242     1.42     0.54  34.0          us   190.0   S
G
484    96  ...  231  1082     1.22     0.53  28.0          us   200.0   S
F
485     2  ...    6     7     1.00     1.00  24.0       Other   195.0   S
F
486    27  ...  158   720     1.43     0.28  20.0          us   192.5   P
```

```
G
487   225  ...  175  1422     0.98     0.44  34.0          us   202.5   P
F
488   197  ...  170   929     1.34     0.60  31.0       Other   207.5
C
489     5  ...    6    17     1.00     0.40  26.0       Other   192.5   S
G

     Weight        BMI
0     81.45  23.798393
1     72.45  22.361111
2     99.00  24.142661
3    106.65  25.377751
4    110.25  26.234384
5    130.05  30.945866
6     99.00  26.035503
7     96.30  25.325444
8    110.25  25.000000
9    117.00  25.910035
10    85.50  22.485207
11   111.60  24.142780
12    99.00  24.142661
13    94.50  24.852071
14   101.25  26.627219
15      NaN        NaN
16   108.00  26.337449
17      NaN        NaN
18    96.75  26.108956
19   125.55  29.159530
20    96.75  25.443787
21    90.00  25.600000
22    94.50  24.226887
23   110.25  25.000000
24    99.00  24.142661
25   117.00  26.530612
26   112.50  27.434842
27      NaN        NaN
28   110.25  27.562500
29   113.85  27.091017
..      ...        ...
460  102.60  25.650000
461   85.95  25.805967
462     NaN        NaN
463   92.25  25.554017
464   87.75  27.851617
465   87.30  24.832000
466  112.50  27.434842
467   83.70  23.185596
468     NaN        NaN
469   99.00  26.035503
470     NaN        NaN
471  108.00  23.916955
472  105.75  26.437500
473   90.00  24.930748
474  100.80  24.581619
475   94.50  26.177285
476   99.00  26.035503
477   90.00  24.930748
478   96.75  24.803717
479   99.00  26.716141
480   78.75  20.710059
```

```
481   83.25  25.694444
482   83.25  25.694444
483   90.00  24.930748
484  101.25  25.312500
485   99.00  26.035503
486   85.05  22.951594
487  117.00  28.532236
488  121.50  28.218900
489   90.00  24.287401

[490 rows x 26 columns]>
```

In [30]:

```
data_encoded = pd.get_dummies(regression, drop_first=True)
data_encoded.head()
```

Out[30]:

	Games_Played	MIN	PTS	FGA	FGP	ThreePA	ThreePP	FTA	FTP	OREB	...	STL_TC
0	26	324	133	137	37.2	57	26.3	24	66.7	6	...	0.5
1	82	1885	954	817	42.1	313	38.7	174	83.3	32	...	0.3
2	47	797	243	208	44.7	48	27.1	61	72.1	46	...	0.5
3	32	740	213	220	41.4	9	11.1	46	65.2	48	...	0.4
4	76	2318	1156	965	53.8	36	30.6	141	75.9	131	...	0.6

5 rows × 29 columns

In [31]:

```
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Age + Height + Weight + BMI + Birth_Place_us + Pos_P
F + Pos_PG + Pos_SF + Pos_SG
```

Feature Selection

The below shows our regressional table with all our variables included in the plot. We will employ backward selection to remove variables with p-values greater than a 0.05 significance level and retrieve a new regressional formula for each iteration.

In [32]:

```
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA
+ ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF
+ EFF + AST_TOV + STL_TOV + Age + Height + Weight + BMI + Birth_Place_u
s + Pos_PF + Pos_PG + Pos_SF + Pos_SG
                            OLS Regression Results
=======================================================================
=======
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 1.
441e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:
-1212.8
No. Observations:                 422   AIC:
2482.
Df Residuals:                     394   BIC:
2595.
Df Model:                          27
Covariance Type:            nonrobust
=======================================================================
===========
                   coef    std err          t      P>|t|      [0.025
0.975]
-----------------------------------------------------------------------
-----------
Intercept        0.8244     70.669      0.012      0.991    -138.112
139.761
Games_Played    -0.0466      0.024     -1.954      0.051      -0.093
0.000
MIN              0.0024      0.001      1.626      0.105      -0.000
0.005
FGA              0.6568      0.003    211.072      0.000       0.651
0.663
FGP              0.1524      0.033      4.587      0.000       0.087
0.218
ThreePA          0.1277      0.003     36.859      0.000       0.121
0.135
ThreePP          0.0524      0.020      2.657      0.008       0.014
0.091
FTA              0.4099      0.005     79.167      0.000       0.400
0.420
FTP             -0.0939      0.014     -6.790      0.000      -0.121
-0.067
OREB            -0.1855      0.008    -24.596      0.000      -0.200
-0.171
DREB            -0.2315      0.006    -39.005      0.000      -0.243
-0.220
REB             -0.4170      0.004   -106.513      0.000      -0.425
-0.409
AST             -0.6512      0.007    -90.021      0.000      -0.665
-0.637
STL             -0.6587      0.018    -37.378      0.000      -0.693
-0.624
BLK             -0.6489      0.014    -48.011      0.000      -0.676
-0.622
TOV              0.6430      0.017     38.457      0.000       0.610
0.676
```

```
PF                 0.0129       0.010       1.304       0.193       -0.007
0.032
EFF                0.6571       0.004     167.382       0.000        0.649
0.665
AST_TOV           -0.6467       0.460      -1.405       0.161       -1.551
0.258
STL_TOV            0.7609       0.849       0.896       0.371       -0.909
2.431
Age                0.0291       0.056       0.517       0.606       -0.082
0.140
Height            -0.0355       0.354      -0.100       0.920       -0.731
0.660
Weight             0.0182       0.347       0.052       0.958       -0.664
0.701
BMI                0.0579       1.384       0.042       0.967       -2.663
2.778
Birth_Place_us     0.5599       0.570       0.982       0.327       -0.561
1.680
Pos_PF            -0.1588       0.865      -0.184       0.854       -1.858
1.541
Pos_PG             1.5718       1.840       0.854       0.393       -2.045
5.189
Pos_SF             0.6418       1.200       0.535       0.593       -1.717
3.001
Pos_SG             1.8263       1.433       1.275       0.203       -0.990
4.643
========================================================================
=======
Omnibus:                      142.463   Durbin-Watson:
2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
838.821
Skew:                           1.315   Prob(JB):                     7.
12e-183
Kurtosis:                       9.386   Cond. No.
1.20e+16
========================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The smallest eigenvalue is 9.59e-24. This might indicate that there
are
strong multicollinearity problems or that the design matrix is singula
r.
```

Removing BMI

In [33]:

```
data_encoded = data_encoded.drop(columns='BMI')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA
+ ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF
+ EFF + AST_TOV + STL_TOV + Age + Height + Weight + Birth_Place_us + Po
s_PF + Pos_PG + Pos_SF + Pos_SG
                            OLS Regression Results
=======================================================================
=======
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 1.
500e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:
-1212.8
No. Observations:                 422   AIC:
2480.
Df Residuals:                     395   BIC:
2589.
Df Model:                          26
Covariance Type:            nonrobust
=======================================================================
===========
                     coef    std err          t      P>|t|      [0.025
0.975]
-----------------------------------------------------------------------
-----------
Intercept          3.7291     13.134      0.284      0.777     -22.093
29.551
Games_Played      -0.0466      0.024     -1.956      0.051      -0.093
0.000
MIN                0.0024      0.001      1.630      0.104      -0.000
0.005
FGA                0.6568      0.003    211.803      0.000       0.651
0.663
FGP                0.1523      0.033      4.598      0.000       0.087
0.217
ThreePA            0.1277      0.003     36.977      0.000       0.121
0.135
ThreePP            0.0523      0.020      2.660      0.008       0.014
0.091
FTA                0.4099      0.005     79.332      0.000       0.400
0.420
FTP               -0.0938      0.014     -6.806      0.000      -0.121
-0.067
OREB              -0.1855      0.008    -24.628      0.000      -0.200
-0.171
DREB              -0.2315      0.006    -39.054      0.000      -0.243
-0.220
REB               -0.4170      0.004   -106.690      0.000      -0.425
-0.409
AST               -0.6512      0.007    -90.369      0.000      -0.665
-0.637
STL               -0.6587      0.018    -37.517      0.000      -0.693
-0.624
BLK               -0.6489      0.013    -48.072      0.000      -0.675
-0.622
TOV                0.6430      0.017     38.510      0.000       0.610
0.676
```

```
PF                  0.0129      0.010      1.307      0.192      -0.007
0.032
EFF                 0.6571      0.004    167.658      0.000       0.649
0.665
AST_TOV            -0.6470      0.460     -1.408      0.160      -1.550
0.257
STL_TOV             0.7638      0.845      0.903      0.367      -0.898
2.426
Age                 0.0291      0.056      0.517      0.605      -0.082
0.140
Height             -0.0501      0.066     -0.764      0.445      -0.179
0.079
Weight              0.0326      0.039      0.831      0.407      -0.045
0.110
Birth_Place_us      0.5605      0.569      0.985      0.325      -0.558
1.679
Pos_PF             -0.1535      0.854     -0.180      0.857      -1.833
1.526
Pos_PG              1.5646      1.829      0.855      0.393      -2.032
5.161
Pos_SF              0.6484      1.188      0.546      0.586      -1.687
2.984
Pos_SG              1.8291      1.429      1.280      0.201      -0.981
4.639
=======================================================================
=======
Omnibus:                      142.331   Durbin-Watson:
2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
837.221
Skew:                           1.314   Prob(JB):                     1.
58e-182
Kurtosis:                       9.380   Cond. No.
1.20e+16
=======================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The smallest eigenvalue is 9.59e-24. This might indicate that there
are
strong multicollinearity problems or that the design matrix is singula
r.
```

Removing Pos_PF

In [34]:

```
data_encoded = data_encoded.drop(columns='Pos_PF')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Age + Height + Weight + Birth_Place_us + Pos_PG + Po
s_SF + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 1.564
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:                -12
12.8
No. Observations:                 422   AIC:                             2
478.
Df Residuals:                     396   BIC:                             2
583.
Df Model:                          25
Covariance Type:            nonrobust
==========================================================================
========
                     coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
--------
Intercept          3.0338     12.536      0.242      0.809     -21.613
27.680
Games_Played      -0.0468      0.024     -1.973      0.049      -0.094
-0.000
MIN                0.0024      0.001      1.631      0.104      -0.000
0.005
FGA                0.6568      0.003    212.152      0.000       0.651
0.663
FGP                0.1535      0.032      4.737      0.000       0.090
0.217
ThreePA            0.1277      0.003     37.037      0.000       0.121
0.135
ThreePP            0.0517      0.019      2.673      0.008       0.014
0.090
FTA                0.4098      0.005     79.644      0.000       0.400
0.420
FTP               -0.0938      0.014     -6.812      0.000      -0.121
-0.067
OREB              -0.1853      0.007    -24.828      0.000      -0.200
-0.171
DREB              -0.2316      0.006    -39.381      0.000      -0.243
-0.220
REB               -0.4169      0.004   -107.038      0.000      -0.425
-0.409
AST               -0.6512      0.007    -90.482      0.000      -0.665
-0.637
STL               -0.6587      0.018    -37.563      0.000      -0.693
-0.624
BLK               -0.6485      0.013    -48.901      0.000      -0.675
-0.622
TOV                0.6433      0.017     38.709      0.000       0.611
0.676
```

```
PF                 0.0129      0.010      1.303      0.193      -0.007
0.032
EFF                0.6571      0.004    167.864      0.000       0.649
0.665
AST_TOV           -0.6485      0.459     -1.413      0.158      -1.551
0.254
STL_TOV            0.7536      0.843      0.894      0.372      -0.903
2.410
Age                0.0301      0.056      0.536      0.592      -0.080
0.140
Height            -0.0478      0.064     -0.744      0.457      -0.174
0.078
Weight             0.0333      0.039      0.854      0.394      -0.043
0.110
Birth_Place_us     0.5466      0.563      0.971      0.332      -0.561
1.654
Pos_PG             1.7503      1.507      1.161      0.246      -1.213
4.714
Pos_SF             0.7928      0.874      0.907      0.365      -0.926
2.511
Pos_SG             1.9966      1.082      1.845      0.066      -0.131
4.124
==========================================================================
====
Omnibus:                      142.743   Durbin-Watson:
2.055
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              84
1.865
Skew:                           1.318   Prob(JB):                    1.55e
-183
Kurtosis:                       9.398   Cond. No.                     1.20
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.59e-24. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Age

In [35]:

```
data_encoded = data_encoded.drop(columns='Age')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Height + Weight + Birth_Place_us + Pos_PG + Pos_SF +
Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 1.632
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:                -12
13.0
No. Observations:                 422   AIC:                             2
476.
Df Residuals:                     397   BIC:                             2
577.
Df Model:                          24
Covariance Type:            nonrobust
==========================================================================
========
                     coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
--------
Intercept          4.5078     12.221      0.369      0.712     -19.518
28.533
Games_Played      -0.0458      0.024     -1.937      0.054      -0.092
0.001
MIN                0.0024      0.001      1.646      0.100      -0.000
0.005
FGA                0.6566      0.003    213.827      0.000       0.651
0.663
FGP                0.1533      0.032      4.737      0.000       0.090
0.217
ThreePA            0.1277      0.003     37.070      0.000       0.121
0.135
ThreePP            0.0511      0.019      2.648      0.008       0.013
0.089
FTA                0.4096      0.005     79.884      0.000       0.400
0.420
FTP               -0.0932      0.014     -6.797      0.000      -0.120
-0.066
OREB              -0.1859      0.007    -25.156      0.000      -0.200
-0.171
DREB              -0.2313      0.006    -39.565      0.000      -0.243
-0.220
REB               -0.4171      0.004   -107.815      0.000      -0.425
-0.410
AST               -0.6515      0.007    -90.886      0.000      -0.666
-0.637
STL               -0.6600      0.017    -37.988      0.000      -0.694
-0.626
BLK               -0.6494      0.013    -49.387      0.000      -0.675
-0.624
TOV                0.6440      0.017     38.912      0.000       0.611
0.677
```

```
PF                  0.0127      0.010      1.287      0.199      -0.007
0.032
EFF                 0.6574      0.004    169.352      0.000       0.650
0.665
AST_TOV            -0.6052      0.451     -1.341      0.181      -1.493
0.282
STL_TOV             0.7571      0.842      0.899      0.369      -0.898
2.412
Height             -0.0525      0.064     -0.826      0.409      -0.177
0.072
Weight              0.0354      0.039      0.914      0.361      -0.041
0.112
Birth_Place_us      0.5543      0.562      0.986      0.325      -0.551
1.660
Pos_PG              1.6347      1.491      1.097      0.273      -1.296
4.565
Pos_SF              0.7674      0.872      0.880      0.379      -0.947
2.482
Pos_SG              1.9269      1.074      1.795      0.073      -0.184
4.037
==========================================================================
====
Omnibus:                      143.572   Durbin-Watson:
2.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              84
9.930
Skew:                           1.325   Prob(JB):                    2.75e
-185
Kurtosis:                       9.427   Cond. No.                     1.20
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.58e-24. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Height

In [36]:

```
data_encoded = data_encoded.drop(columns='Height')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Weight + Birth_Place_us + Pos_PG + Pos_SF + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 1.704
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:                -12
13.3
No. Observations:                 422   AIC:                             2
475.
Df Residuals:                     398   BIC:                             2
572.
Df Model:                          23
Covariance Type:            nonrobust
==========================================================================
========
                     coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
--------
Intercept         -4.9516      4.265     -1.161      0.246     -13.336
3.433
Games_Played      -0.0451      0.024     -1.911      0.057      -0.092
0.001
MIN                0.0024      0.001      1.628      0.104      -0.000
0.005
FGA                0.6567      0.003    214.577      0.000       0.651
0.663
FGP                0.1524      0.032      4.713      0.000       0.089
0.216
ThreePA            0.1277      0.003     37.090      0.000       0.121
0.135
ThreePP            0.0506      0.019      2.625      0.009       0.013
0.089
FTA                0.4099      0.005     80.111      0.000       0.400
0.420
FTP               -0.0942      0.014     -6.899      0.000      -0.121
-0.067
OREB              -0.1854      0.007    -25.179      0.000      -0.200
-0.171
DREB              -0.2317      0.006    -39.775      0.000      -0.243
-0.220
REB               -0.4171      0.004   -107.881      0.000      -0.425
-0.409
AST               -0.6510      0.007    -91.164      0.000      -0.665
-0.637
STL               -0.6601      0.017    -38.014      0.000      -0.694
-0.626
BLK               -0.6508      0.013    -49.930      0.000      -0.676
-0.625
TOV                0.6426      0.016     39.057      0.000       0.610
0.675
PF                 0.0129      0.010      1.312      0.190      -0.006
```

```
0.032
EFF                 0.6574      0.004    169.420      0.000       0.650
0.665
AST_TOV            -0.6051      0.451     -1.341      0.181      -1.492
0.282
STL_TOV             0.8207      0.838      0.980      0.328      -0.827
2.468
Weight              0.0239      0.036      0.661      0.509      -0.047
0.095
Birth_Place_us      0.6398      0.553      1.158      0.248      -0.447
1.726
Pos_PG              2.3001      1.254      1.835      0.067      -0.164
4.765
Pos_SF              0.8669      0.863      1.004      0.316      -0.830
2.564
Pos_SG              2.3155      0.965      2.400      0.017       0.419
4.212
==========================================================================
====
Omnibus:                      142.998   Durbin-Watson:
2.044
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               85
4.002
Skew:                           1.316   Prob(JB):                     3.60e
-186
Kurtosis:                       9.453   Cond. No.                     4.15
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 8.01e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Weight

In [37]:

```
data_encoded = data_encoded.drop(columns='Weight')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Birth_Place_us + Pos_PG + Pos_SF + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.134
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:                -13
99.0
No. Observations:                 490   AIC:                             2
844.
Df Residuals:                     467   BIC:                             2
941.
Df Model:                          22
Covariance Type:            nonrobust
==========================================================================
========
                     coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
--------
Intercept         -0.9852      1.290     -0.763      0.446      -3.521
1.551
Games_Played      -0.0369      0.021     -1.768      0.078      -0.078
0.004
MIN                0.0024      0.001      1.902      0.058   -8.09e-05
0.005
FGA                0.6592      0.003    239.295      0.000       0.654
0.665
FGP                0.1426      0.026      5.399      0.000       0.091
0.195
ThreePA            0.1283      0.003     41.081      0.000       0.122
0.134
ThreePP            0.0505      0.016      3.132      0.002       0.019
0.082
FTA                0.4090      0.005     87.259      0.000       0.400
0.418
FTP               -0.0951      0.012     -7.796      0.000      -0.119
-0.071
OREB              -0.1844      0.007    -27.229      0.000      -0.198
-0.171
DREB              -0.2324      0.005    -43.729      0.000      -0.243
-0.222
REB               -0.4169      0.004   -117.229      0.000      -0.424
-0.410
AST               -0.6528      0.006   -101.644      0.000      -0.665
-0.640
STL               -0.6488      0.016    -41.228      0.000      -0.680
-0.618
BLK               -0.6357      0.012    -53.712      0.000      -0.659
-0.612
TOV                0.6524      0.015     43.129      0.000       0.623
0.682
PF                 0.0038      0.009      0.434      0.664      -0.013
```

```
0.021
EFF                 0.6543      0.003    190.251      0.000       0.647
0.661
AST_TOV            -0.4133      0.383     -1.079      0.281      -1.166
0.339
STL_TOV             0.6039      0.759      0.796      0.426      -0.887
2.095
Birth_Place_us      0.1232      0.466      0.264      0.792      -0.793
1.039
Pos_PG              0.8048      0.811      0.992      0.322      -0.789
2.399
Pos_SF             -0.1095      0.699     -0.157      0.876      -1.483
1.264
Pos_SG              1.0210      0.676      1.511      0.131      -0.307
2.349
==========================================================================
====
Omnibus:                      181.718   Durbin-Watson:
2.008
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             133
9.898
Skew:                           1.417   Prob(JB):                    1.11e
-291
Kurtosis:                      10.590   Cond. No.                     3.90
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 1.01e-24. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Pos_SF

In [38]:

```
data_encoded = data_encoded.drop(columns='Pos_SF')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Birth_Place_us + Pos_PG + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.241
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:15   Log-Likelihood:                -13
99.0
No. Observations:                 490   AIC:                             2
842.
Df Residuals:                     468   BIC:                             2
934.
Df Model:                          21
Covariance Type:            nonrobust
==========================================================================
========
                     coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
--------
Intercept         -1.0155      1.275     -0.797      0.426      -3.520
1.489
Games_Played      -0.0370      0.021     -1.771      0.077      -0.078
0.004
MIN                0.0024      0.001      1.915      0.056   -6.18e-05
0.005
FGA                0.6593      0.003    243.652      0.000       0.654
0.665
FGP                0.1432      0.026      5.485      0.000       0.092
0.195
ThreePA            0.1283      0.003     41.198      0.000       0.122
0.134
ThreePP            0.0500      0.016      3.166      0.002       0.019
0.081
FTA                0.4090      0.005     87.352      0.000       0.400
0.418
FTP               -0.0950      0.012     -7.803      0.000      -0.119
-0.071
OREB              -0.1843      0.007    -27.336      0.000      -0.198
-0.171
DREB              -0.2324      0.005    -43.784      0.000      -0.243
-0.222
REB               -0.4168      0.004   -118.335      0.000      -0.424
-0.410
AST               -0.6527      0.006   -102.536      0.000      -0.665
-0.640
STL               -0.6488      0.016    -41.279      0.000      -0.680
-0.618
BLK               -0.6356      0.012    -53.841      0.000      -0.659
-0.612
TOV                0.6521      0.015     43.459      0.000       0.623
0.682
PF                 0.0040      0.009      0.470      0.638      -0.013
```

```
0.021
EFF                0.6542      0.003    190.774      0.000       0.647
0.661
AST_TOV           -0.4141      0.383     -1.082      0.280      -1.166
0.338
STL_TOV            0.5917      0.754      0.785      0.433      -0.890
2.073
Birth_Place_us     0.1041      0.450      0.232      0.817      -0.779
0.987
Pos_PG             0.8520      0.752      1.133      0.258      -0.626
2.330
Pos_SG             1.0747      0.582      1.847      0.065      -0.069
2.218
==========================================================================
====
Omnibus:                      181.421   Durbin-Watson:
2.009
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             133
5.572
Skew:                           1.414   Prob(JB):                    9.64e
-291
Kurtosis:                      10.577   Cond. No.                    3.93
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.95e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Birth_Place_Us

In [39]:

```
data_encoded = data_encoded.drop(columns='Birth_Place_us')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + PF + EFF
+ AST_TOV + STL_TOV + Pos_PG + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.358
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -13
99.1
No. Observations:                 490   AIC:                             2
840.
Df Residuals:                     469   BIC:                             2
928.
Df Model:                          20
Covariance Type:            nonrobust
==========================================================================
======
                   coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
------
Intercept       -0.9645      1.254     -0.769      0.442      -3.429
1.500
Games_Played    -0.0372      0.021     -1.784      0.075      -0.078
0.004
MIN              0.0024      0.001      1.915      0.056    -6.2e-05
0.005
FGA              0.6593      0.003    244.199      0.000       0.654
0.665
FGP              0.1435      0.026      5.507      0.000       0.092
0.195
ThreePA          0.1283      0.003     41.326      0.000       0.122
0.134
ThreePP          0.0499      0.016      3.165      0.002       0.019
0.081
FTA              0.4091      0.005     87.732      0.000       0.400
0.418
FTP             -0.0951      0.012     -7.815      0.000      -0.119
-0.071
OREB            -0.1844      0.007    -27.423      0.000      -0.198
-0.171
DREB            -0.2323      0.005    -44.055      0.000      -0.243
-0.222
REB             -0.4167      0.004   -118.595      0.000      -0.424
-0.410
AST             -0.6526      0.006   -102.703      0.000      -0.665
-0.640
STL             -0.6485      0.016    -41.381      0.000      -0.679
-0.618
BLK             -0.6356      0.012    -53.899      0.000      -0.659
-0.612
TOV              0.6517      0.015     43.717      0.000       0.622
0.681
PF               0.0040      0.009      0.474      0.635      -0.013
```

```
0.021
EFF               0.6542      0.003    191.453      0.000       0.647
0.661
AST_TOV          -0.4212      0.381     -1.105      0.270      -1.170
0.328
STL_TOV           0.6061      0.750      0.808      0.420      -0.869
2.081
Pos_PG            0.8972      0.726      1.236      0.217      -0.529
2.324
Pos_SG            1.1096      0.561      1.976      0.049       0.006
2.213
==========================================================================
====
Omnibus:                      181.882   Durbin-Watson:
2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             134
4.783
Skew:                           1.417   Prob(JB):                     9.64e
-293
Kurtosis:                      10.605   Cond. No.                     3.92
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.99e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing PF

In [40]:

```
data_encoded = data_encoded.drop(columns='PF')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + EFF + AS
T_TOV + STL_TOV + Pos_PG + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.486
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -13
99.2
No. Observations:                 490   AIC:                             2
838.
Df Residuals:                     470   BIC:                             2
922.
Df Model:                          19
Covariance Type:            nonrobust
==========================================================================
======
                   coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
------
Intercept       -0.9645      1.253     -0.770      0.442      -3.427
1.498
Games_Played    -0.0334      0.019     -1.735      0.083      -0.071
0.004
MIN              0.0025      0.001      2.020      0.044    6.72e-05
0.005
FGA              0.6592      0.003    244.541      0.000       0.654
0.665
FGP              0.1437      0.026      5.519      0.000       0.093
0.195
ThreePA          0.1284      0.003     41.552      0.000       0.122
0.135
ThreePP          0.0495      0.016      3.147      0.002       0.019
0.080
FTA              0.4088      0.005     88.429      0.000       0.400
0.418
FTP             -0.0953      0.012     -7.843      0.000      -0.119
-0.071
OREB            -0.1839      0.007    -27.769      0.000      -0.197
-0.171
DREB            -0.2324      0.005    -44.186      0.000      -0.243
-0.222
REB             -0.4163      0.003   -122.222      0.000      -0.423
-0.410
AST             -0.6529      0.006   -103.102      0.000      -0.665
-0.640
STL             -0.6476      0.016    -41.690      0.000      -0.678
-0.617
BLK             -0.6348      0.012    -54.420      0.000      -0.658
-0.612
TOV              0.6534      0.014     45.155      0.000       0.625
0.682
EFF              0.6540      0.003    192.271      0.000       0.647
```

```
0.661
AST_TOV          -0.4237      0.381     -1.113      0.266      -1.172      
0.324
STL_TOV           0.5913      0.749      0.789      0.430      -0.881      
2.063
Pos_PG            0.8983      0.725      1.239      0.216      -0.527      
2.323
Pos_SG            1.1093      0.561      1.978      0.049       0.007      
2.212
==========================================================================
====
Omnibus:                      180.155   Durbin-Watson:                    
2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             132
0.673
Skew:                           1.404   Prob(JB):                     1.66e
-287
Kurtosis:                      10.537   Cond. No.                     3.98
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.65e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing STL_TOV

In [41]:

```
data_encoded = data_encoded.drop(columns='STL_TOV')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + EFF + AS
T_TOV + Pos_PG + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.626
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -13
99.5
No. Observations:                 490   AIC:                             2
837.
Df Residuals:                     471   BIC:                             2
917.
Df Model:                          18
Covariance Type:            nonrobust
==========================================================================
======
                   coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
------
Intercept       -0.7434      1.221     -0.609      0.543      -3.142
1.656
Games_Played    -0.0320      0.019     -1.667      0.096      -0.070
0.006
MIN              0.0024      0.001      1.955      0.051   -1.17e-05
0.005
FGA              0.6592      0.003    244.706      0.000       0.654
0.664
FGP              0.1421      0.026      5.477      0.000       0.091
0.193
ThreePA          0.1283      0.003     41.574      0.000       0.122
0.134
ThreePP          0.0494      0.016      3.146      0.002       0.019
0.080
FTA              0.4089      0.005     88.523      0.000       0.400
0.418
FTP             -0.0939      0.012     -7.813      0.000      -0.118
-0.070
OREB            -0.1839      0.007    -27.775      0.000      -0.197
-0.171
DREB            -0.2325      0.005    -44.226      0.000      -0.243
-0.222
REB             -0.4164      0.003   -122.295      0.000      -0.423
-0.410
AST             -0.6533      0.006   -103.729      0.000      -0.666
-0.641
STL             -0.6406      0.013    -50.196      0.000      -0.666
-0.616
BLK             -0.6348      0.012    -54.437      0.000      -0.658
-0.612
TOV              0.6509      0.014     46.169      0.000       0.623
0.679
EFF              0.6541      0.003    192.395      0.000       0.647
```

```
0.661
AST_TOV         -0.3221      0.358     -0.900      0.369      -1.026
0.382
Pos_PG           0.7789      0.709      1.098      0.273      -0.614
2.172
Pos_SG           1.0938      0.560      1.952      0.052      -0.007
2.195
==========================================================================
====
Omnibus:                      181.163   Durbin-Watson:
2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             133
8.487
Skew:                           1.411   Prob(JB):                    2.25e
-291
Kurtosis:                      10.589   Cond. No.                    4.05
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.35e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing AST_TOV

In [42]:

```
data_encoded = data_encoded.drop(columns='AST_TOV')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + EFF + Po
s_PG + Pos_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.782
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -13
99.9
No. Observations:                 490   AIC:                             2
836.
Df Residuals:                     472   BIC:                             2
911.
Df Model:                          17
Covariance Type:            nonrobust
==========================================================================
======
                   coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
------
Intercept       -1.0152      1.183     -0.858      0.391      -3.339
1.309
Games_Played    -0.0337      0.019     -1.771      0.077      -0.071
0.004
MIN              0.0024      0.001      1.956      0.051   -1.06e-05
0.005
FGA              0.6590      0.003    245.590      0.000       0.654
0.664
FGP              0.1430      0.026      5.513      0.000       0.092
0.194
ThreePA          0.1285      0.003     41.717      0.000       0.122
0.135
ThreePP          0.0468      0.015      3.031      0.003       0.016
0.077
FTA              0.4088      0.005     88.575      0.000       0.400
0.418
FTP             -0.0947      0.012     -7.900      0.000      -0.118
-0.071
OREB            -0.1838      0.007    -27.775      0.000      -0.197
-0.171
DREB            -0.2328      0.005    -44.376      0.000      -0.243
-0.223
REB             -0.4166      0.003   -122.797      0.000      -0.423
-0.410
AST             -0.6559      0.006   -117.220      0.000      -0.667
-0.645
STL             -0.6414      0.013    -50.374      0.000      -0.666
-0.616
BLK             -0.6351      0.012    -54.503      0.000      -0.658
-0.612
TOV              0.6563      0.013     51.611      0.000       0.631
0.681
EFF              0.6544      0.003    193.342      0.000       0.648
```

```
0.661
Pos_PG           0.5956      0.679      0.877      0.381      -0.739
1.930
Pos_SG           1.0392      0.557      1.866      0.063      -0.055
2.134
==========================================================================
====
Omnibus:                   181.869   Durbin-Watson:
2.006
Prob(Omnibus):               0.000   Jarque-Bera (JB):              135
2.951
Skew:                        1.415   Prob(JB):                     1.62e
-294
Kurtosis:                   10.633   Cond. No.                     4.01
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.52e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Pos_PG

In [43]:

```
data_encoded = data_encoded.drop(columns='Pos_PG')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + EFF + Po
s_SG
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 2.957
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -14
00.3
No. Observations:                 490   AIC:                             2
835.
Df Residuals:                     473   BIC:                             2
906.
Df Model:                          16
Covariance Type:            nonrobust
==========================================================================
======
                   coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
------
Intercept       -0.8600      1.169     -0.736      0.462      -3.157
1.437
Games_Played    -0.0344      0.019     -1.808      0.071      -0.072
0.003
MIN              0.0024      0.001      2.006      0.045    4.97e-05
0.005
FGA              0.6589      0.003    245.884      0.000       0.654
0.664
FGP              0.1401      0.026      5.447      0.000       0.090
0.191
ThreePA          0.1285      0.003     41.721      0.000       0.122
0.135
ThreePP          0.0479      0.015      3.116      0.002       0.018
0.078
FTA              0.4086      0.005     88.668      0.000       0.400
0.418
FTP             -0.0934      0.012     -7.853      0.000      -0.117
-0.070
OREB            -0.1830      0.007    -27.908      0.000      -0.196
-0.170
DREB            -0.2336      0.005    -45.237      0.000      -0.244
-0.223
REB             -0.4167      0.003   -122.851      0.000      -0.423
-0.410
AST             -0.6546      0.005   -121.576      0.000      -0.665
-0.644
STL             -0.6419      0.013    -50.495      0.000      -0.667
-0.617
BLK             -0.6349      0.012    -54.509      0.000      -0.658
-0.612
TOV              0.6562      0.013     51.617      0.000       0.631
0.681
EFF              0.6544      0.003    193.395      0.000       0.648
```

```
0.661
Pos_SG            0.8747      0.524      1.668      0.096      -0.156
1.905
==========================================================================
====
Omnibus:                    184.840   Durbin-Watson:
2.007
Prob(Omnibus):                0.000   Jarque-Bera (JB):             143
7.291
Skew:                         1.428   Prob(JB):
0.00
Kurtosis:                    10.889   Cond. No.                     3.99
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 9.63e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Pos_SG

In [44]:

```
data_encoded = data_encoded.drop(columns='Pos_SG')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ Games_Played + MIN + FGA + FGP + ThreePA +
ThreePP + FTA + FTP + OREB + DREB + REB + AST + STL + BLK + TOV + EFF
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 3.142
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -14
01.8
No. Observations:                 490   AIC:                             2
836.
Df Residuals:                     474   BIC:                             2
903.
Df Model:                          15
Covariance Type:            nonrobust
==========================================================================
======
                  coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
------
Intercept      -0.6324      1.163     -0.544      0.587      -2.918
1.653
Games_Played   -0.0368      0.019     -1.937      0.053      -0.074
0.001
MIN             0.0027      0.001      2.194      0.029       0.000
0.005
FGA             0.6589      0.003    245.449      0.000       0.654
0.664
FGP             0.1381      0.026      5.367      0.000       0.088
0.189
ThreePA         0.1286      0.003     41.707      0.000       0.123
0.135
ThreePP         0.0495      0.015      3.217      0.001       0.019
0.080
FTA             0.4090      0.005     88.728      0.000       0.400
0.418
FTP            -0.0929      0.012     -7.794      0.000      -0.116
-0.069
OREB           -0.1836      0.007    -27.990      0.000      -0.197
-0.171
DREB           -0.2337      0.005    -45.175      0.000      -0.244
-0.224
REB            -0.4174      0.003   -123.785      0.000      -0.424
-0.411
AST            -0.6557      0.005   -122.468      0.000      -0.666
-0.645
STL            -0.6409      0.013    -50.379      0.000      -0.666
-0.616
BLK            -0.6345      0.012    -54.383      0.000      -0.657
-0.612
TOV             0.6568      0.013     51.583      0.000       0.632
0.682
EFF             0.6543      0.003    193.025      0.000       0.648
0.661
```

```
==============================================================================
====
Omnibus:                      189.440   Durbin-Watson:                       
2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             151
2.173
Skew:                           1.463   Prob(JB):                            
0.00
Kurtosis:                      11.094   Cond. No.                        3.88
e+16
==============================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 1.02e-24. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

Removing Games_Played

In [45]:

```
data_encoded = data_encoded.drop(columns='Games_Played')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

```
formula_string_encoded:  PTS ~ MIN + FGA + FGP + ThreePA + ThreePP + FTA +
FTP + OREB + DREB + REB + AST + STL + BLK + TOV + EFF
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 3.347
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:16   Log-Likelihood:                -14
03.7
No. Observations:                 490   AIC:                             2
837.
Df Residuals:                     475   BIC:                             2
900.
Df Model:                          14
Covariance Type:            nonrobust
==========================================================================
====
                 coef    std err          t      P>|t|      [0.025      0.
975]
--------------------------------------------------------------------------
----
Intercept     -0.7573      1.165     -0.650      0.516      -3.046
1.531
MIN            0.0013      0.001      1.290      0.198      -0.001
0.003
FGA            0.6594      0.003    245.960      0.000       0.654
0.665
FGP            0.1285      0.025      5.074      0.000       0.079
0.178
ThreePA        0.1283      0.003     41.543      0.000       0.122
0.134
ThreePP        0.0502      0.015      3.255      0.001       0.020
0.080
FTA            0.4096      0.005     88.781      0.000       0.400
0.419
FTP           -0.0964      0.012     -8.170      0.000      -0.120      -
0.073
OREB          -0.1847      0.007    -28.166      0.000      -0.198      -
0.172
DREB          -0.2334      0.005    -45.006      0.000      -0.244      -
0.223
REB           -0.4181      0.003   -124.422      0.000      -0.425      -
0.411
AST           -0.6560      0.005   -122.228      0.000      -0.667      -
0.645
STL           -0.6398      0.013    -50.197      0.000      -0.665      -
0.615
BLK           -0.6356      0.012    -54.377      0.000      -0.659      -
0.613
TOV            0.6560      0.013     51.399      0.000       0.631
0.681
EFF            0.6553      0.003    194.762      0.000       0.649
0.662
==========================================================================
====
```

```
Omnibus:                      197.327   Durbin-Watson:                  
2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             164
8.948
Skew:                           1.523   Prob(JB):                       
0.00
Kurtosis:                      11.455   Cond. No.                     4.93
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 6.3e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Removing MIN

In [46]:

```
data_encoded = data_encoded.drop(columns='MIN')
formula_string_indep_vars_encoded = ' + '.join(data_encoded.drop(columns='PTS').columns
)
formula_string_encoded = 'PTS ~ ' + formula_string_indep_vars_encoded
print('formula_string_encoded: ', formula_string_encoded)
model_full = sm.formula.ols(formula=formula_string_encoded, data=data_encoded)
###
model_full_fitted = model_full.fit()
###
print(model_full_fitted.summary())
```

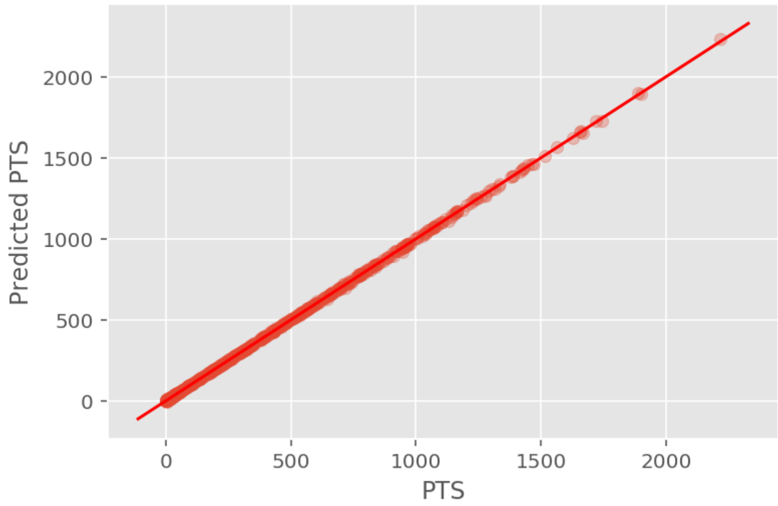
```
formula_string_encoded:  PTS ~ FGA + FGP + ThreePA + ThreePP + FTA + FTP +
OREB + DREB + REB + AST + STL + BLK + TOV + EFF
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                    PTS   R-squared:
1.000
Model:                            OLS   Adj. R-squared:
1.000
Method:                 Least Squares   F-statistic:                 3.600
e+05
Date:                Sun, 27 Oct 2019   Prob (F-statistic):
0.00
Time:                        16:07:17   Log-Likelihood:                -14
04.5
No. Observations:                 490   AIC:                             2
837.
Df Residuals:                     476   BIC:                             2
896.
Df Model:                          13
Covariance Type:            nonrobust
==========================================================================
====
                 coef    std err          t      P>|t|      [0.025      0.
975]
--------------------------------------------------------------------------
----
Intercept     -0.8593      1.163     -0.739      0.460      -3.144
1.426
FGA            0.6608      0.002    269.532      0.000       0.656
0.666
FGP            0.1322      0.025      5.253      0.000       0.083
0.182
ThreePA        0.1298      0.003     45.372      0.000       0.124
0.135
ThreePP        0.0519      0.015      3.379      0.001       0.022
0.082
FTA            0.4085      0.005     89.940      0.000       0.400
0.417
FTP           -0.0953      0.012     -8.092      0.000      -0.118      -
0.072
OREB          -0.1835      0.006    -28.244      0.000      -0.196      -
0.171
DREB          -0.2329      0.005    -45.016      0.000      -0.243      -
0.223
REB           -0.4164      0.003   -134.917      0.000      -0.422      -
0.410
AST           -0.6551      0.005   -123.128      0.000      -0.666      -
0.645
STL           -0.6345      0.012    -52.576      0.000      -0.658      -
0.611
BLK           -0.6346      0.012    -54.370      0.000      -0.658      -
0.612
TOV            0.6567      0.013     51.465      0.000       0.632
0.682
EFF            0.6549      0.003    195.157      0.000       0.648
0.662
==========================================================================
====
Omnibus:                      189.044   Durbin-Watson:
1.999
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             154
2.305
Skew:                           1.451   Prob(JB):
0.00
Kurtosis:                      11.192   Cond. No.                     3.34
e+16
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The smallest eigenvalue is 4.41e-25. This might indicate that there ar
e
strong multicollinearity problems or that the design matrix is singular.
```

In [47]:

```
def plot_line(axis, slope, intercept, **kargs):
    xmin, xmax = axis.get_xlim()
    plt.plot([xmin, xmax], [xmin*slope+intercept, xmax*slope+intercept], **kargs)

# Creating scatter plot
plt.scatter(data_encoded['PTS'], model_full_fitted.fittedvalues, alpha=0.3);
plot_line(axis=plt.gca(), slope=1, intercept=0, c="red");
plt.xlabel('PTS');
plt.ylabel('Predicted PTS');
plt.title('Figure 9: Scatterplot of PTS against predicted PTS', fontsize=15);
plt.show();
```

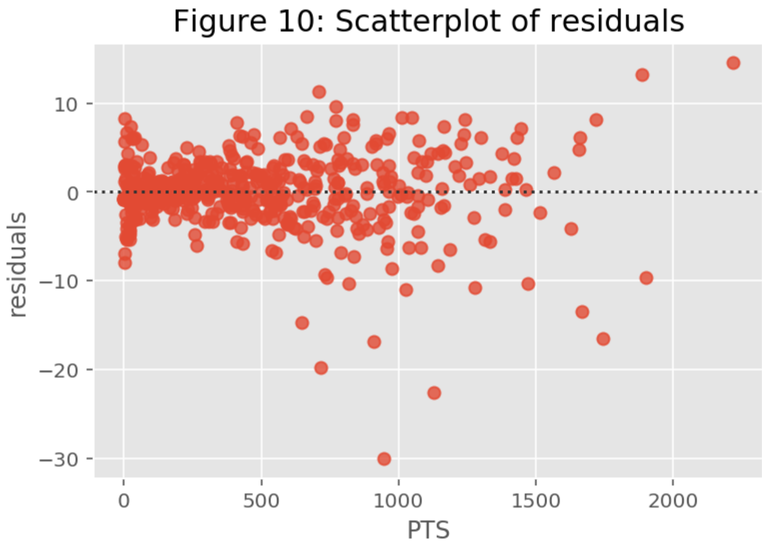
Figure 9: Scatterplot of PTS against predicted PTS



Oddly, the model returned an adjust R-squared of 1.0 which is ominously accurate. This would imply that our model explains 100% of variances. We will check the residuals to check the model

In [48]:

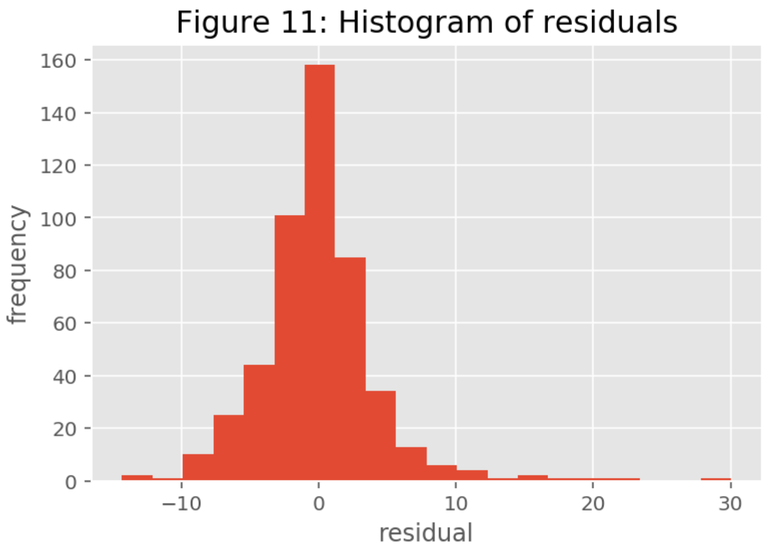
```
sns.residplot(x=data_encoded['PTS'], y=model_full_fitted.fittedvalues);
plt.ylabel('residuals')
plt.title('Figure 10: Scatterplot of residuals', fontsize=15)
plt.show();
```



The residual is fairly random is distributed along the 0 line, implying randomness. There are some deviations for a couple of points the the middle and upper reaches of points scored.

In [49]:

```
residuals = data_encoded['PTS'] - model_full_fitted.fittedvalues
plt.hist(residuals, bins = 20);
plt.xlabel('residual');
plt.ylabel('frequency');
plt.title('Figure 11: Histogram of residuals', fontsize=15);
plt.show();
```



The histrogram suggests that the residuals is normally distributed along the 0 line, backing up our observation from the scatterplot of residuals.

Summary and Conclusion

After reducing down our model to include only necessary variables, we were able to plot a multiple regression plot in which had an initial R-Squared of 1.0. After backward selection with a cut-off point at the significance level of 0.05, we eliminated 14 variables as they had a higher p-value than 0.05. Our final model includes 13 variables that is plotted against our depedent variable of points, where we also received a p-value of 1.0 for our reduced model.

The fact that we received a p-value of 1.0 as well as the fact that it was unwavering suggests that our variables are highly linked, even after removing collinearity.

This adjusted R-squared is high in value, which suggests that the features we used were either good predictors of points in the 2014-2015 NBA seasonal player statistics, or that our model is quite flawed in the sense that it does not actually predict anything.

Even though our regression model depicts the accuracy in such an ominous manner, we will still take it under deliberation with due caution.