Направление Data Analyst, компания P&G

Добро пожаловать на виртуальную стажировку компании Procter&Gamble!

Предлагаем тебе примерить роль аналитика данных в IT-подразделении международной компании сектора FMCG[1], где объединяются бизнес, технологии и инновации.

Выполнение всего блока заданий займет у тебя не более 60-80 минут.

В результате ты научишься:

1. Проводить анализ исходных данных на полноту и оценивать их качество.
2. Искать ассоциативные правила, используя Python.
3. Из всего дата-сета выбирать только те данные, которые нужны для работы.

Рекомендуемый тайминг:

1. 10 минут на первое задание.
2. 15-20 минут на второе задание.
3. 30 минут на третье задание.

Информация о загрузке решения:

Данный проект содержит несколько подзадач. Можно загрузить файл, содержащий решение только части заданий, но по возможности старайся сделать их все.

Желаем удачи!

[1] FMCG (fast –moving consumer goods) — товары повседневного спроса, включающие продукты легкой и пищевой промышленности, а также косметику, предметы личной гигиены, моющие средства и пр.

Задание 1. Проверка данных на полноту

На первом шаге в качестве Data Analyst компании P&G тебе предстоит научиться определять полноту исходных данных.

Утром ты получил письмо от руководителя с инструкцией по выполнению задания.

Привет!

Нам нужно проанализировать данные истории покупок. Сейчас планируем запуск проекта по работе с отсканированными чеками покупателей, которые те загружают для получения кэшбэка и призов по программе лояльности «Кэшбэк 10 %».

Мы хотим **исследовать данные отсканированных чеков на предмет полноты** (дата-сет прикреплен во вложении), то есть **посчитать количество строк, содержащих пустые или пропущенные значения, которые обозначаются как NaN[2] и вывести их долю в процентах от общего количества строк в файле**.

Hints. При работе с большими дата-сетами удобно пользоваться Python. Ниже представлен код, который поможет тебе прочитать файл:

```
df = pd.read_csv('Case Data.csv', sep=",")
print(df)
```

Твоя задача дополнить код так, чтобы он посчитал и напечатал количество строк с NaN-значениями, а также указал в каких столбцах содержится больше всего пустых строк. Для этого используй метод isna() (isnull для версии pandas < 0.21.0). Для подсчета общего числа строк и столбцов можешь воспользоваться методом df.shape.

Оставь несколько строк комментариев в файле кода о том, какие столбцы содержат больше всего NaN-значений.

Наша IT-команда очень рассчитывает на тебя.

Полезные материалы

Статья о качестве данных: Качество данных | Бизнес-Анализ в России (infozone.pro).

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива и включи в него свой документ.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

[2] NaN (Not-a-Number) — одно из особых состояний числа с плавающей запятой; в вычислениях может интерпретироваться как значение, которое не определено или непредставимо.

Задание 2. Поиск ассоциативных правил

После успешного завершения анализа исходных данных на полноту, тебе нужно разобраться с применением Python для составления ассоциативных правил.
На почте ты обнаружил новое письмо от IT-команды с постановкой очередного задания.

Привет!

В качестве второго задания в роли Data Analyst предлагаем тебе познакомиться с ассоциативными правилами.

Обучение ассоциативным правилам, или **Associations Rules Learning (ARL)** происходит на базе правил, помогающих обнаруживать взаимосвязи между транзакциями покупателей по данным истории покупок. В ARL анализируются чеки отдельных потребителей и выявляются правила взаимосвязей. Например, если 80% покупателей, имеющих в чеке пиццу, берут и зубные щетки, то получается следующее правило: «Покупка пиццы является условием для приобретения зубных щеток». При этом 80% — наша уверенность в правиле. Но бывает полезно учитывать не только confidence, но и поддержку. Поддержка — это доля покупателей, которые взяли вместе пиццу и зубные щетки, от общего числа клиентов.

Hints. Для ARL в Python есть библиотека efficient_apriori. Если импортировать оттуда Apriori алгоритм[3], то можно научиться составлять базовые правила. Прежде чем применить apriori на Python, нужно заполнить NaN-значения с помощью метода fillna, а затем сгруппировать теги только для уникальных ID чеков (обрати внимание на groupby). После этого нужно сформировать лист из сгруппированных значений, используя list(map(tuple, сгруппированные_значения)). Наконец, можно обращаться к apriori, однако функция просит указать минимальное значение поддержки и уверенности. Пробуй различные значения для min_support. Помни, что если поддержка высокая, то вероятно речь идет об очевидном правиле. Такие не всегда интересны аналитикам, так как они общеизвестны. Слишком низкая поддержка свидетельствует о нехватке данных для проведения статистического анализа, поэтому правила с низкой поддержкой тоже редко используются. Также не забудь напечатать lift, support, confidence для одного правила, используя цикл с for и print.

Как обычно ждем код с выполненным заданием, где ты должен оставить комментарии по выбранным значениям поддержки и уверенности.

Спасибо!

Полезные материалы

Статья о поиске ассоциативных правил: Association Rule Mining via Apriori Algorithm in Python - Stack Abuse.

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива и включи в него свой документ.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

[3] Apriori алгоритм является одним из наиболее популярных алгоритмов поиска ассоциативных правил.

Задание 3. Найди товары P&G в чеках покупателей

Поздравляем, ты почти освоил работу аналитика данных, впереди осталось только финальное задание.
Тем временем на почте появилось новое письмо с инструкцией, что делать дальше.

```
Привет!

В качестве третьего задания в роли Data Analyst в P&G нужно изучить распределение данных по категориям. Из общего дата-сета выбери только товары P&G, используя столбец «Бренд».

Основные требования к выполнению задания:
   • Найти максимальное количество товаров P&G в дата-сете.
   • Построить график, чтобы понять какие бренды P&G чаще всего есть в чеках покупателей.

Hints. Для нахождения брендов в дата-сете используй следующую функцию:
def count_brand(dataset, brand_name, parser):
    found_rows = df['Бренд'].str.contains(parser).sum()
    print(brand_name + ":", "%0.0f" % df['Бренд'].str.contains(parser).sum())
    return found_rows

Прокомментируй, какие бренды P&G входят в топ-3 по популярности среди покупателей.

Для построения графика советуем использовать plt.barh(). В этом случае названия брендов будут находиться по вертикальной оси и их удобнее читать.

Успехов!
```

Полезные материалы

Для знакомства со всеми брендами компании, посети эту страницу: Бренды | Procter and Gamble.

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.