Data mining

Projekt Grupowy

Wykonali:

Anna Czechowska305124

Oleg Łyżwiński 305158

Warszawa 2024

Tematem projektu była analiza danych dotyczących statystyk zawodników koszykarskiej ligi NBA z sezonu zasadniczego 2023/24. Dane zostały wczytane z internetowego zbioru danych:

https://www.basketball-reference.com/leagues/NBA_2024_totals.html

Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Precious Achiuwa	PF-C	24	TOT	74	18	1624	235	469	.501	26	97	.268	209	372	.562	.529	69	112	.616	191	296	487	97	46	68	83	143	565
1	Precious Achiuwa	C	24	TOR	25	0	437	78	170	.459	13	47	.277	65	123	.528	.497	24	42	.571	50	86	136	44	16	12	29	40	193
1	Precious Achiuwa	PF	24	NYK	49	18	1187	157	299	.525	13	50	.260	144	249	.578	.547	45	70	.643	141	210	351	53	30	56	54	103	372
2	Bam Adebayo	C	26	MIA	71	71	2416	530	1017	.521	15	42	.357	515	975	.528	.529	292	387	.755	159	578	737	278	81	66	162	159	1367
3	Ochai Agbaji	SG	23	TOT	78	28	1641	178	433	.411	62	211	.294	116	222	.523	.483	37	56	.661	74	142	216	83	47	44	64	117	455
3	Ochai Agbaji	SG	23	UTA	51	10	1003	106	249	.426	47	142	.331	59	107	.551	.520	15	20	.750	35	91	126	47	27	29	34	66	274
3	Ochai Agbaji	SG	23	TOR	27	18	638	72	184	.391	15	69	.217	57	115	.496	.432	22	36	.611	39	51	90	36	20	15	30	51	181
4	Santi Aldama	PF	23	MEM	61	35	1618	247	568	.435	106	304	.349	141	264	.534	.528	54	87	.621	72	280	352	138	43	54	69	89	654
5	Nickeil Alexander-Walker	SG	25	MIN	82	20	1921	236	538	.439	131	335	.391	105	203	.517	.560	52	65	.800	35	132	167	204	64	42	76	143	655
6	Grayson Allen	SG	28	PHO	75	74	2513	340	682	.499	205	445	.461	135	237	.570	.649	129	147	.878	48	247	295	227	69	45	95	157	1014
7	Jarrett Allen	C	25	CLE	77	77	2442	519	819	.634	0	6	.000	519	813	.638	.634	233	314	.742	243	568	811	210	53	81	121	147	1271

Kolejne kolumny zawierały informację o:

Rk – numer porządkowy

Player – Imię oraz nazwisko zawodnika

Pos – Pozycja na boisku określona jako:

(PG) point guard

(SG) shooting guard

(SF) small forward

(PF) power forward

(C) center.

Age – wiek zawodnika

Tm – Drużyna, w której gra zawodnik

G – liczba meczy w których zagrał dany zawodnik w czasie sezonu zasadniczego (max 82)

GS - liczba meczy w których zagrał dany zawodnik wychodząc w pierwszym składzie (max 82)

MP – Liczba rozegranych minut w sezonie

FG – Liczba trafionych rzutów

FGA – liczba oddanych rzutów

FG% - Procent trafionych rzutów

3P – Liczba trafionych rzutów za 3 punkty

3PA - liczba oddanych rzutów za 3 punkty

3P% - procent trafionych rzutów za 3 punkty

2P – Liczba trafionych rzutów za 2 punkty

2PA - liczba oddanych rzutów za 2 punkty

2P% - procent trafionych rzutów za 2 punkty

eFG% - statystyka procentowa dotycząca procentu trafień, biorąca pod uwagę czy rzut był z 3 pkt. Czy za 2 pkt.

FT – trafione rzuty wolne

FTA – oddane rzuty wolne

FT% - procent trafionych rzutów wolnych

ORB – Ofensywne zbiórki

DRB – Defensywne zbiórki

TRB – Wszystkie zbiórki

AST – Liczba podań

STL – Przechwyty

BLK – Bloki

TOV – Straty

PF – Faule osobiste

PTS – Zdobyte punkty

Prace rozpoczęto od wczytania danych, po czym przystąpiono do wstępnej obróbki oraz przygotowania do podstawowej analizy danych. W miejsca wartości nan wstawiono wartość 0, ponieważ w tym przypadku wartość nan wynikała z niemożliwości wyznaczania procentu trafień gdy zawodnik nie oddał żadnego rzutu. Usunięto również komórki zawierające same zmienne nan. W zbiorze danych występowały również kilka wierszy dotyczących jednego zawodnika, działo się tak gdy dany zawodnik zmieniał drużynę w czasie sezonu. Podawano wtedy jego statystyki w poszczególnych drużynach oraz całkowite. W zbiorze pozostawiono jedynie statystyki całkowite.

Uzyskano w ten sposób zbiór złożony z 572 obserwacji, zawierający 30 informacji o każdej obserwacji. Dane te uzupełniono o statystyki zawierające informację o średnich na mecz dla statystyk

ORBPG – Ofensywne zbiórki

DRBPG – Defensywne zbiórki

TRBPG – Wszystkie zbiórki

ASTPG – Liczba podań

STLPG – Przechwyty

BLKPG – Bloki

TOVPG – Straty

PFPG – Faule osobiste

PPG – Zdobyte punkty

MPG – Liczba rozegranych minut w sezonie

Oraz zmienne klasyfikacyjne:

Star – określająca czy dany zawodnik jest gwiazdą Ligi na podstawie meczy w których był starterem oraz liczby rozegranych meczy w sezonie

Parametrami określającymi czy dany gracz jest Gwiazdą były:

GS > 2/3G oraz G> 60

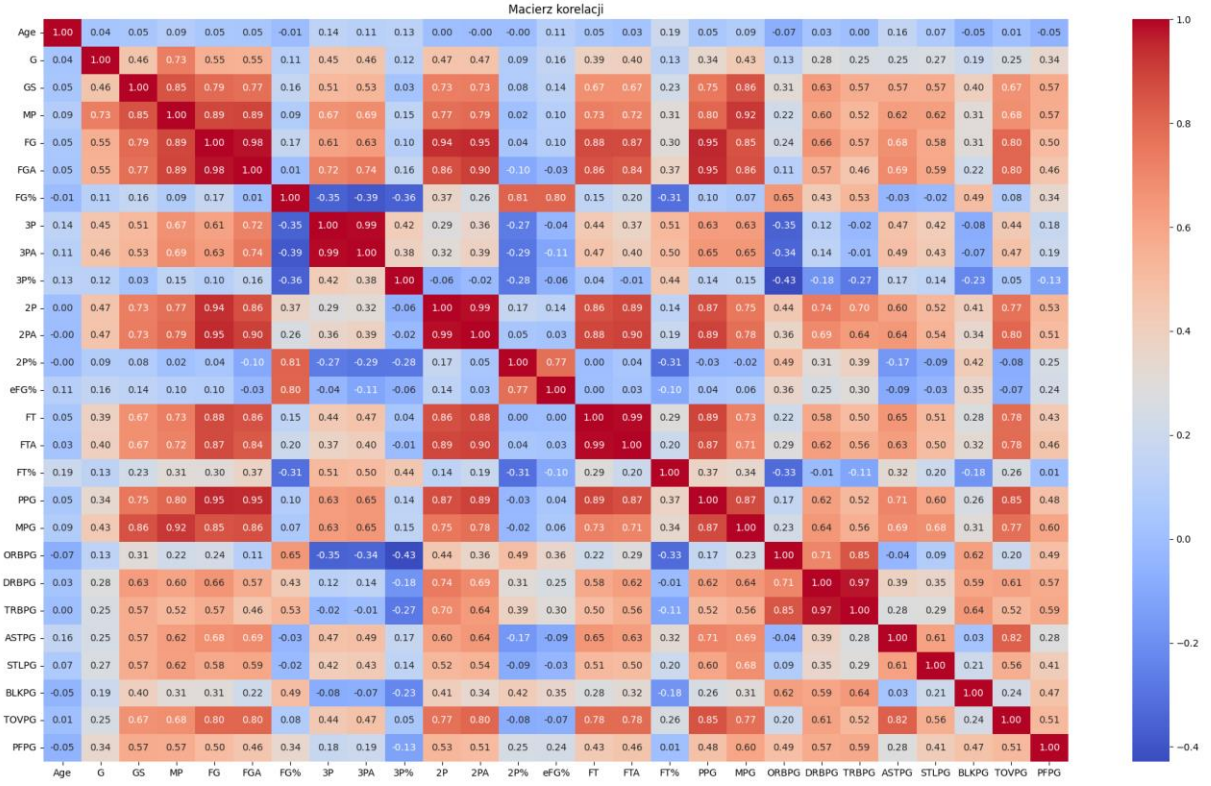
Activ – określająca czy dany zawodnik aktywnie gra w meczach na podstawie średniej liczby minut na boisku oraz liczby rozegranych punktów

Parametrami określającymi czy dany gracz jest Aktywnie grający były:

MPG> 8 oraz G> 40

Uzyskano w ten sposób 42 statystyki dla każdej obserwacji.

Do dalszej analizy i stworzenia modelu wybrano 18 statystyk. Wybrano je na podstawie korelacji Pearsona (pominięto te z największą korelacją) oraz tak by nie powtarzały się parametry od siebie bezpośrednio zależne (np. usunięto 3P%, 2P% oraz FG%, a zostawiono eFG%).



Zmniejszenie liczby zmiennych przyczyni się do poprawienia analizy danych w kolejnych krokach, ponieważ zyskujemy większy stosunek obserwacji do zmiennych. Pozwala to uniknąć tak zwanego przekleństwa wymiarowości.

Analizę chcemy przeprowadzić jedynie dla aktywnych graczy, dlatego zostawiamy obserwacje tam gdzie Active=1. To skutkuje zmianą liczby obserwacji z 572 na 344. Oznacza to że ponad 200 zawodników grało albo krócej niż 8 min na mecz, albo zagrało mniej niż w 40 meczach.

Dla tych danych obliczono podstawowe statystyki jak średnia, min, max i kwantyle:

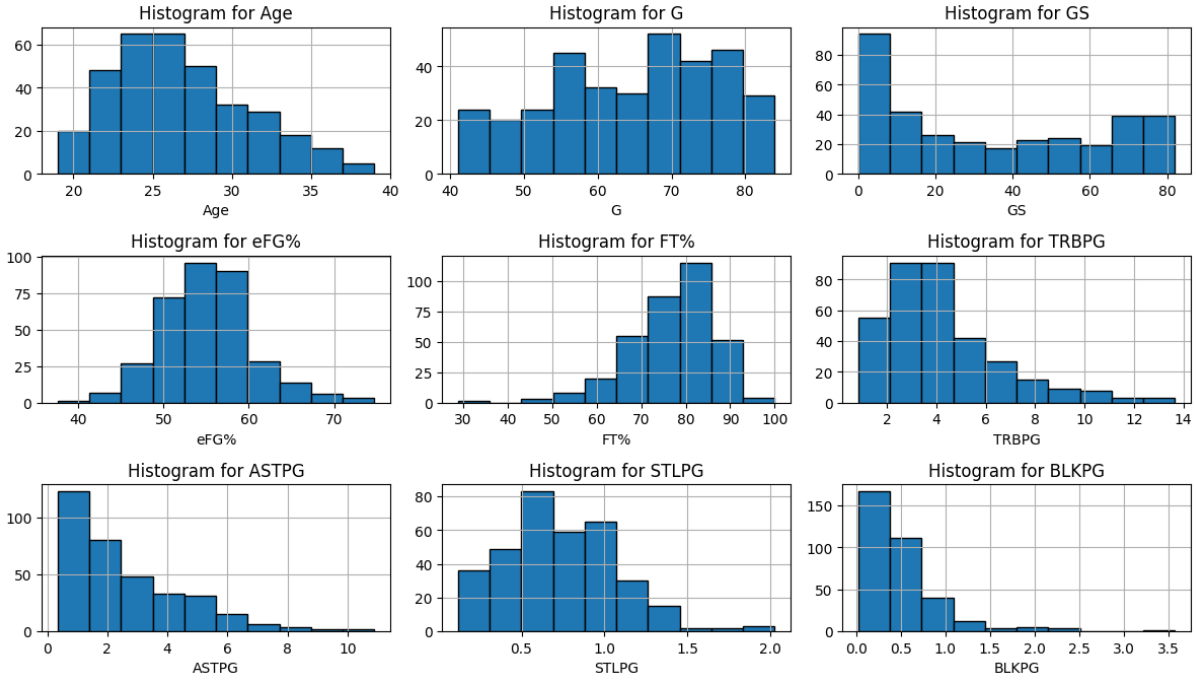
Index	Age	G	GS	eFG%	FT%	TRBPG	ASTPG	STLPG	BLKPG	TOVPG	PFPG	PPG	MPG	Star	Active
count	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344
mean	26	64,544	34,015	55,053	77,017	4,287	2,646	0,734	0,503	1,276	1,827	11,365	23,825	0	1
std	4	11,602	28,024	5,411	9,404	2,384	1,963	0,338	0,444	0,780	0,621	6,613	7,848	0	0
min	19	41,000	0,000	37,600	28,800	0,833	0,341	0,116	0,024	0,205	0,404	2,023	8,955	0	1
25%	23	55,000	7,000	51,600	71,475	2,667	1,192	0,500	0,217	0,670	1,363	6,260	17,143	0	1
50%	26	66,000	28,000	55,000	78,550	3,802	1,960	0,702	0,397	1,101	1,772	9,828	24,789	0	1
75%	29	74,250	60,250	58,100	83,300	5,287	3,630	0,927	0,633	1,660	2,230	15,268	30,524	1	1
max	39	84,000	82,000	74,700	100,000	13,659	10,899	2,027	3,577	4,352	3,561	33,857	37,835	1	1

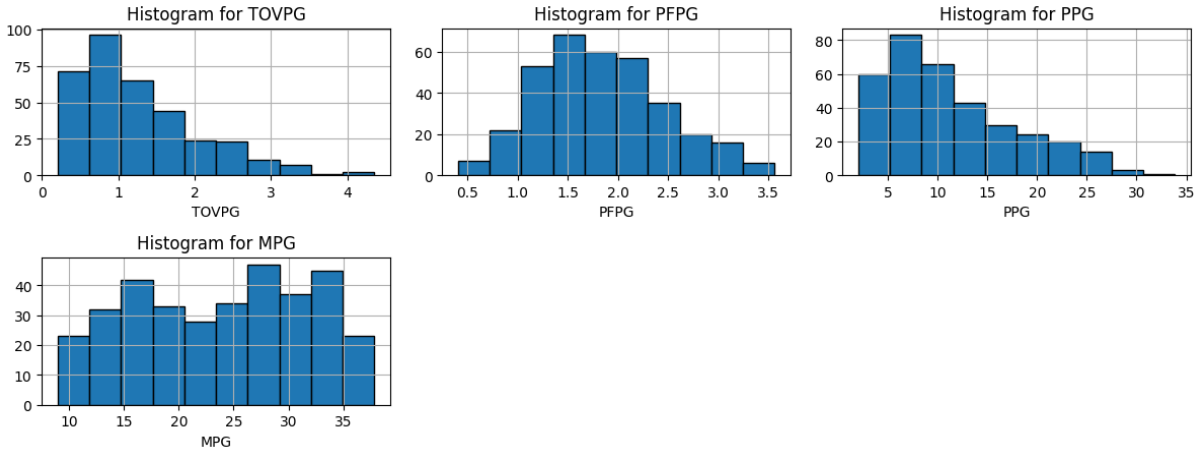
Wyznaczono rozstęp międzykwartylowy i podjęto próbę usunięcia obserwacji znajdujących się poza zakresem +/- IQR*1,5, jednak spowodowało to zmniejszeniem liczby obserwacji z 344 na 247. Jest to zbyt mocna filtracja i usuwa ok.28% danych. Dlatego pominięto ten etap.

Dla wszystkich zamiennych ilościowych obliczono wartość skośności i kurtozy oraz metodami Shapiro-Wilk i Jarque-Bera sprawdzono hipotezy o normalności rozkładu. Jak można zauważyć w poniższej tabeli, wg metod na podstawie p-value, żaden rozkład nie jest normalny. Są to jednak bardzo czułe metody. Patrząc na wartości skośności i kurtozy możemy zauważyć, że różnice wartości nie są większe od 2. Wyjątkiem jest rozkład dla BLKPG (liczba bloków na mecz), gdzie mocno zauważamy skośność prawostronną (A> 2,5) oraz leptokurtyczność (K>10).

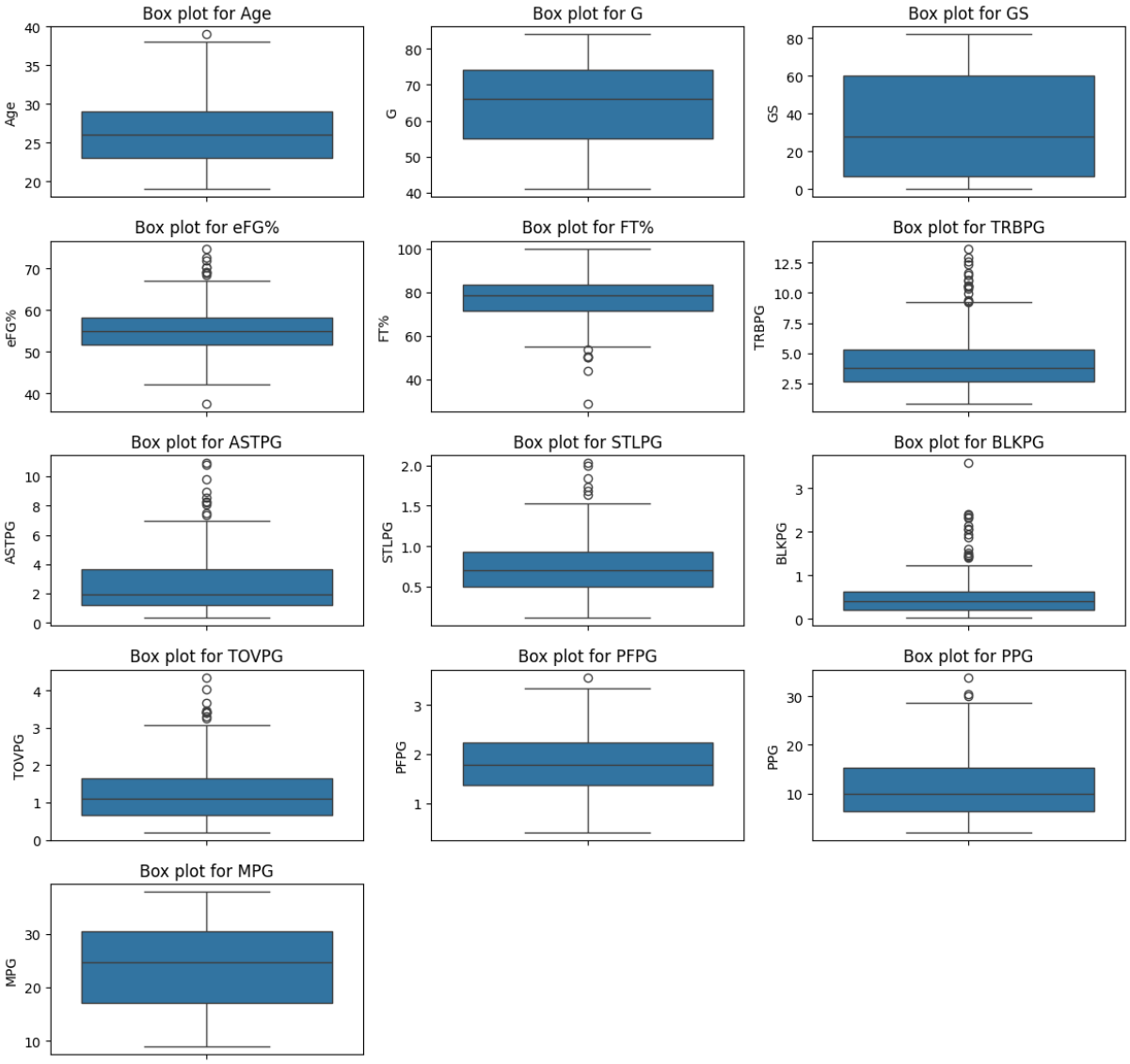
Index	Skośność	Kurtoza	Shapiro-Wilk p-value	Jarque-Bera p-value	Normalność_ S-W	Normalność_ J-B
Age	0,54852	-0,33443	1,19E-07	8,06E-05	Nie normalny	Nie normalny
G	-0,29413	-1,02015	7,18E-09	4,83E-05	Nie normalny	Nie normalny
GS	0,28399	-1,44251	3,71E-15	3,31E-08	Nie normalny	Nie normalny
eFG%	0,42163	0,95925	0,00056867	8,37E-06	Nie normalny	Nie normalny
FT%	-0,90886	1,96808	3,97E-08	4,58E-23	Nie normalny	Nie normalny
TRBPG	1,32121	1,79916	1,28E-14	1,56E-32	Nie normalny	Nie normalny
ASTPG	1,36089	1,81965	1,12E-16	4,34E-34	Nie normalny	Nie normalny
STLPG	0,66732	0,66048	1,28E-06	1,25E-07	Nie normalny	Nie normalny
BLKPG	2,58776	10,09594	1,56E-21	0	Nie normalny	Nie normalny
TOVPG	1,10329	0,96656	3,43E-13	8,66E-19	Nie normalny	Nie normalny
PFPG	0,33779	-0,37641	0,002056136	0,01375399	Nie normalny	Nie normalny
PPG	0,87602	0,02171	2,50E-12	2,78E-10	Nie normalny	Nie normalny
MPG	-0,12755	-1,17089	1,78E-08	3,39E-05	Nie normalny	Nie normalny

Poniżej przedstawiono histogramy, przedstawiające rozkłady dla wszystkich zmiennych. Na podstawie tych charakterystyk graficznych, potwierdzają się wyniki z powyższej tabeli, że nie mamy rozkładów normalnych. Jedynie zmienne Age, eFG%, STLPG oraz PFPG można uznać za w przybliżeniu normalne (wartości skośności i kurtozy <1).

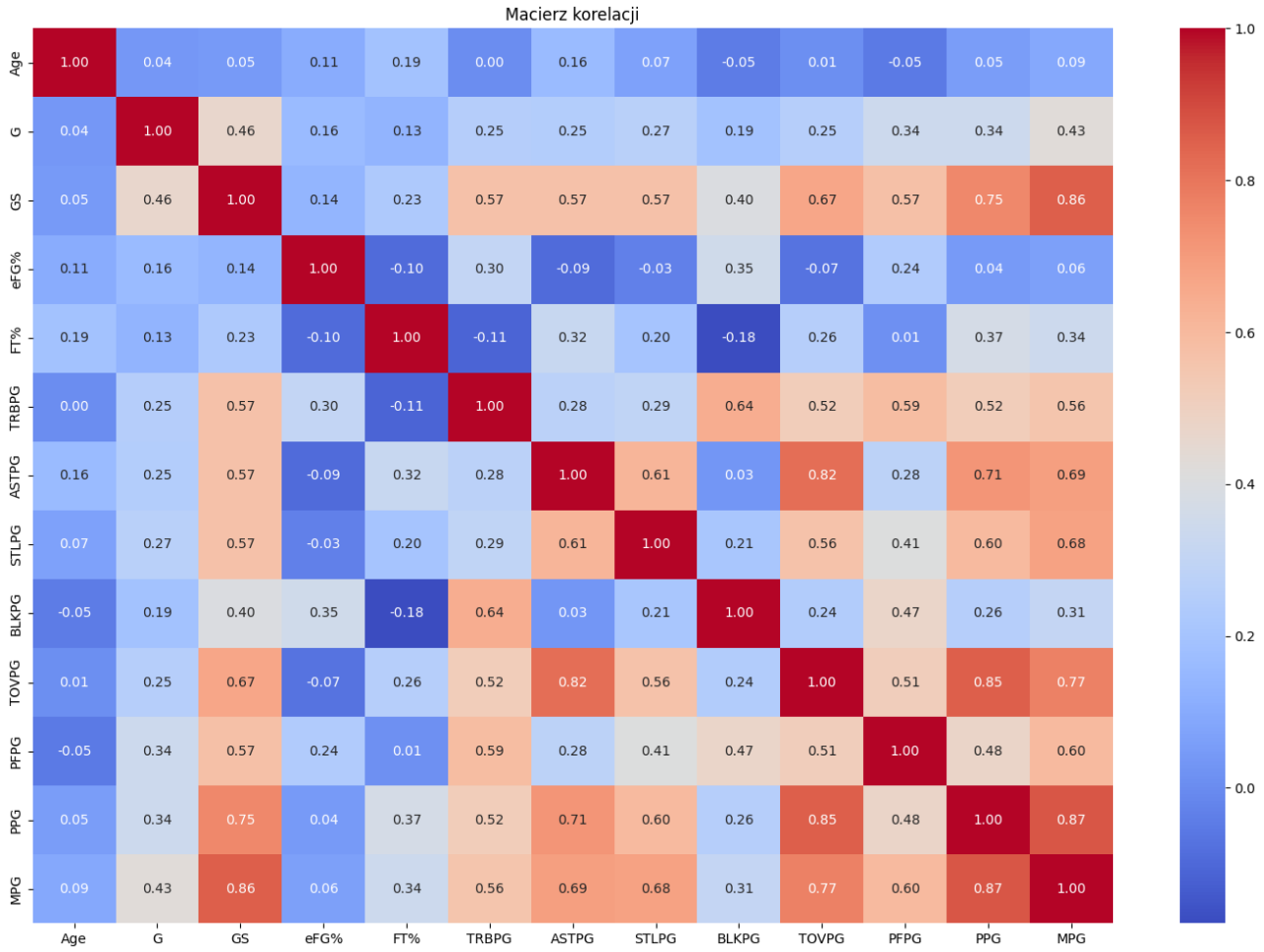




Tak samo dla każdej zmiennej, wygenerowano wykres pudełkowy, gdzie możemy zobaczyć wartości odstające oraz dobrze widoczny jest odstęp międzykwartylowy, zawierający 50% centralnych danych.



Wyznaczono macierz korelacji Pearsona, którą przedstawiono poniżej:



Celem niniejszej analizy było przewidywanie średniej liczby punktów zdobywanych przez zawodnika na podstawie pozostałych zmiennych. Na podstawie macierzy korelacji Pearsona możemy zaobserwować, że zmienna PPG (Punkty na mecz) najbardziej liniowo zależy od:

- MPG (minuty na boisku) jest to oczywista przyczynowość więcej minut na boisku daje możliwość zdobycia większej liczby punktów.
- TOVPG (straty na mecz) jest to również przyczynowość, a nie tylko korelacja, ponieważ żeby zdobywać więcej punktów, trzeba być dłużej na boisku, a będąc dłużej na boisku częściej mamy piłkę i częściej ją tracimy.
- ASTPG (asysty na mecz) obecnie w lidze NBA króluje strategia gry zwana „Small Ball" więc dużo podań i rzutów z dystansu, więc dużo asyst. Strategia ta indukuje, to że zawodnicy grający dużo minut (zdobywający dużo punktów) często asystują innym.
- GS (liczba wyjść w pierwszej piątce) również jest to przyczynowość, ponieważ jeżeli zawodnik częściej jest starterem to zdobywa więcej punktów bo gra więcej.

Do budowania modelu użyto następujące zmienne (zdecydowano o zrezygnowaniu ze zmiennej MPG przy budowaniu modelu. Uznano, że statystyki TOVPG, ASTPG oraz GS są zbyt ważnymi cechami każdego zawodnika, przyjmując tym samym duży wpływ tych wartości na przedyktowaną wartość):

```
X = df.loc[:, ['G','GS', 'eFG%', 'FT%', 'TRBPG', 'ASTPG', 'STLPG', 'BLKPG', 'TOVPG', 'PFPG']]
```

Poniżej przedstawiono statystyki modelu liniowego:

Dep. Variable:	PPG	R-squared:	0.825
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	156.6
Date:	Sat, 01 Jun 2024	Prob (F-statistic):	2.12e-119
Time:	16:00:14	Log-Likelihood:	-838.05
No. Observations:	344	AIC:	1698.
Df Residuals:	333	BIC:	1740.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-11.8630	2.284	-5.195	0.000	-16.355	-7.371
G	0.0155	0.015	1.032	0.303	-0.014	0.045
GS	0.0630	0.009	6.895	0.000	0.045	0.081
eFG%	0.0885	0.032	2.752	0.006	0.025	0.152
FT%	0.1083	0.018	5.981	0.000	0.073	0.144
TRBPG	0.2510	0.105	2.394	0.017	0.045	0.457
ASTPG	-0.3347	0.162	-2.066	0.040	-0.653	-0.016
STLPG	2.2764	0.624	3.649	0.000	1.049	3.504
BLKPG	-0.7473	0.481	-1.553	0.121	-1.694	0.199
TOVPG	5.6215	0.447	12.578	0.000	4.742	6.501
PFPG	-0.9786	0.354	-2.763	0.006	-1.675	-0.282

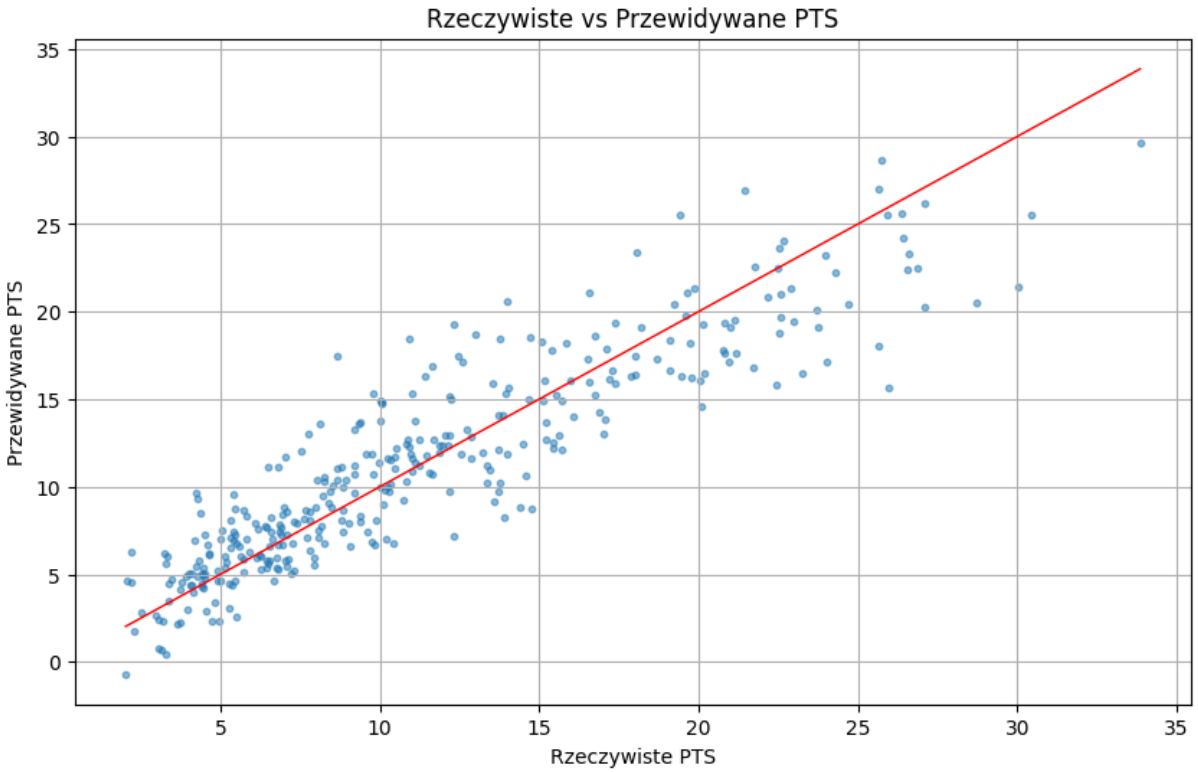
Omnibus:	9.192	Durbin-Watson:	2.055
Prob(Omnibus):	0.010	Jarque-Bera (JB):	13.872
Skew:	0.169	Prob(JB):	0.000972
Kurtosis:	3.924	Cond. No.	1.82e+03

Współczynnik determinacji modelu wyniósł 0,825, co oznacza że 82,5% danych jest poprawnie dopasowana do modelu. Zredukowany współczynnik determinacji wyniósł natomiast 0,819. Różnica jest bardzo mała i wynosi 0,006 co oznacza, że do budowy modelu wykorzystano istotne zmienne.

Dokładne wartości P-value przedstawiono poniżej:

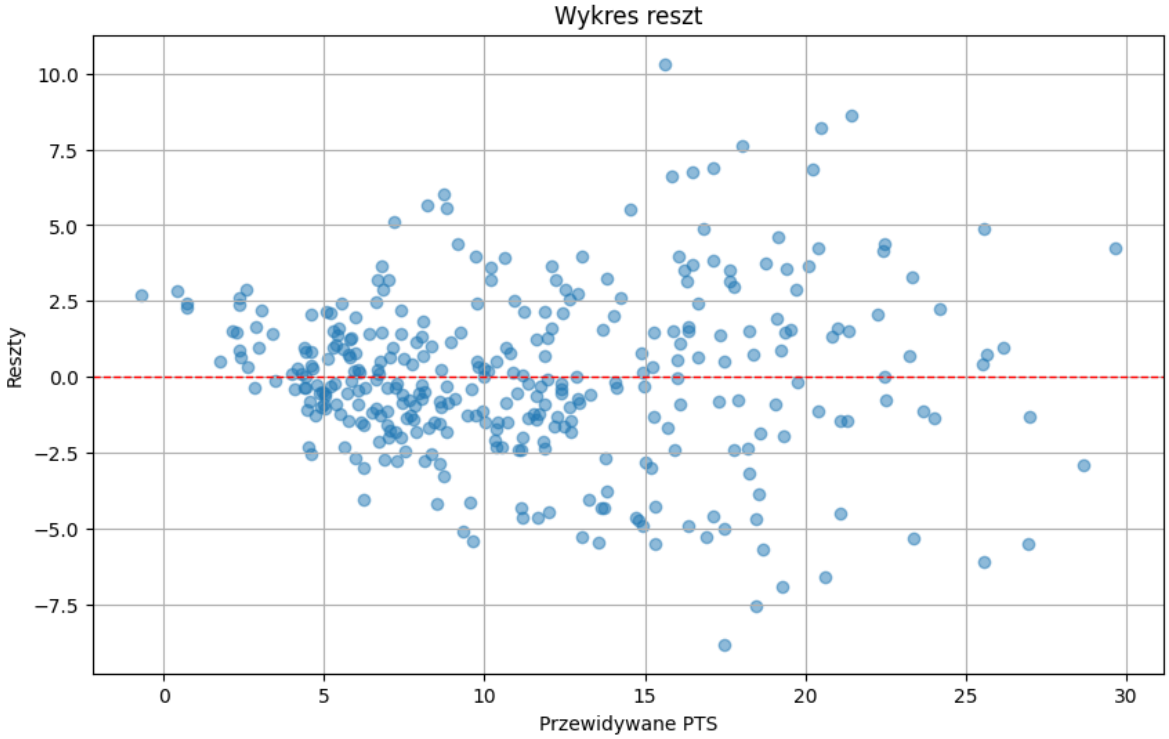
const	3.578e-7
G	0.302793...
GS	0
eFG%	0.006244...
FT%	5.7e-9
TRBPG	0.017211...
ASTPG	0.039641...
STLPG	0.000305...
BLKPG	0.121436...
TOVPG	5.963361...
PFPG	0.006050...

Wykres wartości rzeczywistych od przewidywanych:



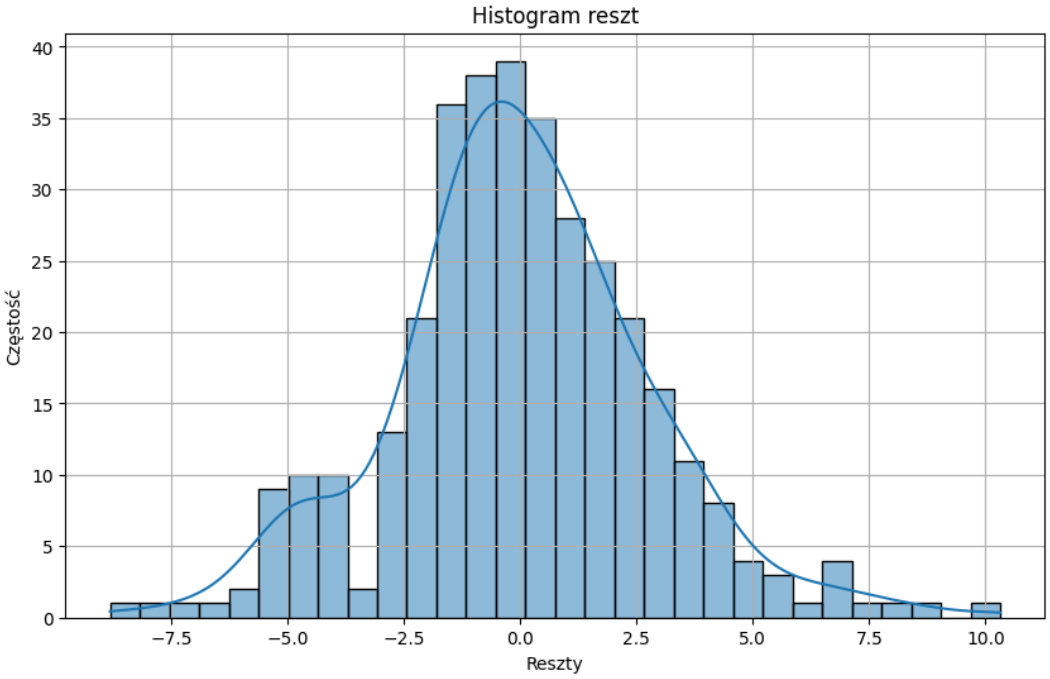
Na podstawie powyższego wykresu, możemy zaobserwować, że zmienne rozkładają się równomiernie po obu stronach prostej. Możemy zaobserwować, jedynie kilka punktów „odsatających".

Wykres reszt:



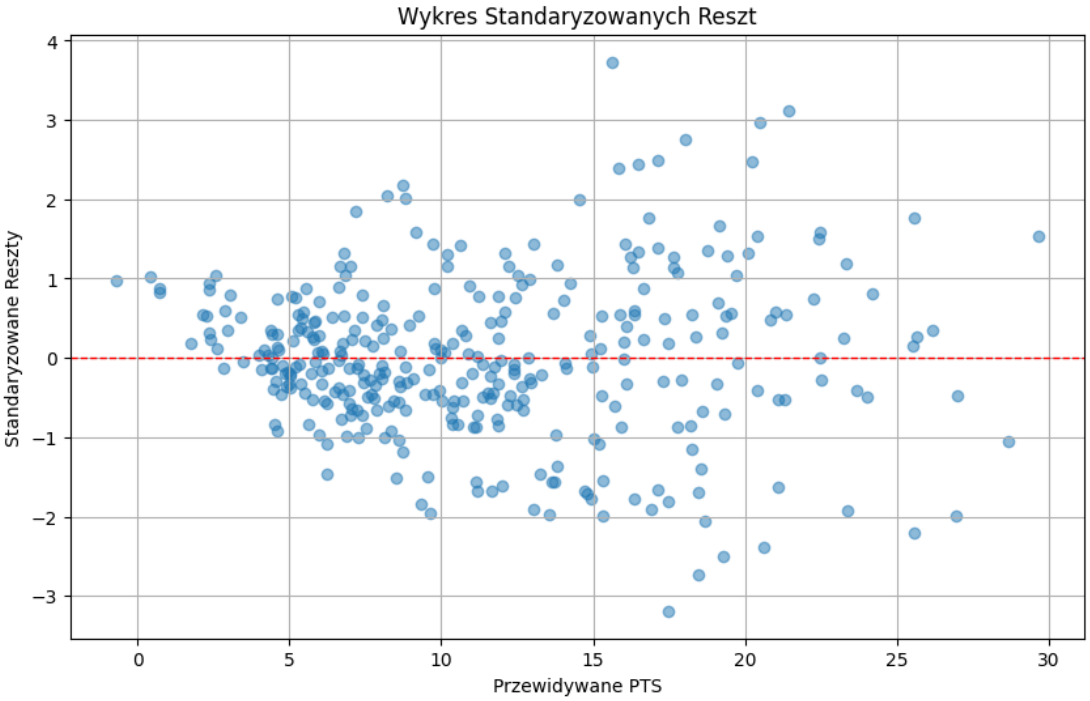
Na wykresie reszt możemy zaobserwować heteroskedastyczność objawiającą się większym zagęszczeniem punktów dla mniejszych wartości przewidywanych.

Histogram reszt:



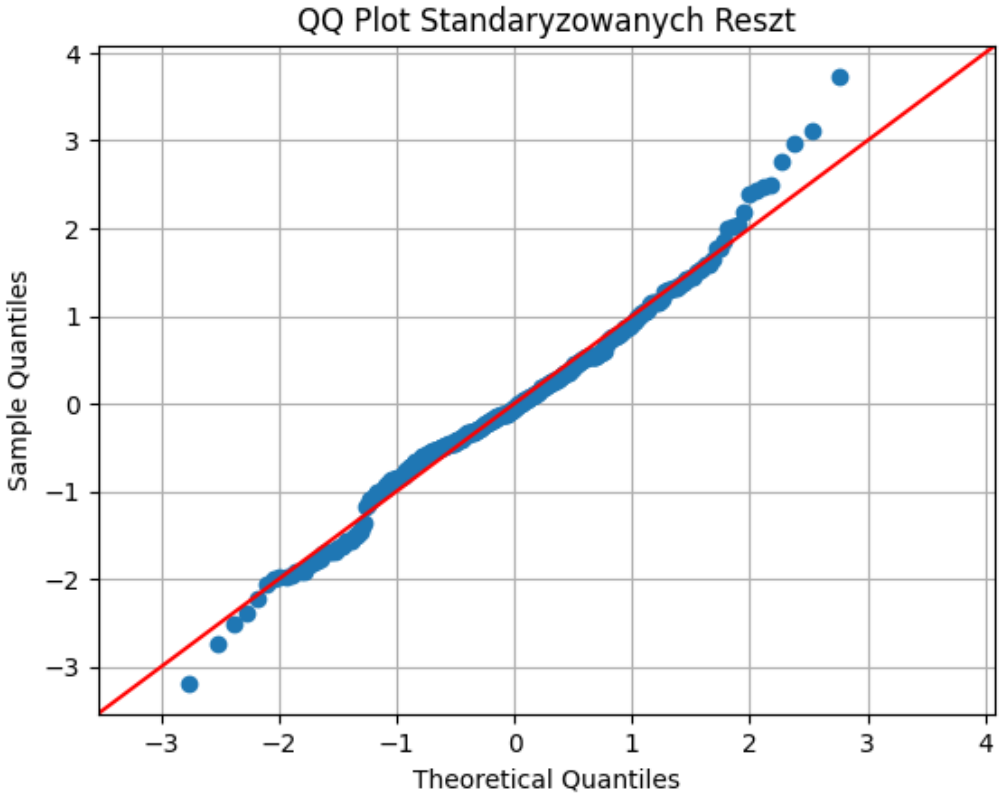
Histogram reszt potwierdza normalność rozkładu reszt.

Wykres reszt standaryzowanych:



Na wykresie reszt standaryzowanych możemy zaobserwować 14. punktów odstających.

Wykres QQ Plot:



Wykres QQ-Plot potwierdza normalność rozkładu reszt.

W celu poprawy modelu, usunięto wartości odstające. Do tego wykorzystano takie metody jak:

- Odległość Cooka- usunięto wartości >4
- DFFITS- usunięto wartości > $2 \cdot \sqrt{\frac{liczba\ zmiennych \cdot liczba\ stopni\ swobody\ modelu}{stopnie\ swobody\ modelu}}$
- Wartość reszt standaryzowanych- usunięto wartości >2

```
# Identify influential observations based on Cook's distance
df['influential_cook'] = np.where(cook_dist > 4 / len, 1, 0)

# Identify influentialobservations based on DFFITS
df['influential_dffits'] = np.where(np.abs(dffits) > 2 * np.sqrt(len * model.df_model / model.df_resid), 1, 0)

'''# Identify influential observations based on DFBETAS for a specific coefficient···

# Oblicz standaryzowane reszty
standardized_residuals = model.get_influence().resid_studentized_internal

# Dodaj standaryzowane reszty do DataFrame
df['standardized_residuals'] = standardized_residuals

# Dodaj kolumnę z wartością 1 jeśli standardized_residuals < 2, inaczej 0
df['residuals_indicator'] = np.where(np.abs(df['standardized_residuals']) > 2, 1, 0)
```

Na podstawie tej filtracji usunięto 32 wartości odstające i na podstawie tak zmienionej grupy danych (312 obserwacji) zbudowano model.

Poniżej przedstawiono statystyki nowego modelu liniowego:

Dep. Variable:	PPG	R-squared:	0.875
Model:	OLS	Adj. R-squared:	0.871
Method:	Least Squares	F-statistic:	211.7
Date:	Sat, 01 Jun 2024	Prob (F-statistic):	9.01e-130
Time:	16:55:34	Log-Likelihood:	-671.90
No. Observations:	312	AIC:	1366.
Df Residuals:	301	BIC:	1407.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-10.2468	1.778	-5.764	0.000	-13.745	-6.749
G	0.0209	0.012	1.740	0.083	-0.003	0.045
GS	0.0565	0.007	7.787	0.000	0.042	0.071
eFG%	0.0871	0.025	3.437	0.001	0.037	0.137
FT%	0.0850	0.014	5.892	0.000	0.057	0.113
TRBPG	0.2525	0.090	2.808	0.005	0.076	0.429
ASTPG	-0.4059	0.132	-3.067	0.002	-0.666	-0.145
STLPG	1.4416	0.514	2.806	0.005	0.431	2.453
BLKPG	-0.8436	0.439	-1.921	0.056	-1.708	0.021
TOVPG	6.3241	0.382	16.542	0.000	5.572	7.076
PFPG	-0.9893	0.294	-3.363	0.001	-1.568	-0.410

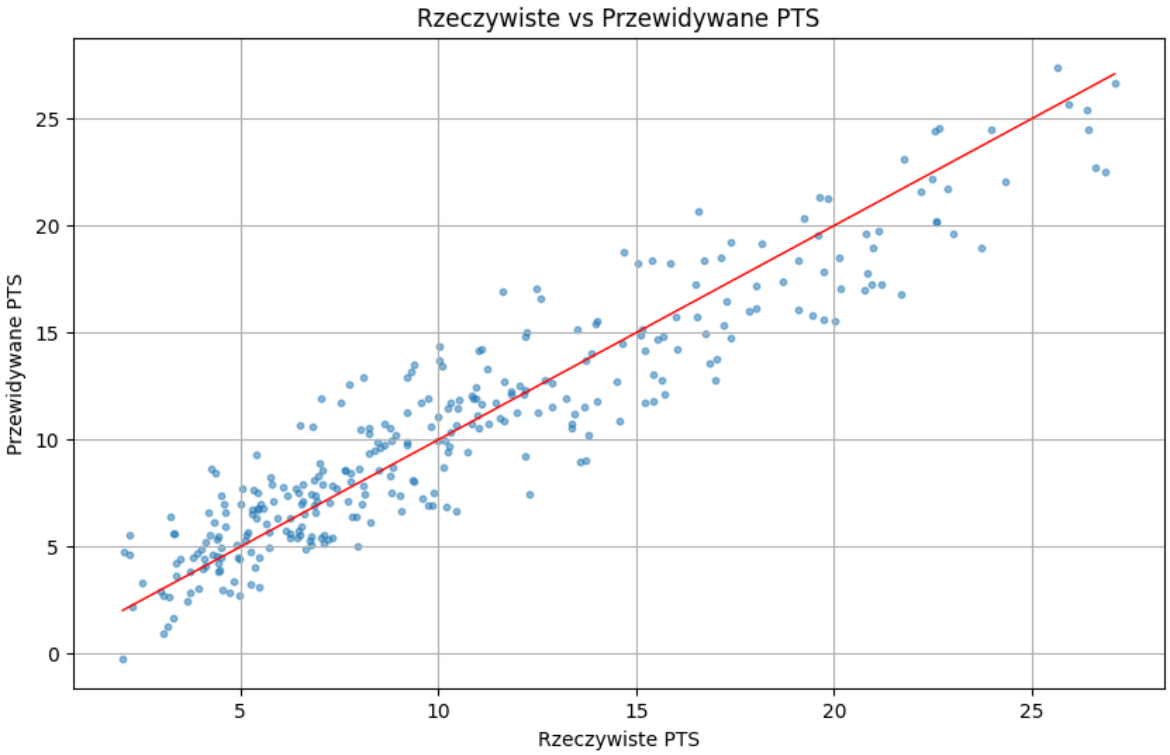
Omnibus:	0.439	Durbin-Watson:	2.083
Prob(Omnibus):	0.803	Jarque-Bera (JB):	0.566
Skew:	-0.010	Prob(JB):	0.754
Kurtosis:	2.792	Cond. No.	1.77e+03

Współczynnik determinacji modelu wyniósł 0,875, co oznacza że 87,5% danych jest poprawnie dopasowana do modelu. Zredukowany współczynnik determinacji wyniósł natomiast 0,871. Różnica jest bardzo mała i wynosi 0,004 co oznacza, że do budowy modelu wykorzystano istotne zmienne. W porównaniu z poprzednim modelem (bez odrzuconych wartości odstających), współczynnik determinacji wzrósł o 5%. Oznacza to, że usunięcie wartości skrajnych poprawiło działanie modelu.

Dokładne wartości P-value przedstawiono poniżej:

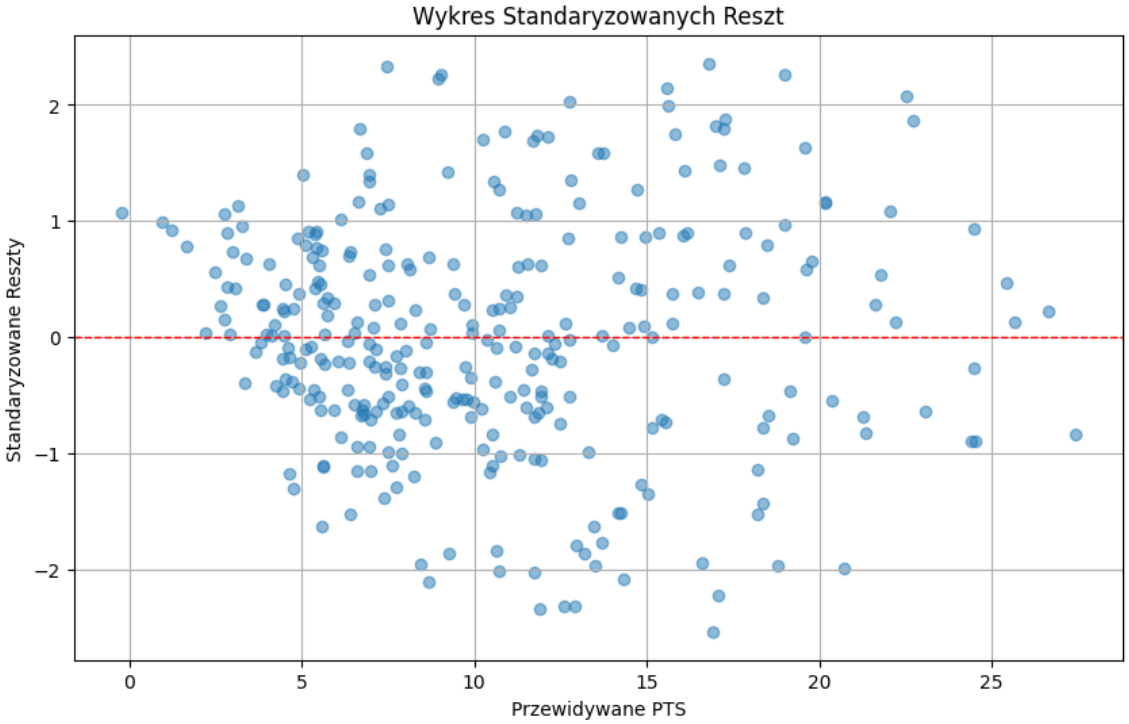
const	2.03e-8
G	0.0828194422
GS	0
eFG%	0.0006707305
FT%	1.02e-8
TRBPG	0.0053112508
ASTPG	0.0023554463
STLPG	0.0053436559
BLKPG	0.0557023543
TOVPG	3.592977758e-44
PFPG	0.0008717512

Wykres wartości rzeczywistych od przewidywanych:

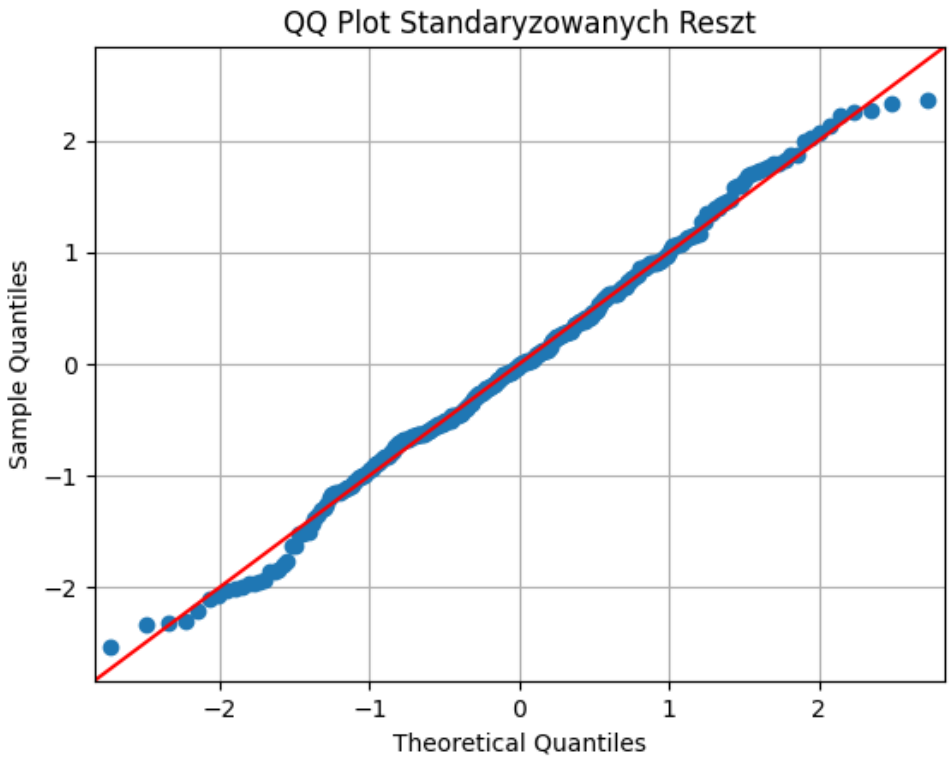


Na wykresie można zauważyć, że zmienne rozkładają się równomiernie po obu stronach prostej i nie są już zauważalne wartości odstające.

Wykres reszt standatyzowanych:



Możemy zaobserwować, że rozkład reszt zmienił się na bardziej homoskedystyczny.



Wykres QQ-Plot potwierdza liniowość oraz normalność rozkładu. Mamy jedynie 4 obserwacje nieznacznie odstające.

Predykcja na podstawie modelu:

```
# Przykładowe statystyki zawodnika
player_stats = {
    'const': [1],
    'G': [82],
    'GS': [82],
    'eFG%': [0.55],
    'FT%': [0.85],
    'TRBPG': [10.0],
    'ASTPG': [5.0],
    'STLPG': [1.5],
    'BLKPG': [3.0],
    'TOVPG': [4.0],
    'PFPG': [3.0]
}
```

```
Przewidywana liczba punktów na mecz: 18.673716654901003
```

Dla TOVPG = 3

```
Przewidywana liczba punktów na mecz: 12.349630693807702
```

Dla TOVPG = 2

```
Przewidywana liczba punktów na mecz: 6.025544732714402
```

Zgodnie z przewidywaniami możemy zaobserwować, że zmiana wartości TOVPG znacząco wpływa na predyktowaną wartość.

Po usunięciu zmiennej TOVPG z modelu uzyskano współczynnik determinacji:

R-squared:	0.741
Adj. R-squared:	0.734

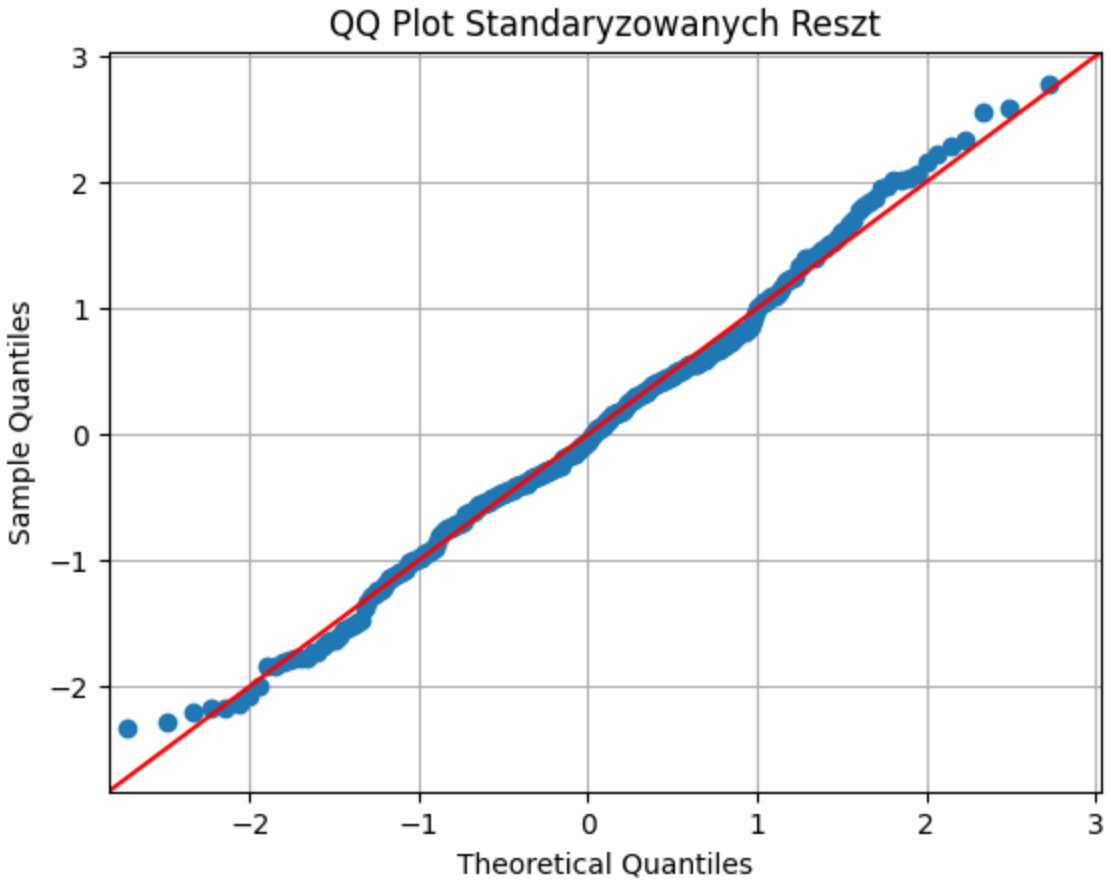
Po usunięciu obserwacji odstających uzyskano następujące parametry modelu:

Dep. Variable:	PPG	R-squared:	0.809
Model:	OLS	Adj. R-squared:	0.804
Method:	Least Squares	F-statistic:	141.1
Date:	Sat, 01 Jun 2024	Prob (F-statistic):	4.36e-102
Time:	18:54:41	Log-Likelihood:	-732.48
No. Observations:	309	AIC:	1485.
Df Residuals:	299	BIC:	1522.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.3915	2.219	-3.782	0.000	-12.758	-4.025
G	0.0148	0.015	0.993	0.322	-0.014	0.044
GS	0.0730	0.009	8.229	0.000	0.056	0.091
eFG%	-0.0060	0.030	-0.197	0.844	-0.066	0.054
FT%	0.1221	0.019	6.598	0.000	0.086	0.159
TRBPG	0.7278	0.104	6.972	0.000	0.522	0.933
ASTPG	1.4052	0.122	11.538	0.000	1.166	1.645
STLPG	0.6263	0.649	0.965	0.335	-0.650	1.903
BLKPG	-0.8267	0.486	-1.700	0.090	-1.784	0.130
PFPG	0.1669	0.345	0.484	0.629	-0.512	0.845

Współczynnik determinacji wyniósł 0,809, natomiast zredukowany współczynnik determinacji wyniósł 0,804. Różniąca ta jest niewielka co świadczy o braku zmiennych niepotrzebnych. Porównując zredukowane współczynniki determinacji do modelu ze zmienną TOVPG (0,871), możemy zaobserwować, że wartość spadła 0,067 więc model opisuje mniejszą część danych jednak nie jest już tak bardzo obciążony. Teraz model najbardziej obciążony jest przez zmienną ASTPG.

const	0.000187751
G	0.3217413522
GS	0
eFG%	0.8438301104
FT%	2e-10
TRBPG	0
ASTPG	1.008582888e-25
STLPG	0.3350972123
BLKPG	0.0901895814
PFPG	0.6285952102



Na podstawie wykresu QQ-Plot możemy potwierdzić hipotezę o liniowości i normalności reszt.

Predykcja na podstawie finalnego modelu:

```
player_stats = {
    'const': [1],
    'G': [82],
    'GS': [82],
    'eFG%': [0.55],
    'FT%': [0.85],
    'TRBPG': [10.0],
    'ASTPG': [5.0],
    'STLPG': [1.5],
    'BLKPG': [3.0],
    'PFPG': [3.0]
}
```

```
Przewidywana liczba punktów na mecz: 12.172017432508328
```

Algorytm KNN

Przeprowadzono klasyfikację danych testowych za pomocą metody KNN, czyli metody najbliższych sąsiadów gdzie parametr k ustawiono na 10. Dla tego parametru osiągnięto najlepsze wyniki. Wybrano 12 zmiennych przedstawionych poniżej:

```
X = df.loc[:, [ 'eFG%', '3P%','2P%','FT%',  'PPG', 'MPG', 'TRBPG',  'ASTPG' ,'STLPG' ,'BLKPG',  'TOVPG'  , 'PFPG']]
```

I na ich podstawie celem było prawidłowe rozpoznanie pozycji gry zawodnika na boisku.

	precision	recall	f1-score	support
C	0.89	0.80	0.84	20
C-PF	0.00	0.00	0.00	1
PF	0.25	0.64	0.36	11
PF-SF	0.00	0.00	0.00	1
PG	0.67	0.61	0.64	23
PG-SG	0.00	0.00	0.00	2
SF	0.36	0.33	0.35	24
SG	0.47	0.32	0.38	22
accuracy			0.50	104
macro avg	0.33	0.34	0.32	104
weighted avg	0.53	0.50	0.50	104

Na podstawie precyzji predykcji wartości ze zbioru testowego, możemy zaobserwować że z największą precyzją rozpoznawani są zawodnicy grający na pozycji centra (89%) oraz rozgrywającego (67%). Dla pozostałych parametrów precyzja zawierała się w przedziale od 25% do 47%, a całkowita trafność wynosiła jedynie 50%. Wyniki te są jednak zrozumiałe. Center jest najbardziej charakterystyczną pozycją na boisku, cechuje go wysoka skuteczność rzutów za 2 punkty (zwykle wsady i rzuty o tablicę z pola trzech sekund) oraz duża liczba zbiórek. Rozgrywający również jest bardzo charakterystyczną pozycją ponieważ cechuje go duża liczba asyst oraz bardzo często wysoka skuteczność za 3 punkty. Pozostałe pozycje na boisku nie są tak bardzo charakterystyczne zawodnicy grający na pozostałych trzech pozycjach, a więc rzucający obrońca, silny skrzydłowy i niski skrzydłowy w obecnej koszykówce nie mają tak silnych cech charakterystycznych. Z zasady:

- Niski skrzydłowy jest wsparciem oraz odciążeniem dla rozgrywającego przy rozgrywaniu i prowadzeniu piłki.
- Silny skrzydłowy to pozycja defensywna. Z pewnością silny skrzydłowy musi doskonale bronić i osłaniać innych zawodników.
- Rzucający obrońca to pozycja, której głównym zadaniem jest obrona, zbieranie piłek i rzucanie do kosza

Możemy uznać więc, że trafność modelu KNN była zadawalająca jak na tak wymagające dane.

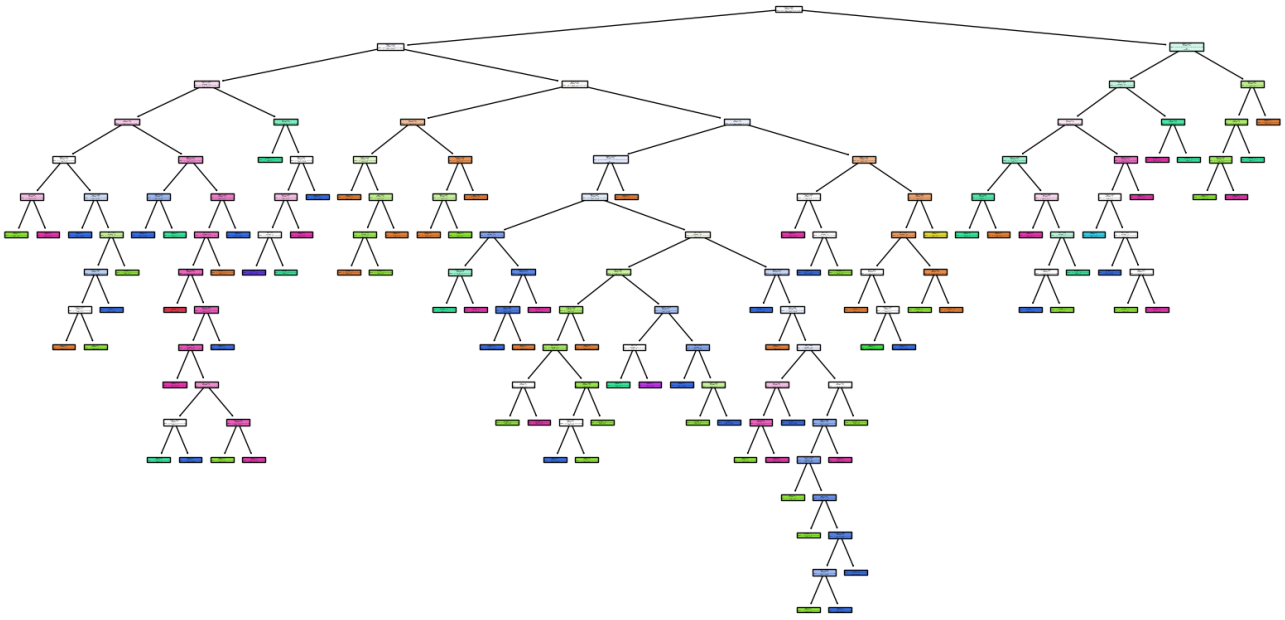
Drzewo decyzyjne

Na podstawie zbioru treningowego zbudowano również drzewo decyzyjne, jednak poprawność rozróżnienia danych wyniosła 42%. Co jest gorszym wynikiem w porównaniu do metody KNN gdzie poprawność wyniosła 50%. Możemy zobaczyć analogię w przypadku poszczególnych pozycji na boisku, ponieważ w tym przypadku również najlepiej rozpoznani zostali zawodnicy na pozycji centra (52%) oraz rozgrywającego (60%).

Classification Report:

	precision	recall	f1-score	support
C	0.52	0.60	0.56	20
C-PF	0.00	0.00	0.00	1
PF	0.25	0.45	0.32	11
PF-SF	0.00	0.00	0.00	1
PG	0.60	0.52	0.56	23
PG-SG	0.00	0.00	0.00	2
SF	0.33	0.25	0.29	24
SF-PF	0.00	0.00	0.00	0
SG	0.43	0.41	0.42	22
accuracy			0.42	104
macro avg	0.24	0.25	0.24	104
weighted avg	0.43	0.42	0.42	104

Graficzna prezentacja drzewa:



Wszystkie problemy modelu predykcji KNN, a więc trudność w rozróżnieniu części pozycji były obecne w modelu drzewa decyzyjnego, co przełożyło się na pozornie niskiej jakości wyniki.