



Утверждено
Протоколом №1 3(э) от 19.12.2024г.
Учебно-методического совета
по довузовскому образованию
Национального исследовательского
университета «Высшая школа экономики»



ВЫСШАЯ ШКОЛА
ЭКОНОМИКИ



Т БАНК

Рекомендации для участников олимпиады «Высшая проба» по профилю «Анализ данных», проходящей в формате Национальной олимпиады по анализу данных DANO

Содержание

1. Аннотация профиля
2. Структура вариантов олимпиадных заданий
3. Перечень и содержание тем
4. Как подготовиться: литература и другие источники

1. Аннотация

DANO — это состязание нового типа, предполагающее современный формат проведения, практические задания, командную и проектную работу участников, открытую образовательную программу для всех желающих.

В рамках олимпиады школьники учатся постановке исследовательского вопроса, формулировке гипотез, поиску причинно-следственных связей в статистических данных из разных областей — от статистики ГИБДД по автомобильным авариям и опросов студентов про их удовлетворенность дистанционной учебой до данных о ценах на жилье в новостройках и о программах кэшбэка. Эти навыки соответствуют профессии аналитика данных — одной из самых востребованных на данный момент.

В отличие от стандартных олимпиад экономического профиля, DANO предполагает значительно большую ориентированность на прикладные задачи: не только решение теоретических вопросов с поиском равновесия и максимизацией прибыли, а еще и формирование конкретных выводов и рекомендаций, подкрепленных реальными данными. DANO при этом не является соревнованием по математике или информатике; математика и программирование для участников олимпиады — это только инструменты решения задачи, а не самоцель.

Участие в DANO будет интересно тем, кто интересуется аналитикой в широком смысле и прикладными задачами, реально решаемыми в настоящих научных и бизнес-исследованиях. Это и участники олимпиад по экономике, особенно те, кто хотел бы уйти от абстрактных модельных примеров в сторону примеров более «жизненных», и те интересующиеся математикой и информатикой школьники, которые хотят применять свои навыки и знания в решении прикладных задач.

Задания этапов DANO охватывают темы на стыке математики, экономики и информатики. Участникам пригодятся знания по этим предметам, однако не лишним будет изучить дополнительные источники и самостоятельно — в этом каждому участнику поможет открытая Библиотека материалов.

2. Структура варианта и типы олимпиадных заданий

Варианты олимпиадных заданий по профилю «анализ данных» одинаковы для всех классов обучения. Задания всех этапов прошлых лет и ответы на них размещены на сайте олимпиады по ссылке dano.hse.ru/tasks.

2.1. Первый (отборочный) этап

Задания могут включать вопросы с выбором одного или нескольких ответов, вопросы с кратким открытым ответом (всего от 20 до 30 вопросов).

Максимальная оценка за задание зависит от его типа и сложности и обычно составляет от 2 до 6 баллов. Все задания оцениваются по системе «все или ничего»: правильный ответ оценивается в полный балл, любой другой — в 0 баллов.

2.2. Второй (отборочный) этап

Задания могут включать вопросы с выбором одного или нескольких ответов, вопросы с кратким открытым ответом (всего от 15 до 25 вопросов).

Максимальная оценка за задание зависит от его типа и сложности и обычно составляет от 2 до 9 баллов. Все задания оцениваются по системе «все или ничего»: правильный ответ оценивается в полный балл, любой другой — в 0 баллов.

2.3. Заключительный этап

Заключительный этап олимпиады состоит из двух туров — задачного и проектного.

Задачный тур выполняется индивидуально в онлайн-системе с прокторингом.

Проектный тур — командный, выполняется при личном присутствии участников на площадке проведения заключительного этапа.

Результат каждого участника — сумма его баллов за задачный тур и баллов его команды за проектный тур. Итоговый результат рассчитывается как 0,5 оценки задачного тура + 0,5 оценки проектного.

Критерии оценки обоих туров прошлых лет размещены на сайте вместе с заданиями. Также можно ознакомиться с работами победителей и призеров: dano.hse.ru/result.

2.3.1. Задачный тур

Вариант состоит из задач, предполагающих развернутый ответ (всего от 4 до 5 задач). Максимальная оценка за задачу зависит от его сложности и обычно составляет от 20 до 40 баллов.

Полным баллом оценивается верное и полное решение. Решения, в которых есть продвижение к правильному ответу, но ответ не достигнут или само решение частично неверное, могут быть оценены в неполный балл в соответствии с критериями, разработанными для каждого задания. Максимально возможный балл участника за задачный тур — 100.

2.3.2. Проектный тур

Задание представляет собой базу данных (датасет) и вопрос или набор вопросов к ней. Решение задания — исследовательский проект на основе представленной базы данных и его защита.

Критерии оценки исследовательского проекта разрабатываются ежегодно и размещаются на сайте олимпиады до начала заключительного этапа. Максимально возможный балл команды за проектный тур — 100.

3. Перечень и содержание тем олимпиадных заданий:

— Основы теории вероятностей:

- Вероятность и её свойства
- Классическое определение вероятности
- Элементы комбинаторики для расчёта вероятностей: перестановки, размещения и сочетания.
- Случайные величины и их моменты (математическое ожидание, дисперсия, эксцесс, асимметрия)
- Оценка степени связи между случайными величинами. Корреляция.
- Основные распределения: нормальное, распределения Стьюдента, Фишера, хи-квадрат.

— Основы математической статистика:

- Генеральная совокупность и выборка
- Методы формирования выборки
- Точечное оценивание параметров распределения
- Интервальное оценивание параметров распределения
- Общая концепция тестирования гипотез: нулевая и альтернативная гипотезы, критическая область, p-value.
- Тестирование гипотез о среднем, доле, равенстве средних и равенстве долей в разных предпосылках.

- Элементы регрессионного анализа: линейная регрессия, тестирование гипотез о значении коэффициентов, оценка качества модели. Интерпретация оценок коэффициентов.
- Основы анализа данных:
 - Типы данных: качественные, количественные. Представление качественных признаков как фиктивных (бинарных/индикаторных) переменных
 - Методы подготовки данных: удаление дубликатов, очистка данных от выбросов, заполнение пропусков
 - Основные описательные статистики. Показатели центра распределения (среднее, мода, медиана) и вариации (дисперсия, стандартное отклонение, интерквартильный размах)
- Методы визуализации данных:
 - Основные типы графиков для визуализации распределения показателей: гистограмма, «ящик с усами», столбчатая диаграмма.
 - Основные типы графиков для визуализации связи между показателями: диаграмма рассеяния, визуализация корреляционной матрицы.
- Основы научного метода: понятия «исследовательский вопрос», «гипотеза», «механизм» гипотезы:
 - Формулировка исследовательского вопроса. Тривиальные и нетривиальные исследовательские вопросы
 - Подбор гипотезы. Различие между гипотезой исследования и статистическими гипотезами.
 - Формулировка механизма. Альтернативные механизмы для подтверждения или опровержения гипотезы.
- Методы и инструменты обработки данных. Допускается использование любых инструментов, например:
 - Python — библиотеки pandas, scipy, numpy, matplotlib;
 - R — встроенные методы для работы с данными, библиотеки dplyr, ggplot2;
 - табличные процессоры — MS Excel и его аналоги.

4. Рекомендуемая литература и другие источники

Для подготовки к олимпиаде можно изучить различные материалы, представленные в [Библиотеке](#): видеоподкасты, образовательные курсы, полезные сайты с базами данных и литературные источники.

Все материалы для подготовки разбиты на категории и помечены специальными тегами, которые помогут сориентироваться, на каком из этапов они пригодятся.

Ниже перечислены несколько литературных источников, которые полезно знать на базовом уровне, а также те, которые будут интересны, если ты уже много знаешь про анализ данных и хочешь изучить что-то для более продвинутого уровня:

