**Модульное домашняя работа 1**

Максимум- 10 баллов (баллы за каждый пункт указаны)

На проверку необходимо прислать два файла: с проведенными расчетами в SPSS и описанием результатов (word или pdf).

Срок сдачи – до 03.04.24

Сформулировать содержательно задачу исследования. Подобрать данные пространственной выборки для проведения предварительного, корреляционного и кластерного анализов. Ограничения: от 5 до 10 переменных (из них допускается не более двух бинарных, все прочие - непрерывные), объем выборки от 80 объектов.

1. ***«Предварительный анализ данных» (3 балла)****:*

Выполнить первичный анализ данных: описать данные, визуализировать их. Нормализовать и стандартизовать данные.

То есть сделать:

* 1. Постановка задачи. Описание показателей. Выдвижение рабочих гипотез.
  2. Графическое представление исходных данных (точечное распределение (Dotplot); листовая диаграмма (Stemplot). Диагностика выбросов (Ящичковая диаграмма Boxplot)
  3. Характеристики положения СВ (mean, median, mode)
  4. Характеристики разброса СВ (размах вариации и коэффициент вариации; дисперсия, стандартное отклонение);
  5. Ранговые характеристики СВ (квартили, децили)
  6. Z-преобразование
  7. - расчет межквартильной разницы (IQR):
  8. - 1,5 и 3 IQR;

- правило 3σ

* 1. Выводы

1. ***«Корреляционный анализ» (3 балла):***

Провести корреляционный анализ, проверить наличие мультиколлинеарных признаков. Обосновать выбор переменных, которые будут участвовать в анализе (от 4 до 7).

То есть сделать:

* 1. Построение полей корреляции для исследования связи между переменными.
  2. Описание связи между переменными (сила, направление связи, наличие аномальных наблюдений)
  3. Построение и интерпретация матрицы парных коэффициентов корреляции (до удаления аномальных наблюдений).
  4. Построение поля корреляции для исследования связи между переменными ПОСЛЕ удаления выбросов. Описание результатов (что изменилось).
  5. Построение и интерпретация матрицы парных коэффициентов корреляции после удаления аномальных наблюдений. Сопоставление коэффициентов корреляции ДО и ПОСЛЕ удаления аномальных наблюдений.
  6. Выводы о связи между признаками.
  7. Построение и интерпретация матрицы частных коэффициентов корреляции,
  8. Сравнение парных и частных коэффициентов корреляции, выводы о характере связей;
  9. Расчет множественного коэффициента корреляции (для переменной, которая будет использована в регрессионном анализе в качестве зависимой переменной). Описание результатов.
  10. Выводы.

1. ***«Кластерный анализ» (4 балла):***

Выбрать и обосновать меру расстояния, которая будет использоваться для проведения иерархической кластеризации. Построить дендограмму. Предположить, на какое количество кластеров разумно разделить выборку на основе полученных результатов.

То есть сделать:

*Построение и анализ дендрограмм:*

1. Рассмотреть 4-5 вариантов разбиения объектов на кластеры: метод ближнего соседа, метод дальнего соседа, центра тяжести, средней связи, и др.
2. Сделать вывод о наиболее предпочтительном числе кластеров (часто исследователь может выдвинуть рабочую гипотезу об альтернативном количестве кластеров, например, 3 или 4 (5)).

\*В отчет поместить только интерпретируемые дендрограммы. Остальные, по желанию, могут быть размещены в приложении.

1. На основании дендрограммы и результатов этого пункта выбрать число кластеров, на которое будет проводится деление методом k средних. Описать полученные результаты.

*Использование метода к-средних для классификации объектов:*

1. Провести кластеризацию методом k-средних по нормализованным и стандартизованным данным. Аргументировать выбор расстояния. В случае значительной разницы в результатах по нормализованным и стандартизованным данным, выбрать наиболее подходящий вариант и обосновать принятое решение.
2. Привести график средних значений кластеров, дать подробное описание кластеров и интерпретацию полученных в работе результатов (необходимо привести состав кластеров, дать им название, обосновать название с помощью графических средств). Сделать выводы