**OPIS PROBLEMU, ROZWIĄZANIE**

Zadanie polegało na analizie danych dotyczących płacy w pewnej firmie. Niektóre dany były podane niepełnie np. brakowało niektórych wartości w tabeli staż pracy. Zdarzały się także wartości, które ewidentnie przekraczały poza normę np. wiek wynosił 200. Pierwszym krokiem było więc zmodyfikowanie pliku z danymi, aby był w pełni poprawny. Aby rozwiązać napotkany problem należy zastąpić wartości odstające na medianę zmiennie. Natomiast w przypadku niepełnych danych należy usunąć wiersze, w których brakuje danych.

**CHARAKTERYSTYKA ZBIORU DANYCH**

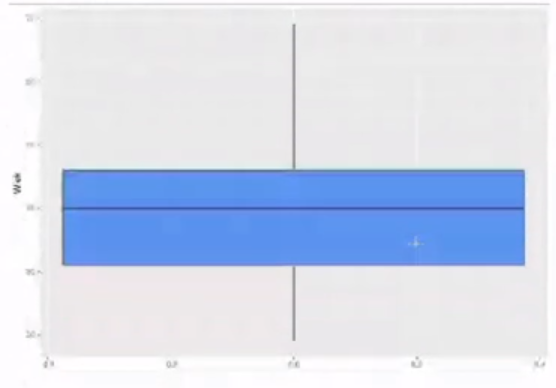
Zbiór charakteryzował pracowników względem zmiennych jakościowych: płeć, wykształcenie, oraz zmiennych ilościowych: wiek, staż pracy i płaca brutto. //kiedy została przeprowadzona analiza, co zawierał zbiór.

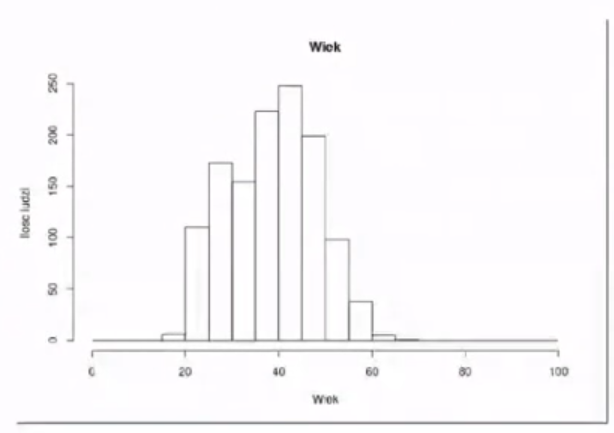
ZMIENNA WYBRANA DO ANALIZY: **wiek**



// średnia mediana moda rozproszenie min max kwartyle percentyle rozstaw wariancja i odychelnie standardowe.

**GRAFICZNA PREZENTACJA ZBIORU**

****



**INTERPRETACJA WYNIKÓW //powyżej górnego kwartyle dane są bardziej rozproszone**

Średni wiek pracowników jest zbliżony bardzo do mediany, wynosi około 40lat. Najczęściej powtarzający się wiek to 43 lata, ponieważ taka jest moda. Pomiędzy 40 a 45 rokiem życia jest najwięcej pracowników: 250 osób, co możemy odczytać z histogramu. 25% pracowników ma poniżej 31 lat i 25% pracowników powyżej 46 lat. Najmłodszy pracownik ma 19lat a najstarszy 69 lat. Rozkład zmiennej jest w miarę symetryczny, nie widać dużej skośności. Podobnie rozproszenie, jest mniej więcej takie samo. //rozproszzenie pomiędzy kwartylem dolnym a medianą jest troszeczkę mniejsze niż pomiędzy medianą a kwartylem górnym.

//+ można jeszcze dołożyć ze średni wiek w podanej próbie wynosił np. 39 lat z odchyleniem +/- 10 lat.

// czy dane sa skupione wokół średniej.

2

Charakterystyka zbioru danych

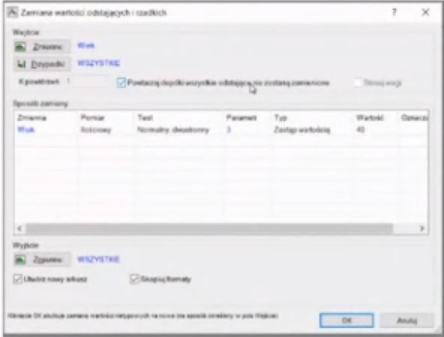
W pliku place zgromadzono dane dotyczące zarobków mieszkańców Warszawy. Badanie zostało przeprowadzone 20.01.1995.



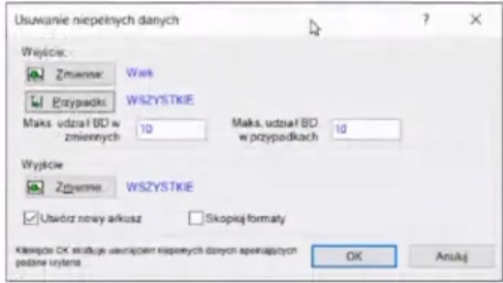
Liczba zmiennych ilościowych: 3

Liczba zmiennych jakościowych: 2

Liczba przypadków 1255



Zastępuję wartości odstające medianą.



Usuwamy niepełne dane

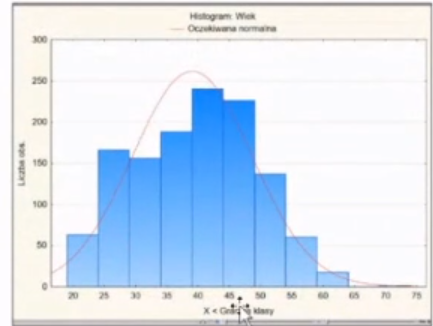
**Miary tendencji centralnej i rozproszenia**

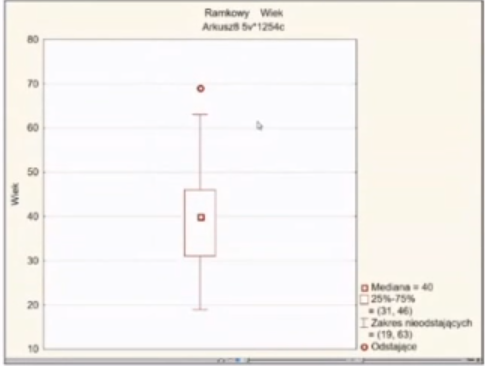
**Miary tendencji centralnej:** moda, mediana, średnia

**Rozproszenia**: rozstęp, odchylenie standardowe, współczynnik zmienności.



**Graficzna prezentacja danych**

****

****

**Interpretacja/Opis wyników**

Wykres jest asymetryczny. Najwięcej osób jest w wieku 40 lat. //na jakiej podstawie wnioskujemy, dolna cześć pudła jest dłuższa więc mamy większe rozproszenie danych w tej części. Natomiast wąs górny jest dłuższy więc z tej strony mamy większe rozproszenie danych.

Na wykresie ramka wąsy znajduje się jedna odstająca wartość równa 69. Górny wąs jest dłuższy od dolnego z czego wynika że dane te są bardziej zróżnicowane niż dane w dolnym wąsie. Mediana nie jest idealnie pośrodku pudełka ze względu na rozproszenie danych. Na histogramie dobrze widać, że rozproszenie pomiędzy kwartylem dolnym a medianą jest większe niż pomiędzy medianą a kwartylem górnym.??? Dlatego też mediana nie znajduje się na środku pudełka, jest bardziej w górnej części, ale za to wąs górny jest dłuższy.

///+Opis tego żę średni wiek wynosi.. z odchyleniem standardowym … . Średnia mówi nam o tym że średni wiek wynosi 39 lat +- 10lat. Mediana wynosi 40 czyli 50% obserwacji znajduje się poniżej mediany i 50% powyżej. Minimalna wartość to 19 czyli pomiędzy 19 a 40 rokiem życia znajduje się dokładnie 50% obserwacji i pomiędzy 40 a 69 znajduje się dokładnie 50% obserwacji.