Projekt indywidualny 2020/2021 SAD
Star Types

Spis treści

1. Sformułowanie celu projektu oraz stworzenie reprezentującego modelu reprezentującego badany problem.

a) Sformułowanie tematu projektu i celu analizy statystycznej

Przedmiotem badania statystycznego będą gwiazdy, możemy je podzielić na 6 głównych typów: brązowe karły, białe karły, czerwone karły, z ciągu głównego (tj. main sequence) są to gwiazdy usadowione na pasie przebiegającym wzdłuż krzywej na diagramie Hertzsprunga-Russella charakteryzują się one podobną masą i składem chemicznym jak nasze słońce, supergiganty oraz hiper-giganty. Naukowcy badają gwiazdy aby określić ich długość życia oraz lepiej zrozumieć nasze słońce. Celem mojej analizy statystycznej jest zbadanie wpływu temperatury gwiazdy na jej jasność.

b) Wytypowanie cech reprezentujących analizowany problem.

Zbiór danych reprezentuje 7 zmiennych , cztery zmienne ilościowe, są nimi : temperatura gwiazdy, jasność gwiazdy obliczana względem jasności słońca, promień gwiazdy obliczany względem promienia słońca i absolutnej emitowanej jasności (jest to miara jasności pozornej gwiazdy ,jaką miała by gwiazda obserwowana z odległości 32,6 lat świetlnych) oraz trzech zmiennych jakościowych, są to: typ gwiazdy wspomniany na początku, kolor gwiazdy oraz klasa widmowa każdej gwiazdy. Wytypowałem do analizy dwie zmienne ilościowe oraz jedną jakościową, są nimi:

Zmienne ilościowe:

- Temperatura (temperature)
- Jasność (luminosity)

Zmienne jakościowe:

- Kolor gwiazdy

c) Pozyskanie danych

Baza danych została stworzona na podstawie szeregu równań astrofizycznych, takich jak:

-prawo Stefana Boltzmanna (w celu obliczenia promieniowania),

-prawo przemieszczania Wienna (określanie temperatury powierzchni gwiazdy za pomocą długości fali) ,

-relacji wielkości względnej,

-efekt parallaxy gwiezdnej (różnice w pozornym położeniu gwiazdy względem dalszych obiektów na sferze niebieskiej)

Baza została stworzona z 240 obserwacji.

Dane wykorzystane do obliczenia powyższych wartości zostały zebrane z sieci, proces ten zajął 3 tygodnie. Brakujące dane zostały obliczone ręcznie, korzystając z powyższych równań astrofizycznych.

Temperatura odnosi się do zmiennej temperaturę, natomiast jasność do luminostiy.

2. Statystyczny opis struktury analizowanych cech, reprezentowanych przez zmienne liczbowe.

a) Podanie wartości statystycznych

Statystyki opisowe (6 class csv (1).sta)

Zmienna	Nważnych	Średnia	Mediana	Moda	Liczność Mody	Minimum	Maksimum	Dolny Kwartyl.	Górny Kwartyl.	Rozstęp	Wariancja	Odch.std	Skośność	Kurtoza
Temperature (K)	240	10497,46	5776,000	3600,000	3	1939,000	40000,00	3343,500	15129,00	38061,00	91248824	9552,425	1,321568	0,877352

Statystyki opisowe (6 class csv (1).sta)

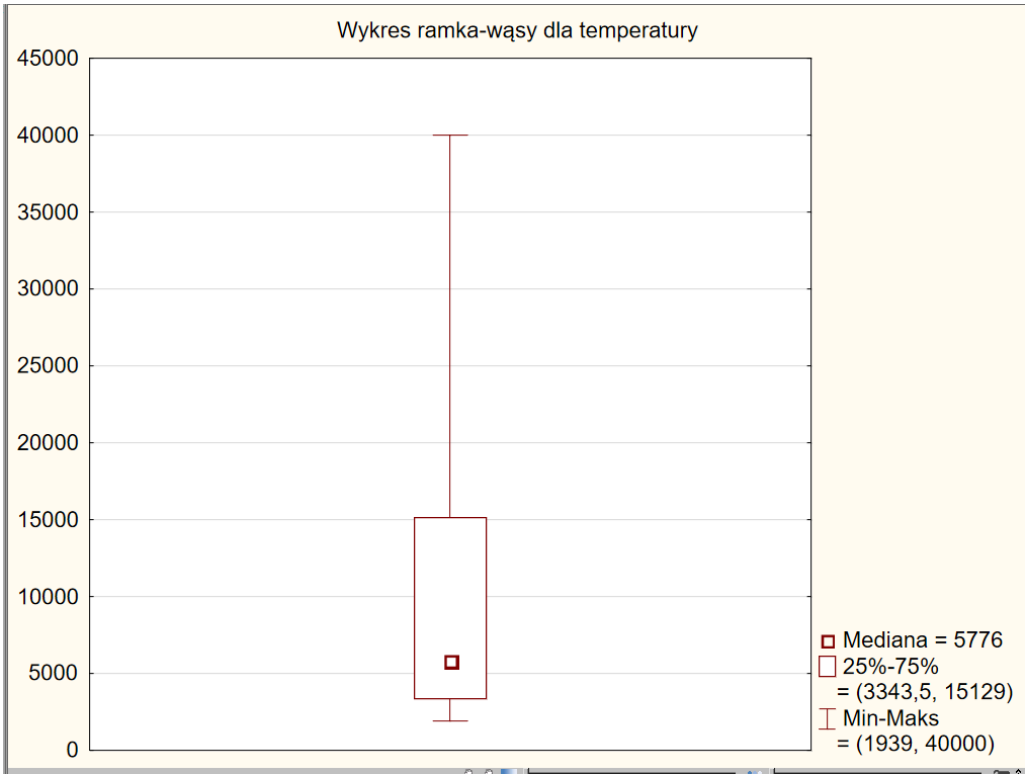
Zmienna	Nważnych	Średnia	Mediana	Moda	Liczność Mody	Minimum	Maksimum	Dolny Kwartyl.	Górny Kwartyl.	Rozstęp	Wariancja	Odch.std	Skośność	Kurtoza
Luminosity(L/Lo)	240	107273,9	198,5000	200000,0	5	29,00000	849420,0	143,5000	198100,0	849391,0	3,217753E+10	179381,0	2,068917	4,469252

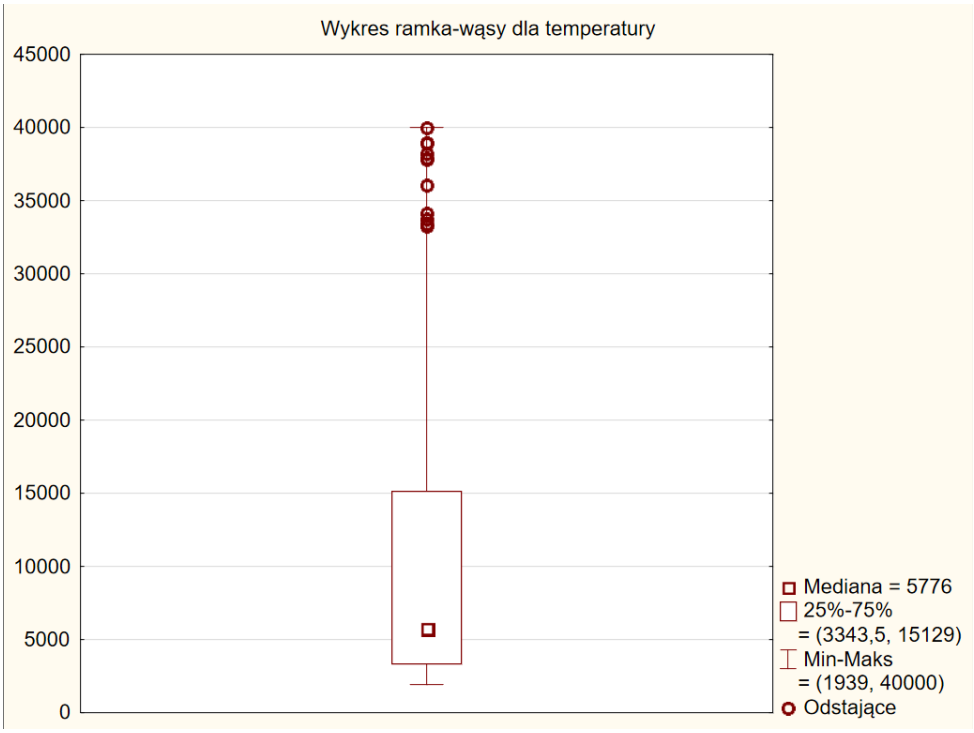
Podsumowanie statystyk opisowych:

Rozstęp dla jasności jest znacznie większy od rozstępu temperatury, tak więc, wskazuję to na dużo większą różnice pomiędzy najmniejsza a największą wartością w przypadku jasności. W obydwu

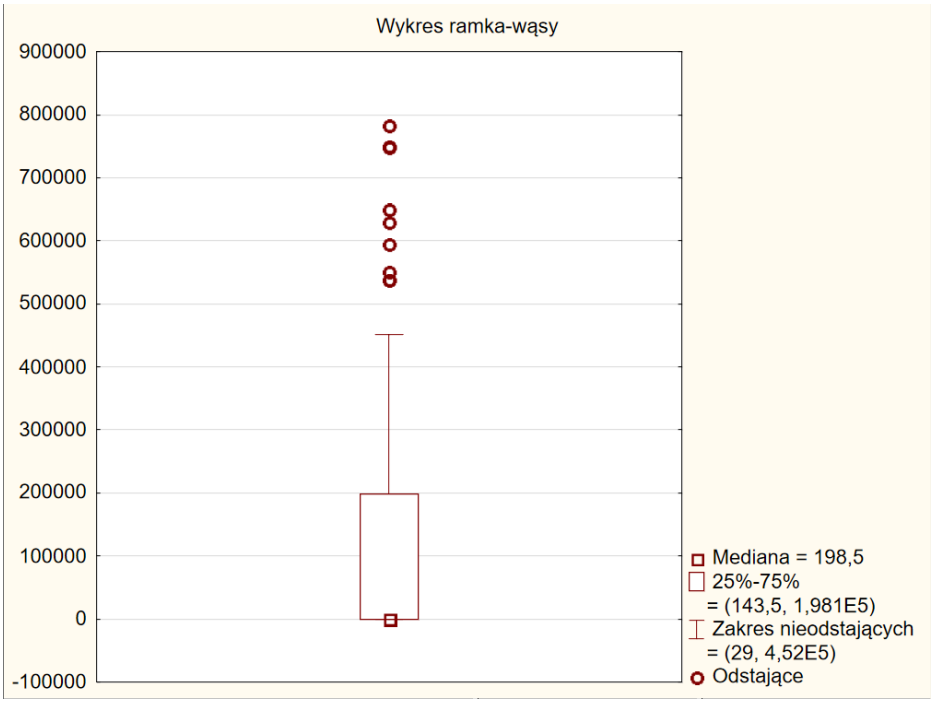
przypadkach kurtoza jest dodatnia, co mówi nam o tym, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Ponadto zarówno dla temperatury oraz jasności skośność jest dodatnia co wskazuje nam w obydwu przypadkach na rozkład prawostronnie skośny. Obserwujemy bardzo dużą wariancje w przypadku jasności, tak więc zróżnicowanie jest duże co potwierdzają wartości minimalna oraz maksymalna.

b) Wykres ramka-wąsy



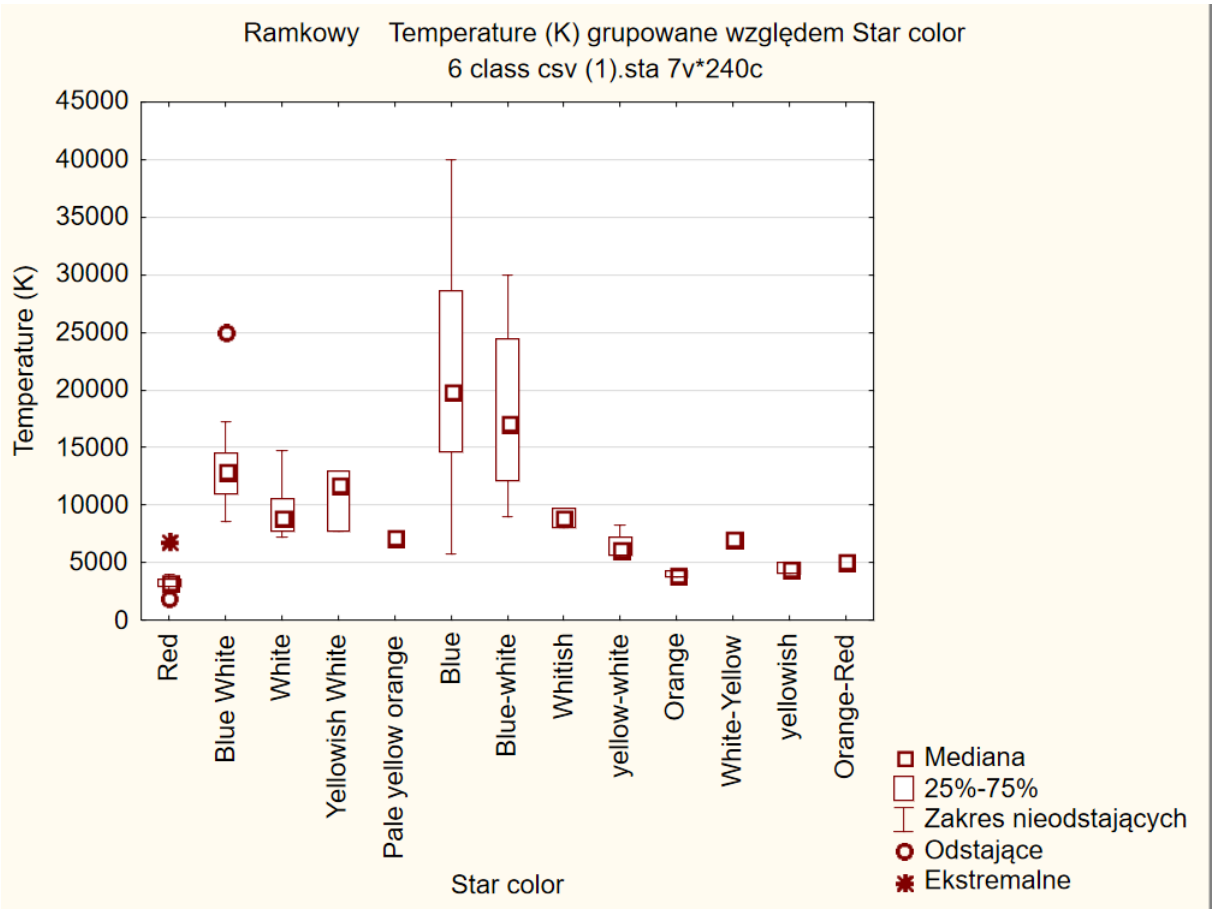


Wykres nie wygląda symetrycznie względem mediany, co można potwierdzić za pomocą wartości statystycznych. Dłuższa część pudełka znajduję się nad medianą co wskazuje na prawostronną skośność wykresu. Wykres jest dość rozproszony, tak więc dane mają rozbieżne wartości.



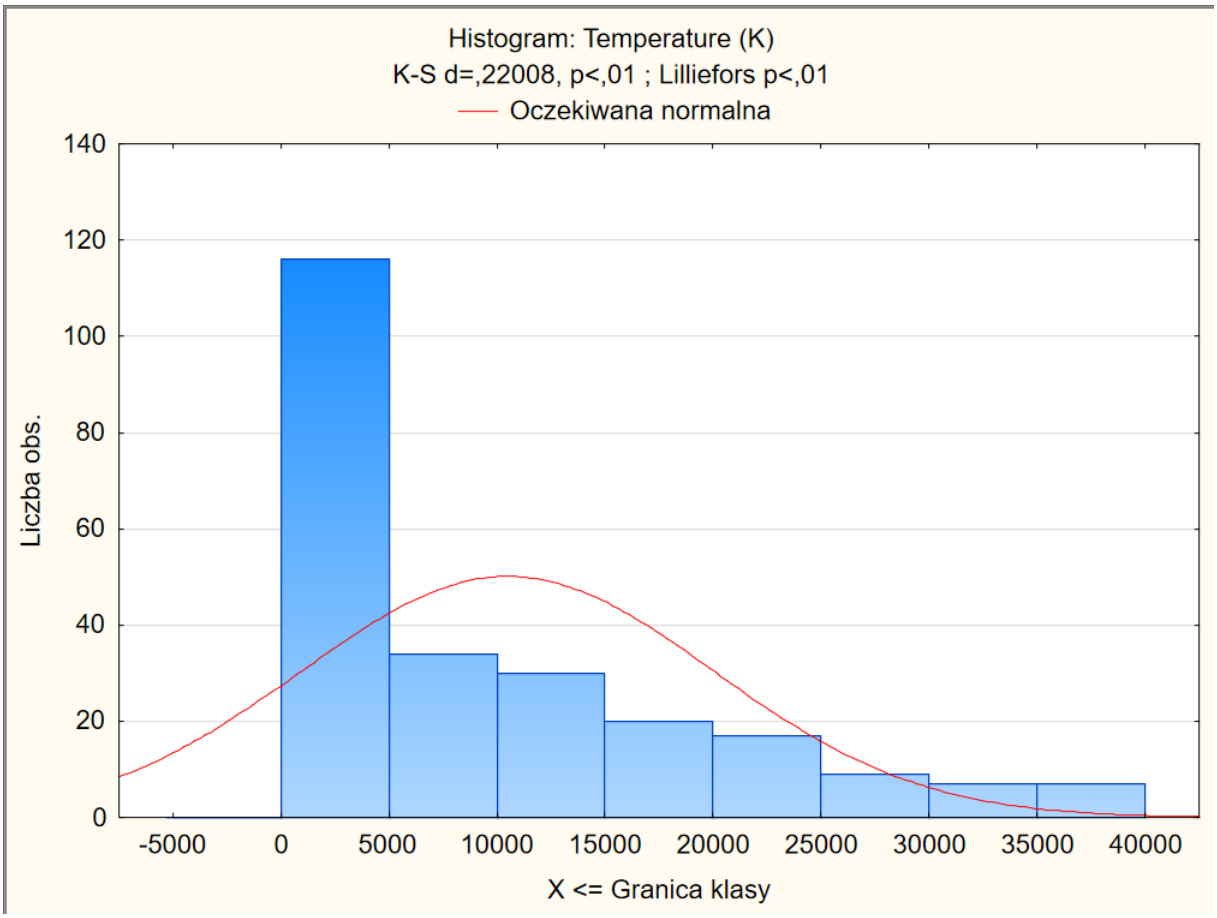
Wykres ramka-wąsy dla jasności nie jest symetryczny względem mediany, dłuższa część pudełka jest nad medianą co mówi nam o rozkładzie prawostronnie skośnym. Posiada on dodatkowo wartości odstające, odbiegające od reszty. Wszystkie wartości mają mniej niż 800000 [J/s].

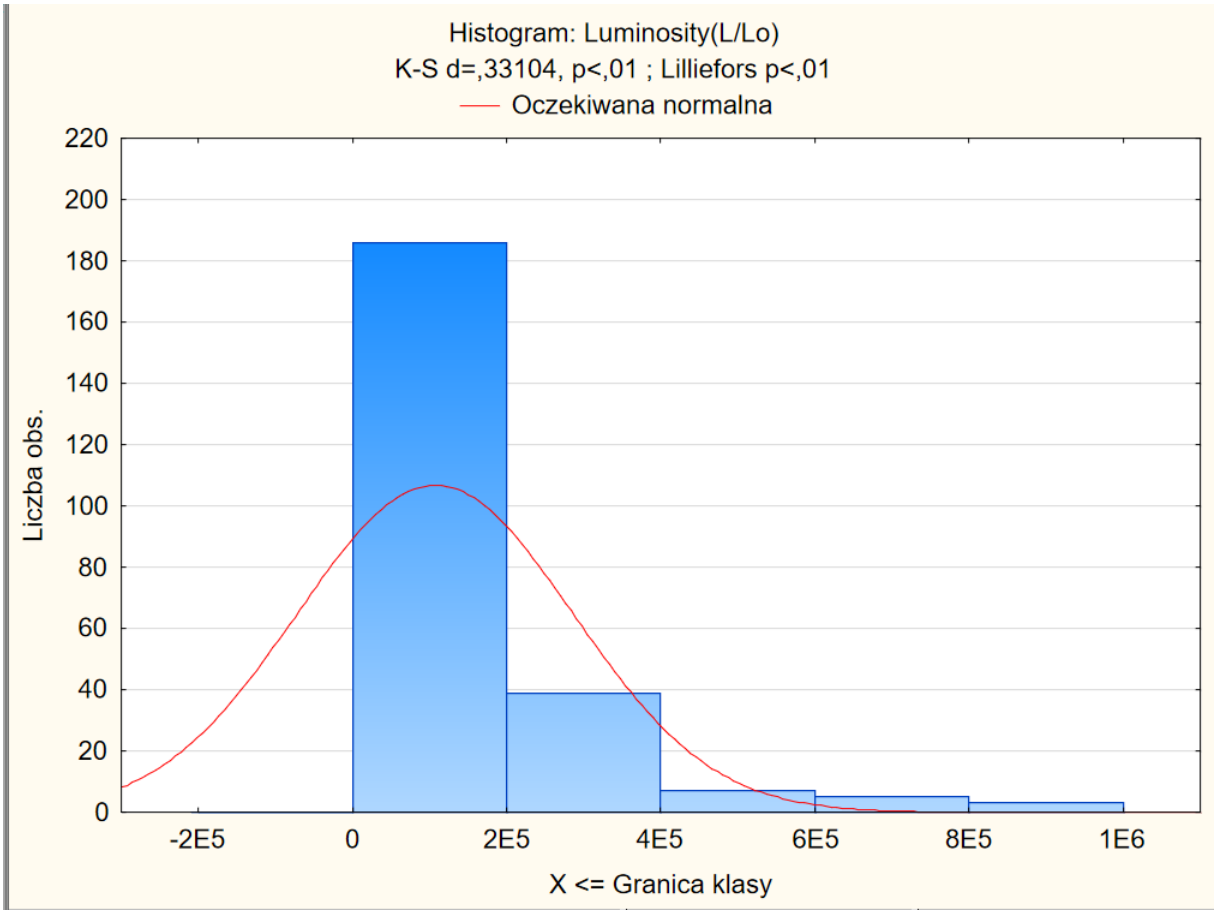
Skategoryzowany wykres ramka wąsy.



Wykres skategoryzowany pokazuje nam, że oprócz odcieni niebieskiego reszta gwiazd posiada ściśle określony kolor dla bardzo ograniczonych wartości temperatury. W przypadku czerwonego koloru występuje wartość ekstremalna, co może być spowodowane tym, że gwiazda znajduje się w stanie przeobrażenia w czerwonego olbrzyma tuż przed swoją „śmiercią”, lecz nie możemy tego wywnioskować z badań statystycznych są to tylko domysły. W przypadku niebieskiego koloru rozstęp jest bardzo duży co wskazuje nam na to, że gwiazdy o tym kolorze posiadają wartości temperatur zawarte w bardzo dużym przedziale.

c) Histogramy


Histogram: Temperature (K)
K-S d=,22008, p<,01 ; Lilliefors p<,01
Oczekiwana normalna
Liczba obs.
140
120
100
80
60
40
20
0
-5000
0
5000
10000
15000
20000
25000
30000
35000
40000
X <= Granica klasy


Histogram: Luminosity(L/Lo)
K-S d=,33104, p<,01 ; Lilliefors p<,01
Oczekiwana normalna
Liczba obs.
220
200
180
160
140
120
100
80
60
40
20
0
-2E5
0
2E5
4E5
6E5
8E5
1E6
X <= Granica klasy

Temperatura gwiazd mieści się w przedziale między 0 -40000 K, obserwacje nie są rozłożone równomiernie co oznacza, że zaobserwowano najwięcej planet z temperaturą mieszczącą się w przedziale 0-5000 K, w tym obszarze jest dominata. Ponadto histogram jest prawostronnie skośny co wynika bezpośrednio z tego, iż średnia temperatury jest większa od jej mediany. W miarę wzrostu temperatury występuje tendencja spadkowa ilości obserwacji. Wnioskami jakie możemy wyciągnąć jest to, iż w kosmosie występuje więcej planet o niskiej temperaturze. Jeśli chodzi o histogram jasności obserwacje również nie są rozłożone równomiernie, dominata występuje w zakresie 0-2e^5 [J/s], co więcej występuje prawostronna skośność. Wnioskiem jaki możemy wysunąć jest to iż, większość gwiazd nie świeci jaśniej niż 2e^5[J/s], powyżej 4e^5[J/s] w żadnym przedziale liczność gwiazd nie przekracza 20 przypadków.

3 Wnioskowanie statystyczne

a) Wyznaczenie przedziału ufności dla wartości oczekiwanych jednej zmiennej.

Zmienna	Statystyki opisowe (6 class csv (1).sta)								
	Nważnych	Średnia	Ufność -95,000%	Ufność 95,000%	Minimum	Maksimum	Odch.std	P. ufności odch. std. -95,000%	P. ufności odch. std. +95,000%
Temperature (K)	240	10497,46	9282,785	11712,14	1939,000	40000,00	9552,425	8767,446	10492,99

Zmienna	Statystyki opisowe (6 class csv (1).sta)								
	Nważnych	Średnia	Ufność -95,000%	Ufność 95,000%	Minimum	Maksimum	Odch.std	P. ufności odch. std. -95,000%	P. ufności odch. std. +95,000%
Luminosity(L/Lo)	240	107273,9	84463,97	130083,8	29,00000	849420,0	179381,0	164640,2	197043,5

Posiadamy 95% pewność, że średnia temperatura gwiazd znajduje się w przedziale (9282,785;11712,14) . Średnia jasność natomiast będzie mieściła się w przedziale (84463,97;130083,8), w obu przypadkach statystyka opisowa potwierdza nam przedział ufności.

b) Weryfikacja hipotezy o zgodności empirycznego rozkładu wybranej cechu z rozkładem normalnym.

Zmienna	Testy normalności (6 class csv (1).sta)					
	N	maks D	K-S p	Lillief. p	W	p
Temperature (K)	240	0,220083	p < ,01	p < ,01	0,793227	0,000000
Luminosity(L/Lo)	240	0,331039	p < ,01	p < ,01	0,659687	0,000000

H0-zmienne posiadają rozkład normalny

H1-brak rozkładu normalnego

$\alpha = 0.05$

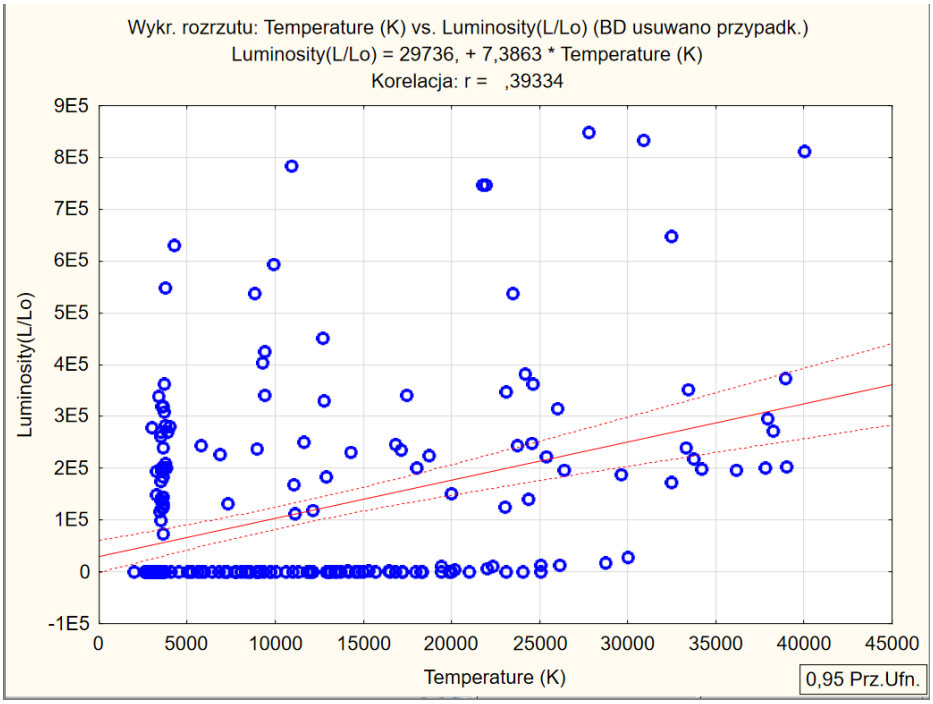
Tak więc $p < \alpha$ brak rozkładu normalnego zarówno w przypadku temperatury jak i jasności.

c) Sprawdzenie czy istnieje związek korelacyjny pomiędzy badanymi zmiennymi oraz jeśli takowy istnieje zbudowanie modelu regresji liniowej.

Jako wstępne przypuszczenie przyjąłem, iż może występować korelacja pomiędzy temperaturą gwiazdy a jasnością jaką emituje.

Korelacje (6 class csv (1).sta)
Oznaczone wsp. korelacji są istotne z p < ,05000
N=240 (Braki danych usuwano przypadkami)

Zmienna	Średnia	Odch.std	Luminosity(L/Lo)	Temperature (K)
Luminosity(L/Lo)	107273,9	179381,0	1,000000	0,393338
Temperature (K)	10497,5	9552,4	0,393338	1,000000



Wykres rozrzutu jak i wartość korelacji wskazują na pewną zależność liniową na poziomie przeciętnym.

Wykres sugeruje, że występuje pewna zależność liniowa między badanymi zmiennymi, nie jest ona zbyt duża, lecz korelacja na poziomie 0,4 jest korelacją przeciętną tak więc zbudowanie modelu regresji jest uzasadnione.

Podsumowanie regresji zmiennej zależnej: Luminosity(L/Lo) (6 class csv (1).sta)
R= ,39333827 R^2= ,15471499 Popraw. R2= ,15116337
F(1,238)=43,562 p<,00000 Błąd std. estymacji: 1653E2

N=240	b*	Bł. std. z b*	b	Bł. std. z b	t(238)	p
W. wolny			29736,13	15868,81	1,873872	0,062173
Temperature (K)	0,393338	0,059595	7,39	1,12	6,600139	0,000000

y=29736,13+7,39*x

Luminosity=29736,13+7,39*Temperature || Jasność=29736,13+7,39*Temperatura

Weyfikacja jakości dopasowania:

statystyka	Stat.podsum.; Zmn. zal.:Luminosity(L/Lo) (6 class csv (1).sta)
	Wartość
R wielorakie	0,393338266
Wielorakie R2	0,154714991
Skorygowane R2	0,151163373
F(1,238)	43,5618371
p	0,000000000264498673
Błąd std. estymacji	165267,852

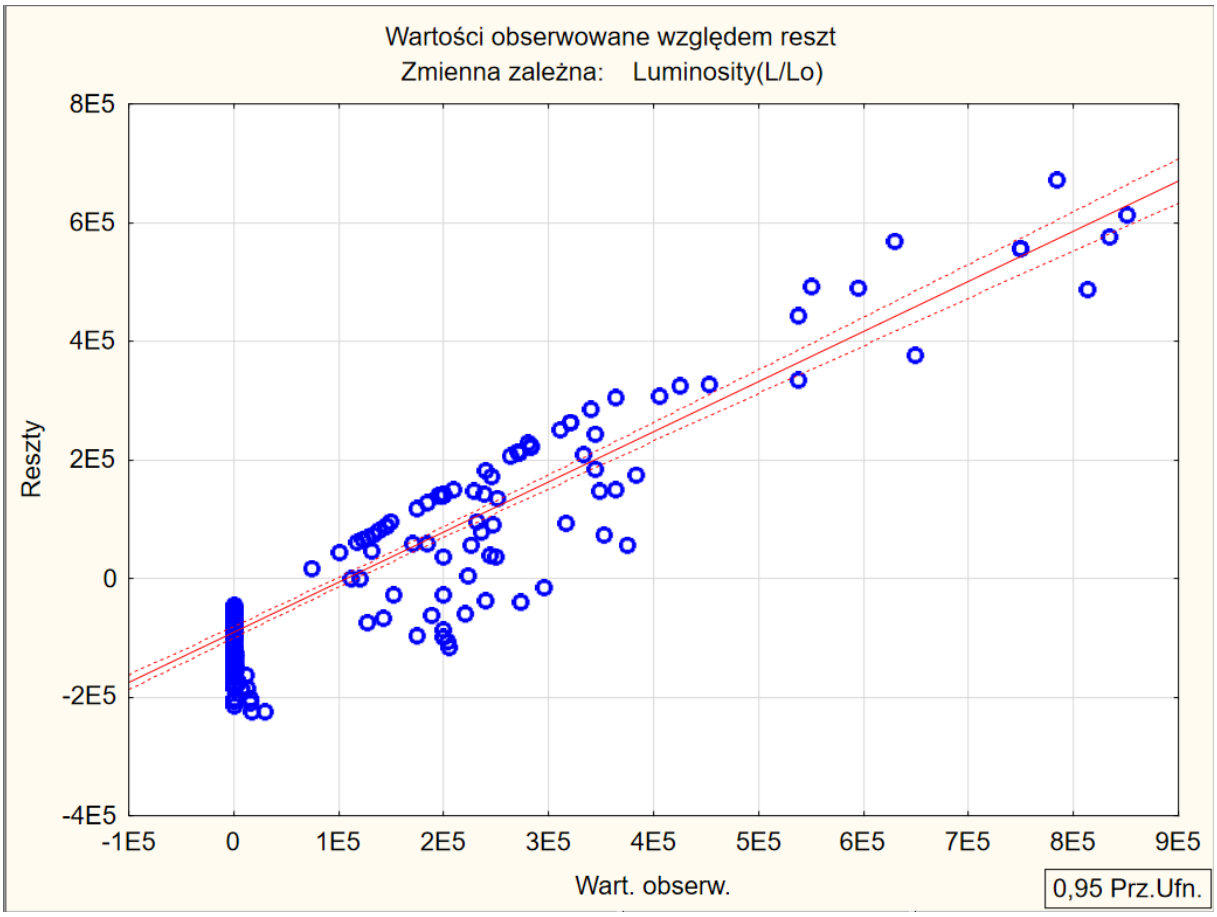
Testujemy hipotezę

h0: m=0 nie ma zależności liniowej pomiędzy zmiennymi,

h1: m ≠ 0 istnieje zależność liniowa.

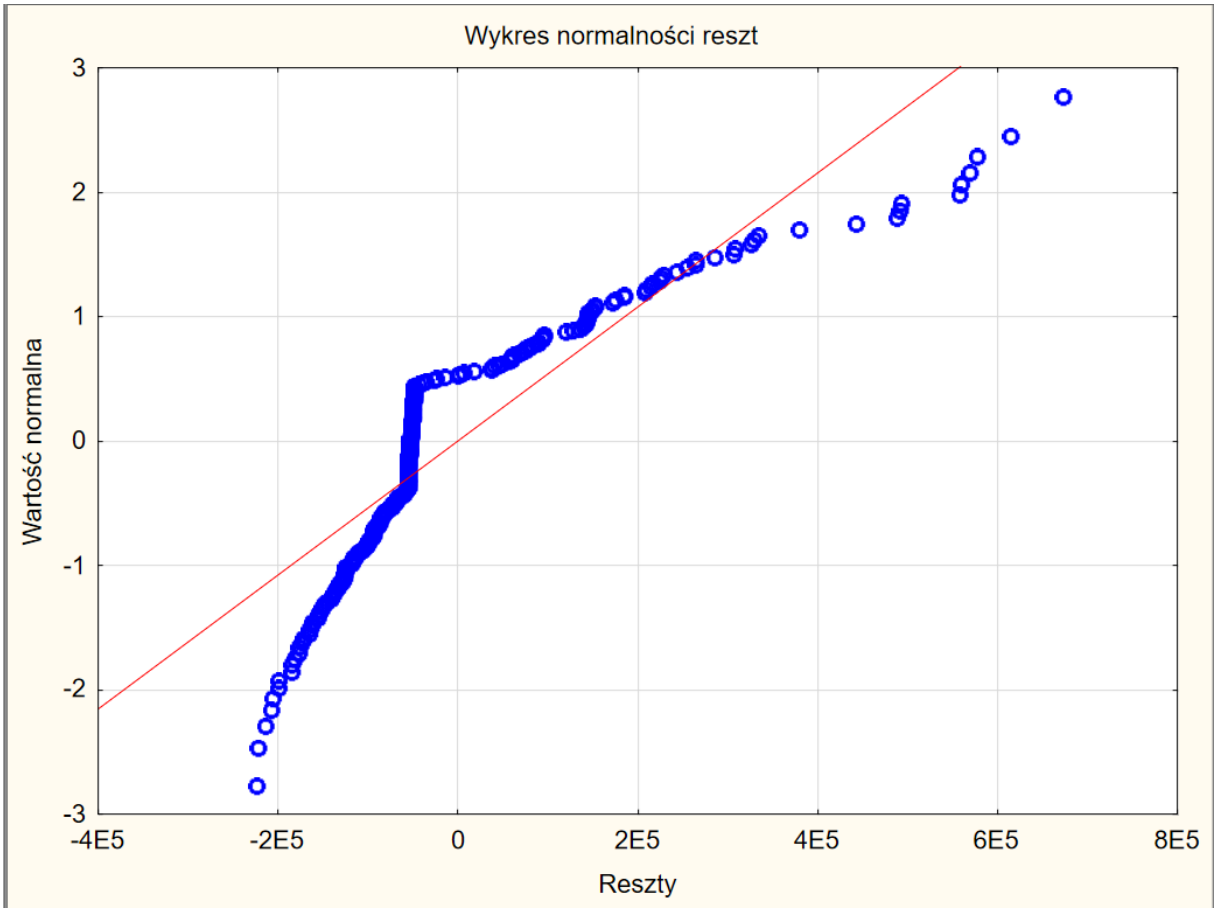
p< α – odrzucono H0, zależność liniowa istnieje i jest istotna statystycznie.

Losowość odchyleń



Założenie heteroscedastyczności nie jest spełnione, gdyż punkty nie są rozproszone równo, większość punktów znajduje się na początku wykresu.

Normalność rozkładu:



Wartości znacznie odstają od krzywej teoretycznej, nie stanowiąc rozkładu normalnego co potwierdza wynik testu Shapiro-Wilk'a

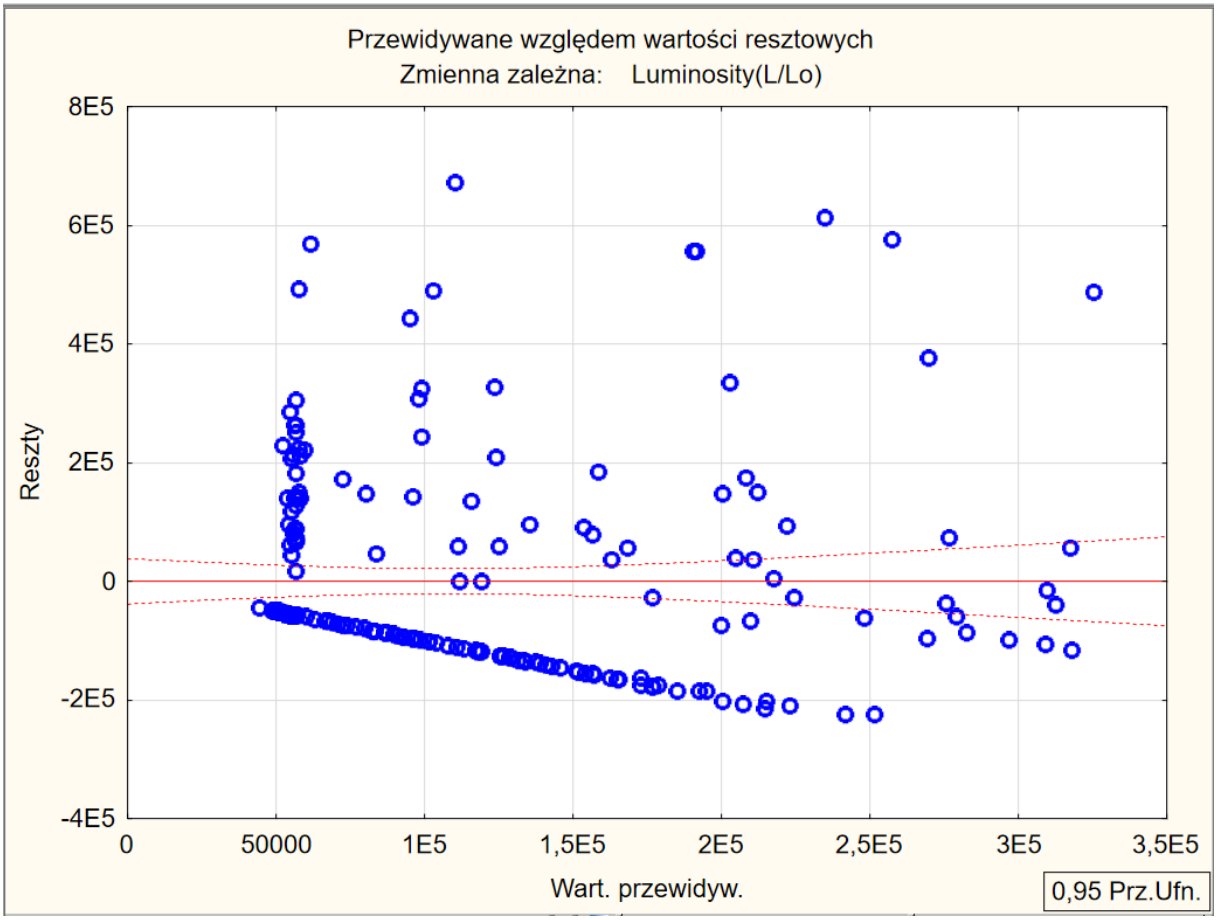
	Testy normalności (Wartości przewidywane i reszty w Skoroszyt2)					
Zmienna	**N**	maks D	K-S p	Lillief. p	W	p
Reszta	240	0,291421	p < ,01	p < ,01	0,801272	0,000000

H0: reszty mają rozkład normalny

H1: reszty nie mają rozkładu normalnego

p < α – zatem odrzucono H0, reszty nie mają rozkładu normalnego

Brak rozkładu normalnego.



Nie występuje stałe rozproszenie reszt, tak więc założenie homoscedastyczności nie jest prawdziwe, występuje heteroscedastyczność.

Interpretacja modelu i wnioski:

Współczynnik korelacji wynosi 0,39, jest to korelacja przeciętna, tak więc jest ona istotna statystycznie . Korelacja jest dodatnia, wraz ze wzrostem temperatury rośnie jasność gwiazdy. Współczynnik determinacji wynosi r2=0,15,co oznacza, że 15 % zmienności zmiennej jasności jaka jest generowana przez gwiazdę wyjaśniono przez model regresji liniowej. Przy temperaturze 0 K jasność wynosi ok.29736 [J/s], wzrost temp. o 1K powoduje wzrost jasności o 7,4 [J/s] .

4.Podsumowanie.

Badanie zależności pomiędzy temperaturą gwiazdy a emitowaną przez nią jasnością ukazuje tendencje wzrostową jasności w miarę wzrostu temperatury gwiazdy, ponadto najwięcej zaobserwowanych gwiazd posiada temperaturę nie przekraczającą 5000 K, tak więc można przypuszczać, że we wszechświecie występuje najwięcej tego typu gwiazd lub okres życia tych z większą temperaturą jest krótszy jest to jednak już tylko przypuszczenie. Jedynie 15% zmienności zmiennej jasność została wyjaśniona modelem regresji, tak więc jest ona zależna od wielu innych czynników.