**RAPORT PROJEKTU Z EKSPLORACJI DANYCH**

**Autor: Edyta Margol, IAD I rok II stopień**

Dane zawierają 7 zmiennych, z czego każda ma 2000 obserwacji (plik Projekt2 arkusz modele). Poniżej przedstawiono fragment zbioru.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Każda z tych zmiennych zawiera braki danych. Poszczególną ich liczbę dla każdej zmiennej przedstawiono poniżej.

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznie

Poniżej natomiast wyznaczono wartości podstawowych statystyk:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

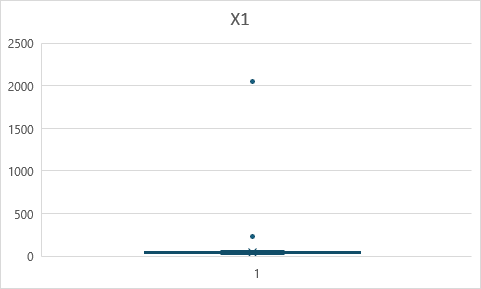
Opis wygenerowany automatycznie

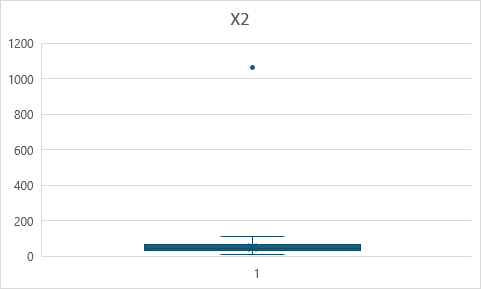
Dominanta oznaczona jako #N/D wskazuje, że nie występuje w analizowanym przedziale, co może być spowodowane brakami danych lub unikalnością wartości. Natomiast dla poszczególnych zmiennych są następujące wnioski:

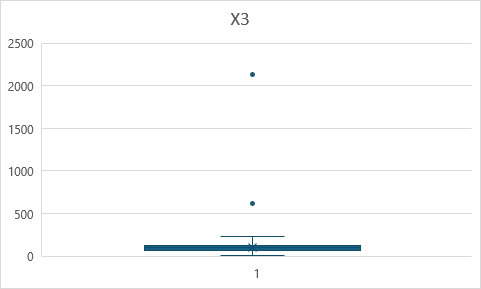
* X1: Średnia i mediana są zbliżone, co wskazuje na symetryczny rozkład większości danych. Wysokie wartości maksymalne znacząco podnoszą odchylenie standardowe, co może sugerować występowanie wartości odstających.
* X2: Rozkład jest symetryczny (średnia ≈ mediana). Wysoka wartość maksymalna (1064,10) sugeruje możliwą obecność skrajnych wartości, które mogą mieć wpływ na średnią i zwiększać odchylenie standardowe.
* X3: Średnia i mediana wskazują na względnie symetryczny rozkład, ale wysokie odchylenie standardowe i wartość maksymalna mogą sugerować duże rozproszenie danych oraz obecność odstających obserwacji.
* X4: Dane w X4 są najbardziej rozproszone (najwyższe odchylenie standardowe). Bardzo duże wartości minimalne i maksymalne wskazują na silną obecność wartości odstających, co może zniekształcać analizy tej zmiennej.
* X5: Średnia i mediana są do siebie zbliżone, a odchylenie standardowe jest relatywnie niskie, co wskazuje na stabilność danych w większości przypadków. Jednak wartość maksymalna jest bardzo wysoka, co może być odstającym punktem w danych.
* X6: Dane w X6 są symetryczne (średnia ≈ mediana), a rozkład jest stosunkowo skupiony (niskie odchylenie standardowe). Wysoka wartość maksymalna sugeruje obecność potencjalnych odstających obserwacji.
* X7: Średnia i mediana są zbliżone, co sugeruje symetrię większości danych. Jednak ekstremalnie niskie i wysokie wartości wskazują na obecność silnych odstających obserwacji, które wpływają na odchylenie standardowe i mogą zakłócać wyniki analiz.

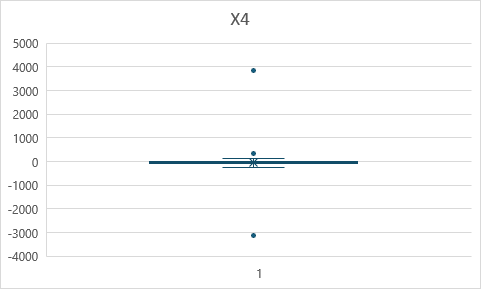
Podsumowując: zmienna X4 wyróżnia się największą zmiennością oraz ekstremalnymi wartościami, co może wymagać dodatkowego przetwarzania (np. usunięcia wartości odstających). Z kolei zmienne X5 i X6 są bardziej stabilne, z niskim rozproszeniem wokół średnich. Wysokie wartości maksymalne w X2, X3, X4, i X7 wskazują na możliwość występowania obserwacji odstających, które mogą znacząco wpłynąć na analizy i modele.

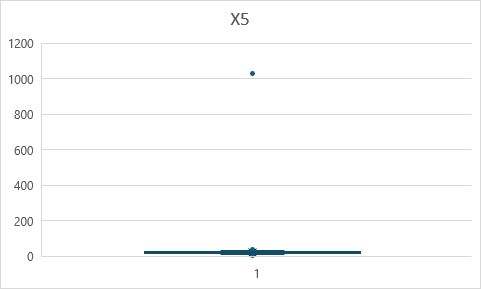
Aby zweryfikować obecność wartości odstających w danych, przygotujemy wykresy pudełkowe dla każdej zmiennej (plik „Projekt2”, arkusz „dane początkowe”). Wartości znajdujące się poza górnym i dolnym wąsem na wykresach będą uznane za wartości odstające.

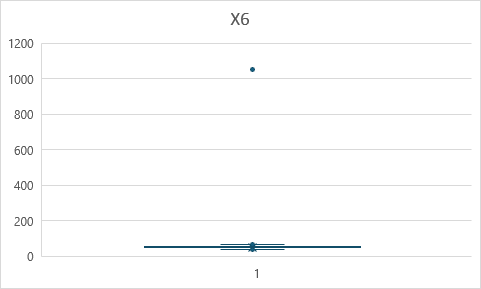


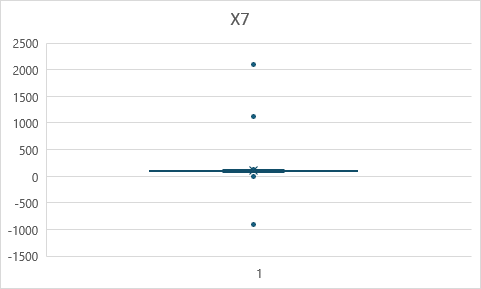












Można zauważyć, że każda ze zmiennych zawiera obserwacje odstające.

Dalsze etapy projektu będą realizowane dwiema metodami.

**PODEJŚCIE 1 – bez usuwania braków danych oraz wartości odstających**

Następnym etapem projektu jest skonstruowanie szeregu rozdzielczego przedziałowego dla zmiennej X1 (plik Projekt2 arkusz szereg).

Na początku wyznaczono podstawowe wartości.

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

Następnie skokonstruowano szereg.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Pierwszy lewy koniec wyznaczono jako wartość minimalną. Prawy koniec określono jako lewy koniec powiększony o długość jednego przedziału. Kolejne lewe końce są równe poprzednim prawym końcom, a prawy koniec każdego następnego przedziału wyznacza się, dodając długość przedziału do lewego końca. Kolumna x\_i stanowi środek przedziału, n\_i – liczbę obserwacji w danym przedziale, x\_i\*n\_i wartość oczekiwaną (liczbę obserwacji w każdym z przedziałów), x\_i^2\*n\_i - drugi moment

Wartości te zostały obliczone w celu wyliczenia średniej, wariancji i odchylenia standardowego z szeregu przedziałowego.

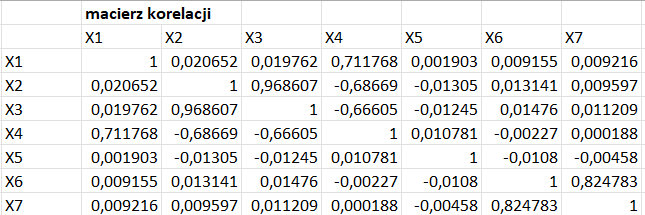
Po skonstruowaniu szeregu zauważono, że obserwacje w całym zbiorze danych są nierównomiernie rozłożone pomiędzy przedziały, przy czym większość obserwacji znajduje się w pierwszym przedziale. Może to sugerować obecność wartości odstających.

Dodatkowo, średnia obliczona z szeregu przedziałowego znacznie różni się od średniej obliczonej bezpośrednio za pomocą funkcji średniej. Taki rozrzut może wskazywać, że wartości odstające mają istotny wpływ na analizę danych.

Kolejnym etapem projektu jest uzupełnienie braków danych.

**1 sposób uzupełniania danych**

W pierwszej kolejności zostanie przeanalizowany najprostszy przypadek – korelacja liniowa pomiędzy zmiennymi. W przypadku występowania takiej zależności możliwe będzie oszacowanie brakujących wartości za pomocą równania regresji liniowej. W związku z tym konieczne jest sprawdzenie, czy między zmiennymi istnieje liniowa zależność (plik: *Projekt2*, arkusz: dane początkowe).



W żadnej parze różnych zmiennych nie zaobserwowano idealnie liniowej korelacji (tj. równej 1). Niemniej jednak bardzo wysoka korelacja występuje, na przykład, między zmiennymi X2 a X3. Analiza wykresu zależności tych dwóch zmiennych sugeruje możliwość istnienia liniowej zależności, choć widoczne są również wartości odstające, które mogą wpływać na wyniki analizy.

Mimo wszystko spróbujemy wyznaczyć równania regresji (plik Projekt2 arkusz Wyznaczanie X2 i X3).

Otrzymujemy, że:

oraz

Zastosowane rozwiązanie nie jest w pełni zadowalające, ponieważ nadal występują przypadki ekstremalnie wysokich błędów. Co istotne, błędy te pojawiają się w obserwacjach, gdzie nie występowały braki danych, a wartości zmiennych X2 i X3 nie były obliczane na podstawie pustych komórek.

Do identyfikacji tych wartości wykorzystano skalę kolorów, w której najwyższe błędy oznaczono kolorem czerwonym, a najniższe kolorem zielonym (szczegóły znajdują się w skoroszycie *Projekt2*, arkusz: *modele*).

**Wniosek:** W pierwszej kolejności należy usunąć wartości odstające.

**2 sposób uzupełniania danych**

Jako drugi sposób uzupełniania braków danych użyto interpolacji liniowej.

Interpolacja liniowa to metoda szacowania brakujących danych poprzez wyznaczanie wartości na podstawie linii prostej między sąsiadującymi znanymi wartościami. Sprawdza się szczególnie dobrze w danych, które zmieniają się w sposób liniowy lub prawie liniowy w krótkich przedziałach. Chociaż charakterystyka podanych danych nie jest znana, warto podjąć próbę zastosowania tej techniki, ponieważ może przynieść zadowalające rezultaty.

**Zasada działania interpolacji liniowej**

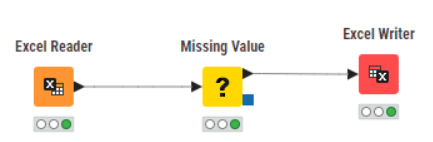
Gdy brakuje wartości y dla pewnego punktu x, interpolacja liniowa wyznacza tę wartość na podstawie równania prostej pomiędzy dwoma sąsiadującymi punktami i

Interpolacja liniowa może być używana do uzupełniania braków danych, zwłaszcza gdy dane mają naturalną tendencję do zmienności liniowej w krótkich odstępach czasu lub przestrzeni.

**Zalety:**

1. **Prostota i efektywność:** jest to szybka i łatwa metoda obliczeniowa.
2. **Zachowanie trendów:** w przypadku małych luk interpolacja liniowa często dobrze odtwarza lokalny trend danych.
3. **Unikanie przesunięć:** interpolacja liniowa nie przesuwa globalnych statystyk danych.

**Zastosowanie interpolacji w KNIME**

****

W KNIME do uzupełniania braków danych można użyć node'a Missing Value, który oferuje różne strategie, w tym interpolację liniową (folder z workspace’em w knime - projekt\_missing\_values). Sposób działania tej metody jest następujący:

1. KNIME analizuje dane kolumnowo, szukając brakujących wartości w danej kolumnie.
2. Dla każdej brakującej wartości:
   * Znajduje najbliższe dostępne wartości powyżej i poniżej brakującego miejsca.
   * Wylicza wartość zgodnie z równaniem interpolacji liniowej.
3. Jeśli brakująca wartość znajduje się na początku lub końcu danych (nie ma jednego z sąsiadów), stosuje inną wybraną strategię (np. wartość domyślną).

Wynikowe zmienne, uzyskane za pomocą interpolacji liniowej, zostały zapisane w pliku odstające\_interpolacja i następnie skopiowane do skoroszytu Projekt2, arkusz: modele. Obliczono różnice (błędy) pomiędzy oryginalnymi zmiennymi a zmiennymi uzyskanymi z modelu. Dla każdej obserwacji, poza tymi, w których występowały braki danych, błędy te wyniosły dokładnie 0.

**PODEJŚCIE 2 – identyfikacja wartości odstających i zastąpienie ich poprzez braki danych**

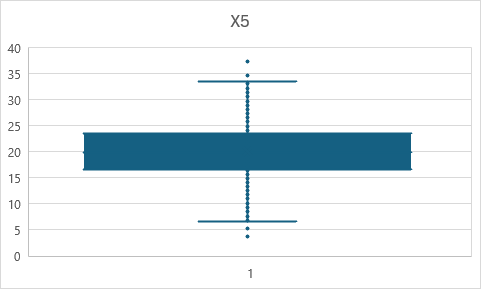
Wartości odstające zostaną usunięte w KNIME poprzez ich zastąpienie brakami danych. Następnie dane te zostaną zapisane do nowego pliku, aby umożliwić dalszą analizę (folder do wglądu – projekt\_odstajace).

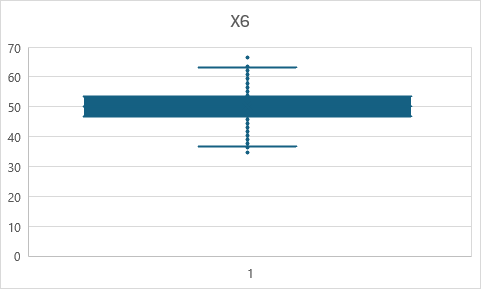
Obraz zawierający tekst, zrzut ekranu

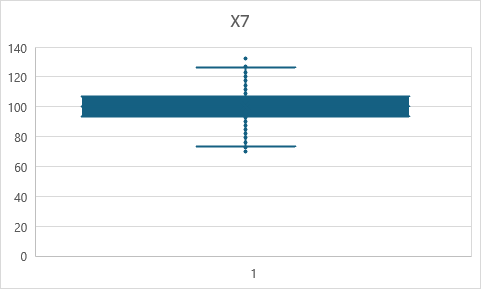
Opis wygenerowany automatycznie

Usunięte zostaną wszystkie obserwacje spoza przedziału

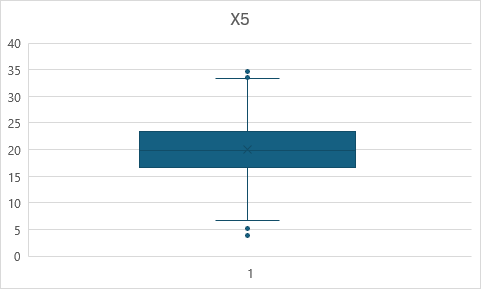
Jednak przy podejściu z wciąż występują obserwacje odstające w zmiennych X5, X6, X7 (plik odstajace3).

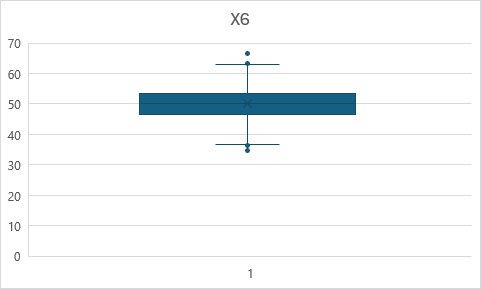


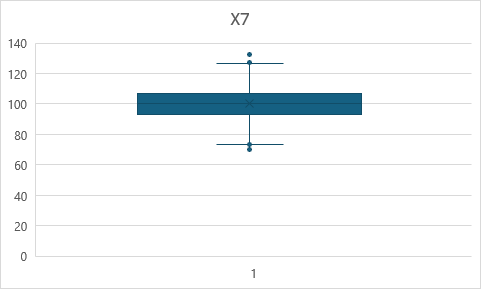




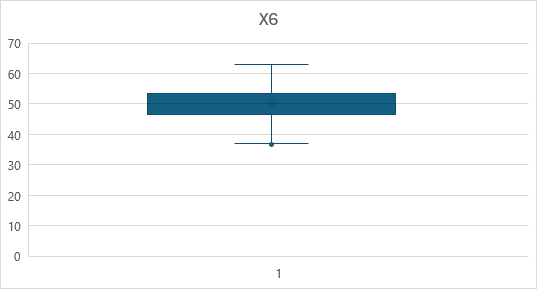
Próba usunięcia wartości odstających za pomocą kryterium 2σ wykazała, że wciąż występują obserwacje odstające, co zostało zapisane w pliku odstajace2.

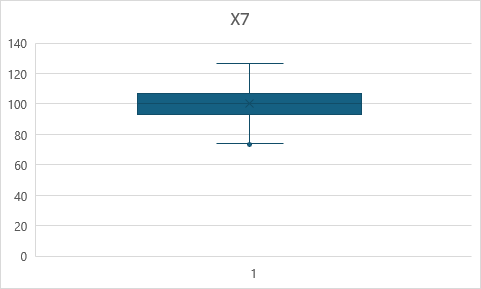






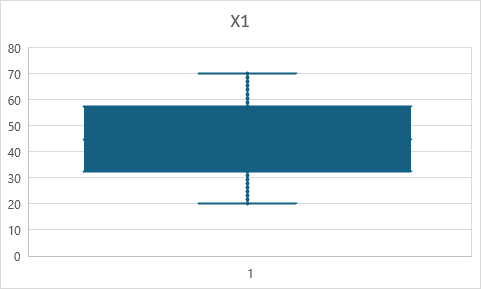
Podjęto zatem próbę usunięcia wartości odstających za pomocą kryterium 1,5σ (plik odstajace1*.5*). W wyniku analizy stwierdzono, że dla zmiennych X6 i X7 pojawiły się obserwacje na pograniczu (widoczne na boxplotach), które jednak nie zostały uznane za wartości odstające. Ostatecznie usunięto 45 obserwacji (2,25%), co stanowi stosunkowo niewielki odsetek. Usunięte obserwacje mogą zawierać błędne pomiary lub inne nieprawidłowe dane, które mogłyby zaburzać model, dlatego ich eliminacja może przyczynić się do poprawy jakości analizy.

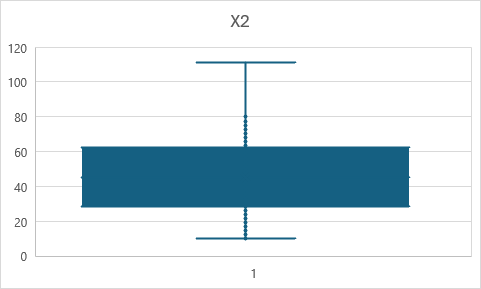


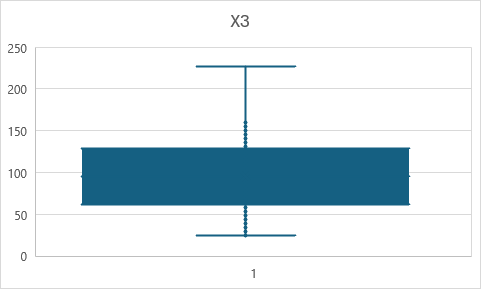


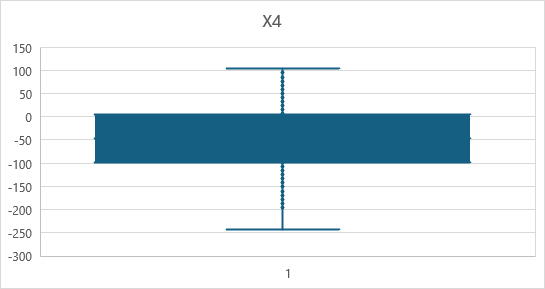
Z histogramów również nie stwierdzono znaczących wartości odstających.

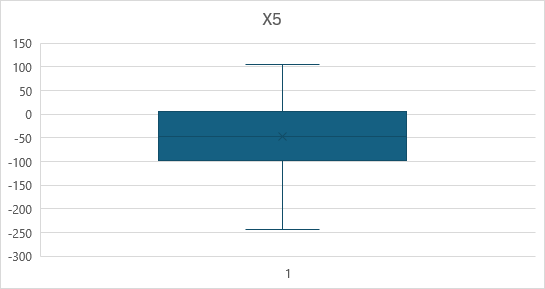
Dla pozostałych zmiennych również nie stwierdzono wartości odstających:



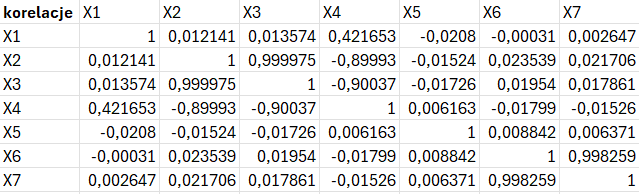








Po usunięciu wartości odstających macierz korelacji prezentuje się lepiej w porównaniu do sytuacji, w której wartości odstające nie zostały usunięte. Zauważalne są wyższe korelacje między zmiennymi oraz większa liczba zmiennych, które można wyznaczyć za pomocą regresji liniowej.



Zmienna X1 wykazuje dość słabe korelacje z pozostałymi zmiennymi, z wyjątkiem umiarkowanej korelacji między X1 a X4. Mimo tego, w programie KNIME udało się wyznaczyć tę zmienną za pomocą równania regresji liniowej, wykorzystując wszystkie pozostałe zmienne jako zmienne niezależne. Proces ten został zrealizowany przy użyciu node'ów ExcelReader oraz LinearRegressionLearner (folder do wglądu – projekt\_uzupelniania).

Obraz zawierający tekst, Czcionka, linia, logo

Opis wygenerowany automatycznie

Na poniższym zrzucie ekranu przedstawiono uzyskane wartości współczynników.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Zmienne X2 i X3 zostały oszacowane za pomocą równania liniowego, ponieważ korelacja między nimi wynosi 0,99, co wskazuje na niemal idealną zależność liniową. Analiza wykresów zależności potwierdza, że taki sposób modelowania jest uzasadniony (szczegóły w pliku „odstające1.5”, arkusz „wyznaczanie X2, X3, X6, X7”).

Równania dla tych zmiennych wyglądają następująco:

oraz

Zmienną X4 wyznaczono w analogiczny sposób jak X1.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Z kolei zmienne X6 i X7 w analogiczny sposób jak X2 i X3.

Zmienna X5 wykazuje bardzo niskie korelacje z pozostałymi zmiennymi, a te korelacje są praktycznie nieistotne. W związku z tym, dla zmiennej X5 zostanie zastosowana prostsza metoda uzupełniania braków danych, mianowicie zastąpienie brakujących wartości medianą wszystkich dostępnych obserwacji.

Po uzupełnieniu braków, błędy w poszczególnych modelach wyglądają następująco (plik do wglądu – odstajace1.5 arkusz modele):

* X1 i X4 – bledy wszedzie wychodza 0 (oprocz oczywiście miejsc, gdzie te zmienne miały braki danych)
* X2 i X3 oraz X6 i X7– wciąż jest ekstremalnie wysokich wartości, jednak one wychodzą dlatego, że zostały policzone na podstawie komórek gdzie były puste wartości, bądż jest akurat liczony brak danych, np.:







* X5 – tutaj nie wyznaczalismy bledu gdyz wszystkie puste komorki zostały po prostu zastąpione mediana

Po uzupełnieniu braków danych, wyznaczone zostały statystyki opisowe, które zostały zapisane w pliku odstajace1*.5*, arkusz: uzupelnionedane.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Dla poszczególnych zmiennych mamy następujace wnioski:

* X1: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany, co sugeruje brak dużych wartości odstających. Większość wartości mieści się w zakresie od 32,36 do 57,22, z pojedynczymi wartościami znajdującymi się poza tym zakresem. Dominanta wynosi 0, co wskazuje, że wartość 0 pojawia się najczęściej w analizowanych danych.
* X2: rozkład danych jest niemal symetryczny, ponieważ średnia jest zbliżona do mediany. Wartości są bardziej rozproszone niż w przypadku zmiennej X1, co sugeruje większe odchylenie standardowe. Dominanta wynosi -2,23, co może wskazywać na specyficzne zjawisko w danych, takie jak częste występowanie ujemnych wartości.
* X3: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są rozproszone, a odchylenie standardowe wynosi 40,38, co stanowi ponad 40% wartości średniej. Często występują wartości małe, a dominanta wynosi 4,46, co może wskazywać na asymetrię rozkładu i sugerować, że rozkład może być przesunięty w stronę niższych wartości.
* X4:rozkład danych jest silnie rozproszony, co wynika z wysokiego odchylenia standardowego. Istnieje prawdopodobieństwo, że rozkład danych jest niesymetryczny, z wyraźnym przesunięciem w stronę wyższych wartości. O tym świadczy różnica między średnią a medianą, a także fakt, że trzeci kwartyl (Q3) przyjmuje wartości dodatnie, co sugeruje przewagę wyższych obserwacji w danych. Brak dominującej wartości (oznaczone jako #N/D) wskazuje na równomierne rozłożenie danych w tej kolumnie.
* X5: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są skupione blisko średniej, co wskazuje na niskie odchylenie standardowe. Często występuje wartość 19,90, która jest dominantą, co oznacza, że ta wartość pojawia się najczęściej i jest równa medianie.
* X6: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są stabilne i mało rozproszone, co wskazuje na niskie odchylenie standardowe (wynoszące około 14% średniej). Często występuje wartość 0,12, która jest dominantą, co może sugerować specyficzne cechy danych, takie jak częste występowanie zerowych lub bardzo małych wartości.
* X7: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są bardziej rozproszone niż w przypadku zmiennej X6, ale nadal stosunkowo stabilne. Często występuje wartość 0,10, która jest dominantą, co, podobnie jak w przypadku zmiennej X6, może wskazywać na specyficzne cechy danych, takie jak częste występowanie tej wartości w zbiorze danych.

Podsumowując: kolumny X1, X2, X3, X5, X6 i X7 mają średnią zbliżoną do mediany, co sugeruje symetryczność rozkładu. Wyjątkiem jest X4, która może mieć asymetryczny rozkład. Zmienne X4 i X3 mają największe rozproszenie (wysokie odchylenie standardowe), co wskazuje na dużą zmienność w wartościach. Zmienne X5 i X6 są najbardziej stabilne, z najmniejszym odchyleniem standardowym. Kolumny X3 i X4 mają najszersze zakresy wartości (Q0 do Q4), co potwierdza ich dużą zmienność.

Następnie skonstruowano szereg rozdzielczy przedziałowy (plik odstajace 1.5 arkusz szereg).

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

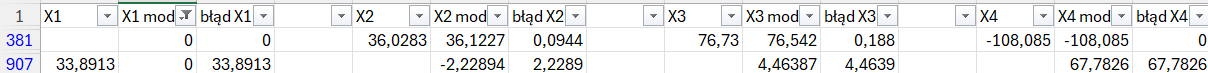
Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

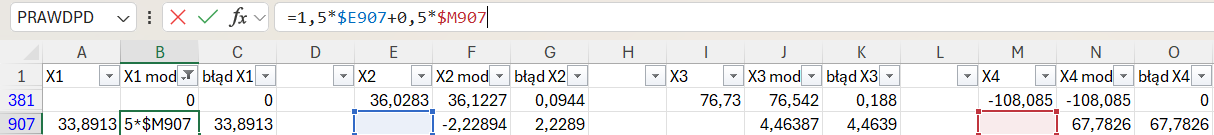
Opis wygenerowany automatycznie

Po uzupełnieniu braków danych zauważono poprawę w rozłożeniu danych pomiędzy poszczególne przedziały oraz w zgodności średniej z szeregu z dokładną średnią. Mimo to, obserwacje nadal nie są równomiernie rozłożone, a średnia z szeregu i średnia dokładna wciąż się różnią.

Dodatkowo, zauważono, że prawdopodobnie wciąż istnieje wartość odstająca, ponieważ w przedziałach nr 7, 8 i 9 nie ma żadnych obserwacji, a w przedziale 10 znajduje się tylko jedna obserwacja. W związku z tym zindentyfikowano tą wartość – jest to po prostu wartość największa,a następnie zastąpiono tę pojedynczą obserwację medianą.

Ponadto, jedna obserwacja została usunięta, ponieważ zmienna X1 = 0 została obliczona na podstawie braków danych (dotyczy to dokładnie wiersza 907, który został uznany za podejrzany w arkuszu modele).





Po tych zmianach szereg wygląda następująco:

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Wartość X1=0 prawdopodobnie również stanowi obserwację odstającą. Świadczy o tym fakt, że w pierwszym przedziale znajduje się tylko jedna obserwacja, podczas gdy kolejna większa liczba obserwacji występuje dopiero w przedziale trzecim. W związku z tym wartość ta została zastąpiona medianą.

Po wprowadzeniu tych zmian szereg przedstawia się w bardziej uporządkowanej formie (arkusz szereg(2)).

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Dane są równomiernie rozłożone a średnie z szeregu i dokładna są do siebie zbliżone.

**Wnioski z projektu**

Analiza danych pokazała, jak istotne znaczenie mają wartości odstające, ponieważ mogą one znacząco wpływać na wyniki badań. Możliwość przeprowadzenia bardziej precyzyjnej analizy byłaby większa, gdyby znany był charakter badanych danych. Niestety, w tym przypadku losowy charakter danych ograniczył takie możliwości. Warto jednak zauważyć, że bardziej zaawansowane metody, takie jak interpolacja, często okazują się skuteczniejsze, nawet jeśli dane nie spełniają wszystkich formalnych wymagań do ich zastosowania.