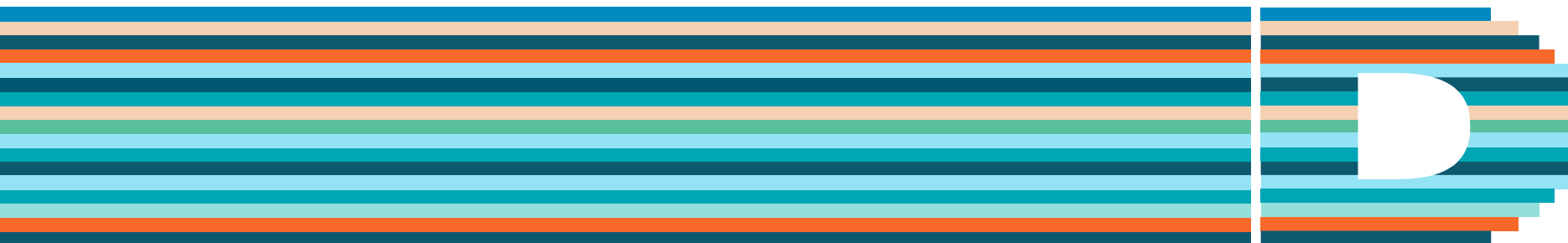


---

# Итоговый проект по программе «Data-аналитик: старт карьеры»

Поток № ДА-102  
группа 1

Выполнил:  
Горохов Дмитрий Евгеньевич



Национальный  
исследовательский  
Томский  
государственный  
университет

2023 г.

## data-diving

академия аналитики данных  
при Томском государственном  
университете



## Приветствие

---




**Добрый день, уважаемые члены комиссии!**

**Вашему вниманию представляется проект по теме аналитического исследования:  
«Анализ товарного предложения персональных компьютеров в онлайн-магазинах»**

---

**Название проекта:**

«Анализ товарного предложения персональных компьютеров в онлайн-магазинах»

-  **Бизнес-цель заказчика:** Достигнуть поступления денежной выручки от интернет-продаж компьютеров в размере 5 млн. руб. за 6 мес.
-  **Объект исследования:** Интернет-продажи компьютеров конкурирующих организаций (продавцов на онлайн-площадке).
-  **Предмет исследования:** Технические характеристики компьютеров с наибольшим объемом интернет-продаж: Тип и количество ядер процессора, Тип и объем оперативной памяти, жесткого диска, тип видеопроцессора.

**Цель анализа:** Выявление технических параметров компьютеров, с наибольшим спросом на интернет-площадке для планирования сборки востребованных комплектаций в соответствии с бизнес-целью.



## Выбор методологии

Проект выполнен с использованием методологии **CRISP-DM**

### 1. Работа с бизнес-требованиями:

-  - Бизнес-анализ (Определение бизнес-цели, определение цели аналитики, подготовка плана проекта)

### 2. Работа с данными:

-  - Анализ данных (Описание данных, изучение данных, проверка качества данных)
- Подготовка данных (Выборка данных, очистка данных, генерация данных, интеграция данных, форматирование данных)

### 3. Разработка и внедрение:

-  - Моделирование (Выбор алгоритмов, статистических методов, оценка качества моделей)
- Оценка решения (Оценка процесса, оценка результатов)
- Внедрение (Подготовка отчета, ревью проекта)



## Этапы исследования

# Этапы проведенного исследования

Разработка  
аналитического решения

Предобработка данных

Разведочный анализ

Статистическое  
исследование данных

Интерпретация данных



## Этапы исследования

### Разработка аналитического решения.

#### Определение :



- Бизнес-цели заказчика,
- Цели анализа данных,
- Объекта предмета исследования,
- Источников данных,
- Типов данных и способов их получения,
- Методологии и этапов исследования,
- Методов анализа,
- Требований к результату анализа.



Описание проекта	
<b>Название проекта:</b> «Анализ товарного предложения персональных компьютеров в онлайн-магазинах»	
	<b>Бизнес-цель заказчика:</b> Достигнуть поступления денежной выручки от интернет-продаж компьютеров в размере 5 млн. руб. за 6 мес.
	<b>Объект исследования:</b> Интернет-продажи компьютеров конкурирующих организаций (продавцов на онлайн-площадке).
	<b>Предмет исследования:</b> Технические характеристики компьютеров с наибольшим объемом интернет-продаж: Тип и количество ядер процессора, Тип и объем оперативной памяти, жесткого диска, тип видеопроцессора.
<b>Цель анализа:</b> Выявление технических параметров компьютеров, с наибольшим спросом на интернет-площадке для планирования сборки востребованных комплектаций в соответствии с бизнес-целью.	



## Этапы исследования



## Предобработка данных:

Задача	Примененные методы
Обзор файла и описание проблем	(head(), tail(), sample(), shape(), info(), copy())
Выбор признаков для анализа	(drop(columns=))
Проверка типов данных, приведение данных в необходимый тип	(dtypes, astype())
Проверка и обработка пустых значений	info(), isna(), isnull(), dropna(), fillna()
Проверка на дубликаты, их обработка	uplicated(), drop_duplicates()
Распаковка данных	apply(), функции, lambda-функции, работа со словарями
Обзор и очистка значений признаков	unique(), apply(), функции, lambda-функции, isna(), notna(), re.sub(), loc-конструкции, str.contains()
Кодирование признаков,	apply(), функции, lambda-функции, isna(), notna()
Приведение таблицы в наглядный формат	reset_index(), rename(columns=)



## Этапы исследования



## Разведочный анализ:

Задача	Примененные методы
Анализ числовых признаков	<code>describe()</code> , <code>min()</code> , <code>max()</code> , <code>mean()</code> , <code>median()</code> , <code>quantile()</code> , <code>hist()</code> , <code>boxplot()</code>
Анализ взаимосвязи числовых признаков	<code>plot()</code> , <code>PairGrid()</code>
Анализ категориальных признаков	<code>describe()</code> , <code>mode()</code> , <code>value_counts()</code> , <code>bar()</code> , <code>barh()</code> , <code>pie()</code>
Анализ взаимосвязи числовых и категориальных признаков	<code>boxplot()</code>
Срезы, фильтрация, группировка таблиц	<code>loc</code> , <code>iloc</code> , <code>[]</code> , операторы сравнения, <code>str.contains</code> , <code>isin()</code> , <code>groupby()</code> , <code>sort_values()</code>



# 1

## Описание проекта



### Этапы исследования

#### Разведочный анализ данных:

В ходе выполнения кейса выполнены неграфический и графический статистический



анализ взаимосвязи:

- Количественных, порядковых и категориальных признаков, а также их влияние на количественный факторный целевой показатель «Цена», «Продажи»,

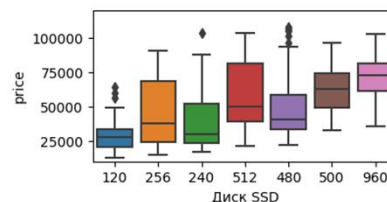


- Выполнены срезы, фильтрация, группировка таблиц,

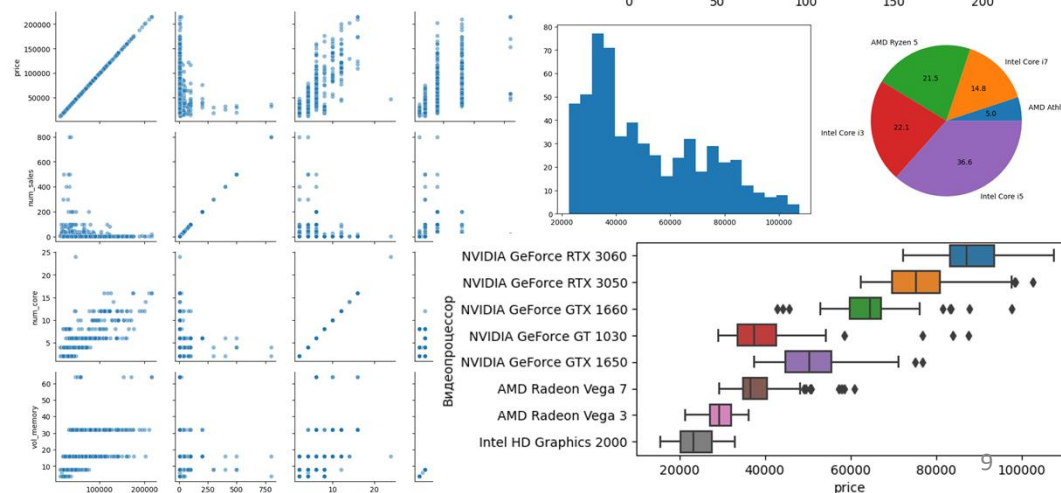
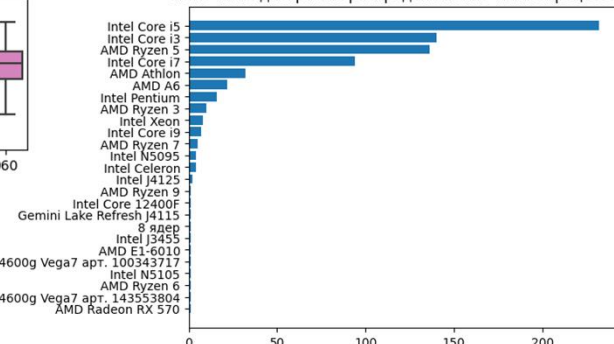


- Исследование проведено с применением языка программирования Python, с подключением и использованием библиотек numpy, pandas, matplotlib, seaborn.

	title	price	num_sales	type_proc	num_core	type_memory	vol_memory	Диск HDD	Диск SSD	Видеопроцессор
29	Офисный Компьютер Robotcomp Секретарь V1	21355	20	AMD Athlon	2	DDR 4	8	0	240	AMD Radeon Vega 3
30	Офисный Компьютер Robotcomp Секретарь V2	29482	5	AMD Athlon	2	DDR 4	8	0	240	AMD Radeon Vega 3
31	Офисный Компьютер Robotcomp Секретарь V3	34788	10	AMD Athlon	2	DDR 4	8	0	240	AMD Radeon Vega 3
176	Игровой компьютер Roo24 AMD Ryzen 3000G/A320/D...	26999	5	AMD Athlon	2	DDR 4	8	0	480	AMD Radeon Vega 3
...	...	...	...	...	...	...	...	...	...	...
645	Игровой компьютер Системный блок ПК CompDay Ma...	27136	5	AMD Athlon	2	DDR 4	16	0	240	AMD Radeon Vega 3
653	Игровой компьютер Системный блок ПК CompDay Дрон	31901	5	AMD Athlon	2	DDR 4	16	0	480	AMD Radeon Vega 3
663	Игровой Компьютер Robotcomp Star V1	29209	5	AMD Athlon	2	DDR 4	16	0	240	AMD Radeon Vega 3
697	Игровой ПК AMD Athlon X4 880K/8GB/240GB/GTX 16...	45600	5	AMD Athlon	4	DDR 3	8	0	240	NVIDIA GeForce GTX 1660



Столбчатая диаграмма распределения по типам процессоров





Этапы исследования



Статистический анализ:

Задача	Примененные методы
Анализ числовых признаков	describe(), min(), max(), mean(), std(), median(), skew(), kurtosis(), shapiro(), hist(), boxplot()
Анализ взаимосвязи числовых признаков	corr(), pearsonr(), spearmanr(), plot(), PairGrid(), heatmap()
Анализ категориальных признаков	describe(), mode(), value_counts(), bar(), barh()
Анализ взаимосвязи числовых и категориальных признаков	boxplot(), shapiro(), ttest_ind(), mannwhitneyu(), kruskalwallis()
Срезы, фильтрация, группировка таблиц	loc, iloc, [], операторы сравнения, str.contains, isin(), groupby(), sort_values()



## Этапы исследования

### Статистическое исследование данных:

- В ходе исследования ко всем наборам значений применены методы описательных статистик, проверки нормальности данных по критерию Шапиро-Уилка,
- Для выполнения анализа влияния категориальных признаков на количественные показатели применен метод сравнения групп,
- Для анализа связи количественных и порядковых признаков, а также анализа категориальных признаков между собой, применен корреляционный анализ.



```
1 dfs[:2]
```

	price	num_sales	processor	num_core	type_memory	vol_memory	disc	videoprocessor
0	55625	40	0	6	0	16	1	4
1	48386	30	1	4	0	16	1	4

Проверка нормальности по критерию Шапиро-Уилка

```
1 stats.shapiro(dfs['price'])
2 #выборка противоречит нормальному закону распределения
```

ShapiroResult(statistic=0.9425963163375854, pvalue=3.034856871017566e-12)

```
1 stats.shapiro(dfs['num_sales'])
2 #выборка противоречит нормальному закону распределения
```

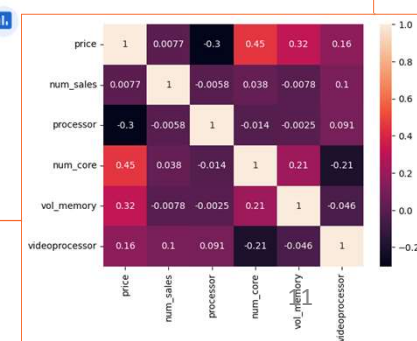
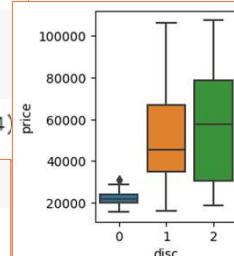
ShapiroResult(statistic=0.47980356216430664, pvalue=2.7583904776443576e-34)

```
2 stats.mstats.kruskalwallis(df_d1, df_d2, df_d3)
```

KruskalResult(statistic=51.593064517983294, pvalue=6.2619325838755975e-12)

```
1 dfs[['price', 'num_sales', 'processor', 'num_core', 'vol_memory', 'videoprocessor']].corr(method='spearman') # ранговая корреляция Спирмена
```

	price	num_sales	processor	num_core	vol_memory	videoprocessor
price	1.000000	-0.017848	-0.327616	0.400327	0.457358	0.168868
num_sales	-0.017848	1.000000	0.010488	0.030995	0.022203	0.073010
processor	-0.327616	0.010488	1.000000	0.031868	-0.016261	0.032041
num_core	0.400327	0.030995	0.031868	1.000000	0.339092	-0.211606
vol_memory	0.457358	0.022203	-0.016261	0.339092	1.000000	-0.004824
videoprocessor	0.168868	0.073010	0.032041	-0.211606	-0.004824	1.000000



## Этапы исследования

### Интерпретация данных:

- Промежуточные выводы по блокам, разделам в процессе анализа,
- Итоговый вывод, включающий основные выводы по разделам,
- Рекомендации заказчику.

#### Выводы по разделу

Выводы по разделу "Предобработка данных".

1. Размер датасета составляет 4500 наблюдений и 16 признаков.
2. В датасете выявлено наличие полных дубликатов в количестве 490 строк, которые были удалены на этапе предобработки.
3. Из датасета исключены 2 столбца с указанием габаритов товара как не несущие полезной информации для целей последующего анализа.
4. При подготовке количественных признаков для предстоящих расчетов выявлено наличие пустых значений в столбцах "sales" в количестве 2075 строк и "price" в количестве 1 строки. Наличие пустых значений в столбце "sales" свидетельствует об

#### Итоговый вывод по кейсу.

1. Для целей исследования представлены данные по продажам компьютеров на ведущих интернет площадках. Вышеуказанный датасет содержит 4500 наблюдений по 16 признакам.
2. После предварительной подготовки данных на этапах очистки и разведочного анализа для целей исследования сформирован датасет, данные которого приведены к метрической и неметрической шкалам, отдельные признаки кодифицированы. Таким образом, к началу проведения статистического анализа датасет содержит 454 наблюдений и 8 признаков, из которых 2 количественных признака, 2 категориальных и 4 порядковых.
3. В ходе выполнения кейса выполнены неграфический и графический статистический анализ взаимосвязи количественных, порядковых и категориальных признаков, а также их влияние на факторный целевой количественный показатель "Цена".
4. Исследование проведено с применением инструментов языка программирования Python, с подключением и использованием

### Интерпретация данных

#### Рекомендации для Заказчика:



Согласно результатам полученного анализа наиболее устойчивые продажи компьютеров находятся в ценовом диапазоне от 23169 до 74872 руб., а в пятерку лидеров по продажам входят компьютеры с такими наиболее часто встречающимися техническими характеристиками:

- тип процессора: 'Intel Core i5', 'Intel Core i3', 'AMD Ryzen 5', 'Intel Core i7', 'AMD Athlon',
- с количеством ядер от 2 до 6,
- тип памяти DDR4,
- объем оперативной памяти от 8 до 16 ГБ,
- преимущественно с жестким диском SSD,
- объемом дискового пространства от 240 до 480 ГБ,
- видеопроцессорами: 'AMD Radeon Vega 7', 'NVIDIA GeForce RTX 3050', 'NVIDIA GeForce GT 1030', 'Intel HD Graphics 2000', 'NVIDIA GeForce RTX 3060'.

Принимая во внимание вышеизложенные рекомендации, основанные на методах статистического анализа, Заказчик будет последовательно и устойчиво следовать к реализации заявленной бизнес-цели.



## Выработка гипотезы

### Гипотеза исследования:



**Совокупное товарное предложение персональных компьютеров в онлайн-магазинах не имеет статистически значимых отличий по цене и техническим характеристикам.**

### Методы проверки гипотезы:



**Анализ взаимного влияния показателей (технических характеристик) между собой и на факторный целевой показатель (цена) проводился методом постановки и проверки статистических гипотез.**



## Проверка гипотезы



### Методы проверки гипотезы:

Наименование	Критерии оценки
Проверка гипотезы о нормальности	Критерий Шапиро-Уилка (Шапиро-тест)
Анализ влияния категориальных признаков на количественный целевой показатель: - для двух независимых совокупностей - для трех и более независимых совокупностей	Критерий сравнения групп Манна-Уитни Критерий сравнения групп Краскала-Уолиса
Анализ связи количественных и порядковых признаков	Ранговая корреляции Спирмена
Проверка гипотез о значимости коэффициентов корреляции: - для количественных признаков - для порядковых признаков	Статистика Пирсона Статистика Спирмена
Проверка наличия связи категориальных признаков	Критерий «Хи-квадрат»



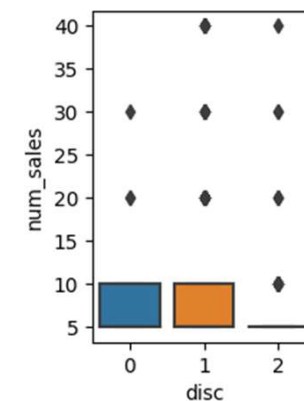
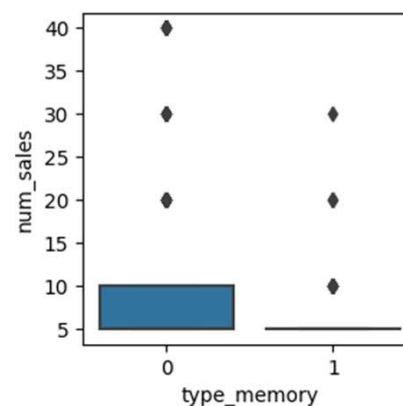
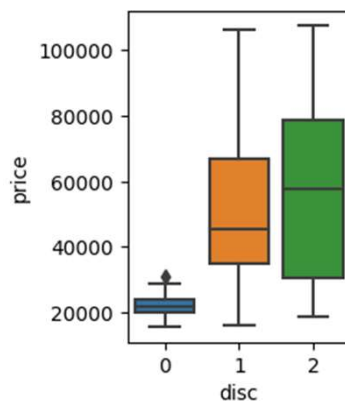
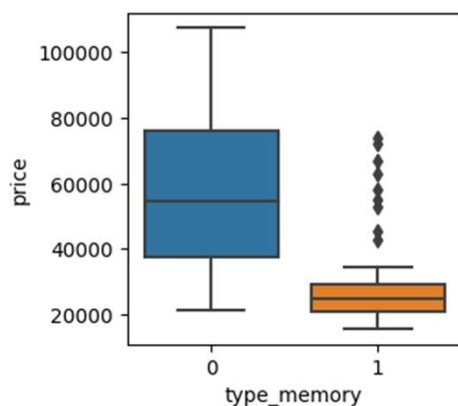
## Описание результатов

### Результаты проверки гипотезы (графическое представление)

#### 1. Анализ влияния категориальных признаков на количественные целевые показатели:



-рvalue исследуемых признаков много меньше уровня значимости (0,05), следовательно, нулевая гипотеза о равенстве отклонена и различия показателей признаны **статистически значимыми**, т.е. установлено влияние категориальных показателей «Тип памяти», «Диск» на факторные целевые показатели «Цена», «Количество продаж».







## Описание результатов

### Результаты проверки гипотезы (графическое представление)



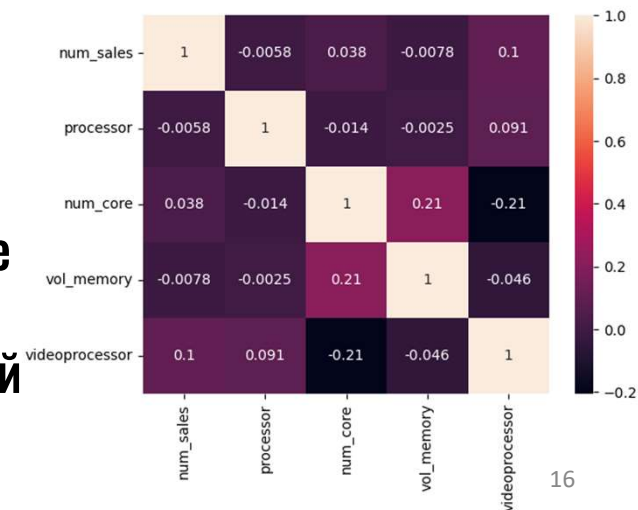
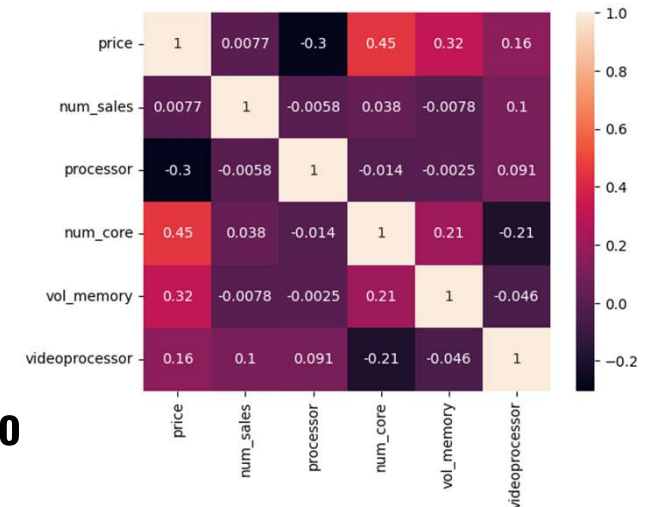
## 2. Анализ связи количественных и порядковых признаков:

### Факторный целевой показатель «Цена»:

-При расчете статистики Спирмена во всех случаях  $p\text{value} < 0,05$ , т.е. **гипотеза о незначимости отклонена**. Значение связи показателя «Цена» и показателей «Тип процессора», «Количество ядер процессора», «Объем оперативной памяти», «Тип видеопроцессора» **статистически значимо**.

### Факторный целевой показатель «Количество продаж»:

-При расчете статистики Спирмена во всех случаях получается  $p\text{value} > 0,05$ , т.е. **гипотеза о незначимости не отклонена**. Значение связи показателя «Количество продаж» и показателей «Тип процессора», «Количество ядер процессора», «Объем оперативной памяти», «Тип видеопроцессора» **статистически незначимо**.







## Описание результатов

### Результаты проверки гипотезы (графическое представление)

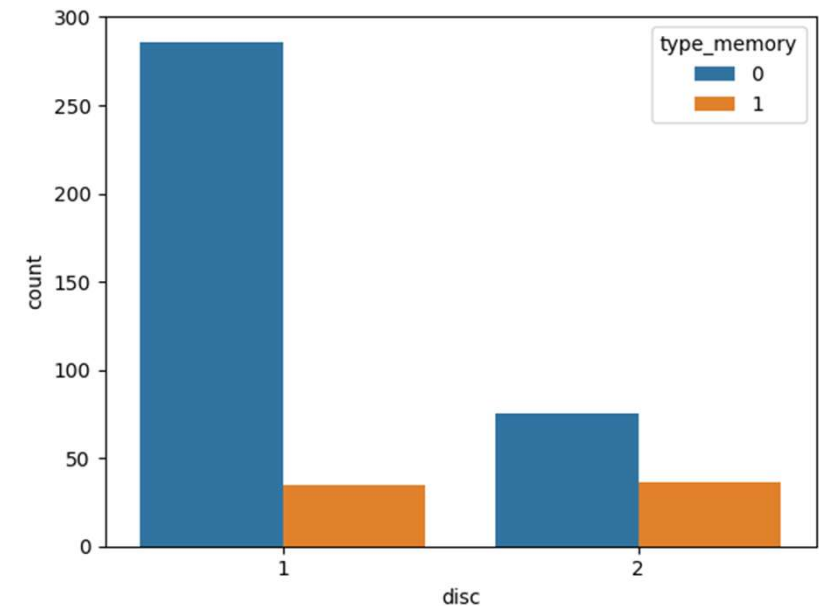
### 3. Проверка наличия связи категориальных признаков

Получено  $p$ -value ниже уровня значимости (0,05), следовательно, **имеется основание отклонить**



**нулевую гипотезу** о том, что частоты примерно равны и, следует вывод о том, что **связь между показателями «Диск», «Тип памяти» существует.**

Пропорция между типами дисков и памяти компьютеров неодинакова.



Согласно Шапиро-тесту все исследуемые выборки противоречат нормальному закону распределения.

**Полученные результаты взаимозависимости показателей с целевым показателем «Цена» :**

Наименование	Критерии оценки	Значение pvalue	Гипотеза $H_0$	Статистическая значимость	Влияние/связь
Анализ влияния категориальных признаков на количественный целевой показатель: - для двух независимых совокупностей - для трех и более независимых совокупностей	Манна-Уитни Краскала-Уолиса	<0,05	Отклонена	Различия статистически значимы	Установлено влияние категориальных показателей "Тип памяти", «Диск» на факторный целевой показатель «Цена»
Анализ связи количественных и порядковых признаков, Проверка гипотез о значимости коэффициентов корреляции: - для количественных признаков - для порядковых признаков	Ранговая корреляции Спирмена, Критерии Пирсона, Спирмена	<0,05	Отклонена	Различия статистически значимы	Установлена связь между целевым показателем «Цена» и «Количество продаж», «Тип процессора», «Количество ядер процессора», «Объем оперативной памяти», «Видеопроцессор»
Проверка наличия связи категориальных признаков	«Хи-квадрат»	<0,05	Отклонена	Различия статистически значимы	Установлена связь между категориальными показателями «Диск» и «Тип памяти»

**Согласно Шапиро-тесту  
все исследуемые выборки противоречат нормальному закону распределения.**

**Полученные результаты взаимозависимости показателей с целевым показателем «Продажи» :**

Наименование	Критерии оценки	Значение pvalue	Гипотеза $H_0$	Статистическая значимость	Влияние/связь
Анализ влияния категориальных признаков на количественный целевой показатель: - для двух независимых совокупностей - для трех и более независимых совокупностей	Манна-Уитни Краскала-Уолиса	<0,05	Отклонена	Различия статистически значимы	Установлено незначительное влияние категориальных показателей "Тип памяти", «Диск» на факторный целевой показатель «Количество продаж»
Анализ связи количественных и порядковых признаков, Проверка гипотез о значимости коэффициентов корреляции: - для количественных признаков - для порядковых признаков	Ранговая корреляция Спирмена, Критерии Пирсона, Спирмена	>0,05	Не отклонена	Различия статистически не значимы	Отсутствует связь между целевым показателем «Количество продаж» и «Цена», «Тип процессора», «Количество ядер процессора», «Объем оперативной памяти», «Видеопроцессор»

Рекомендации для Заказчика:

**Цена компьютера, связана с техническими характеристиками компьютеров и наиболее чувствительна к таким как:**



### Рекомендации для Заказчика:

**Не имеют связи с количеством продаж, т.е. не оказывают никакого влияния на популярность и, как следствие, объем реализации компьютеров следующие технические характеристики:**



### Рекомендации для Заказчика:



Согласно результатам полученного анализа наиболее устойчивые продажи компьютеров находятся в ценовом диапазоне от 23169 до 74872 руб., а в пятерку лидеров по продажам входят компьютеры с такими наиболее часто встречающимися техническими характеристиками:



- тип процессора: 'Intel Core i5', 'Intel Core i3', 'AMD Ryzen 5', 'Intel Core i7', 'AMD Athlon',
- с количеством ядер от 2 до 6,
- тип памяти DDR4,
- объем оперативной памяти от 8 до 16 ГБ,
- преимущественно с жестким диском SSD,
- объемом дискового пространства от 240 до 480 ГБ,
- видеопроцессорами: 'AMD Radeon Vega 7', 'NVIDIA GeForce RTX 3050', 'NVIDIA GeForce GT 1030', 'Intel HD Graphics 2000', 'NVIDIA GeForce RTX 3060'.

Принимая во внимание вышеизложенные рекомендации, основанные на методах статистического анализа, Заказчик будет последовательно и устойчиво следовать к реализации заявленной бизнес-цели.



**Спасибо  
за внимание!**

**Горохов Дмитрий Евгеньевич**

**Поток Но ДА-102 группа 1**