**Исследовательский проект по НИС «Анализ данных в Python»**

**Часть 1**

1. Сформируйте с помощью веб-скрейпинга или найдите любой готовый набор данных, например, на портале Kaggle.com, содержащий не менее 200 наблюдений (строк) и не менее 10 переменных (столбцов). Файл должен содержать как метрические, так и категориальные переменные.

2. Опишите датасет. Укажите количество строк и столбцов. По каждой переменной укажите:

- название и метки к переменной и к значениям, которые она принимает (если есть),

- что она измеряет или отражает, в каких единицах измерения,

- тип переменной: метрическая, категориальная (номинальная/порядковая/бинарная),

- тип данных в колонке (string, integer, пр.),

- число пустых значений.

3. Измените тип данных в столбцах на более подходящий (если применимо).

4. Рассчитайте новый столбец с данными на основе существующего(их).

5. Перекодируйте любую метрическую переменную в категориальную. Обоснуйте выбранные интервалы/группы.

6. Чётко и детально сформулируйте цель и задачи исследования.

7. Сформулируйте не менее 10 исследовательских гипотез, которые будут проверены в ходе анализа данных. По каждой гипотезе уточните и обоснуйте выбор переменных и методов анализа данных, которые будут использованы для её проверки.

8. Постройте не менее 5 графиков по имеющимся данным. У графиков должны быть подписи осей, названия и легенда. Каждый график должен отражать некоторые тенденции или взаимосвязи, наблюдаемые в данных. Под каждым графиком нужно написать комментарий, описывающий тенденции или взаимосвязи, которые удалось выявить в результате его построения.

9. Постройте не менее 3-х сводных таблиц. Под каждой таблицей нужно написать комментарий, описывающий тенденции или взаимосвязи, которые удалось выявить в результате её построения. Проведите тест Хи-квадрат, чтобы оценить наличие взаимосвязи между рассмотренными переменными (проверьте подходят ли данные для проведения теста, сформулируйте гипотезы для теста, сделайте выводы по результатам проведения теста).

10. Рассчитайте описательные статистики, характеризующие центральную тенденцию и разброс значений относительно меры центральной тенденции по всей выборке и по отдельным группам наблюдений. Под каждым выводом напишите комментарий, описывающий тенденции или взаимосвязи, которые удалось выявить в результате расчётов.

11. Посчитайте корреляции между числовыми данными и интерпретируйте полученные результаты. Сделайте выводы о специфике взаимосвязей между рассмотренными переменными.

12. Постройте модель линейной регрессии. В модели должно быть не менее 5 предикторов. Детально опишите полученную модель. Сформулируйте содержательные выводы по результатам проведенных расчётов. Нужно уточнить особенности взаимосвязей, наблюдаемых между каждым предиктором и зависимой переменной.

13. Постройте модель бинарной регрессии. В модели должно быть не менее 5 предикторов. Детально опишите полученную модель. Сформулируйте содержательные выводы по результатам проведенных расчётов. Нужно уточнить особенности взаимосвязей, наблюдаемых между каждым предиктором и зависимой переменной.

14. Проведите кластерный анализ данных. Сформулируйте содержательные выводы по результатам проведенных расчётов. Подробно опишите каждый из кластеров, придумайте для них названия. Если кластерный анализ не применим к Вашему набору данных, то нужно объяснить почему он не применим. В этом случае кластерный анализ можно не проводить.

15. Сделайте выводы относительно того, какие исследовательские гипотезы нашли подтверждение в ходе исследования, а какие были отвергнуты и почему.

16. Приведите доводы в пользу того, что цель и задачи исследования были реализованы. Сформулируйте общие выводы по результатам анализа данных. Какие выводы показались вам ожидаемыми, а какие неожиданными и почему?

**Часть 2**

Необходимо собрать данные из Интернета и сохранить их в таблице MS Excel. Задачу нужно придумать самостоятельно и подробно описать. Нужно собрать данные не менее чем со 100 разных веб-страниц. Файл Excel должен содержать не менее 10 разных колонок с данными и, соответственно, не менее 100 строк, каждая из которых содержит данные из одной веб-страницы.

**Условия выполнения и сдачи обеих частей работы**

Задание может быть выполнено индивидуально или в группах по 2 человека. Защиты работ состоятся в конце февраля – начале марта. Дедлайн – 20 февраля 23:55.

Додедлайналюбому члену группы необходимо загрузить в Smart LMS следующие файлы:

1. Jupyter Notebook, содержаний код, на основе которого выполнялись обе части задания;
2. Презентация PowerPoint, в которой отражены основные результаты проведённого анализа данных;
3. Файл Excel, содержащий данные, собранные при выполнении второй части задания.

В названии файлов нужно указать фамилии всех членов группы.

Длительность выступления одной группы – 10-15 минут. Оба участника группы должны присутствовать на защите и презентовать часть работы. То есть, не один человек презентует всю работу, а оба члена группы. Порядок выступления участников группа определяет самостоятельно. Во время презентации должна быть включена веб-камера. По итогам доклада могут быть заданы уточняющие вопросы.