Домашнее задание 3

Муханько Артем

25.03.2024

Содержание

Введение

Использовался датасет, который представляет собой информацию о клиентах и их взаимодействии с компанией, которая занимается доставкой еды.

Описание набора данных:

Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5th campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on *gold* products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase

Table 1: Meta-data table

В ходе работы будет проведена подготовка данных к исследованию, разведочный анализ данных и визуализация, выдвижение и статистическая проверка гипотез.

1)Материалы и методы

1.1) Подготовка данных

В датасете 2205 строк и 39 столбцов данных.

```
df.shape
```

```
(2205, 39)
```

Пропущенные значения отсутствуют.

```
df.isnull().any().any()
```

```
False
```

Исходно категориальные переменные были закодированы с использованием one hot encoding. Это можно увидеть на примере переменной об образовании.

```
df.iloc[:,31:36].head()
```

	education_2n Cycle	education_Basic	education_Graduation	education_Master	education_PhD
0	0	0	1	0	0
1	0	0	1	0	0
2	0	0	1	0	0
3	0	0	1	0	0
4	0	0	0	0	1

Признаки:

```
df.columns
```

```
Index(['Income', 'Kidhome', 'Teenhome', 'Days since last purchase', 'MntWines',
       'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts',
       'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases',
       'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
       'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1',
       'AcceptedCmp2', 'Complain', 'Z_CostContact', 'Z_Revenue', 'Response',
       'Age', 'Customer_Days', 'marital_Divorced', 'marital_Married',
       'marital_Single', 'marital_Together', 'marital_Widow',
       'education_2n Cycle', 'education_Basic', 'education_Graduation',
       'education_Master', 'education_PhD', 'MntTotal', 'MntRegularProds',
       'AcceptedCmpOverall'],
      dtype='object')
```

Можно создать несколько отдельных полезных переменных, которые позволят лучше понять набор данных и раскрыть ценную информацию.

Я выделил отдельно и создал переменные:

“Spending” – итоговая сумма, потраченная по всем 6 категориям продуктов.

“Martial_situation” – сгруппировал семейное положение в 2 категории (одинокий, в паре)

“Has_child” – имеет ли клиент детей (1 – да, 2 - нет)

“Educationnal_years” – общее количество лет обучения человека в соответствии с дипломом.

Также в новый набор данных включил возраст.

Будем анализировать полученный датасет.

```
df.head()
```

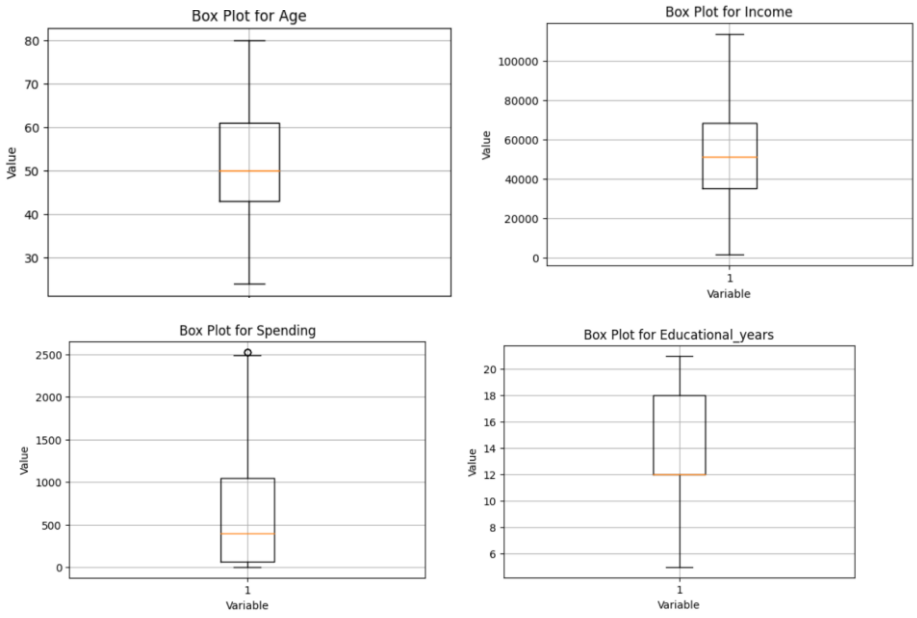
	Age	Income	Spending	Marital_Situation	Has_child	Educational_years	Education
0	63	58138.0	1617	Alone	No	12	Graduation
1	66	46344.0	27	Alone	Yes	12	Graduation
2	55	71613.0	776	In couple	No	12	Graduation
3	36	26646.0	53	In couple	Yes	12	Graduation
4	39	58293.0	422	In couple	Yes	21	PhD

1.2) Разведочный анализ данных

Общая статистика по признакам:

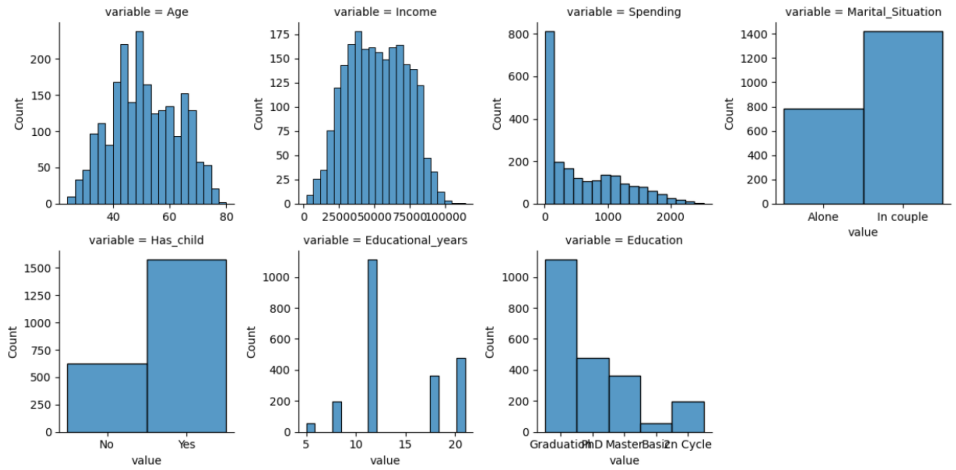
	Age	Income	Spending	Educational_years
count	2205.000000	2205.000000	2205.000000	2205.000000
mean	51.095692	51622.094785	606.821769	14.402721
std	11.705801	20713.063826	601.675284	4.505712
min	24.000000	1730.000000	5.000000	5.000000
25%	43.000000	35196.000000	69.000000	12.000000
50%	50.000000	51287.000000	397.000000	12.000000
75%	61.000000	68281.000000	1047.000000	18.000000
max	80.000000	113734.000000	2525.000000	21.000000

Построим boxplot для каждого числового признака.

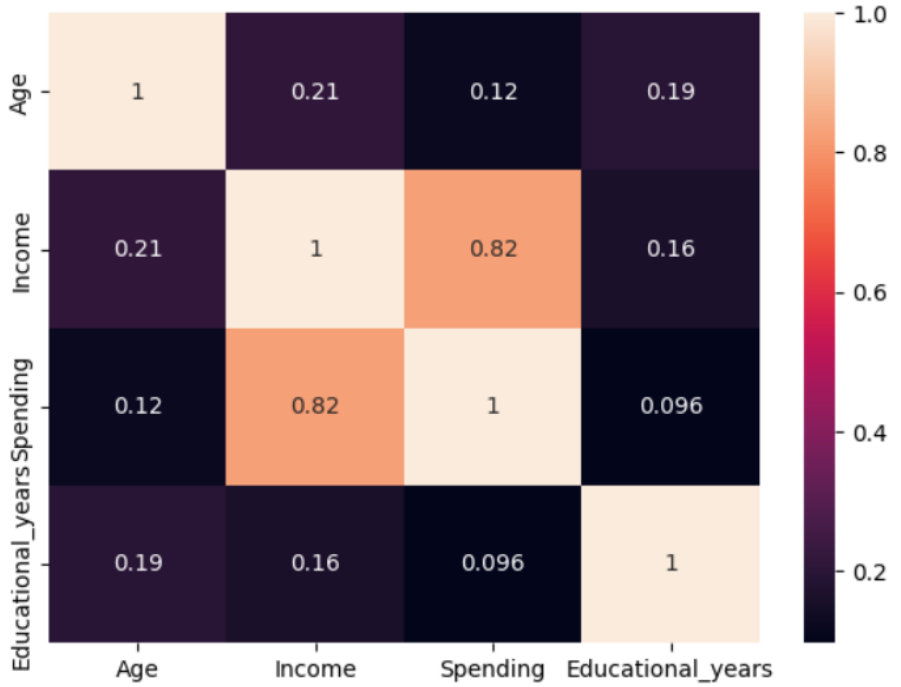


Можно обратить внимание, что отсутствуют значительные выбросы.

Гистограммы распределений:

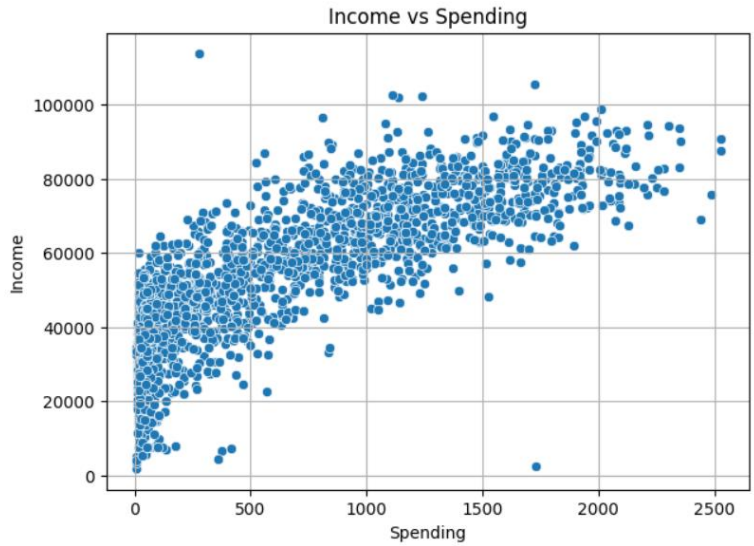


Построим корреляционную матрицу между признаками:



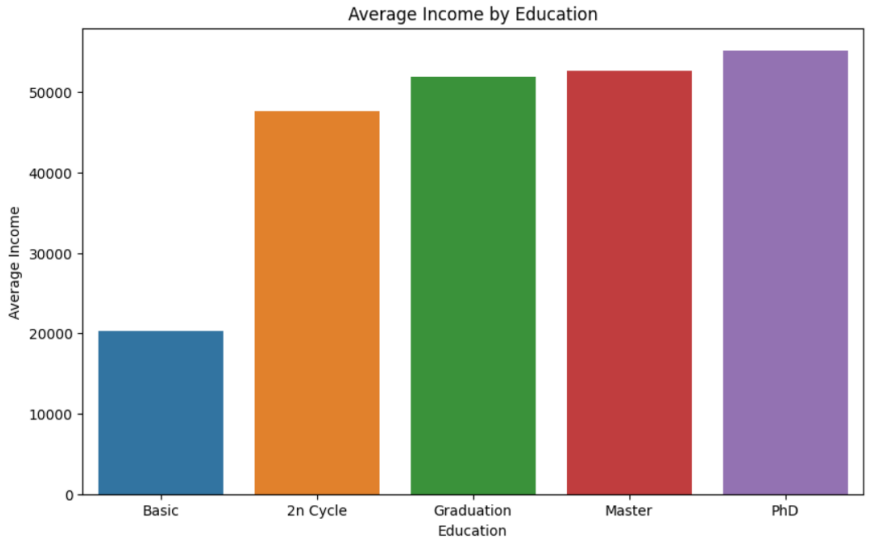
Наибольшая корреляция между итоговой потраченной суммой и доходом клиента.

Построю диаграмму рассеивания. Ось X – итоговая сумма, потраченная по всем категориям, Y – доход клиента.



Можно будет проверить статистическим тестом является ли корреляция между годовым доходом и суммой расходов статистически значимой.

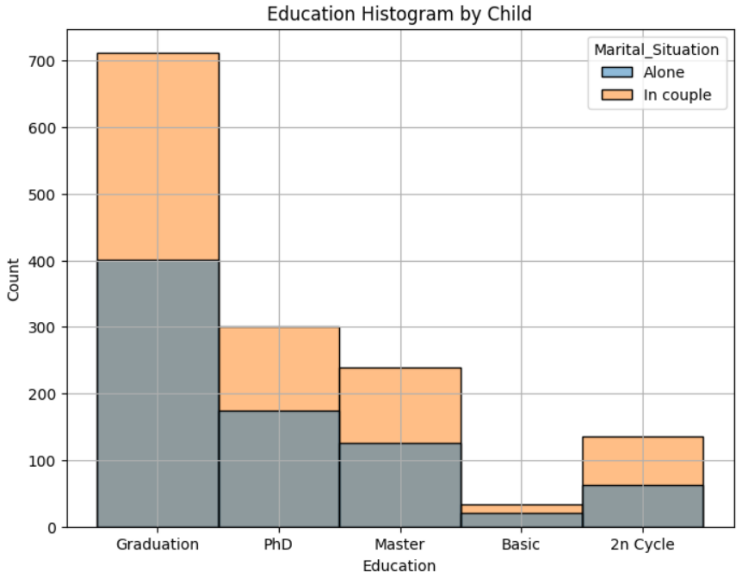
Постим гистограмму распределения уровня образования и среднего дохода.



Чем выше уровень образования, тем больше средняя заработная плата.

Можно выдвинуть гипотезу: отличается ли статистически средняя заработная плата людей с уровнем образования Master от PhD.

Также можно рассмотреть гистограмму уровня образования и семейного положения.



По данным можно заметить, что распределение по уровню образования похоже для 2 групп населения.

Статистическим тестом можно будет проверить есть ли связь между образованием и семейным положением.

1.3) Формирование вопросов

По результатам разведочного анализа данных были выделены 3 предположения, которые можно проверить статистическими тестами.

1. Корреляция между годовым доходом и суммой расходов статистически значима?

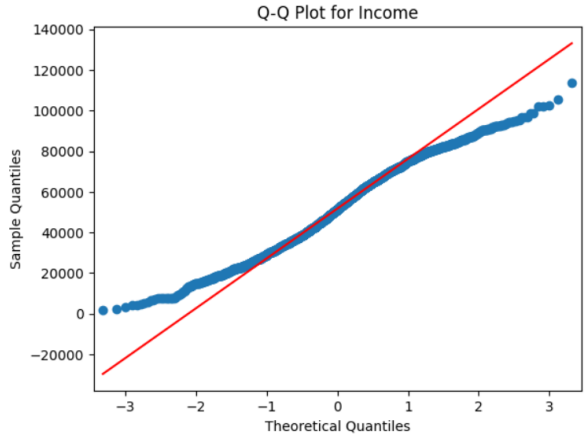
2. Отличается ли статистически средняя заработная плата людей с уровнем образования Master от людей с уровнем образования PhD?

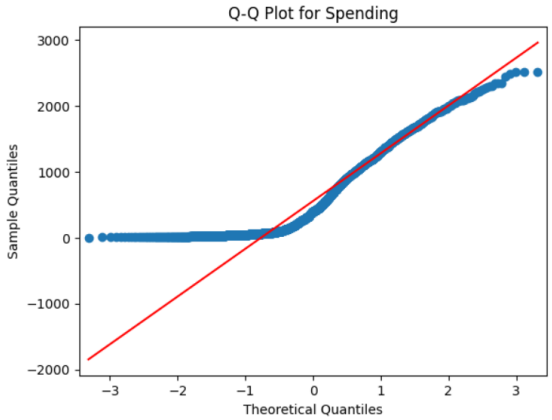
3. Есть ли связь между уровнем образования и семейным положением?

2) Результаты (статистическая проверка гипотез)

2.1) 1 вопрос

Необходимо проверить распределения на нормальность. Построю графики qq-plot.





По графикам можно увидеть, что точки отклоняются от предполагаемого распределения и распределения не являются нормальными.

Буду использовать тест ранговой корреляции Спирмена.

Сформулируем гипотезы:
H0: нет монотонной связи между доходом и суммой расходов

H1: есть монотонная связь между доходом и суммой расходов

Уровень значимости используется как 5%.

```
from scipy.stats import spearmanr
cor, pvalue = spearmanr(df[['Spending']], df[['Income']])
print("Spearman correlation test : correlation coefficient : %.4f, pval: %.4f" % (cor, pvalue))
```

```
Spearman correlation test : correlation coefficient : 0.8604, pval: 0.0000
```

Так как p-value < 0.05 можно отвергнуть нулевую гипотезу. Коэффициент корреляции 0.86 говорит о довольно сильной положительной корреляции.

2.2) 2 вопрос

Проверка на нормальность

Для проверки на нормальность можно использовать и другой способ, критерий Шапиро-Уилка.

```
Phd_graduate=Diploma[df['Education']=='PhD']
Master_graduate=Diploma[df['Education']=='Master']

stat, p = shapiro(Phd_graduate.Income)
stat1, p1 = shapiro(Master_graduate.Income)
print('Statistics=%.3f, p=%.3f' % (stat, p))
print('Statistics=%.3f, p=%.3f' % (stat1, p1))
```

```
Statistics=0.991, p=0.005
Statistics=0.988, p=0.003
```

p-value < 0.05, отвергается нулевая гипотеза о нормальности распределений.

Так как распределения не являются нормальными буду использовать непараметрический тест: критерий Манна-Уитни(U -test).

H0: средние ранги 2 групп равны

H1: средние ранги 2 групп не равны

```
from scipy.stats import mannwhitneyu
print('PhD: median = %.0f stdv = %.1f' % (np.median(Phd_graduate.Income), np.std(Phd_graduate.Income)))
print('Master: median = %.0f stdv = %.1f' % (np.median(Master_graduate.Income), np.std(Master_graduate.Income)))

stat, p =  mannwhitneyu(Phd_graduate.Income, Master_graduate.Income)
print(stat,p)
```

```
PhD: median = 55005 stdv = 18343.9
Master: median = 50920 stdv = 19392.6
93667.5 0.04350343114442508
```

p-value < 0.05, отвергаем нулевую гипотезу.

Средняя зарплата людей с уровнем образования PhD отличается от средней зарплаты людей с уровнем образования Master.

2.3) 3 вопрос

Буду использовать тест независимости Хи-квадрат, так как оба признака категориальные.

H0: уровень образования и семейное положение независимы

H1: уровень образования и семейное положение не независимы

Таблица с наблюдаемыми значениями:

```
crosstab = pd.crosstab(df["Education"], df["Marital_Situation"])
crosstab
```

Marital_Situation	Alone	In couple
Education		
2n Cycle	62	136
Basic	20	34
Graduation	401	712
Master	125	239
PhD	175	301

Можно получить таблицу с ожидаемыми значениями

```
from scipy.stats import chi2_contingency, chi2

stat, p, dof, expected = chi2_contingency(crosstab)
print('Degress of freedom = %d' % dof)
```

```
Degress of freedom = 4
```

```
print(expected)
```

```
[[ 70.31020408 127.68979592]
 [ 19.1755102   34.8244898 ]
 [395.22857143 717.77142857]
 [129.25714286 234.74285714]
 [169.02857143 306.97142857]]
```

Число степеней свободы – 4

```
print("stat=",round(stat,3), "p-value=", round(p,3))
```

```
stat= 2.253 p-value= 0.689
```

В результате получили p-value > 0.05.

Нет оснований отклонить нулевую гипотезу. Можем предположить, что семейное положение не зависит от наличия диплома с уровнем достоверности 95%.

Выводы

Была проведена предварительная обработка данных, анализ, визуализация и поиск зависимостей. С использованием статистических тестов были получены ответы на три ключевых вопроса в области бизнеса.

Используя тест ранговой корреляции Спирмена была установлена статистически значимая корреляция между доходом клиента и общими тратами в сервисе.

При помощи критерия Манна-Уитни было выявлено, что средний доход клиентов с уровнем образования PhD отличается от среднего дохода клиентов с уровнем образования Master с уровнем достоверности 95%.

Наконец, результаты теста Хи-квадрат не позволили сделать вывод о наличии зависимости между семейным положением клиента и его уровнем образования.