

# Badanie ogłoszeń używanych samochodów - Raport

**401039, Michał Szczurek, czwartek 14<sup>40</sup>**

*AGH, Wydział Informatyki Elektroniki i Telekomunikacji  
Rachunek prawdopodobieństwa i statystyka 2020/2021*

Kraków, 24 stycznia 2021

*Ja, niżej podpisany własnoręcznym podpisem deklaruję, że przygotowałem przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.*

Michał Szczurek

## Spis treści

<b>1</b>	<b>Streszczenie raportu</b>	<b>1</b>
<b>2</b>	<b>Opis danych</b>	<b>2</b>
<b>3</b>	<b>Analiza danych</b>	<b>2</b>
3.1	Uniwersalne funkcje uzyte do analizy danych . . . . .	2
3.2	Analiza ceny sprzedawanych aut . . . . .	3
3.3	Analiza liczby zdjec na ogłoszeniach . . . . .	7
3.4	Analiza liczby przejechanych kilometrow sprzedawanych aut . . .	10
3.5	Analiza lat produkcji sprzedawanych aut . . . . .	13
<b>4</b>	<b>Badanie wybranych rozkładów i estymatorów przedzialowych</b>	<b>17</b>
4.1	Wprowadzenie . . . . .	17
4.2	Metody testowania rozkładów . . . . .	18
4.3	Badanie estymatorów przedzialowych . . . . .	19
4.4	Rozkład przejechanych kilometrów dla aut marki Porsche . . . .	19
4.5	Rozkład przejechanych kilometrów dla aut marki Chery . . . .	21
4.6	Rozkład cen dla aut marki SsangYong . . . . .	24
<b>5</b>	<b>Zależności między zmiennymi</b>	<b>27</b>
<b>6</b>	<b>Wnioski</b>	<b>39</b>

## 1 Streszczenie raportu

Raport powstał w oparciu o analizę danych dotyczących ogłoszeń sprzedaży używanych samochodów na Białorusi. Głównym celem analizy było sprawdzenie jakie cechy mają wpływ na cenę sprzedawanych samochodów. Analizy dokonano przy pomocy języka R i programu RStudio. W badaniu wykorzystano następujące cechy:

- manufacturer name - nazwa marki,
- color - kolor,
- odometer value - liczba przejechanych kilometrów,
- year produced - data produkcji samochodu,
- engine capacity - pojemność silnika (w litrach),
- body type - typ samochodu (np. suv, sedan),
- price usd - cena w dolarach,
- number of photos - liczba zdjęć w ogłoszeniu,
- days listed - liczba dni od wystawienia ogłoszenia

Analizując dane szczególną uwagę przywiązyano do zależności między ceną a liczbą przejechanych kilometrów i rokiem produkcji. Wpływ marek i kolorów samochodów nie został jednak pominięty (jak i innych mniej istotnych cech). Stworzony podczas badania model wskazał, że dla wszystkich cech ilościowych za wyjątkiem liczby przejechanych kilometrów mają pozytywny (wraz ze wzrostem badanych wartości cena rośnie) wpływ na ceny samochodu.

Podczas badań odkryto również, że wbrew intuicji większość cech najprawdopodobniej nie pochodzi z rozkładu normalnego. Prawdopodobnie jest to spowodowane, tym że na dane takie jak ilość zdjęć, cena ludzie mają bezpośredni wpływ.

## 2 Opis danych

Dane zostały zebrane z różnych stron internetowych 2 stycznia 2019 i umieszczone na stronie internetowej pod adresem

<https://www.kaggle.com/lepkhenkov/usedcarscatalog>.

Dane są bardzo bogatym zbiorem informacji dotyczących zarówno stanu sprzedawanych aut jak i samych ogłoszeń (dla każdego ogłoszenia zebrano 30 cech), w związku z czym nie wszystkie cechy zostaną przeanalizowane w niniejszym raporcie. Głównymi obiektami analizy były:

- Cena samochodów
- Liczba zdjęć zawarta w ogłoszeniach

- Liczba przejechanych kilometrów
- Lata produkcji sprzedawanych samochodów

Plik zawierający dane został wobec powyższego odpowiednio przetworzony tak, by nie zawierać danych nieujętych w analizie. Dodatkowo dane zostały przefiltrowane pod kątem poprawności - część rekordów była uszkodzona.

### 3 Analiza danych

Poniżej zamieszczono wyniki uzyskane wskutek analizy danych.

#### 3.1 Uniwersalne funkcje uzyte do analizy danych

Ponizej znajduje sie kod funkcji uzytych podczas badan

1. Funkcja obliczajaca dominante

```
> mode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
```

2. Funkcja wypisujaca podstawowe informacje o wybranej cesze

```
> get_basic_stats <- function(value){
+   cat("Wartość średnia:",formatC(mean(value), digits = 2, format = "f"),"\n")
+   cat("Medianą:",formatC(median(value), digits = 2, format = "f"),"\n")
+   cat("Odchylenie standardowe:",formatC(sd(value), digits = 2, format = "f"),"\n")
+   cat("Wariancja:",formatC(var(value), digits = 2, format = "f"),"\n")
+   cat("Minimum:",formatC(min(value), digits = 2, format = "f"),"\n")
+   cat("Maksimum:",formatC(max(value), digits = 2, format = "f"),"\n")
+   cat("Dominanta:",formatC(mode(value), digits = 2, format = "f"),"\n")
+   cat("Skośność:",formatC(skewness(value), digits = 2, format = "f"),"\n")
+   cat("Kurtoza:",formatC(kurtosis(value), digits = 2, format = "f"),"\n")
+   cat("Rozstęp międzykwartylowy:",formatC(IQR(value), digits = 2, format = "f"),
+        "\n")
+ }
```

#### 3.2 Analiza ceny sprzedawanych aut

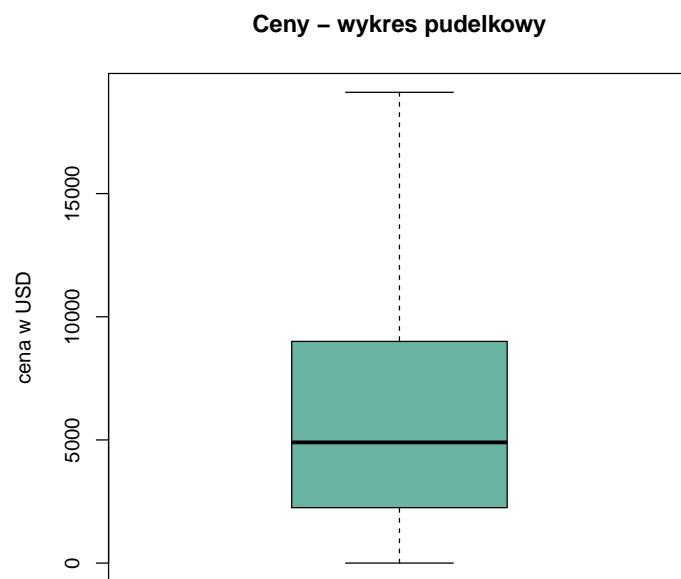
Poniżej przedstawiono informacje dotyczące cen podanych w ogłoszeniach sprzedaży samochodów. Cena wyrażona została w Dolarze Amerykańskim [USD].

Wartość średnia: 6740.37  
 Medianą: 4900.00  
 Odchylenie standardowe: 6444.72  
 Wariancja: 41534376.29

Minimum: 1.00  
Maksimum: 50000.00  
Dominanta: 1500.00  
Skośność: 2.23  
Kurtoza: 10.24  
Rozstęp miedzykwartylowy: 6750.00

Sporządzono również następujące wykresy dotyczące ceny samochodów:

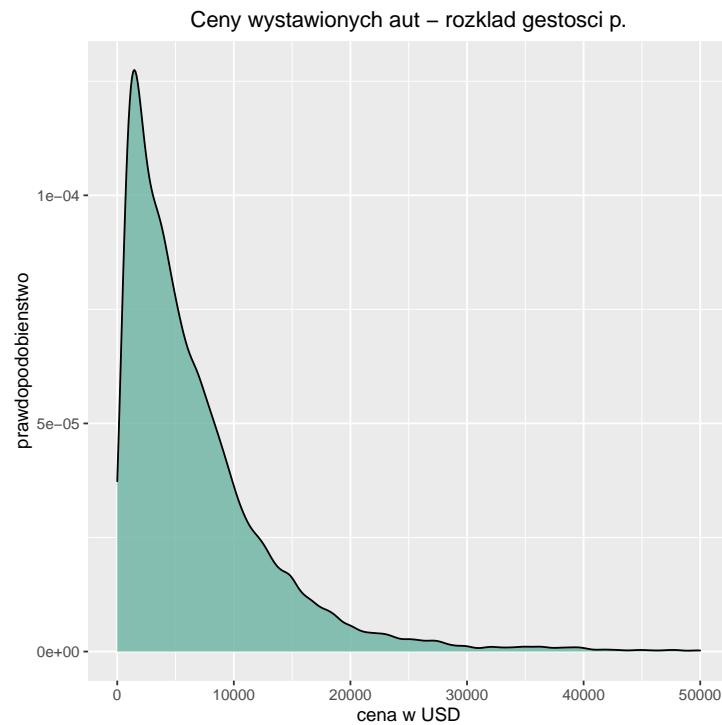
1. Wykres pudelkowy Powyzszy wykres dobrze ilustruje informacje podane



Rysunek 1: Wykres pudelkowy cen samochodow

powyzej. Na wykres nie naniesiono punktow odstajacych "outliers" poniewaz ze wzgledu na duza liczbe danych takich punktow bylo wiele, co pogarszalo czytelosc.

## 2. Wykres rozkładu gęstości prawdopodobieństwa



Rysunek 2: Wykres gestosci rozkladu cen samochodow

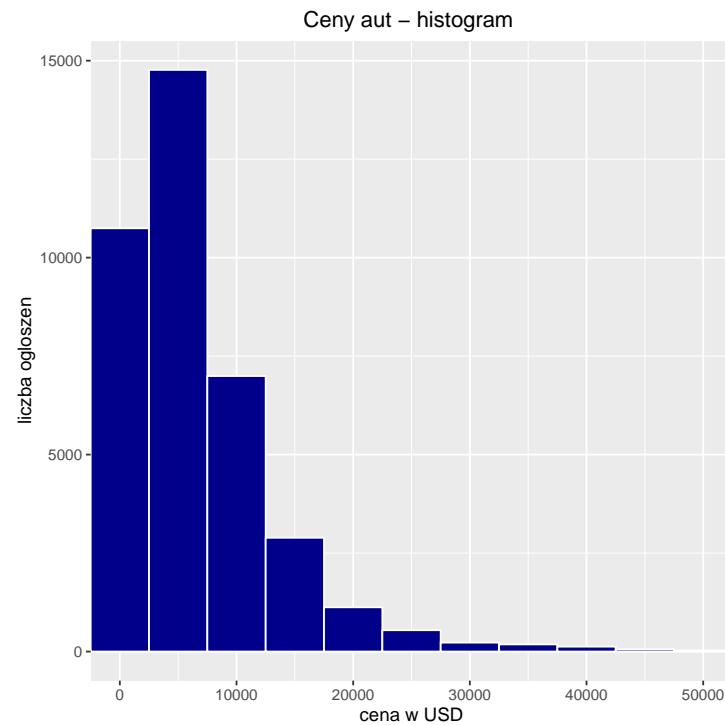
Powyzszy wykres umozliwia dokonanie obserwacji:  
Znaczna czesc aut wystawiona jest za cene nie przekraczajaca 10000 [USD]. Dokladniej jest to

**78.92 % wszystkich pojazdow**

Znikoma czesc aut wystawiana jest za cene przekraczajaca 30000 [USD].  
Dokladniej jest to

**1.24 % wszystkich pojazdow**

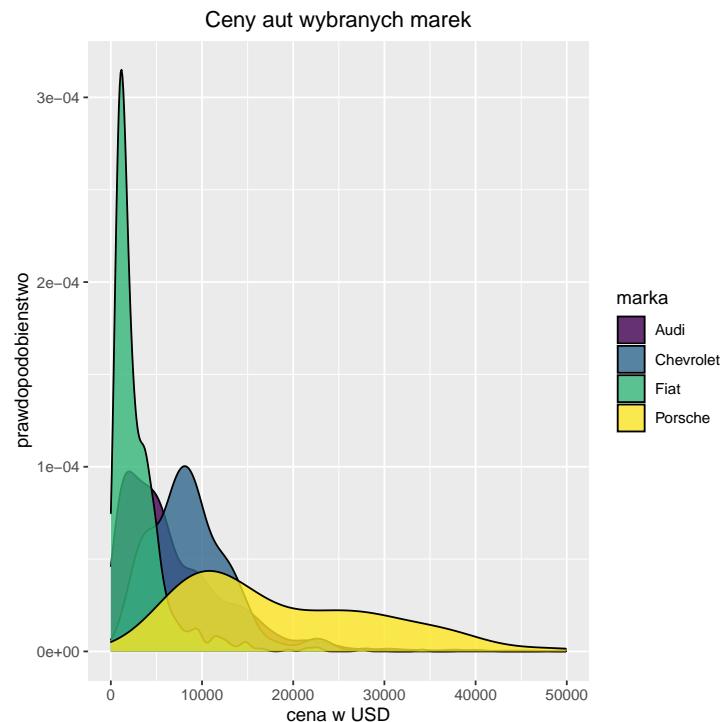
3. Histogram z przedziałem 5000 [USD]



Rysunek 3: Histogram cen samochodow z przedziałem 5000 [USD]

Na powyższym wykresie widać, że najpopularniejszym przedziałem cenowym używanych samochodów jest 2500 - 7500 [USD].

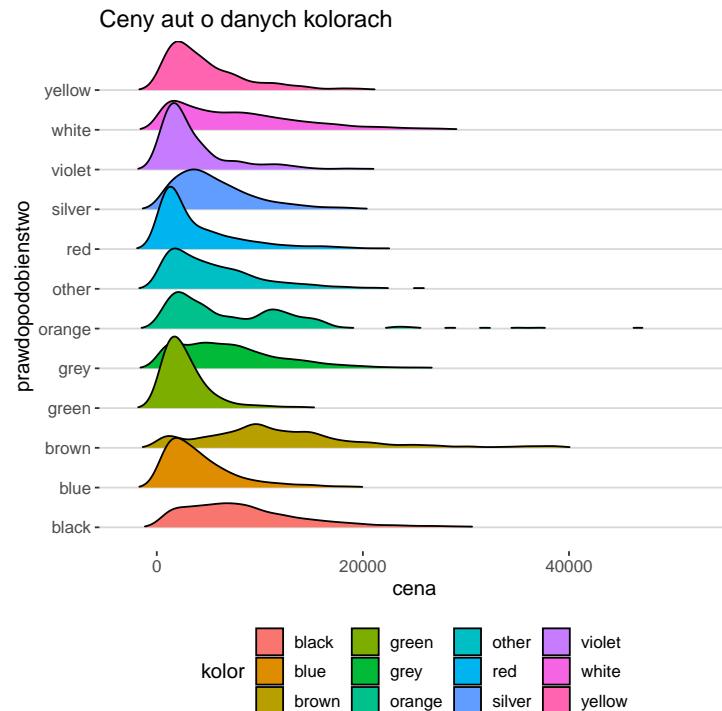
4. Wykres gęstości cen dla konkretnych marek samochodów (ze względu na dużą liczbę marek na wykresie przedstawiono tylko kilka wybranych)



Rysunek 4: Wykres rozkładu gęstości cen samochodów dla wybranych marek

Na powyższym wykresie możemy dostrzec wyraźna różnicę w cenie w zależności od marki samochodu. Szczególna uwagę przyciąga wykres gęstości cen Fiata - przytaczająca większość aut tej marki sprzedaje się poniżej 5000 [USD]. Zupełnie inaczej sytuacja wygląda w sytuacji Porche. Ceny tej marki są znacznie bardziej równomiernie rozłożone z dominanta w okolicach 10000 [USD], czyli około 4 razy drożej niż Fiat. Wykresy cen Audi i Chevroleta nie są tak spektakularne, jednak warto zaznaczyć, że w obu przypadkach funkcja gęstości jest skoncentrowana bardziej na prawo niż w przypadku ogólnej gęstości samochodów - są to marki drożej wystawiane niż przeciętne.

- Wykres gęstości cen w dla konkretnych kolorów samochodów (Należy zwrócić uwagę na fakt, że kolor wykresu niekoniecznie pokrywa się z kolorem pojazdu)



Rysunek 5: Wykres rozkładu gęstości cen samochodów dla wybranych kolorów

W wiekszości przypadków wykres przypomina ogólny wykres gęstości rozkładu cen. Na szczególną uwagę zasługują auta czarne, szare i brązowe. Ceny tych kolorów są bardziej równomiernie rozłożone, co może oznaczać, że auta w tych kolorach są chętniej kupowane, więc cena może być wyższa niż cena tego samego samochodu w innym kolorze.

### 3.3 Analiza liczby zdjęć na ogłoszeniach

Poniżej przedstawiono informacje dotyczące liczby zdjęć aut na ogłoszeniach

Wartość średnia: 9.70  
 Medianą: 8.00  
 Odchylenie standardowe: 6.10  
 Wariancja: 37.21  
 Minimum: 1.00  
 Maksimum: 86.00

Dominanta: 6.00

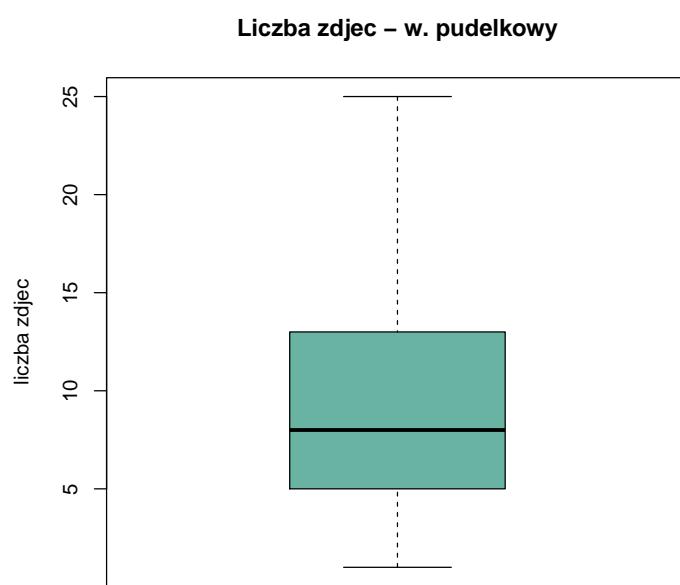
Skośność: 1.59

Kurtoza: 7.91

Rozstęp międzykwartylowy: 8.00

Sporządzono również następujące wykresy dotyczące liczby zdjęć samochodów:

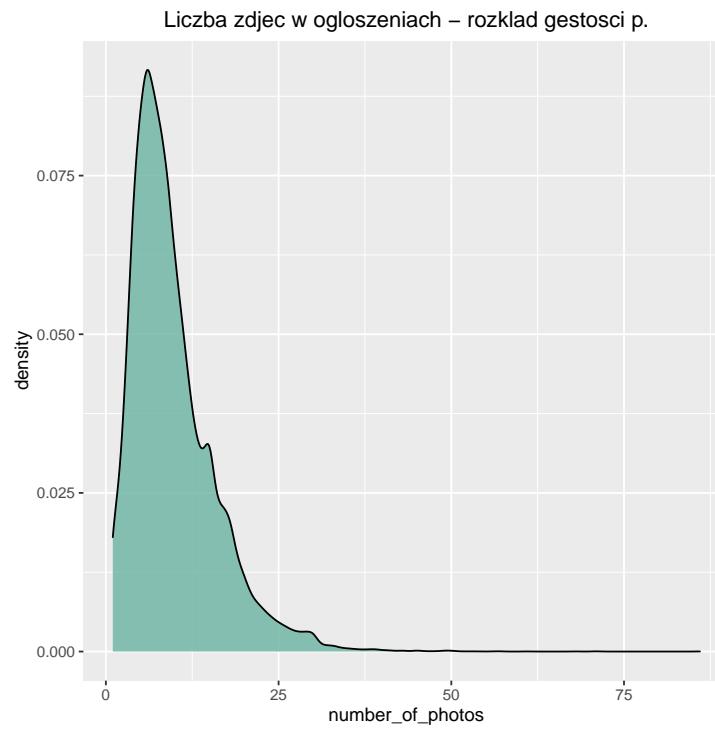
1. Wykres pudelkowy Powyższy wykres dobrze ilustruje informacje podane



Rysunek 6: Wykres pudelkowy liczby zdjęć na ogłoszeniach

powyżej. Na wykres nie naniesiono punktów odstających ”outliers” ponieważ ze względu na dużą liczbę danych takich punktów było wiele, co pogarszało czytelność.

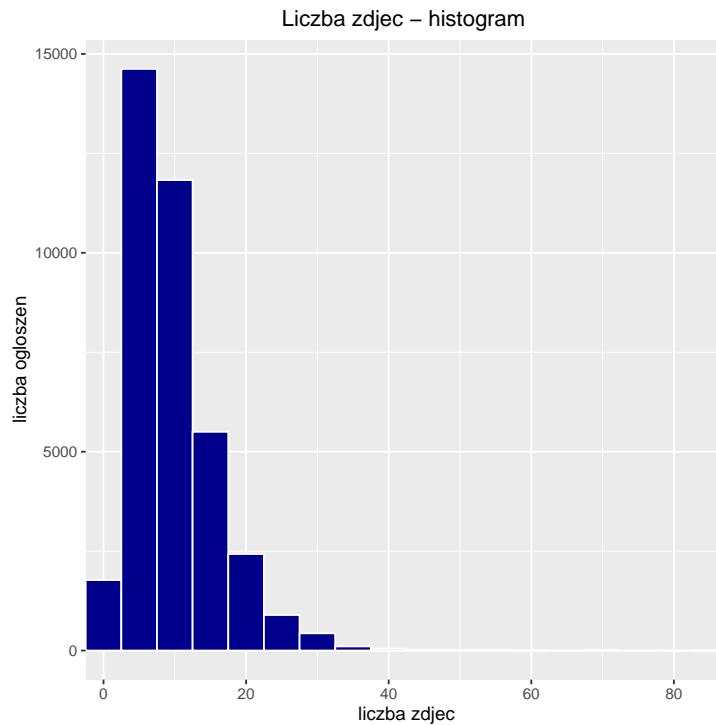
2. Wykres rozkładu gęstości liczby zdjęć



Rysunek 7: Wykres rozkładu gęstości liczby zdjęć

Na powyższym wykresie można zauważyc, że do znacznej części ogłoszeń dodawane jest mniej niż 10 zdjęć, a tylko nieliczni sprzedawcy decysują się na pokazaniu na ogłoszeniu więcej niż 25 zdjęć.

### 3. Histogram z przedziałem 5 [zdjec]



Rysunek 8: Histogram liczby zdjec z przedziałem 5 [zdjec]

Powyzszy wykres pokazuje, ze najwiecej osob dodaje do ogłoszeń od 3 do 12 zdjec (przedzial pierwszy ma srodek w zerze). Wykres ten pokazuje tez wyrazie (lepiej niz poprzedni), jak niewiele osob decyduje sie na dodanie 2 lub mniej zdjec.

### 3.4 Analiza liczby przejechanych kilometrow sprzedawanych aut

Ponizej przedstawiono informacje dotyczące liczby przejechanych kilometrow aut na ogłoszeniach

Wartość średnia: 251067.32

Mediana: 250000.00

Odchylenie standardowe: 134317.94

Wariancja: 18041309028.49

Minimum: 0.00

Maksimum: 1000000.00

Dominanta: 300000.00

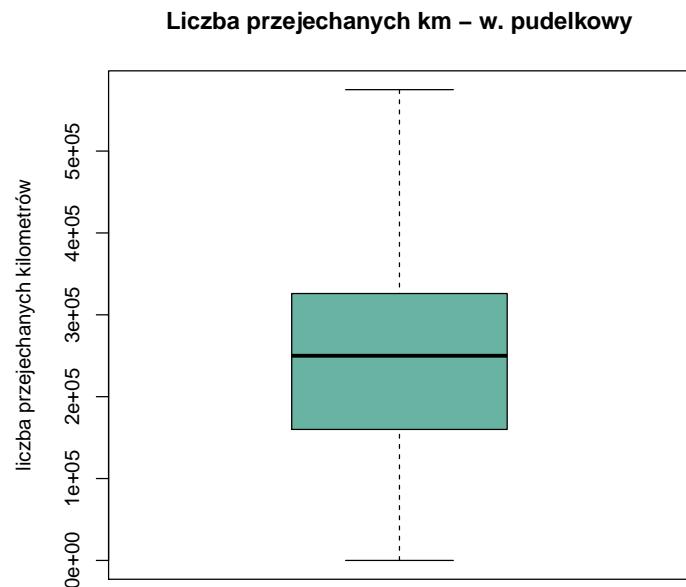
Skośność: 1.13

Kurtoza: 7.81

Rozstęp międzykwartylowy: 166000.00

Sporządzono również następujące wykresy dotyczące liczby przejechanych kilometrów:

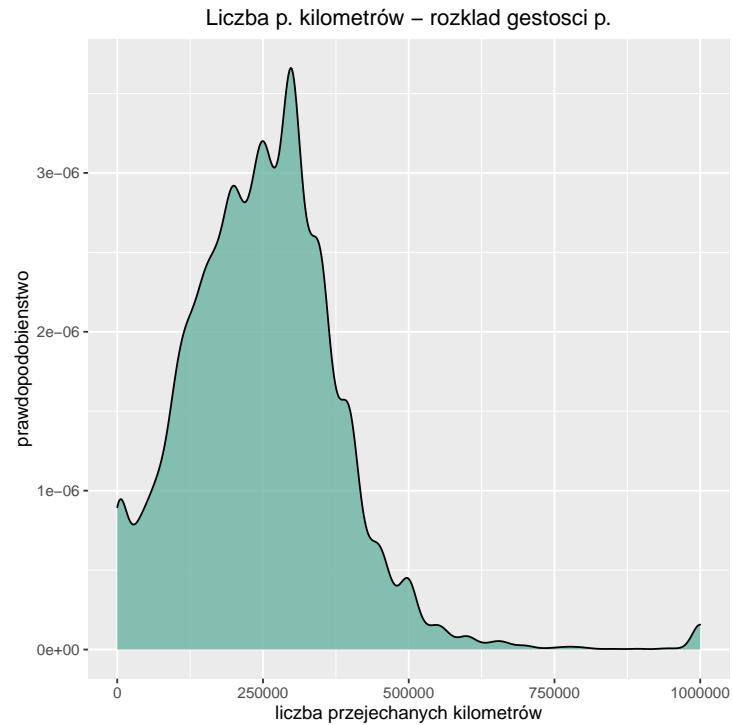
1. Wykres pudelkowy



Rysunek 9: Wykres pudelkowy liczby przejechanych kilometrów

Powyzszy wykres dobrze ilustruje informacje podane powyzej. Na wykres nie naniesiono punktow odstajacych "outliers" poniewaz ze wzgledu na duza liczbe danych takich punktow bylo wiele, co pogarszalo czytelnosc.

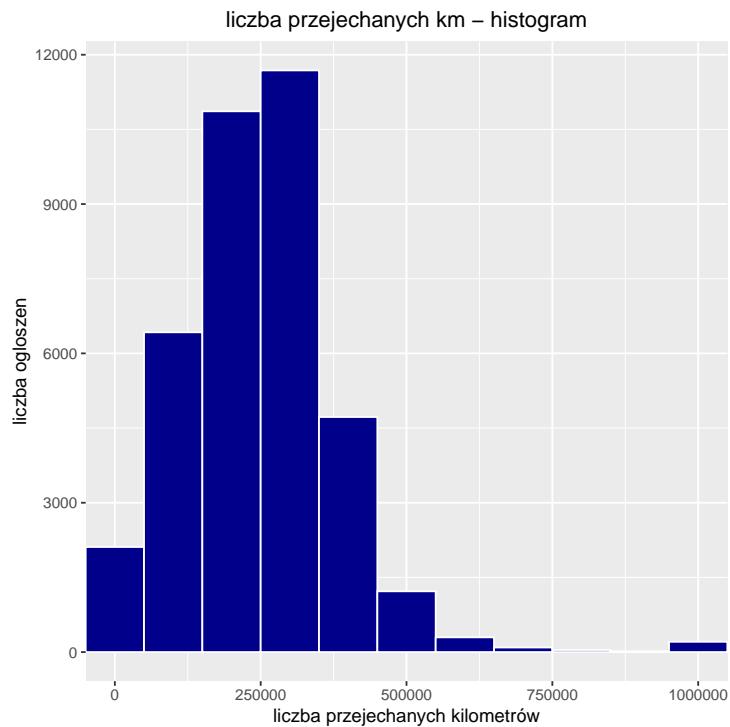
## 2. Wykres rozkładu gęstości prawdopodobieństwa



Rysunek 10: Wykres rozkładu gęstości przejechanych kilometrow

Rozkład jest w prawdziwie daleki od tego by mozna go bylo nazwac normalnym, mozna jednak zaobserwować charakterystyczny (choćiaż mocno znieksztalcony) kształt dzwonu. Na wykresie dwie właściwości szczególnie zwrociły moja uwagę - Pierwsza z nich były nagle, skokowe wzrosty gestosci w kilku miejscach. Skok taki mozna zaobserwować na przykład w okolicach 250000 [km]. Może to wskazywać na skłonność ludzi do przejezdania pewnej ustalonej liczby kilometrów (np. tak "ładnej" jak 250000) przed sprzedazem auta. Inna możliwość jest to, że sprzedawcy zaokrąglają wartości w ogłoszeniach do "okrągłych" liczb. Druga ciekawa zależność był nagły wzrost sprzedazy w okolicach miliona kilometrów.

### 3. Histogram z przedziałem 100000 [km]



Rysunek 11: Histogram przejechanych kilometrow z podzialem 100000 [km]

Na wykresie widac, ze najczesciej sprzedawane sa samochody majace na licznikach wartosc w przedziale 100 - 350 tys. km.

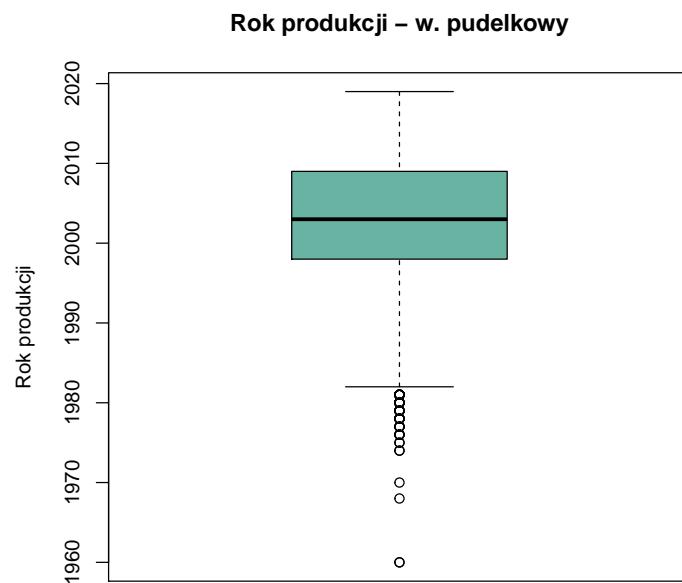
### 3.5 Analiza lat produkcji sprzedawanych aut

Ponizej przedstawiono informacje dotyczące lat produkcji aut wystawionych w badanych ogłoszeniach.

Wartość średnia: 2003.18  
Medianą: 2003.00  
Odchylenie standardowe: 7.72  
Wariancja: 59.56  
Minimum: 1960.00  
Maksimum: 2019.00  
Dominanta: 1998.00  
Skośność: -0.14  
Kurtoza: 2.57  
Rozstęp miedzykwartylowy: 11.00

Sporządzono również następujące wykresy dotyczące roku produkcji:

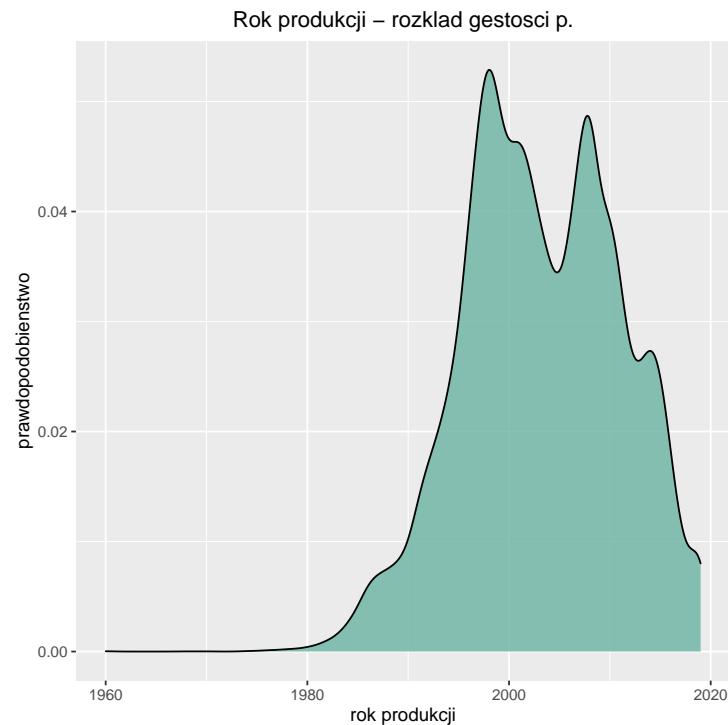
1. Wykres pudełkowy



Rysunek 12: Wykres pudelkowy lat produkcji sprzedawanych aut

Powyzszy wykres dobrze ilustruje informacje podane powyzej. Wykres dobrze ilustruje fakt, ze znaczna czesc sprzedawanych aut pochodzi z przelomu wieku XIX i XX.

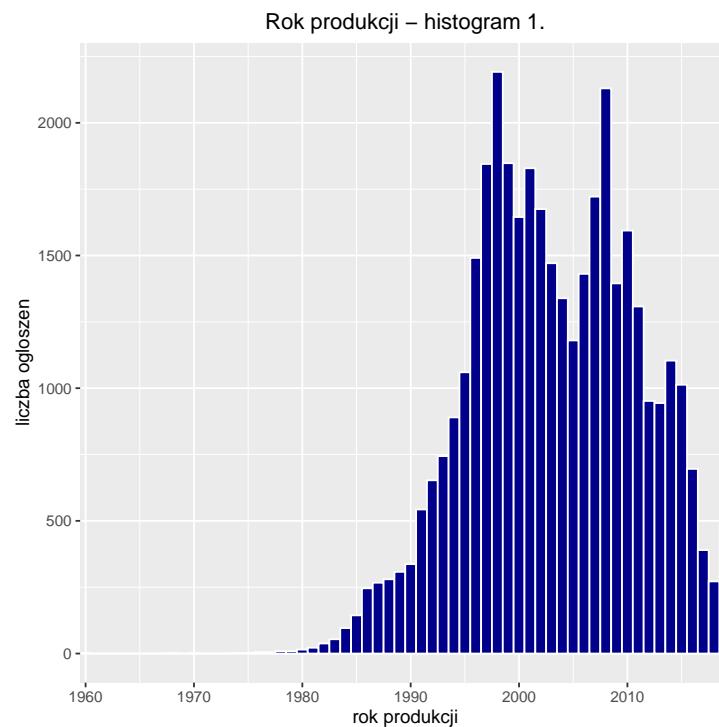
## 2. Rozkład gestosci lat produkcji [lat]



Rysunek 13: Rozkład gestosci lat produkcji

Lata produkcji nie mają rozkładu ciąglego. Pomimo to pozwolilem sobie na pozostawienie wykresu gestosci w tej formie - jest on przejrzysty.  
Postanowilem jednak dodatkowo stworzyć histogram o przedziale 1 roku, który dokładniej oddaje zależności w badanych danych. Wykres potwierdza obserwacje, które były widoczne już na wykresie pudelkowym.

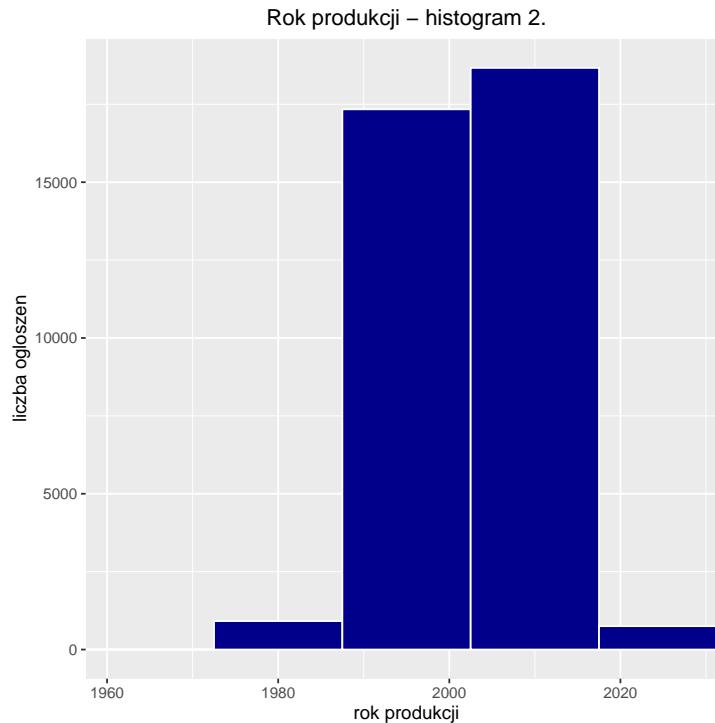
3. Histogram z przedziałem 1 [roku]



Rysunek 14: Histogram lat produkcji z przedziałem 1 [roku]

Ciekawym zjawiskiem obserwowanym na wykresie jest znaczne zmniejszenie się liczby sprzedawanych aut z okolic roku 2005 w stosunku do aut z lat 2000 i 2008. Niestety pomimo dokonania analizy kursów walut i wydarzeń ogólnoswiatowych jak i tych na Białorusi nie udało mi się ustalić przyczyny tego fenomenu.

4. Histogram z przedziałem 15 [lat]



Rysunek 15: Histogram lat produkcji z przedziałem 15 [lat]

Wykres bardzo wyraznie przedstawia omawiana już zależność.  
Przytaczająca większość pojazdów została wyprodukowana w latach  
1995 - 2015.

## 4 Badanie wybranych rozkładów i estymatorów przedzialowych

### 4.1 Wprowadzenie

Podczas analizowania danych w poprzednim podpunkcie nie zauważylem, zeby rozkłady gestosci przypominaly ktore z popularniejszych rozkładów. Postanowiłem wobec tego wykorzystać fakt, że populacja badanych ogłoszeń jest bardzo liczna (ok. 39 tys.). Wobec tego postanowiłem przeanalizować rozkłady wybranych cech dla podgrup mających jakąś wspólną cechę. Wyniki tych analiz można zaobaczyć poniżej.

## 4.2 Metody testowania rozkladow

Rozklady zmiennych badana przywiazujac uwage glownie do tego, czy moga one pochodzi z rozkladu normalnego - taki fakt umozliwilby wykonanie innych ciekawszych testow.

### Test Shapiro-Wilka

Test Shapiro - Wilka nie wymaga spełnienia szczególnych założen. Pozwala on sprawdzić, czy zmienna pochodzi z rozkładu normalnego. Hipoteza zerowa i alternatywna w testie mają postać:

**Hipoteza zerowa:** Próba pochodzi z populacji o rozkładzie normalnym

**Hipoteza alternatywna:** Próba nie pochodzi z populacji o rozkładzie normalnym  
Rozkład uznaje się za normalny, gdy  $p$  - wartość testu jest większa od 0.05.

### Wykres kwantyl - kwantyl

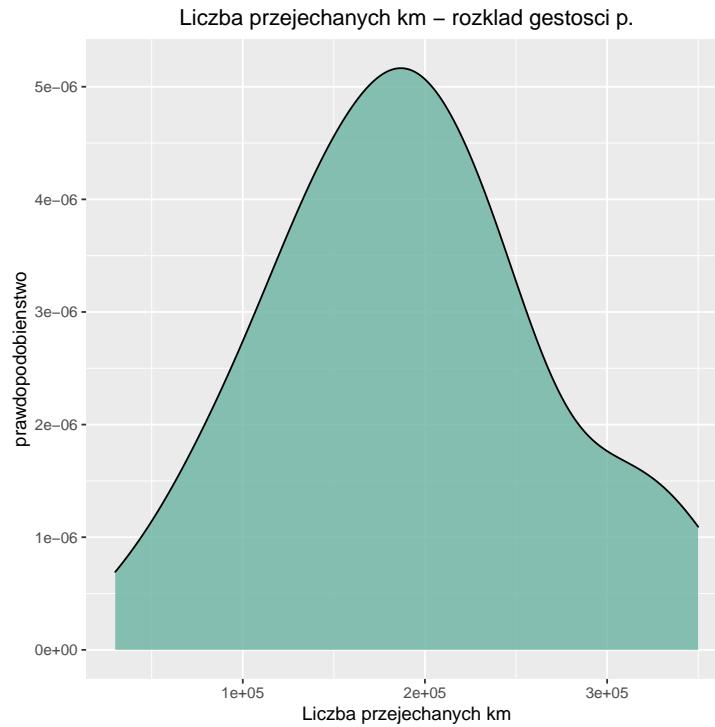
Wykres pokazuje jak kwantyle badanego rozkładu mają się do teoretycznych. Rozkład uznaje się za normalny, gdy punkty na wykresie leżą wzduż krzywej.

## 4.3 Badanie estymatorów przedzialowych

Do badania estymatora wartości oczekiwanej rozkładu użyto  $t$  - testu (Wersji dla jednej próbki). Test ten pozwala wyznaczyć przedział ufności wartości oczekiwanej i jako hipoteza zerowa przyjmuje fakt, że rzeczywista wartość oczekiwana populacji jest równa podanej wartości. Ważnym założeniem tego testu jest konieczność tego, aby zmienna pochodziła z rozkładu normalnego. Podczas testowania nie używamy testu w celu sprawdzenia, czy dana wartość jest rzeczywista wartość oczekiwana, a raczej sprawdzalem jakie są przedziały ufności na danym poziomie istotności.

## 4.4 Rozkład przejechanych kilometrów dla aut marki Porsche

Jednym z rozkładów, który zwrocił moja uwagę jako wyjątkowo podobny do rozkładu normalnego był rozkład przejechanych kilometrów dla aut marki Porsche.



Rysunek 16: Rozkład przejechanych kilometrów dla aut marki Porsche

Pierwszym krokiem było sporządzenie wykresu Q-Q.

Jak widać na wykresie nie wszystkie punkty leżą idealnie na prostej, jednak wciąż są one na tyle blisko że zasadne jest przeprowadzenie testu Shapiro - Wilka

```
> print(shapiro.test(porsche$odometer_value))
```

```
Shapiro-Wilk normality test
```

```
data: porsche$odometer_value
W = 0.98547, p-value = 0.6846
```

P - wartość testu znaczaco przekroczyła 0.05. Można więc założyć, że zmienna pochodzi z rozkładu normalnego.

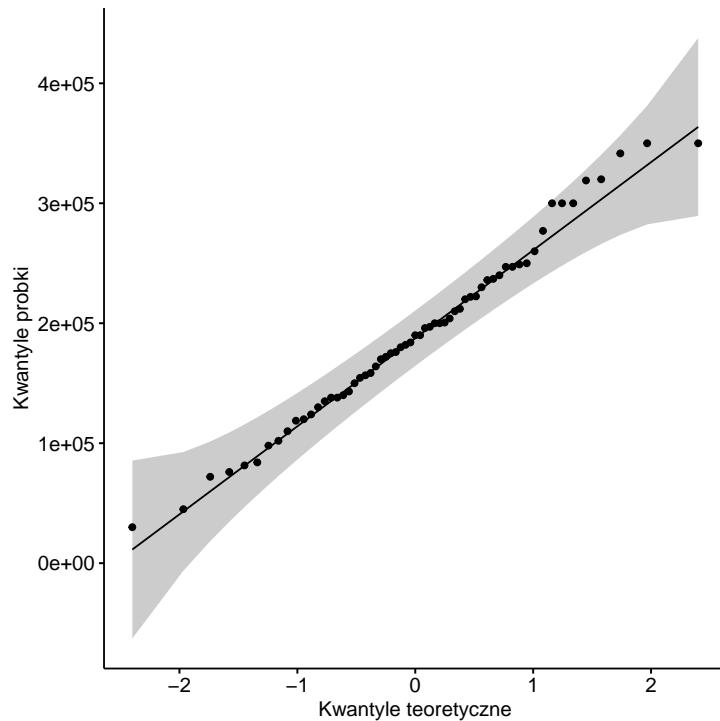
Wobec tego że rozkład zmiennej jest normalny, warto jest dowiedzieć się o niej czegos więcej. Wykonajmy więc standardową analizę:

Wartość średnia: 190581.54

Medianą: 190000.00

Odchylenie standardowe: 75291.37

Wariancja: 5668791109.02



Rysunek 17: Wykres Q-Q dla rozkładu przejechanych kilometrów Porsche

Minimum: 30000.00  
 Maksimum: 350000.00  
 Dominanta: 300000.00  
 Skośność: 0.19  
 Kurtoza: 2.64  
 Rozstęp miedzykwartylowy: 99000.00

Dzieki temu, ze założylismy normalność rozkładu, możliwe jest wykonanie t - testu:

```

> t.test(porsche$odometer_value)

One Sample t-test

data: porsche$odometer_value
t = 19.77, df = 60, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 171298.5 209864.6
sample estimates:

```

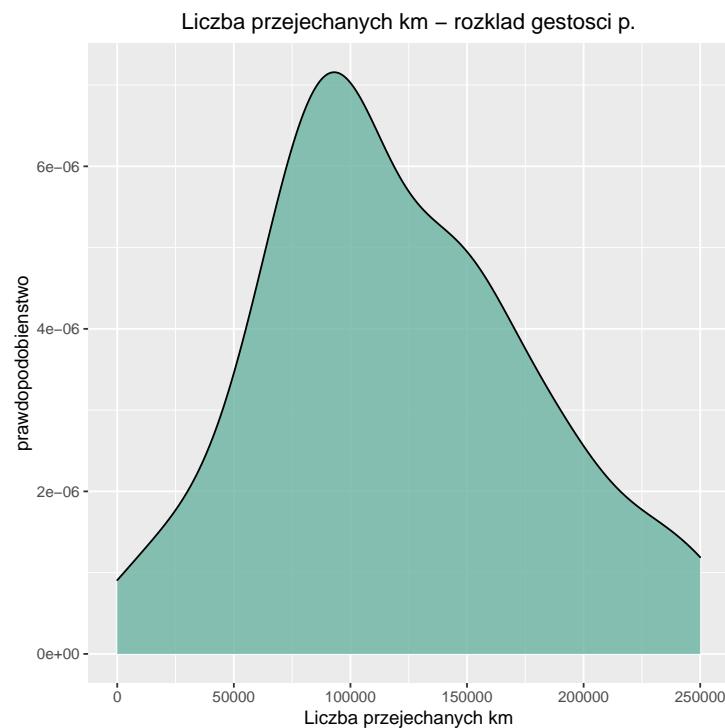
mean of x

190581.5

Test wskazuje na to, ze srednia liczba przejechanych kilometrow dla sprzedawanych Porsche mieści się w przedziale [171298.5 209864.6] z 95% pradopodobienstwem. Hipoteza zerowa zerowa została oczywiście odrzucona - wartość srednia populacji nie jest równa 0. Sprawdzanie hipotezy zerowej nie było jednak prawdziwym celem testu - tym było poznanie przedziału ufności.

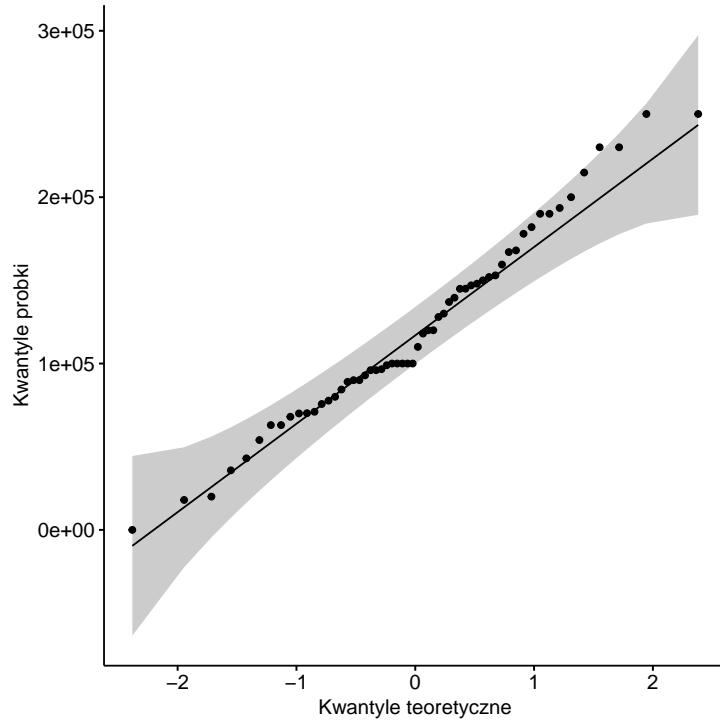
#### 4.5 Rozkład przejechanych kilometrów dla aut marki Chery

Znalezienie znanych rozkładów w badanych danych było sporym wyzwaniem. Jednak wartość przejechanych kilometrów wydala się być zmienną, która najczęściej miała rozkład przypominający rozkład normalny. Być może spowodowane, jest to "losowość" tego czynnika - przykładowo ceny zwykle są dobierane z pewną premedytacją. Wobec powyższego zdecydowalem się przenalizować liczby przejechanych kilometrów po raz kolejny - tym razem dla aut marki Chery. Ponizej znajduje się jego wykres:



Rysunek 18: Rozkład przejechanych kilometrów dla aut marki Chery

Nastepnym krokiem bylo sporzadzenie wykresu Q-Q.



Rysunek 19: Wykres Q-Q dla rozkladu przejechanych kilometrow Porsche

Podobnie jak poprzednim razem - wykres nie przedstwia idealnej krzywej, daje jednak nadzieje na pozytywny wynik testu Shapiro - Wilka.

```
> print(shapiro.test(chery$odometer_value))
```

```
Shapiro-Wilk normality test
```

```
data: chery$odometer_value  
W = 0.97754, p-value = 0.3555
```

P - wartosc testu przekroczyła 0.05. Mozna wiec zalozyc, ze zmienna pochodzi z rozkladu normalnego. Zalozenie to jest jednak bardziej kontrowersyjne niz w przypadku Porsche.

Wobec tego ze rozklad zmiennej uznalismy jako normalny, warto jest dowiedziec sie o niej czegos wiecej. Wykonajmy wiec standardowa analize:

```
Wartość średnia: 120491.84  
Mediana: 105000.00  
Odchylenie standardowe: 57976.45
```

```
Wariancja: 3361268485.05
Minimum: 0.00
Maksimum: 250000.00
Dominanta: 100000.00
Skośność: 0.33
Kurtoza: 2.66
Rozstęp miedzykwartylowy: 71654.00
```

Ponownie zbadam rowniez srednia wartosc przy pomocy t- testu.

```
> t.test(chery$odometer_value)
```

```
One Sample t-test
```

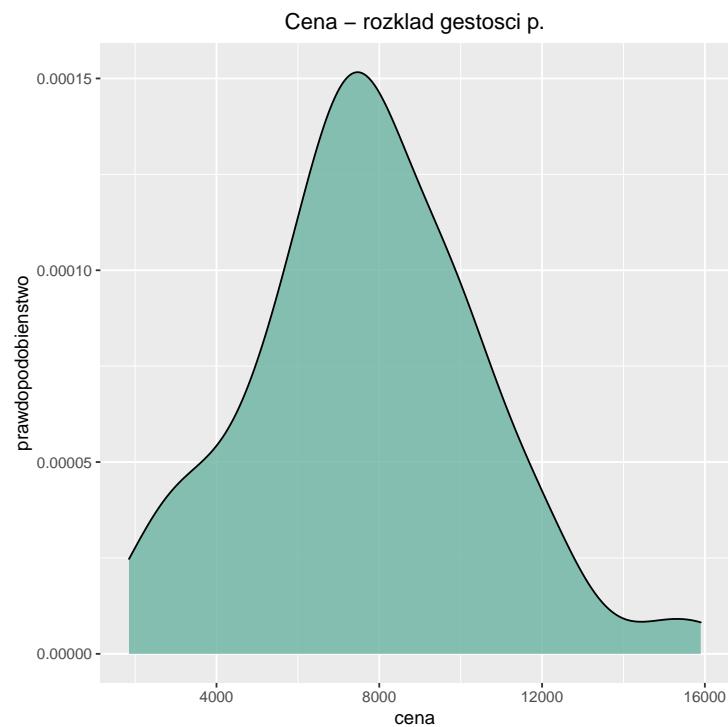
```
data: chery$odometer_value
t = 15.828, df = 57, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 105247.7 135736.0
sample estimates:
mean of x
 120491.8
```

Test wskazuje na to, ze srednia liczba przejechanych kilometrow dla sprzedawanych Chery miesci sie w przedziale  $[105247.7, 135736.0]$  z 95% pradopodobienstwem.

#### 4.6 Rozklad cen dla aut marki SsangYong

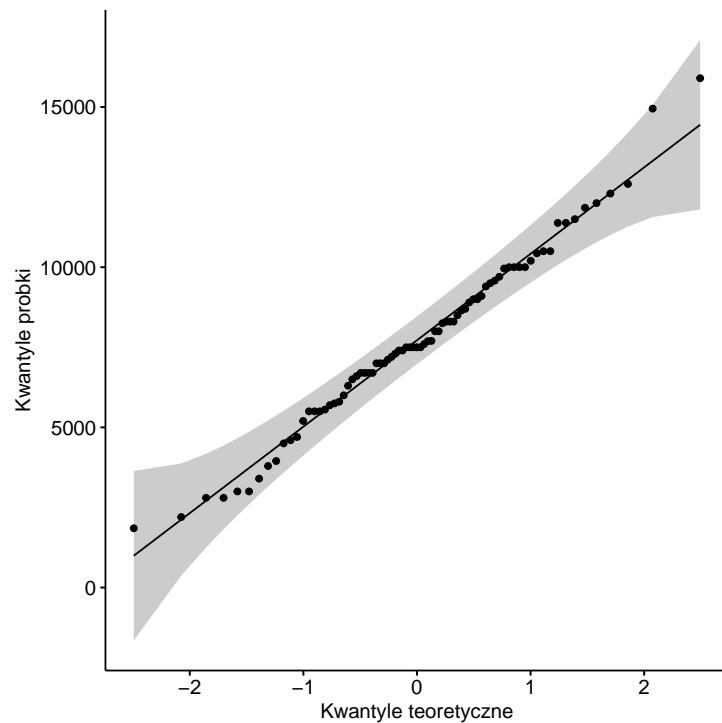
Poprzednio badne rozklady nie satysakcjonowały mnie w pełni. Głównym założeniem projektu było badanie cen a nie liczby przejechanych kilometrów. Niestety znalezienie pozdzbioru aut, dla których rozkład cen przypominałby jakiś znanego rozkład nie było proste. Wobec tego napisalem funkcję, która automatycznie wykonuje test Shapiro - Wilka dla wszystkich podzbiorów aut należących do tej samej marki.

Dzieki temu udało mi się odkryć, że rozkład najbliższy normalnemu posiada wykres cen aut marki SSangYong (p wartość testu wynosiła 0.5061). Ponizej znajduje się ilustracja przedstawiająca ten rozkład.



Rysunek 20: Rozkład cen dla aut marki SsangYong

Sporządzono również wykres Q-Q, aby upewnić się, co do normalności rozkładu:



Rysunek 21: Wykres Q-Q dla cen SsangYong

Ponizej znajdują się podstawowe informacje o cenie aut tej marki:

Wartość średnia: 7719.32  
 Mediana: 7500.00  
 Odchylenie standardowe: 2788.96  
 Wariancja: 7778294.28  
 Minimum: 1850.00  
 Maksimum: 15900.00  
 Dominanta: 6700.00  
 Skośność: 0.24  
 Kurtoza: 3.30  
 Rozstęp miedzykwartylowy: 3639.50

Test wykazał, że z 95% prawdopodobienstwem średnia cena aut tej marki znajduje się w przedziale [7094.630, 8344.015] USD. Skoro rozkład można uznać za normalny, to można policzyć również estymator przedzialowy dla średniej korzystając z t-testu.

One Sample t-test

```

data: sy$odometer_value
t = 18.231, df = 78, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
204779.6 254986.1
sample estimates:
mean of x
229882.8

```

**T - test dla dwóch zmiennych** Korzystajac z normalnosci rozkladu cen samochodow tej marki, postanowilem rowniez wykonac T - test w wersji dla dwóch probek. Ten test pozwala ocenic czy prawdziwe wartosci sredniej dwóch probek sa rozne. W tym celu podzielilem auta marki SsangYong na dwa podzbiory - posiadajace mniej/ wiecej niz 250 tys. km na liczniku (mediana dla wszystkich samochodow). Okazalo sie ze rozklad cen w podzbiorach rowniez moze zostac uznany za normalny - p - wartosc testu dla mniej eksplotowanych aut wynosila 0.697, a dla bardziej eksplotowanych - 0.0495 (co pozwolilem sobie zaokraglic do 0.05). Liczebosc obu pozdzbiorow byla podobna (48:32 na korzysc mniej eksplotowanych pojazdow). Zalozenia testu sa wiec spełnione. Ponizej znajduje sie kod dzielacy auta na podzbiory i dodajacy nowa kolumnę symbolizujaca eksplotacje auta na kopopii danych.

```

> less_exploded_half <- sy[sy$odometer_value<=250000,]
> more_exploded_half <- sy[sy$odometer_value>=250000,]
> exploded <- vector()
> sy_copy <- sy
> for (row in 1:nrow(sy_copy)){
+   odometer <- sy_copy[row, "odometer_value"]
+   exploded = append(exploded, odometer >=250000)
+ }
> sy_copy$is_exploded = exploded

```

Ponizej znajduja sie wyniki t - testu:

```
> t.test(less_exploded_half$price_usd, more_exploded_half$price_usd)
```

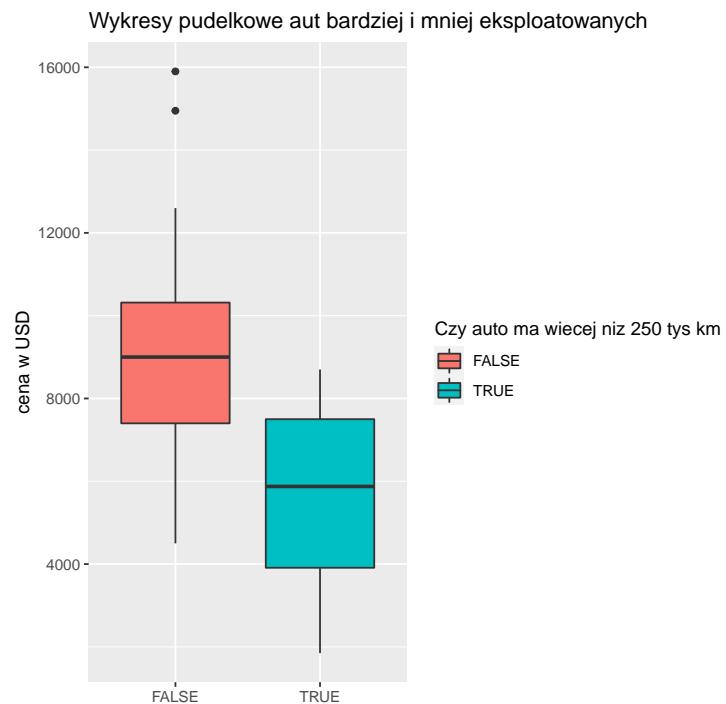
```
Welch Two Sample t-test
```

```

data: less_exploded_half$price_usd and more_exploded_half$price_usd
t = 5.923, df = 75.298, p-value = 8.829e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2063.930 4155.642
sample estimates:
mean of x mean of y
8894.245 5784.459

```

Test wskazuje na wyrazna roznice w wartosci oczekiwanej dla obu probek (Hipoteza zerowa oznacza zeroa roznice w medianach, a p- wartosc jest bardzo niska). Ponizej znajduje sie wykres pudelkowy cen dla tych dwoch zbiorow ilustrujacy roznice w cenach.



Rysunek 22: Wykres porównujacy ceny aut w zaleznosci od ekspolatacji

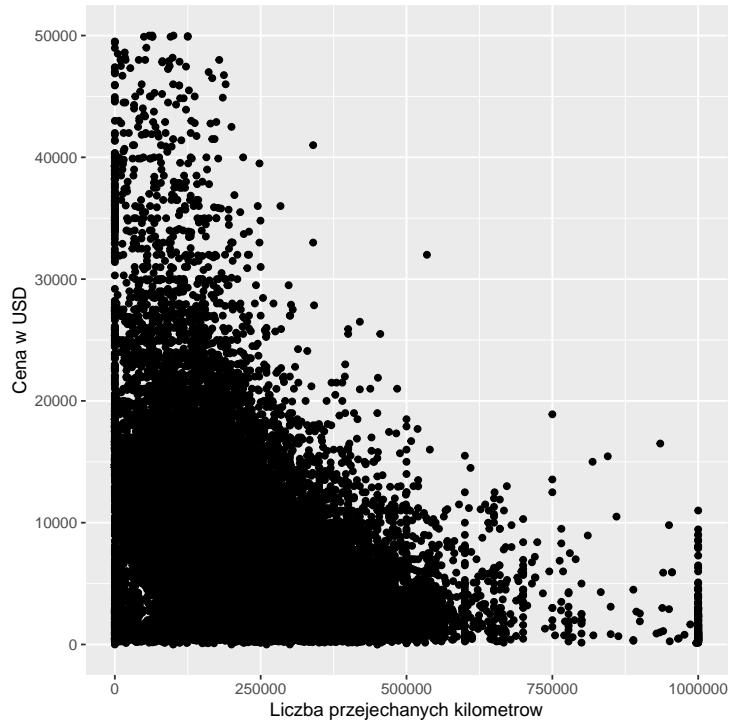
Celem testu jest wprowadzeniem do nastepnej sekcji, gdzie badana bedzie zaleznosc miedzy zmiennymi, dla ogólnego przypadku.

## 5 Zaleznosci miedzy zmiennymi

Głównym zagadnieniem, które chciałem zbadac byla zaleznosc ceny wystawianych samochodow w zaleznosci od ich cech (jak i chech samego oglaszenia). W tym celu stworzone zostały cztery modele.

### Model 1- przejechane kilometry

Pierwszym pomyslem było sprawdzenie, czy istnieje zaleznosc miedzy cena, a liczba przejechanych kilometrow. Istnienie zjawiska cofania licznikow wskazywaloby, na to ze zaleznosc powinna byc dosyć znaczna. W tym celu stworzono wykres ceny w zaleznosci od liczby przejechanych kilometrow.



Rysunek 23: Wykres zaleznosci ceny od przejechanych kilometrow

Wykres nie jest zbyt czytelny, co jest spowodowane spora liczba danych (około 39 tysiecy). Mimo to, mozna zauwazyc, ze cena zdaje sie malec wraz z liczba przejechanych kilometrow. Nastepnym krokiem bylo zbadanie korelacji tych dwoch zmiennych.

```
> cor(cars$odometer_value, cars$price_usd)
```

```
[1] -0.4441862
```

Korealcja miedzy zmiennymi jest mniejsza niz sie tego spodziewalem.  
Mimo tego zdecydowalem sie na stworzenie modelu:

```
> model1 <- lm(cars$price_usd ~ cars$odometer_value)
> summary(model1)
```

Call:

```
lm(formula = cars$price_usd ~ cars$odometer_value)
```

Residuals:

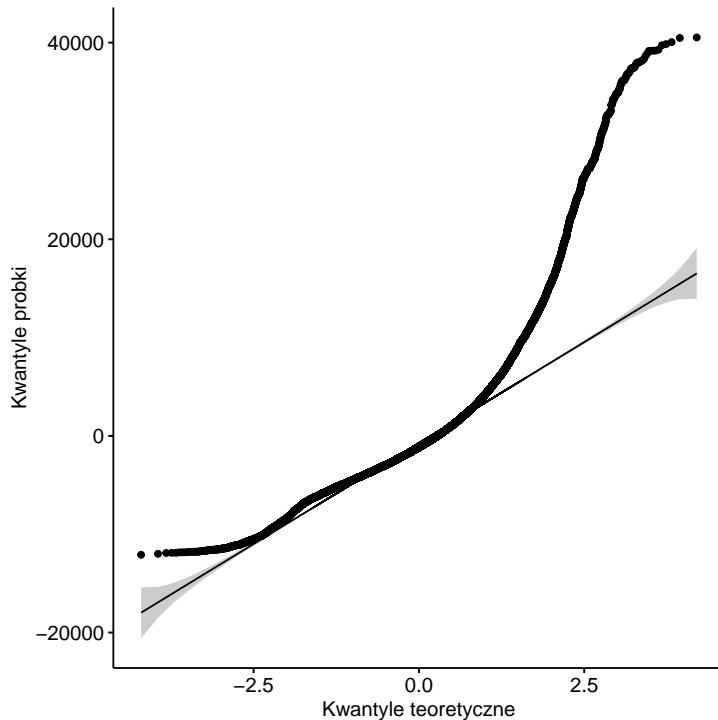
Min	1Q	Median	3Q	Max
-12090	-3495	-1155	2044	40523

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.209e+04	6.306e+01	191.74	<2e-16 ***
cars\$odometer_value	-2.131e-02	2.215e-04	-96.23	<2e-16 ***
---				
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'

Residual standard error: 5774 on 37677 degrees of freedom  
Multiple R-squared: 0.1973, Adjusted R-squared: 0.1973  
F-statistic: 9261 on 1 and 37677 DF, p-value: < 2.2e-16

P - wartosc hipotezy oznaczajacej niezaleznosc tych dwoch parametrow jest bardzo mala. Oznacza to, ze naprawdopodobniej warto pozostawic ta zmienna w modelu. Wartosc  $R^2$  jest stanowczo za mala, by uznać model za dobry. Sprawdzilem jeszcze jak wyglada rozklad roznicy. Jak widac na wykresie rozklad roznicy



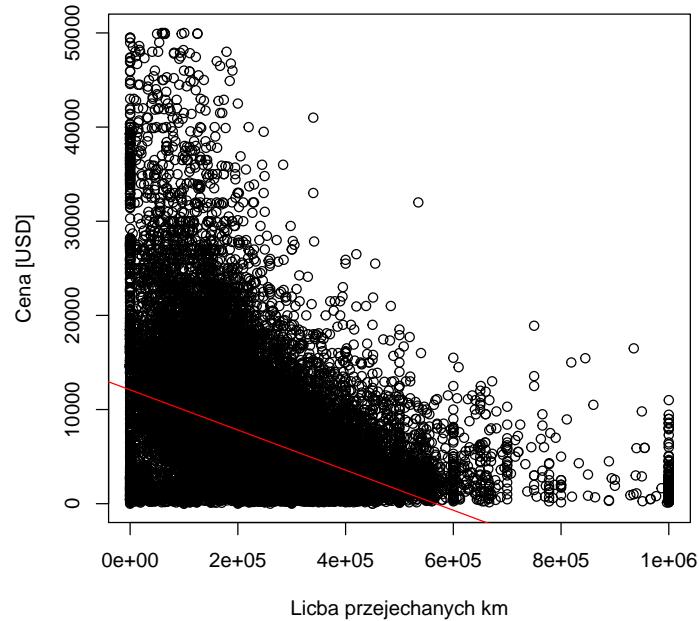
Rysunek 24: Q - Q Wykres dla roznicy w modelu 1.

jest niestety znaczaco rozny od rozkladu normalnego szczegolnie dla wyzszych kwantyli. Ostatecznie model prezentuje sie nastepujaco:

```

> plot(cars$odometer_value, cars$price_usd, xlab = "Licba przejechanych km",
+       ylab = "Cena [USD]")
> abline(model1, col="red")

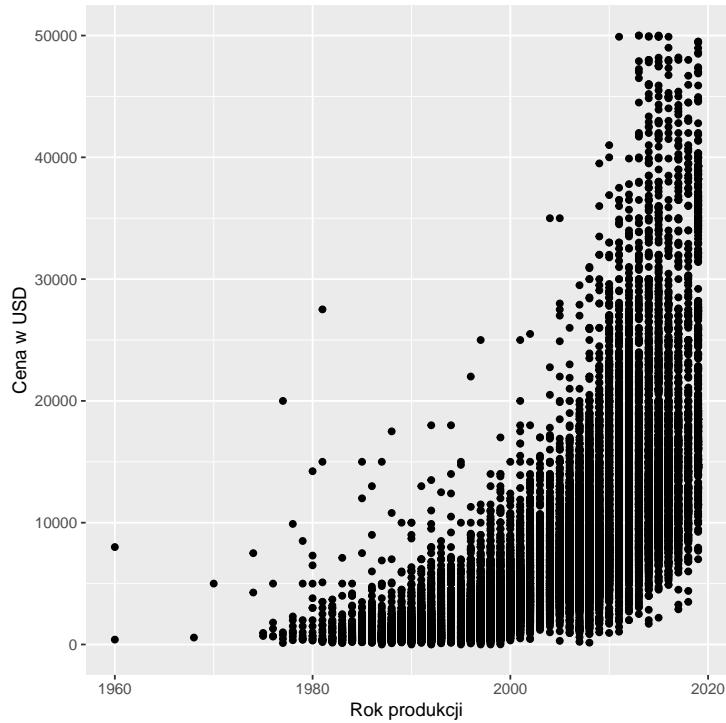
```



Rysunek 25: Krzywa uzyskana w modelu 1.

### Model 2- Rok produkcji

Nastepnym pomyslem bylo sprawdzenie czy cena jest uzaleziona od roku produkcji. Skoro dilerzy samochodowi co roku robia wyprzedaż starego rocznika, to moze faktycznie auta blyskawicznie traca na wartosci? Pierwszym krokiem bylo sporzadzenie wykresu:



Rysunek 26: Wykres zaleznosci ceny od roku produkcji

Wykres zdaje sie dawac nadzieje na znalezienie koleracji. Ponownie nie jest on jednak zbyt czytelny.

```
> cor(cars$year_produced, cars$price_usd)
[1] 0.7265187
```

Tym razem koleracja miedzy zmiennymi byla pozytywnie zaskakujaca. Mozna stwierdzic ze zmienne sa istotnie zalezne  
Nastepnie stworzylem model

```
> model2 <- lm(cars$price_usd ~ cars$year_produced)
> summary(model2)

Call:
lm(formula = cars$price_usd ~ cars$year_produced)

Residuals:
    Min      1Q  Median      3Q     Max 
-12235   -2423   -902   1295  38417
```

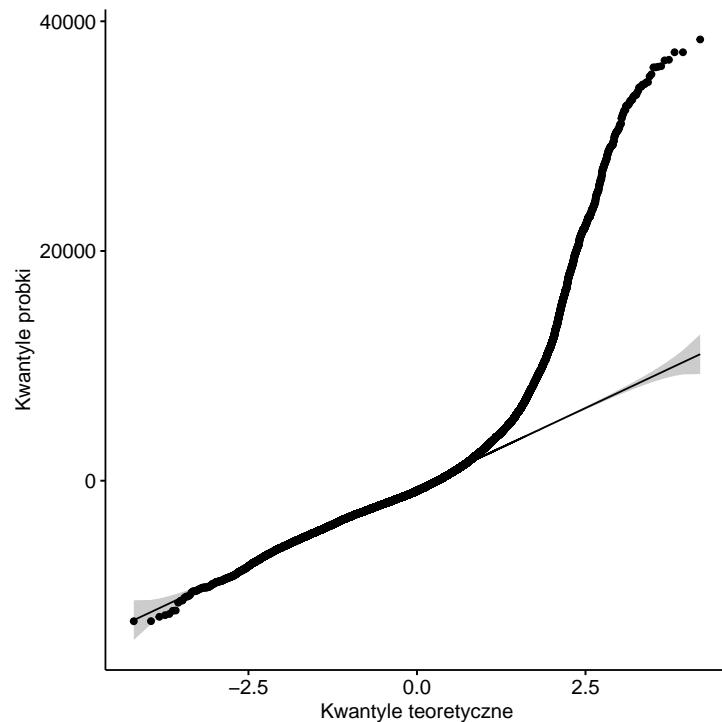
```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.209e+06 5.922e+03 -204.1 <2e-16 ***
cars$year_produced 6.067e+02 2.956e+00 205.2 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4429 on 37677 degrees of freedom
Multiple R-squared: 0.5278, Adjusted R-squared: 0.5278
F-statistic: 4.212e+04 on 1 and 37677 DF, p-value: < 2.2e-16

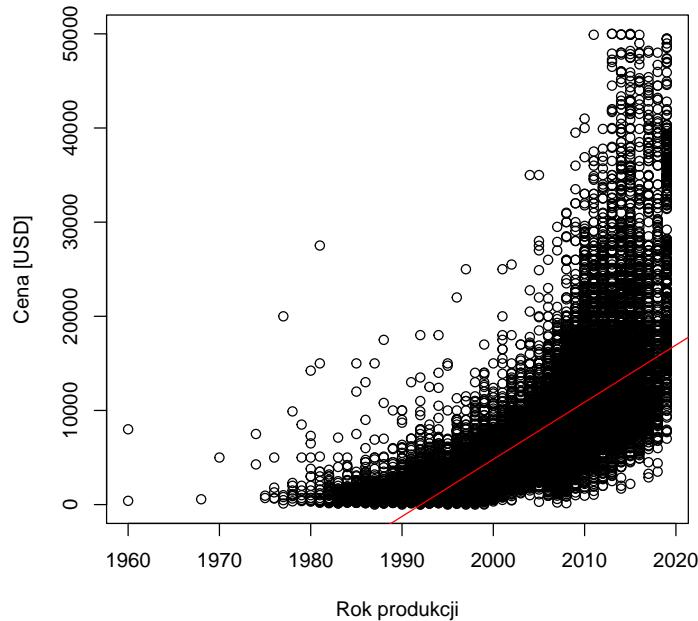
```

P - wartosc hipotezy oznaczajacej niezaleznosc tych dwoch parametrow jest bardzo mala. Oznacza to, ze najprawdopodobniej warto pozostawic ta zmienna w modelu. Wartosc  $R^2$  jest lepsza niz w poprzednim modelu, jednak wciaz nie jest satysfakcjonujaca. Kolejnym krokiem bylo sprawdzenie, jak wyglada rozklad roznicy. Jak widac na wykresie rozklad nie do konca pokrywa sie z teoretycznym



Rysunek 27: Q - Q Wykres dla roznicy w modelu 2.

rozkladem normalnym, szczegolnie dla wiekszych wartosci. Ostatecznie model prezentuje sie nastepujaco:



Rysunek 28: Krzywa uzyskana w modelu 2.

### Model 3- Eksperyment

Dotychcza kazda zmienna wykazywala zaleznosc na tyle istotna by znajdowac sie w modelu. Postanowilem wiec zobaczy, co stanie sie, gdy model uzaleznie od wszystkich wartosci liczbowych, ktorymi dysponuje.

```
> model3 <- lm(cars$price_usd ~ cars$odometer_value + cars$year_produced +
+                  cars$engine_capacity + cars$number_of_photos + cars$duration_listed)
> summary(model3)

Call:
lm(formula = cars$price_usd ~ cars$odometer_value + cars$year_produced +
    cars$engine_capacity + cars$number_of_photos + cars$duration_listed)

Residuals:
    Min      1Q Median      3Q     Max 
-12656   -2088   -604    1217   33849 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10000.000  1000.000  10.000 0.0000000 ***
cars$odometer_value  100.000   10.000  10.000 0.0000000 ***
cars$year_produced  -100.000   10.000 -10.000 0.0000000 ***
cars$engine_capacity  100.000   10.000  10.000 0.0000000 ***
cars$number_of_photos  100.000   10.000  10.000 0.0000000 ***
cars$duration_listed  100.000   10.000  10.000 0.0000000 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```

(Intercept) -1.081e+06 6.439e+03 -167.942 < 2e-16 ***
cars$odometer_value -4.873e-03 1.810e-04 -26.929 < 2e-16 ***
cars$year_produced 5.404e+02 3.207e+00 168.511 < 2e-16 ***
cars$engine_capacity 2.846e+03 3.040e+01 93.605 < 2e-16 ***
cars$number_of_photos 1.032e+02 3.442e+00 29.977 < 2e-16 ***
cars$duration_listed 1.083e+00 1.812e-01 5.979 2.26e-09 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3900 on 37663 degrees of freedom
(10 observations deleted due to missingness)
Multiple R-squared: 0.6336, Adjusted R-squared: 0.6335
F-statistic: 1.302e+04 on 5 and 37663 DF, p-value: < 2.2e-16

```

Okazalo sie, ze kazda wartosc ma znaczący wpływ na ceny. Nawet tak niepozorne czynniki jak liczba zdjęć sa powiązane z ceną jakiej sprzedający chce za samochód. Dodatkowo model ten jest najlepszym ze wszystkich dotychczasowych. Poziom skomplikowania jednak drastycznie wzrósł. Warto również zwrócić uwagę, na fakt, że w każdym przypadku oprócz liczby przejechanych kilometrów współczynnik był dodany, tak więc cena rosła wraz ze wzrostem wartości cechy. Jest to wyjaśnialne dla wszystkich cech. Najmniej trywialne wyjaśnienie tego stanu rzeczy dotyczy czasu od wystawienia ogłoszenia. Można założyć, że auta tanie szybciej się sprzedają, wobec czego dotyczące ich ogłoszenia nie są już dostępnego - więc im dłużej auto jest wystawione, tym jest droższe. Ta cecha ma jednak najmniejszy wpływ na model.

Pozostaje jeszcze sprawdzić, czy nastąpiła zmiana w rozkładzie reszt: Niestety ten wykres wydaje się być bez zmian niezależnie od dobranego modelu.

#### **Model 4 - model wykorzystujący wszystkie znane cechy**

Skoro poprzedni model okazał się być lepszy od poprzednich, to może słuszne będzie uzycie wszystkich znanych cech podczas budowania modelu? Postanowimy to sprawdzić.

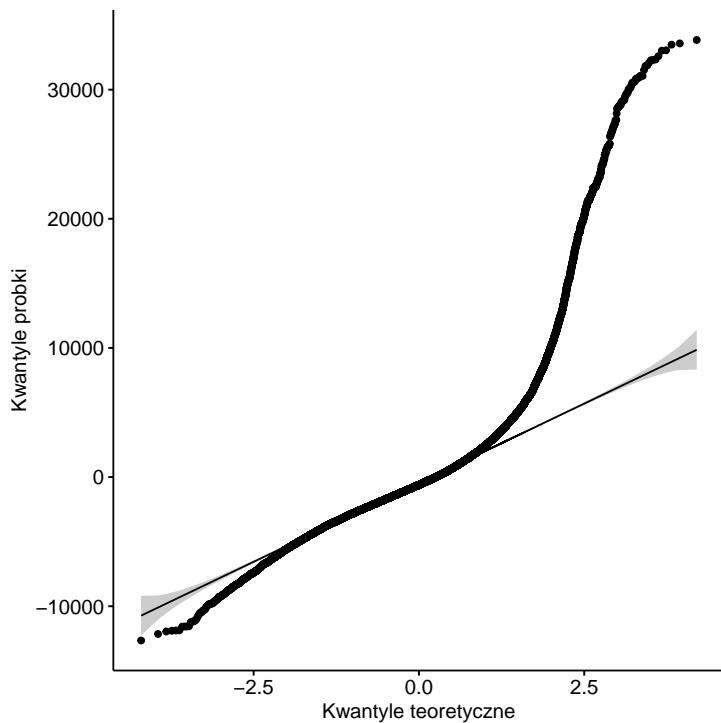
```

> model4 <- lm(cars$price_usd ~ cars$odometer_value + cars$year_produced +
+                  cars$engine_capacity + cars$number_of_photos + cars$manufacturer_name +
+                  cars$color + cars$body_type)
> summary(model4)

Call:
lm(formula = cars$price_usd ~ cars$odometer_value + cars$year_produced +
   cars$engine_capacity + cars$number_of_photos + cars$manufacturer_name +
   cars$color + cars$body_type)

Residuals:
    Min      1Q Median      3Q     Max 
-13499  -1739   -290   1103  33001 


```



Rysunek 29: Q - Q Wykres dla roznicy w modelu 3.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.010e+06	6.183e+03	-163.354	< 2e-16 ***
cars\$odometer_value	-5.174e-03	1.633e-04	-31.684	< 2e-16 ***
cars\$year_produced	5.085e+02	3.063e+00	165.992	< 2e-16 ***
cars\$engine_capacity	1.298e+03	3.478e+01	37.330	< 2e-16 ***
cars\$number_of_photos	7.239e+01	2.987e+00	24.232	< 2e-16 ***
cars\$manufacturer_nameAlfa Romeo	-1.986e+03	4.767e+02	-4.167	3.09e-05 ***
cars\$manufacturer_nameAudi	1.248e+03	4.203e+02	2.970	0.002984 **
cars\$manufacturer_nameBMW	1.298e+03	4.192e+02	3.095	0.001969 **
cars\$manufacturer_nameBuick	-3.935e+03	6.440e+02	-6.111	1.00e-09 ***
cars\$manufacturer_nameCadillac	-2.953e+03	6.616e+02	-4.463	8.12e-06 ***
cars\$manufacturer_nameChery	-7.729e+03	6.067e+02	-12.740	< 2e-16 ***
cars\$manufacturer_nameChevrolet	-3.563e+03	4.453e+02	-8.001	1.27e-15 ***
cars\$manufacturer_nameChrysler	-2.655e+03	4.477e+02	-5.931	3.04e-09 ***
cars\$manufacturer_nameCitroen	-2.238e+03	4.250e+02	-5.265	1.41e-07 ***
cars\$manufacturer_nameDacia	-4.620e+03	6.043e+02	-7.644	2.15e-14 ***
cars\$manufacturer_nameDaewoo	-3.725e+03	4.743e+02	-7.854	4.14e-15 ***



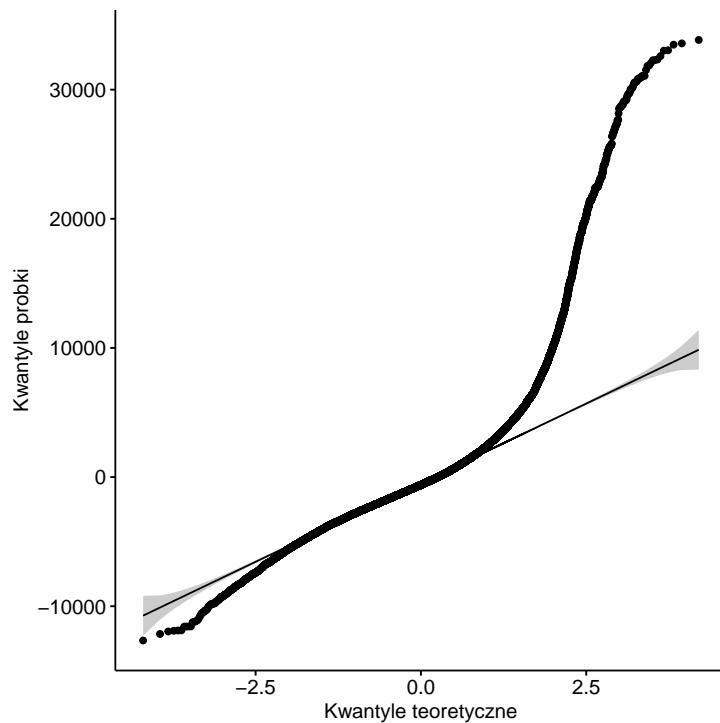
```

cars$colorviolet          6.438e+01  1.642e+02  0.392  0.695060
cars$colorwhite           3.630e+02  6.897e+01  5.264  1.42e-07 ***
cars$coloryellow          -4.492e+01  2.121e+02  -0.212  0.832260
cars$body_typecoupe       -1.972e+03  4.209e+02  -4.686  2.79e-06 ***
cars$body_typehatchback   -3.816e+03  4.023e+02  -9.486  < 2e-16 ***
cars$body_typerliftback   -3.902e+03  4.274e+02  -9.129  < 2e-16 ***
cars$body_typelimousine   -2.195e+03  1.129e+03  -1.944  0.051922 .
cars$body_typeminibus     -3.900e+02  4.118e+02  -0.947  0.343551
cars$body_typeminivan     -2.804e+03  4.043e+02  -6.936  4.09e-12 ***
cars$body_typepickup       1.148e+03  5.058e+02  2.269  0.023274 *
cars$body_typesedan        -3.503e+03  4.009e+02  -8.738  < 2e-16 ***
cars$body_typesuv          3.479e+02  4.032e+02  0.863  0.388202
cars$body_typeuniversal    -3.659e+03  4.030e+02  -9.081  < 2e-16 ***
cars$body_typevan          -1.890e+03  4.220e+02  -4.478  7.55e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3358 on 37593 degrees of freedom
(10 observations deleted due to missingness)
Multiple R-squared:  0.7288,      Adjusted R-squared:  0.7282
F-statistic:  1347 on 75 and 37593 DF,  p-value: < 2.2e-16

```

Okazuje sie, ze model jest dokladniejszy. Jest to rowniez pierwszy model, dla ktorego wartosc  $R^2$  przekroczyla 0.70. Niewatpliwa wada tego modelu wysoki poziom skomplikowania. Warto zauwazyc, ze teraz nie wszystkie cechy sa juz istotne w modelu (chociaz znaczna czesc nadal jest). Ponadto, ze zmienne jakosciowe sa teraz rozpatrywane jako wartosci prawda/falsz okreslajace, czy dany element nalezy do odpowiedniego podzbioru. Postanowilem rowniez sprawdzic rozklad reszt, jednak tym razem rowniez nie zauwazylem istotnych zmian.



Rysunek 30: Q - Q Wykres dla roznicy w modelu 4.

## 6 Wnioski

Wnioski płynące z przeprowadzonej analizy, są następujące:

- Na cene samochodow ma wpływ wiele czynników,
- liczba przejechanych kilometrow nie jest wcale najwazniejszym czynnikiem determinujacym cene,
- zaskakujaco istotny (w kontekscie ceny) jest rok produkcji samochodu,
- niektore kolory (czarny, brazowy, bialy) zdaja sie byc wystawiane drozej niz inne,
- zgodnie z oczekiwaniemi marka ma duzy wpływa na cene samochodu, co widac na przykladzie Porsche i Fiata,
- najwiecej wystawianych aut pochodzi z lat 1985 - 2015.