Школа аналитики
Пример приема 2023/24

Обзор заданий

Уважаемые заявители,

Поздравляем с успешной сдачей вступительного испытания!

Следующим шагом в процессе поступления является решение тематического исследования. Это задание должно дать вам краткое представление о типичных задачах и обязанностях, с которыми аналитик данных сталкивается каждый день при принятии решений и рекомендаций на основе данных для решения реальных бизнес-задач.

Пожалуйста, ознакомьтесь с описанием дела и дальнейшими инструкциями

ниже. Удачи!

Искренне Ваш,

Команда Школы Аналитики

Основные пометки

Тематическое исследование предназначено для проверки вашей способности определять ключевые особенности и атрибуты данного набора данных, определять основные взаимосвязи и применять процесс анализа данных для получения значимых выводов.

Ожидается, что успешные кандидаты продемонстрируют сильные аналитические навыки и навыки решения проблем, а также хорошее владение фундаментальными инструментами анализа данных.

Требования к программному обеспечению

Хотя для выполнения задания не существует особых требований к программному обеспечению, мы настоятельно рекомендуем использовать следующие приложения:

- MS Excel (предварительный анализ данных)
- Python/R/Stata (продвинутый уровень)[1]анализ данных)
- MS PowerPoint (презентация результатов)

[1]Обратите внимание, что использование языков программирования не является обязательным, поскольку базовый анализ данных можно выполнить в MS Excel. Однако было бы преимуществом продемонстрировать свои технические навыки, выполнив более сложный анализ данных с помощью Python, R или Stata.

Описание тематического исследования

Для основного задания ознакомьтесь с описанием, описанием проблемы и дополнительной информацией кейса «Маркетинговая аналитика».

Кейс содержит следующие разделы:

- Введение
- Описание проблемы
- Данные и дополнительная информация
- Краткое описание задач и вопросов
- Метаданные

Маркетинговая аналитика

Введение

Вы работаете аналитиком данных в компании по благоустройству дома среднего и премиум-класса. Основной регион продаж компании — Москва и Московская область, а недавно был добавлен новый вариант доставки, позволяющий охватить клиентов по всей стране.

Описание проблемы

Согласно недавнему анализу отзывов клиентов, снизилась доля клиентов среднего возраста (35–59 лет), которые, как правило, являются наиболее прибыльными клиентами для компании. Анализируя структуру предпочтений клиентов этой возрастной категории, ваша задача как аналитика данных — предоставить рекомендации для успешной рекламной кампании, ориентированной на эту конкретную демографическую группу.

Данные и дополнительная информация

Компания провела 4 опроса клиентов, чтобы собрать информацию об общем опыте и предпочтениях клиентов при совершении онлайн-покупок. Детали опроса изложены ниже.

Опрос	Описание
Обзор товаров	• пол • возраст • первая покупка («Да» или «Нет») • рейтинг сервиса • рейтинг продукта
Служба доставки обзор	• пол • возраст • рейтинг сервиса • рейтинг продукта
Любимые продукты	• пол • возраст • постельное белье, товары для ванной, одеяла, товары для кухни, парфюмерия и декор, одежда для дома («Да» – покупали товар раньше, «Нет» – никогда не покупали товар)
Категория	• пол • возраст • спальня, ванная, кухня, посуда, домашняя одежда, шторы (1 – «Маловероятно, что куплю товар», 5 – «Великовероятно, что куплю товар»)

Краткое описание задач и вопросов

1. Определите бизнес-задачу
2. Сформулируйте цель и опишите, как вы ее достигнете.
3. Опишите данные, которые будут использоваться в анализе. Какие дополнительные источники информации можно использовать?
4. Определить методы анализа данных
5. Проанализируйте данные:
    a. Опишите данные. Какие выводы можно сделать при планировании рекламной кампании?

6. Определить потребности целевой аудитории. На что вам нужно обратить внимание?

в. Визуализируйте данные, используя различные методы (линейная диаграмма, столбчатая диаграмма и т. д.). Пожалуйста, используйте как минимум 3 разных типа диаграмм.

6. Создайте презентацию PowerPoint, чтобы обобщить и продемонстрировать основные выводы.

Метаданные

Файл**«market_poll1.csv»**содержит данные, относящиеся к обзору продукта:

#	Имя	Описание
1	секс	*пол интервьюируемого*
2	возраст	*возраст интервьюируемого*
3	first_purchase	*1, если это первая покупка, иначе 0 оценка сервиса от 1*
4	Grade_service	*(очень плохо) до 5 (очень хорошо) оценка товара от 1*
5	сорт_продукта	*(очень плохо) до 5 (очень хорошо)*

Файл**«market_poll2.csv»**содержит данные, относящиеся к обзору службы доставки:

#	Имя	Описание
1	секс	*пол интервьюируемого*
2	возраст	*возраст интервьюируемого*
3	first_purchase	*1, если это первая покупка, иначе 0 оценка сервиса от 1*
4	Grade_service	*(очень плохо) до 5 (очень хорошо) оценка товара от 1*
5	сорт_продукта	*(очень плохо) до 5 (очень хорошо)*

Файл**«market_poll3.csv»**содержит данные, связанные с предпочтениями клиента в отношении продукта:

#	Имя	Описание
1	секс	*пол интервьюируемого*
2	возраст	*возраст интервьюируемого*
3	постельное белье	*1, если интервьюируемый купил постельное белье, в противном случае 0*
4	ванна	*1, если респондент купил туалетные принадлежности, в противном случае 0*
5	одеяло	*1, если респондент купил одеяла, в противном случае 0*
6	набор	*1, если опрашиваемый купил товары для кухни, иначе 0 1,*
7	декор	*если опрашиваемый купил парфюмерию и декор, иначе 0*
8	ткань	*1, если опрашиваемый купил одежду для дома, иначе 0*

Файл**«market_poll4.csv»**содержит данные, связанные с предпочтениями категорий продуктов:

#	Имя	Описание
1	секс	*пол интервьюируемого*
2	возраст	*возраст интервьюируемого*
3	кровать	*вероятность покупки товаров для спальни от 1 (низкая вероятность) до 5 (высокая вероятность)*

4	ванна	*вероятность покупки товаров для ванной от 1 (низкая вероятность) до 5 (высокая вероятность)*
5	набор	*вероятность покупки кухонных товаров от 1 (низкая вероятность) до 5 (высокая вероятность)*
6	блюдо	*вероятность покупки блюд от 1 (низкая вероятность) до 5 (высокая вероятность)*
7	ткань	*вероятность покупки домашней одежды от 1 (низкая вероятность) до 5 (высокая вероятность)*
8	шторы	*вероятность покупки штор от 1 (низкая вероятность) до 5 (высокая вероятность)*

Дополнительное присвоение кредита

Введение

Чтобы заработать дополнительные баллы к общей оценке за практический пример, кандидатам предлагается выполнить дополнительное задание.

Это потребует знания и применения более продвинутых методов анализа данных, включая обработку естественного языка (NLP).

Описание проблемы

Основная цель вашего исследования — определить, можно ли использовать описание розничного продавца в качестве вывода о ценовой категории. Другими словами, вам необходимо изучить взаимосвязь между тем, как описывается конкретный ритейлер, и какой ценовой сегмент он представляет.

Данные и дополнительная информация

Набор данных содержит информацию примерно о 2740 ритейлерах, работающих в России. Данные были удалены с российского сайта, собирающего информацию о российской торговой недвижимости и торговых центрах.

Краткое описание задач и вопросов

1. Импортируйте все модули Python, которые вам могут понадобиться: pandas как pd, numpy как np, nltk, re, sklearn.
2. Импортируйте данные из файла «russian_retail.csv»: pd.read_csv().
3. Чтобы преобразовать тексты в наборы признаков с помощью метода «мешка слов», необходимо сначала отфильтровать текст, оставив только значимые слова. Загрузить набор малозначащих слов из корпуса для целей фильтрации командой nltk.download() (в появившемся окне, вкладка Corpora, пункт стоп-слова)
4. Импортируйте стоп-слова из nltk.
5. Определите функции фильтра
6. Скопируйте данные с помощью функции copy() в новую переменную и используйте функцию фильтра. . применить(лямбда s: smth_to_words(s)). Распечатайте первые три строки
7. Визуализируйте регионы с помощью WordCloud
8. Визуализируйте описание с помощью WordCloud
9. Визуализируйте другие столбцы с помощью любых диаграмм. Почему вы использовали этот тип визуализации?

Более подробное описание задачи и пошаговое руководство можно найти в шаблоне Jupyter Notebook.**«SoA – Пример поступления – Присвоение дополнительных кредитов».**

Метаданные

Файл**«russian_retail.csv»**содержит 7 столбцов, которые идентифицируют каждого ритейлера по специализации, ценовой категории, регионам присутствия и подробному описанию его бизнеса.

Имя	Тип данных	Описание
имя	объект	*# название продавца*
страна_происхождение	объект	*# страна происхождения продавца*
домен	объект	*# специализация бизнеса*
цена_категория	объект	*# ценовая категория*
основан	int64	*# год основания компании*
присутствие_регионы	объект	*# города и регионы России, в которых есть магазины ритейлера*
описание	объект	*# описание продавца*

Рекомендации по подаче

Ниже вы можете ознакомиться с общими требованиями к подаче задания по тематическому исследованию. Ваша заявка должна включать следующие файлы:

Основное задание

1. Презентация MS PowerPoint конвертирована в PDF-файл.
    - Имя файла: «Фамилия Имя – Презентация – Маркетинговая аналитика – 2»
    - Презентация должна иметь**БОЛЬШЕ НЕ НАДО**более 2–3 слайдов[2]
2. Таблица MS Excel.**ИЛИ**Файл Python/R/Stata
    - Имя файла: «Фамилия Имя – Анализ данных – Маркетинговая аналитика – 2»

Дополнительное присвоение кредита

1. Блокнот Jupyter с кодом Python
    - Имя файла: «Фамилия Имя – Дополнительное присвоение кредита – 2»

Все файлы должны быть отправлены не позднее**23:55 29 сентября**. Материалы, полученные после указанного срока, не принимаются и не отмечаются.

[2]Обратите внимание, что любая информация, выходящая за пределы максимального лимита слайдов, не будет рассматриваться и засчитываться в итоговую оценку.

Рубрика оценивания

Критерии выставления оценок по основному заданию приведены в таблице 1.

Таблица 1. Рубрики оценки тематического исследования

#	Критерии	Описание	Точки
		ОСНОВНОЕ ЗАДАНИЕ	
1	Бизнес-проблема	Бизнес-задача четко сформулирована и соответствует описанию проблемы.	2
2	Дизайн исследования	Цель и процесс исследования четко обозначены и соответствуют очертаниям проблемы.	2
3	Описание данных	Ключевые особенности набора данных правильно определены и описаны.	2
4	Методология	Приведено подробное описание с перечнем методов анализа данных.	3
5	Анализ данных	Проводится предварительный анализ данных: • Импорт данных • Агрегация и преобразование данных • Визуализация данных • Интерпретация данных	4
6	Полученные результаты	Результаты четко сформулированы и правильно выведены из проведенного анализа данных.	3
7	Рекомендации	Рекомендации четко изложены и соответствуют результатам.	2
8	Стиль и форматирование	Представленный файл имеет четкую структуру и соответствует общим стандартам оформления документа: • Анализ данных в Excel имеет единый формат для всей таблицы (числа, шрифты, выравнивание и т. д.). **ИЛИ** • Анализ данных, выполненный на Python, R или Stata, соответствует руководствам по стилю для конкретного языка программирования.	1
9	Презентация	Представленный файл имеет четкую структуру и соответствует общим стандартам оформления документа: • Оформление слайдов в MS PowerPoint (цифры, шрифты, выравнивание и т.д.) • Слайды MS PowerPoint лаконичны и не перегружены информацией. • Презентация MS PowerPoint имеет четкую и логичную подачу информации.	1
	Общий		**20**

Критерии выставления оценок за дополнительное задание приведены в Таблице 2.

Таблица 2. Рубрики оценки тематического исследования

#	Критерии	Описание	Точки
		ДОПОЛНИТЕЛЬНОЕ ПРЕДОСТАВЛЕНИЕ КРЕДИТА	
1	Разведочные данные Анализ (ЭДА)	Проводится предварительный анализ данных: • Импорт данных • Очистка данных • Агрегация и преобразование данных • Визуализация данных • Интерпретация данных	2
2	Естественный язык Обработка (НЛП)	Соответствующие библиотеки Python для обработки естественного языка используются для анализа текста и получения выводов на основе данных.	4
3	Визуализация	Все необходимые визуализации выполняются с использованием библиотеки Wordcloud Python и других типов визуализации.	3
4	Стиль и форматирование	Представленный файл имеет четкую структуру и соответствует общим стандартам оформления документа: • Анализ данных выполняется на Python, и решение отправляется в виде блокнота Jupyter. • Код тщательно прокомментирован и соответствует рекомендациям по стилю PIP 8.	1
	Общий		**10**
	Тотал Инк. дополнительный кредит		**30**