**Ответьте на вопросы по данным из датасета cacao\_flavors.xls.**

|  |  |
| --- | --- |
| **1. Первичный осмотр данных** | |
|  | |
|  | |
|  | |
| 1.1 Сколько в датасете числовых столбцов? |  |
| 1.2 Сколько в датасете строковых столбцов? |  |
| 1.3 Переименуйте столбцы | |
| Новые названия: | |
| title\_col=['bar\_id', 'company', 'specific\_origin', 'ref', 'review\_date',  'cocoa\_percent', 'company\_location', 'rating', 'bean\_type', 'broad\_origin'] | |
|  | |
| **1.4. Анализ данных столбцов** | |
|  | |
| 1.4.1. Данные **вar\_id.** Сколько уникальных значений? График | |
|  | |
|  | |
|  | |
| 1.4.2. Данные **сompany.** | |
| Сколько компаний? |  |
| Отфильтровать лишние данные и вывести только те компании, у которых в рейтинге больше 10 продуктов. Сколько таких? |  |
|  | |
| 1.4.3 Данные **specific\_origin.** | |
| Сколько уникальных регионов в датасете? |  |
| Сколько значений, встретившихся в столбце более 10 раз? |  |
|  | |
| 1.4.4 Данные **ref.** Построить график распределения. Выводы | |
| Есть аномалии в данных? пропуски? |  |
|  | |
|  | |
| 1.4.5 Данные **review\_date**. Построить график распределения. Выводы | |
| Есть аномалии в данных? пропуски? |  |
| Содержит ли этот столбец выбросы? |  |
|  | |
| 1.4.6 Данные **cocoa\_percent.** Предварительно преобразовать данные в вещественные числа. | |
| Есть аномалии в данных? пропуски? |  |
| Содержит ли этот столбец выбросы? |  |
| Построить график распределения. Выводы | |
|  | |
| Отфильтруйте данные руководствуясь здравым смыслом | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
| 1.4.7 Данные **company\_location.** | |
| Сколько стран? |  |
| Сколько стран, у которых больше 10 ревью? |  |
|  | |
|  | |
| 1.4.8 Данные **rating** - будет учиться предсказывать будущая модель | |
| Есть выбросы? |  |
|  | |
| Постройте график распределения оценок после выявления выбросов | |
|  | |
| Отфильтруйте данные | |
|  | |
|  | |
|  | |
| 1.4.9 Данные **bean\_type.** | |
| Есть пропуски? |  |
| Сколько значений, встретившихся в столбце более 10 раз |  |
| Отфильтруйте данные. | |
|  | |
|  | |
|  | |
| 1.4.10 Данные **broad\_origin.** | |
| Есть пропуски? |  |
| Отфильтруйте данные. | |
| Сколько значений, встретившихся в столбце более 10 раз |  |
| Сколько уникальных значений? |  |
|  | |
| В столбце много повторяющихся значений с различным написанием. Напишите функцию, исправляющую эту ситуацию. | |
|  | |
| Как изменятся данные | |
| Сколько значений, встретившихся в столбце более 10 раз |  |
| Сколько уникальных значений? |  |
|  | |
|  | |
| 2. **Корреляционный анализ числовых данных** | |
| Построить матрицу корреляций. Сделать выводы | |
|  | |
|  | |
|  | |
| **3. Анализ номинативных переменных.** | |
| Построить графики распределения | |
|  | |
|  | |
| **4. Общие выводы** | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
| **5. Построить линейную регрессионную модель для прогнозирования рейтинга. Посчитать характеристики качества модели** | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |