Opis danych:

Zbiór danych zawiera następujące informacje o studentach:

- Numer_indeksu: unikalny numer indeksu studenta (liczba sześciocyfrowa).
- Płeć: Płeć studenta, tj. "Kobieta" lub "Mężczyzna".
- Kierunek_studiów: kierunek, na którym studiuje student. Możliwe kierunki to: "Informatyka", "Ekonomia", "Psychologia", "Biologia".
- Wiek: wiek studenta.
- Czas_nauki: ilość godzin nauki tygodniowo.
- Średnia_ocen: średnia ocen studenta.

Zadania do wykonania w projekcie:

1. Przygotowanie danych:

- Załaduj dane i przekształć zmienne jakościowe na typ factor.
- Wyświetl podsumowanie danych. Na podstawie tego określ, które zmienne są jakościowe, a które ilościowe.

2. Obsługa braków danych:

Sprawdź, czy w zestawie danych występują wartości NA. Jeśli tak:

- Dla zmiennych jakościowych zastąp brakujące wartości najczęściej występującą kategorią.
- Dla zmiennych ilościowych zastąp brakujące wartości średnią. Pamiętaj, aby dla zmiennej "wiek" średnia była zaokrąglona do pełnych wartości.
- Jeżeli NA pojawia się w kolumnie Numer_indeksu, usuń tę obserwację ze zbioru danych.

3. Funkcja wyznaczająca kwotę stypendium:

Napisz funkcję, która na podstawie średniej określi, czy danej osobie należy się stypendium, a jeśli tak, to jakiej wysokości.

Progi otrzymania stypendium:

- jeżeli średnia ≥ 4.5, to stypendium wynosi 750 zł,
- jeżeli średnia ≥ 4.7, to stypendium wynosi 850 zł,
- jeżeli średnia ≥ 4.9, to stypendium wynosi 950 zł.

Argumentem funkcji powinna być ramka danych.

Funkcja powinna zwrócić listę złożoną z 4 elementów:

- liczby przyznanych stypendiów,
- macierzy/ramki danych z numerami indeksów stypendystów wraz z wyliczonymi kwotami,
- miesięcznego kosztu stypendium (sumy wszystkich przyznanych stypendiów).

4. Wykres słupkowy dla kierunku studiów i płci:

- Używając biblioteki ggplot, stwórz wykres słupkowy (pionowe słupki) porównujący kierunek studiów z płcią. Wykres powinien mieć odpowiednio podpisane osie, tytuł, legendę oraz zmienione kolory słupków i tła.
- Na podstawie otrzymanego wykresu, opisz, jakie wnioski można wyciągnąć na temat zależności między płcią a kierunkiem studiów w zbiorze danych.

5. Histogram dla średniej ocen:

- Stwórz histogram dla średniej ocen, dodając linię reprezentującą jądrowy estymator gęstości oraz przerywane linie oznaczające średnią, modę i medianę. Oznacz osie, nadaj tytuł, zmień kolory histogramu oraz linii i dodaj legendę (legendy powinny uwzględniać styl linii, np. przerywana linia w legendzie odpowiada przerywanej linii na wykresie).
- Określ charakterystykę rozkładu: modalność, skośność i symetryczność.

6. Wykresy pudełkowe dla czasu nauki a kierunek studiów:

- Przy użyciu ggplot, narysuj wykresy pudełkowe dla czasu nauki w podziale na kierunek studiów. Wykres powinien mieć podpisane osie, tytuł oraz zmienione kolory.
- Na podstawie wykresu określ, czy rozkłady czasu nauki różnią się w zależności od kierunku. Który kierunek wymaga średnio najwięcej czasu, a który najmniej? Który kierunek charakteryzuje się największą wariancją? Czy są widoczne wartości odstające i jak je zidentyfikować?

7. Analiza rozkładu czasu nauki dla najliczniejszego kierunku studiów:

- Wybierz czas nauki dla najliczniejszego kierunku studiów. Przeprowadź analizę rozkładu: przetestuj i porównaj co najmniej dwa różne rozkłady, wyestymuj ich parametry i oceń dopasowanie.

8. Wykres zależności średniej ocen od czasu nauki dla kierunku "Biologia":

- Stwórz wykres punktowy, przedstawiający zależność średniej ocen od czasu nauki dla kierunku "Biologia". Wykres powinien mieć podpisane osie, tytuł, zmienione kolory oraz legendę.
- Na podstawie wykresu określ, jaki rodzaj zależności występuje między zmiennymi. Oblicz korelację między średnią ocen a czasem nauki. Sprawdź założenia regresji liniowej, wyestymuj współczynniki prostej regresji i dodaj ją na wykres. Przeprowadź również diagnostykę modelu za pomocą odpowiednich wykresów.