ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ст. преподаватель |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №2 |
| ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4216 |  |  |  | С.Д. Комолова |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2024

**Цель лабораторной работы:** изучение связи между признаками двумерного набора данных, визуализация данных.

**Индивидуальное задание:**

Вариант 6:

Задание 1: Использовать seaborn. По группировке - количество клиентов по количеству детей и по каждому семейному статусу (marital\_status) построить гистограмму

Задание 2: Использовать pandas и plot. По сводной таблице (pivot\_table) - отобразить минимальный доход по образованию. Отобразить маркеры в виде круга красного цвета размером 18.

Задание 3: Использовать matplotlib. Отфильтровать клиентов по year\_birth больше 1980. Построить круговую диаграмму, которая отображает процент клиентов каждого семейного положения.

**Ход работы:**

Ссылка на репозиторий github: <https://github.com/sufferix/data-analysis>

В данной лабораторной работе необходимо составить различные диаграммы на основе нового файла `clients2.csv`. Была определено, что данный файл идентичен с файлом в лабораторной работе №1, поэтому были совершены такие же действия для предварительной обработки данных: были удалены лишние объекты, такие как дубликаты и строки с нулевыми значениями, и изменены названия столбцов и тип данных некоторых из них. Также были испортированы библиотеки `matplotlib` и `seaborn`, которые понадобятся для создания диаграмм

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv('clients2.csv',sep=';')

df = df.rename(columns= {'Kidhome':'Children'})

df = df.rename(columns= {'NumDealsPurchases':'Purchases'})

df = df.rename(columns= {'Year\_Birth': 'Birth\_Year'})

df = df.rename(columns= {'Dt\_Customer': 'Reg\_Date'})

df = df.dropna()

df['Children'] = df['Children'].astype(int)

df['Purchases'] = df['Purchases'].astype(int)

df['Reg\_Date'] = pd.to\_datetime(df['Reg\_Date'], format='%d.%m.%Y')

dubles = ['SINGL','Alone']

df['Marital\_Status'] = df['Marital\_Status'].replace(dubles,'Single')

df['Marital\_Status'] = df['Marital\_Status'].replace('MARRIED','Married')

df = df.drop\_duplicates()

**Построение диаграмм:**

Для того, чтобы понять, на основе каких данных можно составлять диаграммы, можно вывести маленькую часть датасета с помощью метода `df.head()` (см. Рисунок 1)

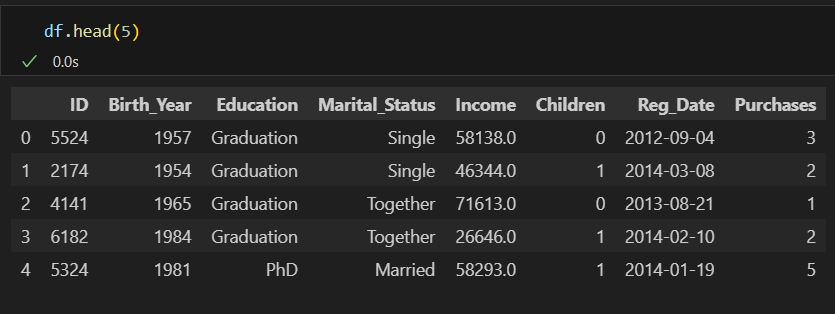


Рисунок 1 – Метод df.head()

В данном датасете большинство столбцов имеет числовые значения и только столбцы `Education` и `Marital\_Status` содержат в себе текстовые значения. На основе столбцов с числовыми данными будут построены следующие диаграммы: `scatter`и `heatmap`

Диаграмму рассеяния можно создать с помощью метода `scatter`библиотеки `matplotlib`, которая сокращена до `plt`. Внутри данного метода необходимо поместить столбцы, на основе которых будет создана диаграмма, также метод может содержать другие параметры, к примеру, цвет и размер точек, а также цветовые карты.

Была построена диаграмма рассеяния на основе столбцов `Income` и `Birth\_Year`, так как эти данные могут визуализировать полезную для анализа информацию (см. Рисунок 2)

plt.scatter(df['Income'],df['Birth\_Year'],alpha=0.5)

plt.title('Диаграмма рассеяния возраста от дохода')

plt.show()



Рисунок 2 – Диаграмма рассеяния возраста от дохода

По данной диаграмме можно увидеть, что в датасете есть несколько выбивающихся точек, а остальные значения находятся близко друг к другу.

Также с помощью метода `pairplot` библиотеки `seaborn` была выведена матрица диаграмм, где показана взаимосвязь всех числовых столбцов. Были выведены все столбцы, кроме столбца `ID`, он был удалён с помощью метода `drop()` (см. Рисунок 3).

df\_corr = df.drop('ID',axis=1)

sns.pairplot(df\_corr)

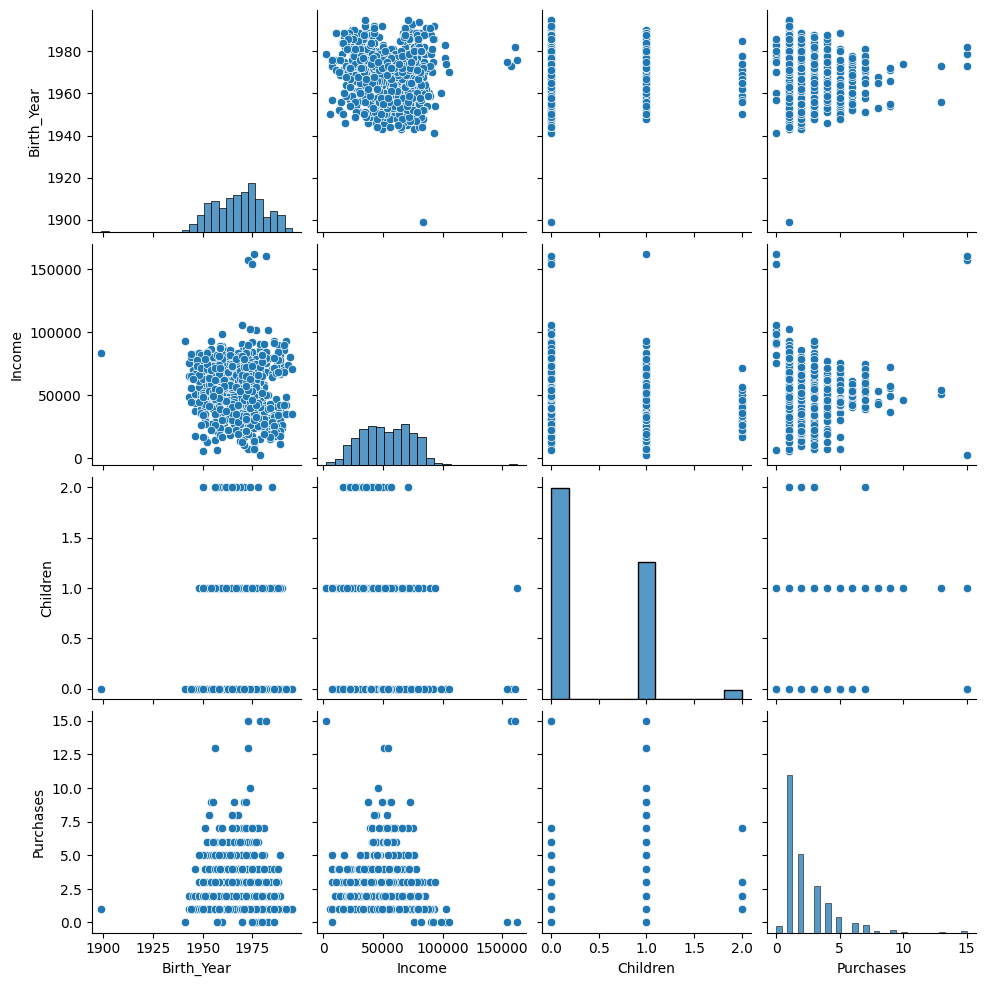


Рисунок 3 – Матрица диаграмм рассеяния

По данной матрице диаграмм можно увидеть, что линейной зависимости столбцов друг с другом практически нет. Графики по диагонали показывают, каких значений больше всего по всему датасету. К примеру, можно увидеть, что у большинства клиентов нет детей, а также большинство клиентов сделали только 1 заказ

Была составлена матрица ковариаций с помощью метода `cov()` (см. Рисунок 4). Ковариация показывает, как переменные взаимодействуют друг с другом. Например, если ковариация положительная, переменные имеют тенденцию расти вместе, если отрицательная — одна переменная увеличивается, когда другая уменьшается.

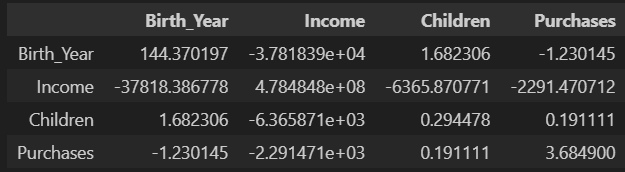


Рисунок 4 – Матрица ковариаций

Наиболее сильная связь наблюдается между возрастом и доходом, а также между доходом и количеством детей — обе эти зависимости отрицательные. Это говорит о том, что более молодые клиенты зарабатывают меньше, а также клиенты имеют больший доход, если у них меньше детей

Далее была выведена таблица корреляции всех числовых столбцов с помощью метода `df.corr()` (см. Рисунок 5). Корреляция показывает уровень зависимости числовых столбцов друг от друга.

correl = df\_corr.corr(numeric\_only= True)

display(correl)

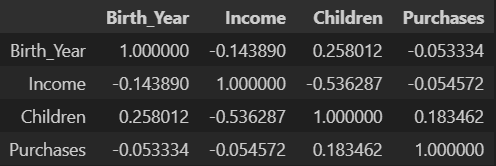


Рисунок 5 – Матрица корреляций

По данной таблице можно понять, что взаимосвязь столбцов друг с другом практически нулевая, то есть корреляция между ними не наблюдается. Также была построена диаграмма `heatmap`, которая показывает уровень корреляции между столбцами графически с помощью метода `heatmap` библиотеки `seaborn`, сокращённо `sns` (см. Рисунок 6).

sns.heatmap(correl,annot=True, cmap='coolwarm')

plt.title('Тепловая карта корреляций')

plt.show()

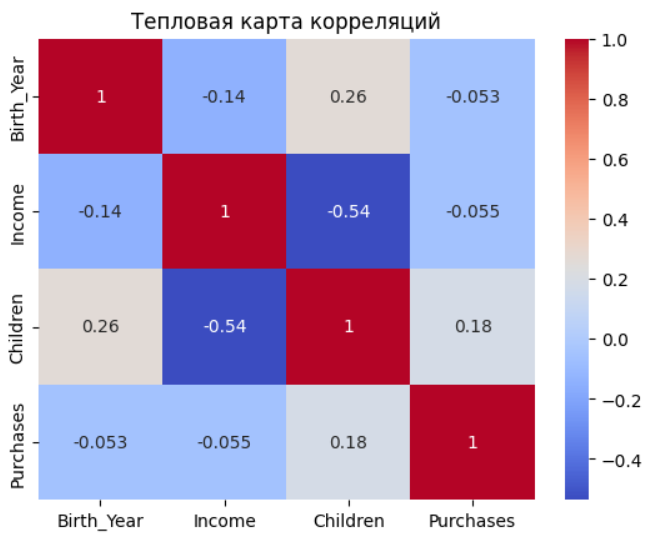


Рисунок 6 – Тепловая карта корреляций

Значения между всеми столбцами имеют низкую корреляцию, кроме корреляции столбцов `Income` и `Children`

**Задание 1:**

Для того чтобы сделать гистограмму, необходимо сгруппировать датасет по количеству детей и по семейному положению и создать из этого датафрейм с помощью метода `pd.DataFrame()` (см. Рисунок 7).

un = df.groupby(['Marital\_Status','Children'])

df\_hist = un['ID'].count()

hist = pd.DataFrame(df\_hist)

display(hist)



Рисунок 7 – Группировка по количеству детей и по семейному положению

Из получившейся таблицы можно сделать гистограмму. Создаётся гистограмма с помощью метода `histplot` библиотеки `seaborn` (см. Рисунок 8). Данный метод имеет множество параметров, были использованы параметры `data` - данные, по которой будет создаваться гистограмма, `x` - что будет находиться на оси абсциссы, `weights` - по какому значению будет строиться гистограмма, `hue` - какие значения будут окрашиваться в разные цвета, `multiple` - параметр, отвечающий за расположение различных показателей, `kde` - линия тренда, принимает бинарное значение.

sns.histplot(data=hist,x="Marital\_Status",weights="ID",hue="Children", multiple="dodge",kde=False)

plt.xlabel('Семейное положение')

plt.ylabel('Количество клиентов')

plt.title('Диаграмма количества клиентов по семейному статусу и числу детей')

plt.show()

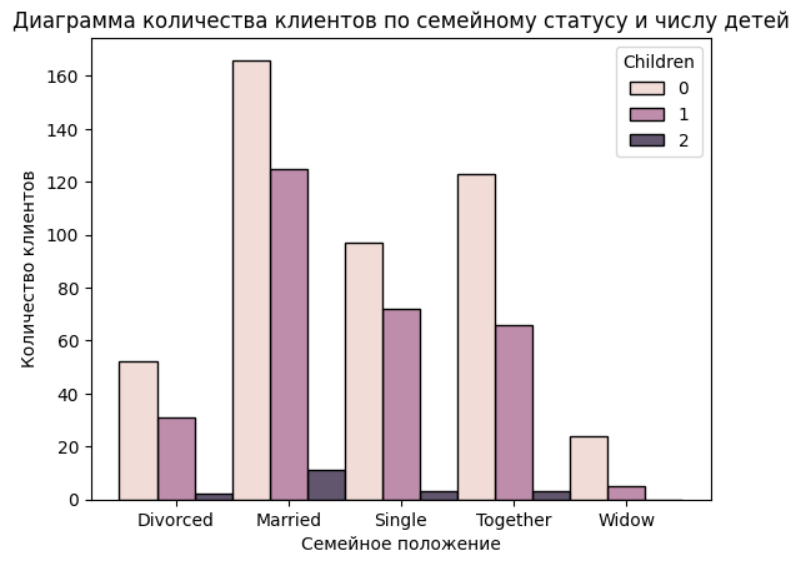


Рисунок 8 – Гистограмма количества клиентов по семейному статусу и числу детей

По данной гистограмме можно сделать следующие выводы:

* Больше всего клиентов не имеют детей
* Замужние/женатые клиенты составляют крупнейшую группу
* Меньше всего клиентов с двумя детьми
* Категории `Single` и `Together` имеют схожие распределения
* Вдовцы составляют самую маленькую группу клиентов, ни у одного вдовца нет 2 детей

**Задание 2:**

Была создана сводная таблица минимального дохода по уровню образования с помощю метода `pivot.table()` (см. Рисунок 9).

pivot\_min = df.pivot\_table(values="Income", index="Education", aggfunc="min")

display(pivot\_min)

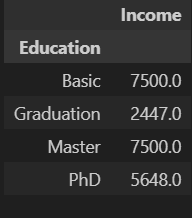


Рисунок 9 – Сводная таблица минимального дохода по образованию

Далее была создана диаграмма рассеяния по данной таблице с помощью метода `plot` библиотеки `pandas`. На вход были приняты такие параметры как столбцы `Income` и `Education`, вид диаграммы `scatter` и изменение размера и цвета точек на диаграмме (см. Рисунок 10).

pivot\_min = pivot\_min.reset\_index()

sctr = pivot\_min.plot(x='Education', y='Income', kind='scatter',s=180,c=['red'])

plt.xlabel('Уровень образования')

plt.ylabel('Минимальный доход')

plt.title('График минимального дохода по образованию')

plt.grid(True)

plt.show(sctr)

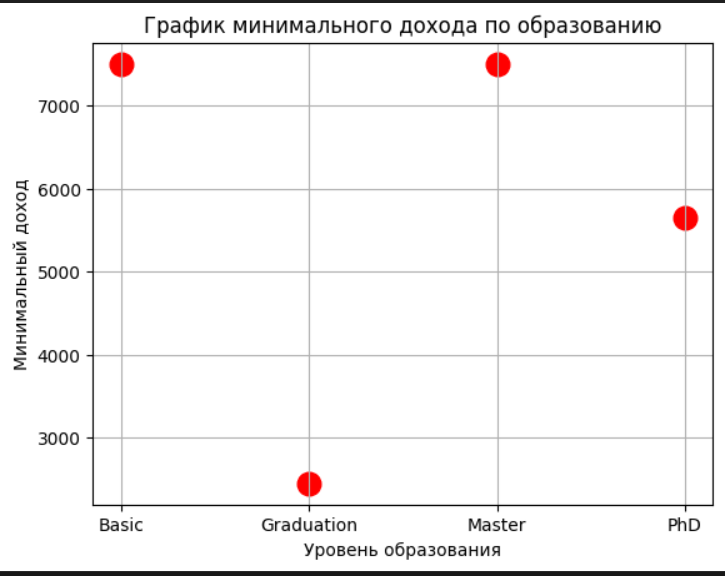


Рисунок 10 – График минималного дохода по образованию

Выводы по данной диаграмме можно сделать следующие:

* Наивысший минимальный доход для уровня образования `Basic` и `Master`
* Наименьший минимальный доход для уровня `Graduation`
* Минимальный доход для `PhD` ниже, чем у `Basic` и `Master`

**Задание 3:**

Для того, чтобы создать круговую диаграмму, датасет был отсортирован по дате рождения больше 1980 года. Была создана диаграмма с помощю метода `pie` библиотеки `matplotlib` (см. Рисунок 11). В параметрах были занесены столбцы, по которым будет создаваться диаграмма, а также отображение процентного соотношения

df\_filter = df.query('Birth\_Year > 1980')

filter\_pivot = df\_filter.pivot\_table(values='ID', index='Marital\_Status', aggfunc="count")

filter\_pivot = filter\_pivot.reset\_index()

plt.pie(filter\_pivot['ID'],labels=filter\_pivot['Marital\_Status'],autopct='%1.1f%%')

plt.title('Распределение клиентов по семейному положению')

plt.legend(bbox\_to\_anchor=(1, 1))

plt.show()

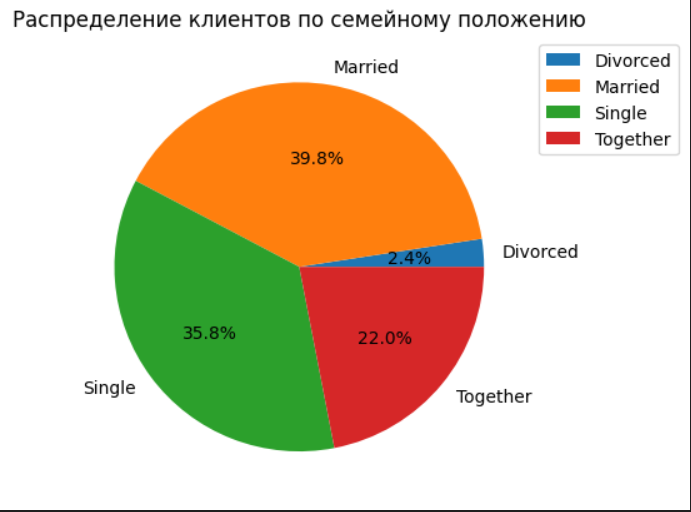


Рисунок 11 – Распределение клиентов по семйному положению

Из диаграммы можно сделать вывод, что большинство клиентов младше 1980 года находятся в браке или не состоят в официальных отношениях. Люди, живущие вместе, составляют около пятой части клиентов, а разведенные составляют небольшую часть клиентской базы.

Вывод:

В данной лабораторной работе были созданы несколько графиков по заданному датасету: гистограммы, диаграммы рассеяния, heatmap, круговая диаграмма. Были использованы библиотеки pandas, matplotlib и seaborn для создания всех диаграмм. Также были рассчитаны корреляция и ковариация данных.

В ходе выполнения лабораторной работы было выяснено, что данные в данном датасете практически никак не взаимосвязаны друг с другом и имеют слабую корреляцию кроме зависимости дохода от количества детей: чем меньше детей, тем больше достаток, а также коэффициент ковариации показал сильную отрицательную взаимосвязь дохода с годом рождения.

По графикам, созданным в индивидуальном варианте, можно сделать вывод, что большая часть клиентов находится в браке, а также у большинства клиентов по каждому семейному положению нет детей. Больший заработок имеют люди, которые окончили школу и магистратуру.