

## **Лабораторная работа** «Корреляционный и дисперсионный анализ данных»

В файле **insurance.csv** представлены данные о стоимости медицинской страховки в зависимости от комбинации различных факторов.

Исследовались переменные:

age: возраст (количественная)  
sex: пол (категориальная: male/female)  
bmi: индекс массы тела (количественная)  
children: количество детей/иждивенцев (дискретная количественная)  
smoker: курит ли (категориальная: yes/no)  
region: регион (категориальная: southwest/southeast/northwest/northeast)  
charges: стоимость страховки.

**1.** Установить три переменные, имеющие наибольшую корреляционную связь со стоимостью медицинской страховки (charges).

*Для количественных переменных сравнить значения различных коэффициентов корреляции. Проверить значимость коэффициентов корреляции.*

*Построить корреляционную матрицу для всех количественных переменных.*

**2.** Установить наличие корреляционной связи между полом (sex) и курением (smoke).

**3.** На основе однофакторного дисперсионного анализа выяснить, влияют ли:

- 1) фактор курения,
- 2) индекс массы тела,
- 3) возраст (разделить на 3 категории)

на стоимость медицинской страховки?

*Сделать визуализацию распределения данных по группам на основе boxplot.*

*Проверить выполнение требований для данных, которые необходимы для проведения классического дисперсионного анализа. Если требования не выполнены, все равно провести дисперсионный анализ.*

**4.** На основе двухфакторного дисперсионного анализа выяснить, влияет ли эффект курения на стоимость медицинской страховки по-разному для мужчин и женщин?

Отчет устный по результатам выполненной работы.