Кадарметов Дмитрий

Группа: БИВТ-ВП-23

Лабораторная работа 1.

Тема. Сбор и подготовка данных.

Цель: освоить навык предварительной подготовки данных.

Этапы выполнения работы:

1. Определить и описать кратко предметную область. Сформулировать гипотезу или цель анализа, которые можно в дальнейшем проверить на основе анализа выбранных данных.
2. Для выбранной предметной области зафиксировать анализируемый объект и его характеристики (2 или 3). Найти соответствующие данные для анализа. Указать источник данных.
3. Описать генеральную совокупность. Определить и описать способ построения выборки из генеральной совокупности объекта исследования.
4. Представить данные выборки в табличном виде: строки - экземпляры объекта, столбцы - данные характеристик.
5. Выполнить и зафиксировать в отчете результат необходимых операций предобработки данных (в случае отсутствия необходимости выполнения конкретной операции, указать - не требуется):
    - разделение данных,
    - приведение данных к одинаковым единицам измерения,
    - преобразование к унифицированной лексике,
    - объединение данных из разных источников,
    - соединение данных из разных источников,
    - агрегация данных,
    - заполнение отсутствующих числовых значений,
    - очистка данных.
6. В соответствии с имеющимися данными построить график для каждого набора данных (для каждой характеристики объекта).
7. Построить диаграмму рассеяния (точечная диаграмма) для набора данных.
8. Сделать вывод.
9. Подготовить и загрузить отчет.

ОТЧЕТ О ПРОДЕЛАННОЙ РАБОТЕ

1. Предметная область – продажи кофейных напитков в кофейнях. Она включает в себя процессы, связанные с приготовлением и реализацией кофе, чая и других напитков, а также продажей сопутствующих товаров (выпечка, десерты) в условиях кофейни.

Цель работы – получить базовое представление о динамике продаж кофейных напитков в течение года чтобы определить наиболее благоприятное время для открытия кофейни.

2. В рамках настоящей работы мы будем анализировать только продажи кофейных напитков, поскольку они являются основным товаром, приносящей предприятию прибыль.

Кофейные напитки бывают разных видов (американо, капучино, и др.), могут приобретаться в разное время суток (утро, день, вечер, ночь), в разное время года (осень, зима, весна, лето), с помощью разных типов оплаты (наличные, карта), за разную стоимость.

Источник данных для анализа – сайт Kaggle.

Ссылка на набор данных: https://www.kaggle.com/datasets/navjotkaushal/coffee-sales-dataset/data

3. Генеральная совокупность – все кофейные напитки, проданные за все время в кофейнях. Вид выборки: простая случайная - статистический метод, при котором каждое наблюдение в большой совокупности имеет равные шансы попасть в выборку.

4. Данные выборки:

hour_of_day	cash_type	money	coffee_name	Time_of_Day	Weekday	Month_name	Weekdaysort	Monthsort	Date	Time
10	card	38,70	Latte	Morning	Fri	Mar	5	3	01.03.2024	10:15:51
12	card	38,70	Hot Chocolate	Afternoon	Fri	Mar	5	3	01.03.2024	12:19:23
12	card	38,70	Hot Chocolate	Afternoon	Fri	Mar	5	3	01.03.2024	12:20:18
13	card	28,90	Americano	Afternoon	Fri	Mar	5	3	01.03.2024	13:46:33
13	card	38,70	Latte	Afternoon	Fri	Mar	5	3	01.03.2024	13:48:15
15	card	33,80	Americano with Milk	Afternoon	Fri	Mar	5	3	01.03.2024	15:39:48
16	card	38,70	Hot Chocolate	Afternoon	Fri	Mar	5	3	01.03.2024	16:19:03
18	card	33,80	Americano with Milk	Night	Fri	Mar	5	3	01.03.2024	18:39:04
19	card	38,70	Cocoa	Night	Fri	Mar	5	3	01.03.2024	19:22:02
19	card	33,80	Americano with Milk	Night	Fri	Mar	5	3	01.03.2024	19:23:16
19	card	33,80	Americano with Milk	Night	Fri	Mar	5	3	01.03.2024	19:29:17
10	card	28,90	Americano	Morning	Sat	Mar	6	3	02.03.2024	10:22:07
10	card	33,80	Americano with Milk	Morning	Sat	Mar	6	3	02.03.2024	10:41:41
11	card	33,80	Americano with Milk	Morning	Sat	Mar	6	3	02.03.2024	11:59:45
14	card	28,90	Americano	Afternoon	Sat	Mar	6	3	02.03.2024	14:38:36

5. Предобработка данных:
   - разделение данных – не требуется
   - приведение данных к одинаковым единицам измерения – не требуется
   - преобразование к унифицированной лексике – не требуется
   - объединение данных из разных источников – не требуется
   - соединение данных из разных источников – не требуется
   - агрегация данных – не требуется
   - заполнение отсутствующих числовых значений – не требуется
   - очистка данных – удаление столбцов, дублирующих информацию, и столбцов с данными, не относящимися к цели анализа, а именно: час дня, тип оплаты, время дня (утро, день, вечер), день недели, номер дня недели, номер месяца, точная дата и точное время совершения покупки.

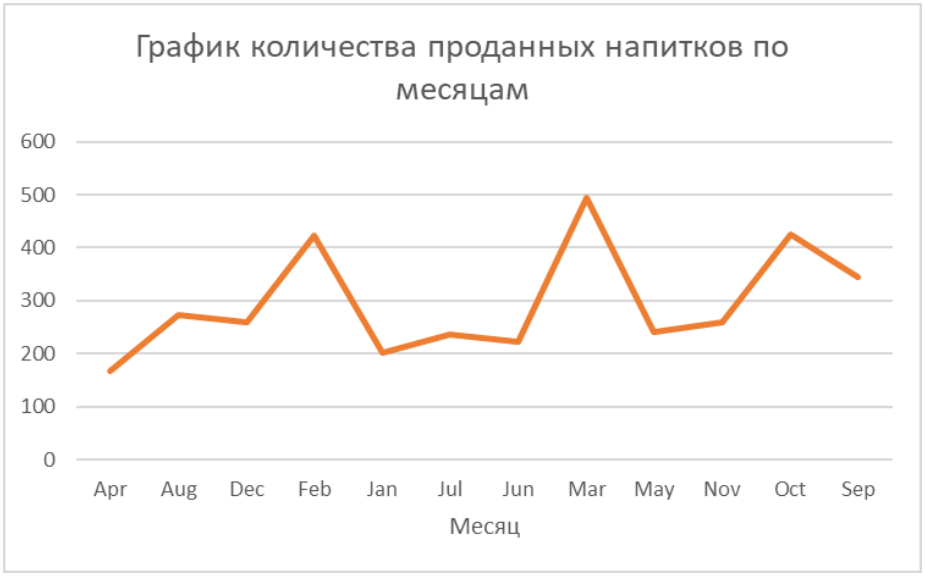
Таблица после предобработки:

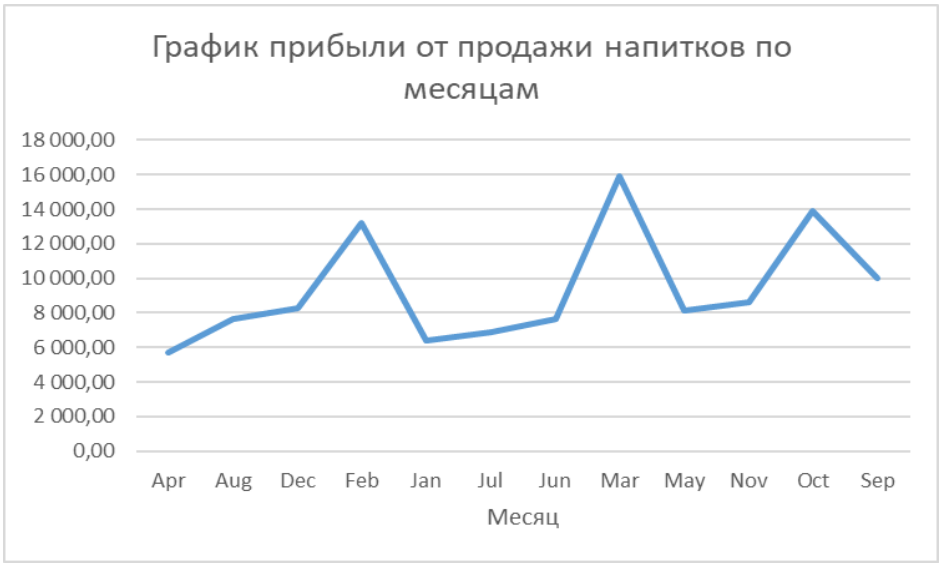
money	coffee_name	Month_name
38,7	Latte	Mar
38,7	Hot Chocolate	Mar
38,7	Hot Chocolate	Mar
28,9	Americano	Mar
38,7	Latte	Mar
33,8	Americano with Milk	Mar
38,7	Hot Chocolate	Mar
33,8	Americano with Milk	Mar
38,7	Cocoa	Mar
33,8	Americano with Milk	Mar
33,8	Americano with Milk	Mar
28,9	Americano	Mar
33,8	Americano with Milk	Mar
33,8	Americano with Milk	Mar
28,9	Americano	Mar

После предобработки на основе полученной таблицы была построена сводная таблица:

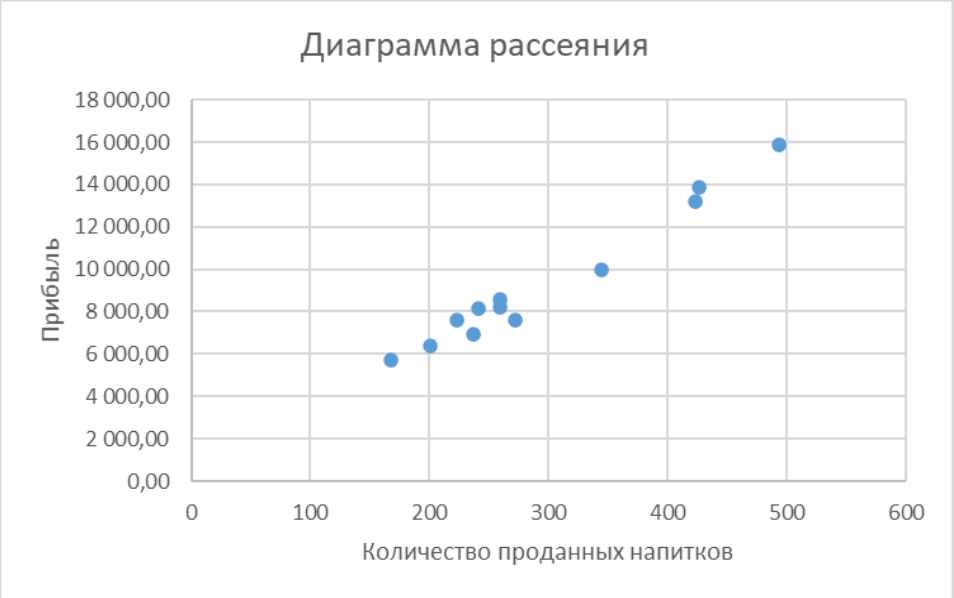
Названия строк	Количество по полю coffee_name	Сумма по полю money
Apr	168	5 719,56
Aug	272	7 613,84
Dec	259	8 237,74
Feb	423	13 215,48
Jan	201	6 398,86
Jul	237	6 915,94
Jun	223	7 617,76
Mar	494	15 891,64
May	241	8 164,42
Nov	259	8 590,54
Oct	426	13 891,16
Sep	344	9 988,64
Общий итог	**3547**	**112 245,58**

6. На основе построенной сводной таблицы были построены графики:





7. Диаграмма рассеяния для набора данных



8. Вывод:

В ходе работы были проведены сбор и предварительная подготовка данных для анализа, после чего по обработанным данным были построены графики и диаграмма рассеяния.

По графикам и диаграмме можно заметить что:

1) Пик продаж кофейных напитков – это месяцы: февраль, март, октябрь и сентябрь, причем в марте количество продаж особенно отличается от остальных месяцев, что можно видеть на диаграмме рассеяния.

2) Графики количества проданных напитков по месяцам и прибыли от продаж по месяцам практически совпадают по своей форме, из чего можно сделать вывод, что соотношение между количеством приобретаемых напитков по виду напитка остается достаточно постоянным в течение года. Небольшие отличия в форме графика наблюдаются в июле и августе.