Arkusz pytań

Nazwisko: _____

Imię: _____

Kod badania: _____

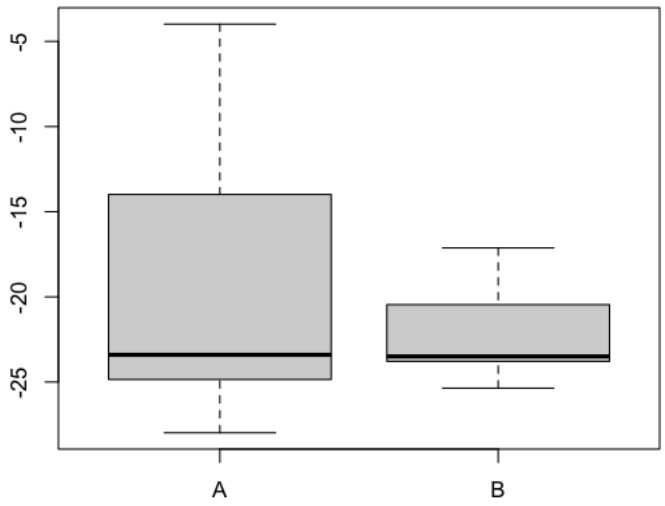
Podpis: _____

Jak prawidłowo zaznaczać?
Ten formularz odpowiedzi będzie automatycznie zeskanowany. Nie składaj i nie poplam go. Używaj długopisu z czarnym wkładem do zaznaczania pól. Jeśli chcesz dokonać korekty, zamaluj całkowicie korygowany kwadracik. Pole te zostanie zinterpretowane jako niezaznaczone.

1)

Na poniższym rysunku rozkład zmiennej dla dwóch prób (A oraz B) został zaprezentowany w formie równoległych wykresów ramka-wąsy.

Które z poniższych stwierdzeń są prawdziwe? *(Uwaga: Stwierdzenia mogą być prawdziwe bądź zupełnie fałszywe.)*



a) Parametry położenia dla obu rozkładów są mniej więcej równe.
b) W przypadku obu rozkładów nie zauważamy istotnych wartości odstających.
c) Rozrzut w próbie A jest wyraźnie większy niż w B.
d) (A)symetria dla obu rozkładów jest na podobnym poziomie.
e) Rozkład B jest prawoskośny.

2) Jaki jest główny cel operacji "data wrangling"?

a) Tworzenie modeli statystycznych na danych.
b) Przekształcanie danych w formę odpowiednią do analizy, czyli ich czyszczenie, łączenie i formatowanie.
c) Wykonywanie testów statystycznych na danych.
d) Wizualizacja danych.

3) Czym jest "data tidying" w analizie danych?

a) Przekształcanie danych z różnych źródeł w jeden spójny format.
b) Usuwanie duplikatów z danych.
c) Zmiana danych z postaci tekstowej na numeryczną.
d) Formatowanie danych do postaci łatwej do analizy, czyli standaryzacja i uporządkowanie danych.

4) Jakie jest główne różnice między poleceniami git merge a git rebase?

a) git merge łączy gałęzie, tworząc nowy commit, a git rebase przenosi commity z jednej gałęzi na drugą
b) git merge przenosi zmiany do zdalnego repozytorium, a git rebase wykonuje operacje tylko lokalnie
c) git merge łączy gałęzie bez zmian w historii, a git rebase zmienia historię commitów
d) git merge usuwa konflikty, a git rebase je generuje

5) W jaki sposób zespół może efektywnie współpracować nad projektem analitycznym przy użyciu Git i GitHub, aby uniknąć konfliktów podczas pracy nad tym samym plikiem R?

a) Zespół powinien pracować na różnych gałęziach (branches) i łączyć je za pomocą pull requestów, aby uniknąć konfliktów.
b) Zespół powinien pracować na jednej gałęzi i regularnie przesyłać zmiany bez łączenia ich z innymi, aby uniknąć problemów.
c) GitHub nie obsługuje pracy nad plikami R, więc zespół musi pracować w innych narzędziach.
d) Zespół powinien tworzyć tylko jeden plik, nad którym wszyscy pracują jednocześnie, aby uniknąć rozbieżności.

6) Jakie techniki używa się w EDA, aby sprawdzić, czy dane mają rozkład normalny?

a) Test t-Studenta i analiza regresji.
b) Test Shapiro-Wilka, histogramy i wykresy Q-Q.
c) Test Chi-kwadrat i test ANOVA.
d) Wykresy pudełkowe i wykresy rozrzutu.

7) Jakie techniki można wykorzystać do rozpoznawania wzorców braków danych w zbiorze danych?

a) Analiza wykresu rozrzutu (scatter plot) dla każdej pary zmiennych
b) Użycie funkcji missing w językach programowania jak R lub Python
c) Wykorzystanie narzędzi do analizy danych takich jak macierz braków, wykresy ciepła (heatmapy) oraz testy statystyczne na rozkład braków danych (np. test Little's MCAR)
d) Przeprowadzenie analizy regresji, aby przewidzieć brakujące dane

8) Jakie są potencjalne wady imputacji brakujących danych przy użyciu średniej (mean imputation)?

a) Może to prowadzić do zaniżenia zmienności w danych i zaburzenia rozkładu
b) Jest czasochłonne i może znacząco wydłużyć proces analizy
c) Może prowadzić do przesunięcia w rozkładzie danych, ale nie wpływa na wyniki modelu
d) Nie ma żadnych wad – to najbezpieczniejsza technika imputacji

9) W jaki sposób różnią się podejścia R i Pythona do analizy danych?

a) R ma wbudowane funkcje do analizy danych, podczas gdy Python wymaga dodatkowych bibliotek, takich jak Pandas, NumPy, czy SciPy.
b) Python ma wbudowaną funkcję do analizy danych, a R wymaga dodatkowych bibliotek.
c) R i Python mają identyczne podejście do analizy danych, różnią się tylko składnią.
d) R jest szybszy w analizie danych niż Python, ponieważ jest zaprojektowany specjalnie pod kątem statystyki.

10) Marek pracuje nad analizą wyników testów matematycznych uczniów w szkole. Chce zobaczyć, jak rozkładają się wyniki testów wśród uczniów, a także sprawdzić, czy wśród najstarszych uczniów wyniki są bardziej zróżnicowane niż wśród młodszych. Postanowił zastosować wykres, który pozwoli zobaczyć rozkład wyników oraz potencjalne wartości odstające.

Jaki wykres będzie najlepszy do wizualizacji rozkładu wyników testów, w tym wartości odstających, i porównania ich między grupami wiekowymi uczniów?

a) Wykres słupkowy (bar chart)
b) Wykres pudełkowy (box plot)
c) Histogram
d) Wykres punktowy (scatter plot)

11) Anna pracuje w firmie zajmującej się analizą sprzedaży online. Ma dane dotyczące miesięcznych wyników sprzedaży w różnych regionach przez ostatnie 3 lata. Chciałaby zobaczyć, jak w czasie zmieniają się wyniki sprzedaży w każdym z regionów, aby zidentyfikować, czy któreś regiony mają wyraźny trend wzrostu lub spadku.

Jakim wykresem Anna powinna się posłużyć, aby zobaczyć zmiany w czasie, porównać wyniki sprzedaży w różnych regionach i zidentyfikować ewentualne trendy?

a) Wykres słupkowy (bar chart)
b) Wykres punktowy (scatter plot)
c) Wykres liniowy (line chart)
d) Wykres kołowy (pie chart)

12) W jakim przypadku należy zastosować test chi-kwadrat?

a) Kiedy porównujemy średnie dwóch grup.
b) Kiedy analizujemy zależność między dwiema zmiennymi kategorycznymi.
c) Kiedy mamy dane liczbowe i chcemy obliczyć średnią.
d) Kiedy porównujemy częstości obserwacji w różnych kategoriach.

13) Które z poniższych stwierdzeń najlepiej opisuje zastosowanie rozstępu międzykwartylowego (IQR) jako miary rozproszenia?

a) IQR jest wrażliwy na wartości odstające, ponieważ bierze pod uwagę wszystkie wartości w zbiorze danych.
b) IQR jest odporny na wartości odstające, ponieważ mierzy rozproszenie tylko między 25. a 75. percentylem.
c) IQR jest miarą rozproszenia, ale wymaga, aby dane były rozkładem normalnym.
d) IQR jest używany tylko w przypadku danych kategorycznych, gdy nie możemy używać średniej ani wariancji.

14) Która z poniższych metod jest najbardziej odpowiednia do walidacji danych numerycznych w celu wykrycia wartości odstających?

a) Test Chi-kwadrat
b) Analiza skupień (cluster analysis)
c) Wykres pudełkowy (box plot)
d) Test t-Studenta
e) Test Grubbsa

15) W jakich sytuacjach brakujące dane w zestawie danych mogą być uznane za "brudne" dane?

a) Kiedy brakujące dane są całkowicie losowe i nie wpływają na ogólną jakość analizy.
b) Kiedy brakujące dane są systematyczne i mogą wprowadzać błędy w analizie, takie jak niereprezentatywność próby.
c) Kiedy brakujące dane są rzadkie i mają minimalny wpływ na końcowe wyniki analizy.
d) Kiedy brakujące dane są usuwane w procesie oczyszczania danych.

16) Co oznacza "forkowanie" projektu na GitHubie?

a) Tworzenie kopii repozytorium w celu wprowadzenia zmian, które później mogą być połączone z oryginalnym repozytorium.
b) Tworzenie nowego repozytorium z istniejącego w celu jego udostępnienia innym użytkownikom.
c) Tworzenie nowego brancha w tym samym repozytorium, aby uniknąć konfliktów.
d) Używanie zewnętrznych bibliotek i narzędzi w projekcie.

17) Jakie są zalety używania Markdown w porównaniu do tradycyjnego HTML?

a) Markdown jest łatwiejszy do napisania i odczytania, jest bardziej przejrzysty i nie wymaga znajomości HTML
b) Markdown jest szybszy do renderowania w przeglądarkach internetowych niż HTML
c) Markdown umożliwia tworzenie zaawansowanych animacji na stronach
d) Markdown jest bardziej skomplikowany, co daje większe możliwości niż HTML

18) Co oznacza "Multiple Imputation" (imputacja wielokrotna) i kiedy jest stosowana?

a) Proces wielokrotnego usuwania brakujących danych w różnych próbach
b) Technika imputacji, która polega na generowaniu wielu zestawów imputowanych danych, a następnie łączeniu wyników analiz w celu uzyskania bardziej wiarygodnych wniosków
c) Proces, w którym brakujące dane są imputowane kilkoma różnymi metodami, a następnie porównuje się wyniki każdej z metod
d) Imputacja danych przy użyciu wielu wartości średnich, mediana i mode, aby uzyskać najbardziej dokładne dane

19) Przeprowadzono badanie na grupie 50 pacjentów, którzy otrzymali różne dawki leku. Celem było sprawdzenie, jak dawka leku wpływa na zmniejszenie ciśnienia krwi. Wyniki ciśnienia krwi przed i po terapii zostały zapisane dla każdego pacjenta.

Pytanie:
Jaką analizę statystyczną lub wykres należy zastosować, aby porównać zmiany ciśnienia krwi przed i po terapii?

a) Wykres rozrzutu (scatter plot) z wynikami przed i po terapii na osiach x i y.
b) Test t-Studenta dla prób zależnych (paired t-test), aby porównać średnie zmiany ciśnienia krwi.
c) Wykres pudełkowy (boxplot), aby zobrazować zmiany ciśnienia krwi przed i po terapii.
d) Wykres słupkowy porównujący średnie zmiany ciśnienia w zależności od dawki leku.

20) Które z poniższych stwierdzeń dotyczy praktycznej istotności?

a) Praktyczna istotność zależy od p-wartości testu.
b) Praktyczna istotność jest związana z wielkością efektu i jego znaczeniem w kontekście badania.
c) Praktyczna istotność mówi, czy wynik jest statystycznie istotny.
d) Praktyczna istotność nie ma nic wspólnego z rzeczywistym efektem w populacji.