Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____Радиотехнический (РТ) _____

КАФЕДРА Кафедра «Автоматизированные системы обработки информации и управления»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

_____Анализ базы данных супермаркета_____ ___

Студент _____ _____ **____Робертс Д. А.___**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ _____
(Подпись, дата) (И.О.Фамилия)

Консультант _____ _____
(Подпись, дата) (И.О.Фамилия)

20 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой _______
(Индекс)

_____ _____
(И.О.Фамилия)

« _____ » _____ 20 ____ г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме _____Анализ базы данных супермаркета _____

_____

Студент группы __РТ5-51Б_____

_____Робертс Даниил Александрович_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

_____

Источник тематики (кафедра, предприятие, НИР) _____

График выполнения НИР: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Техническое задание __ Сбор, анализ и обработка исходных данных для выдвижения гипотез и их проверки, путем анализа данных_____

_____

_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на ___ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

_____

_____

_____

_____

Дата выдачи задания « ___ » _____ 20__ г.

Руководитель НИР _____ _____
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

ВВЕДЕНИЕ 4
Основная часть 5
Определение данных для анализа 5
Цель 5
Анализ полученных данных 5
Формулирование гипотез 6
Обоснование сформулированных гипотез 6
Проверка 1 гипотезы 6
Проверка 2 гипотезы 7
Проверка 3 гипотезы 8
Проверка 4 гипотезы 9
Корреляционный анализ 10
ЗАКЛЮЧЕНИЕ 11
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 12

ВВЕДЕНИЕ

Научно-исследовательская работа (НИР) непосредственно ориентирована на подготовку студентов бакалавриата к работе в реальных условиях профессиональной деятельности.

Цель НИР – приобретение обучающимися знаний о систематике и методологии научных исследований, формирование практических навыков ведения самостоятельной научной работы, основным результатом которой является написание и успешная защита бакалаврского диплома.

Задачи:

– Освоение современных средств сбора и обработки информации;
– Формулирование актуальности, проблемных ситуаций, целей и задач исследования;
– Обучение методике и способам самостоятельного решения научно-исследовательских задач;
– Освоение методов исследования и проведения экспериментальных работ;
– Обработка полученных результатов, анализ и представление их в виде законченных научно-исследовательских разработок;
– получение данных для написания выпускной бакалаврской работы.
– Определение данных.
– Формулирование гипотез.
– Загрузка данных в Python.
– Проверка данных.
– Очистка данных.
– Преобразование данных.
– Выбор данных для анализа.
– Агрегирование данных.
– Визуализация данных.
– Подтверждение или опровержение поставленных гипотез.

Основная часть

Определение данных для анализа

В большинстве густонаселенных городов растет число супермаркетов, и конкуренция на рынке также высока. Набор данных представляет собой один из исторических данных о продажах компании-супермаркета, который был зарегистрирован в 3 разных филиалах за 3 месяца. Данные были взяты из открытого источника Kaggle.

Цель

Исследовать базу данных и изучить данные о продающихся продуктов, выявить зависимость количества продаваемых товаров различных категорий от характеристик покупателей, от времени суток, а также провести анализ цен на разные категории товаров.

Анализ полученных данных

Полученный для анализа набор данных содержит в себе 1000 записей, имеет 17 колонок.

Значение колонок следующее:

1) Invoice ID - индивидуальный айди, сгенерированный компьютером

2) Branch - ветвь супермаркета (всего 3 ветви, обозначенные буквами A, B и C)

3) City - местонахождение отделения

4) Customer type - Тип клиентов, Members - обладатели бонусной карты, Normal - обычные покупатели

5) Gender - пол

6) Product line - Общие группы категорий товаров - Electronic accessories - Электронника, Fashion accessories- Модные аксессуары, Food and beverages - Еда и напитки, Health and beauty - Здоровье и красота, Home and lifestyle - Дом и образ жизни, Sports and travel - Спорт и путешествия.

7) Unit price - цена за единицу товара

8) Quantity - количество купленных товаров

9) Tax 5% - налог

10) Total - цена с налогом

11) Date - дата покупки

12) Time - время покупки

13) Payment - способ оплаты (3 возможных: Cash - нал, Credit card - кредитной картой, Ewallet - электронный кошелёк)

14) cogs - Стоимость проданных товаров

15) gross margin percentage - процент валовой маржи

16) gross income - Валовой маржинальный доход

17) Rating - рейтинг клиентов по их общему опыту совершения покупок (по шкале от 1 до 10)

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1

Формулирование гипотез

– Женщины тратят на покупки больше мужчин (Более 5% в среднем)

– Покупатели с бонусной картой тратят больше денег в супермаркете (Более 5% в среднем)

– Больше всего прибыли супермаркетам приносит продажа товаров из категории "еда и напитки"

– Больше всего денег покупатели тратят днём (Время с 12:00 - 18:00)

Обоснование сформулированных гипотез

Предположении о том, что женщины тратят на покупки больше, чем мужчины было выдвинуто, потому что основным покупателем супермаркетов выступают женщины, на них ориентирована чаще всего маркетинговая компания.

Гипотеза о покупателях с бонусной картой была выдвинута по тем же причинам, что и в первой гипотезе.

Предположение, что больший доход супермаркетам приносят товары из категории «еда и напитки» было выдвинуто, потому что это товары первой необходимости и спрос на них есть всегда.

Большинство людей активны днём и ходят в супермаркеты именно днём. Кто-то просыпается только в обед, из-за чего не может сходить утром в магазин, кто-то работает до полудня. А вечер – это время, которое хочется провести с семьей, друзьями и другими близкими людьми.

Последняя гипотеза объясняется тем, что стоимость производства товаров из категории «электроника» достаточна высока. В электроники используется медь, золото для создания микросхем. Также процесс сборки достаточно дорогой.

Проверка 1 гипотезы

Данную гипотезу можно проверить в данной выборке, так как количество мужчин и женщин в выборке совпадает. Выборка сбалансирована для исследования по гендеру.

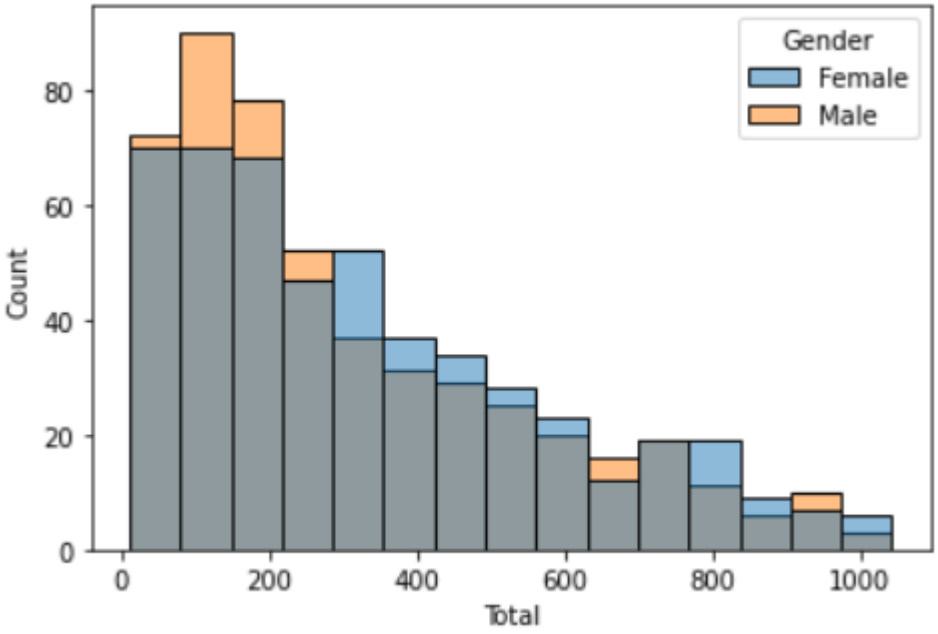
Для проверки первой гипотезы воспользуемся командой группировки по полю Gender и вычислим среднее значение каждого поля:

data.groupby('Gender').mean()

Gender	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
Female	55.263952	5.726547	15.956936	335.095659	319.138723	4.761905	15.956936	6.964471
Male	56.081944	5.292585	14.799487	310.789226	295.989739	4.761905	14.799487	6.980962

Также построим следующую гистограмму

sns.histplot(data=data, x="Total", hue="Gender")



По гистограмме видно, что мужчины совершают гораздо больше покупок на небольшие суммы, в то время, как девушки лидеры по количеству покупок в более крупных покупках. Также в таблице выборки, где были вычислены средние показатели по каждому полю, в зависимости от гендера, видно, что средний чек (Total) у женщин больше на 8%.

Гипотезу можно считать подтверждённой.

Проверка 2 гипотезы

В исходных данных одинаковое количество обладателей карты и обычных покупателей, так что выборка сбалансирована для данной гипотезы.

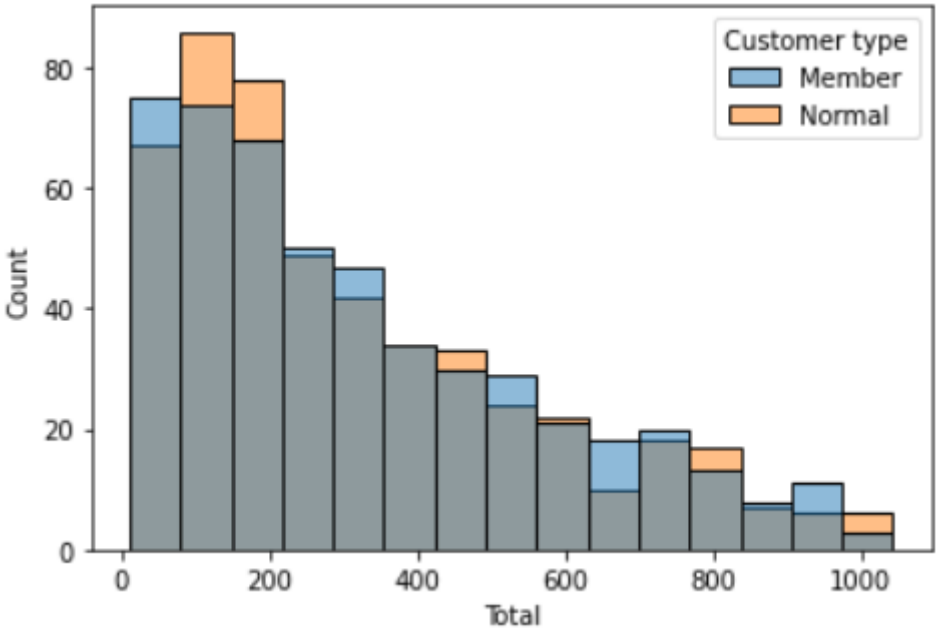
Для проверки следующей гипотезы воспользуемся командой группировки по полю Customer type и вычислим среднее значение каждого поля:

data.groupby('Customer type').mean()

Customer type	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
Member	56.206986	5.558882	15.609110	327.791305	312.182196	4.761905	15.609110	6.940319
Normal	55.135130	5.460922	15.148707	318.122856	302.974148	4.761905	15.148707	7.005210

Построим гистограмму:

sns.histplot(data=data, x="Total", hue="Customer type")



Как видно из графика, держатели бонусной карты и обычные покупатели не обладают заметным различием по сумме чека. Из данных выборки можно посчитать, что сумма среднего чека у держателей карты на 3% выше, что незначительно для подтверждения гипотезы. Ожидаемые данные были выше.

Гипотезу можно считать неподтверждённой.

Проверка 3 гипотезы

Строим круговую диаграмму:

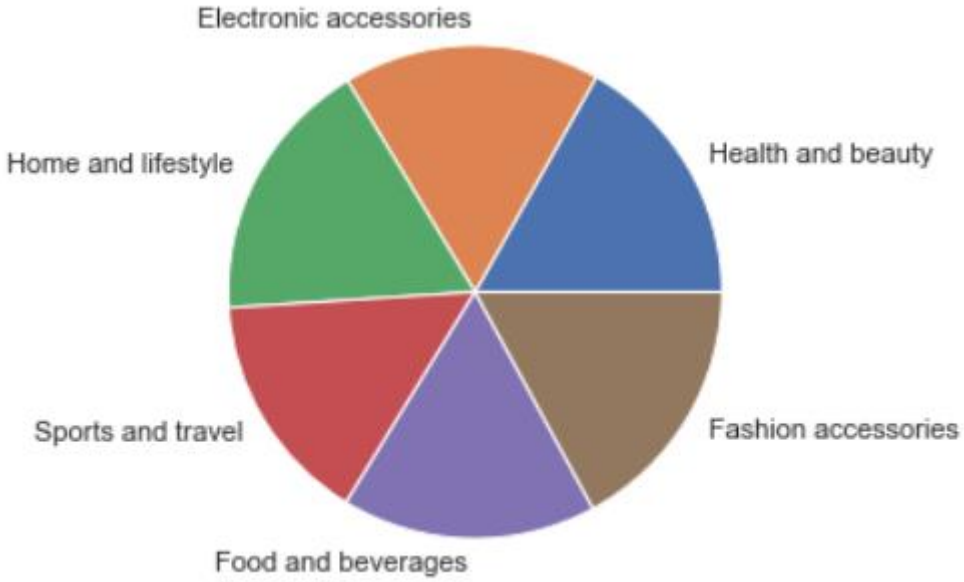
```
x=data['Product line'].unique()

y=data.groupby('Product line').sum()["Total"]

fig, ax = plt.subplots()

ax.pie(y, labels=x)

ax.axis("equal")
```



По диаграмме видно, что траты покупателей в категориях равномерно распределены. Гипотеза не подтвердилась.

Проверка 4 гипотезы

Для проверки этой гипотезы, необходимо добавить к нашим данным новое поле, которое будет отвечать за разделение на «Утро», «День», «Вечер», «Ночь». Для этого делаем следующие преобразования:

Преобразуем исходное поле в наборе данных в пандовское время для дальнейшей работы с ним.

data['time']=pd.to_datetime(data.Time)

В зависимости от времени проводим разделения на «Утро», «День», «Вечер», «Ночь» по следующему принципу:

Утро – 6:00-12:00
День – 12:00-18:00
Вечер – 18:00-24:00
Ночь - 24:00-6:00

data.loc[(data.time<'2021-12-21 6:00:00','day_part')]='Night'

data.loc[(data.time>'2021-12-21 6:00:00','day_part')]='Morning'

data.loc[(data.time>'2021-12-21 12:00:00','day_part')]='Day'

data.loc[(data.time>'2021-12-21 18:00:00','day_part')]='Evening'

Видоизменённый набор данных:

City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating	time	day_part
Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1	2021-12-21 13:08:00	Day
Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6	2021-12-21 10:29:00	Morning
Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4	2021-12-21 13:23:00	Day
Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4	2021-12-21 20:33:00	Evening
Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3	2021-12-21 10:37:00	Morning

Группируем по полю «day_part» и считаем количество сделок
data.groupby("day_part").count()

day_part	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating	time
Day	530	530	530	530	530	530	530	530	530	530	530	530	530	530	530	530	530	530
Evening	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279
Morning	191	191	191	191	191	191	191	191	191	191	191	191	191	191	191	191	191	191

Строим диаграмму:

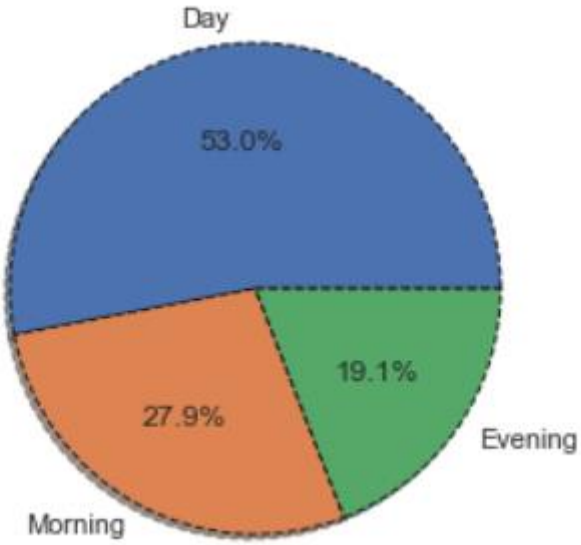
x=data['day_part'].unique()

y=data.groupby('day_part').count().Total

fig, ax = plt.subplots()

ax.pie(y, labels=x, shadow=True ,autopct='%1.1f%%',wedgeprops={'lw':1, 'ls':'--','edgecolor':"k"})

ax.axis("equal")



По данным видно, что большая часть покупок действительно совершается в периоде с 12:00-18:00, то есть днём.

Гипотеза подтвердилась.

Корреляционный анализ

```
numeric_col = ['Total', 'Rating', 'Quantity']

corr = data.loc[:,numeric_col].corr()

corr
```

	Total	Rating	Quantity
Total	1.000000	-0.036442	0.705510
Rating	-0.036442	1.000000	-0.015815
Quantity	0.705510	-0.015815	1.000000

Подобные поля были выбраны для исследования влияния рейтинга у клиента по опыту совершения покупок на такие параметры, как сумма потраченных денежных средств (Total) и количество купленного товара ('Quantity'). Корреляция последних двух параметров ожидаема. Чем больше товаров купит клиент, тем выше будет их цена.

Сильной зависимости количества покупаемых товаров и рейтингом клиента не замечена. Зависимость суммы потраченных средств от рейтинга также незначительна.

ЗАКЛЮЧЕНИЕ

Мной была проделана работа по анализу данных сети супермаркетов. Были сформулированы гипотезы. Часть из них подтвердилась, часть была опровергнута. Из проделанных исследований были сделаны выводы, что на количество покупок сильно влияет время суток, мало влияет пол покупателя и почти не влияет наличие у покупателя бонусной карты.

Корреляционный анализ показал отсутствие влияния рейтинга покупателя на количество товаров, которое он покупает и на их стоимость. Из чего можно поставить под сомнение пользу данной оценки для анализа данных и моделирования различных ситуаций.

Во время выполнения исследовательской работы мною были приобретены практические знания для сбора, обработки и анализа информации. Навыки в аналитическом мышлении повышены. Опыт в постановке гипотез получен. Данная работа помогла улучшить навыки в работе с данными.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Kaggle [https://www.kaggle.com/aungpyaeap/supermarket-sales]
2. Cyberleninka [https://cyberleninka.ru]
3. Документация Pandas [https://pandas.pydata.org/]
4. Документация MatPlotLib [https://matplotlib.org]
5. Документация Seaborn [https://seaborn.pydata.org/]