Министерство цифрового развития, связи и массовых коммуникаций
Российской Федерации
Ордена Трудового Красного Знамени федеральное государственное
бюджетное образовательное учреждение высшего образования
«Московский технический университет связи и информатики»
Кафедра «МКиИТ»

Лабораторная работа №1(часть 2)

по дисциплине «Data mining»

Москва 2023

О задании

Задание состоит из двух разделов, посвященных работе с табличными данными с помощью библиотеки pandas и визуализации с помощью matplotlib. В каждом разделе вам предлагается выполнить несколько заданий.

Задание направлено на освоение jupyter notebook (будет использоваться в дальнейших заданиях), библиотекам pandas и matplotlib.

0. Введение

Сейчас мы находимся в jupyter-ноутбуке (или ipython-ноутбуке). Это удобная среда для написания кода, проведения экспериментов, изучения данных, построения визуализаций и других нужд, не связанных с написаем production-кода.

Ноутбук состоит из ячеек, каждая из которых может быть либо ячейкой с кодом, либо ячейкой с текстом размеченным и неразмеченным. Текст поддерживает markdown-разметку и формулы в Latex.

Для работы с содержимым ячейки используется *режим редактирования* (*Edit mode*, включается нажатием клавиши **Enter** после выбора ячейки), а для навигации между ячейками искользуется *командный режим* (*Command mode*, включается нажатием клавиши **Esc**). Тип ячейки можно задать в командном режиме либо с помощью горячих клавиш (**y** to code, **m** to markdown, **r** to edit raw text), либо в меню *Cell -> Cell type*.

После заполнения ячейки нужно нажать *Shift + Enter*, эта команда обработает содержимое ячейки: проинтерпретирует код или сверстает размеченный текст.

In [ ]:
```
# ячейка с кодом, при выполнении которой появится output
2 + 2
```

А это ***ячейка с текстом***.

Ячейка с неразмеченым текстом.

Попробуйте создать свои ячейки, написать какой-нибудь код и текст какой-нибудь формулой.

In [ ]:
```
# your code
```

Здесь находится ~~не~~большая заметка о используемом языке разметки Markdown. Он позволяет:

1. Составлять упорядоченные списки
2. #Делать ##заголовки ###разного уровня
3. Выделять *текст* ~~при~~ **необходимости**
4. Добавлять ссылки

- Составлять неупорядоченные списки

Делать вставки с помощью LaTex:

$$\begin{cases} x = 16\sin^3(t) \\ y = 13\cos(t) - 5\cos(2t) - 2\cos(3t) - \cos(4t) \\ t \in [0, 2\pi] \end{cases}$$

1. Табличные данные и Pandas

Pandas — удобная библиотека для работы с табличными данными в Python, если данных не слишком много и они помещаются в оперативную память вашего компьютера. Несмотря на неэффективность реализации и некоторые проблемы, библиотека стала стандартом в анализе данных. С этой библиотекой мы сейчас и познакомимся.

Основной объект в pandas это DataFrame, представляющий собой таблицу с именованными колонками различных типов, индексом (может быть многоуровневым). DataFrame можно создавать, считывая таблицу из файла или задавая вручную из других объектов.

В этой части потребуется выполнить несколько небольших заданий. Можно пойти двумя путями: сначала изучить материалы, а потом приступить к заданиям, или же разбираться "по ходу". Выбирайте сами.

Материалы:

1. Pandas за 10 минут из официального руководства
2. Документация (стоит обращаться, если не понятно, как вызывать конкретный метод)
3. Примеры использования функционала

Многие из заданий можно выполнить несколькими способами. Не существуют единственно верного, но попробуйте максимально задействовать арсенал pandas и ориентируйтесь на простоту и понятность вашего кода. Мы не будем подсказывать, что нужно использовать для решения конкретной задачи, попробуйте находить необходимый функционал сами (название метода чаще всего очевидно). В помощь вам документация, поиск и stackoverflow.

In [81]:
```
%matplotlib inline
import pandas as pd
```

Данные можно скачать отсюда.

1. Откройте файл с таблицей (не забудьте про её формат). Выведите последние 10 строк.

Посмотрите на данные и скажите, что они из себя представляют, сколько в таблице строк, какие столбцы?

In [82]:
```
data = pd.read_csv('./data.csv')
data.tail(10)
```

Out[82]:

	order_id	quantity	item_name	choice_description	item_price
4612	1831	1	Carnitas Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Rice,...	$9.25

	order_id	quantity	item_name	choice_description	item_price
4613	1831	1	Chips	NaN	$2.15
4614	1831	1	Bottled Water	NaN	$1.50
4615	1832	1	Chicken Soft Tacos	[Fresh Tomato Salsa, [Rice, Cheese, Sour Cream]]	$8.75
4616	1832	1	Chips and Guacamole	NaN	$4.45
4617	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Sour ...	$11.75
4618	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Sour Cream, Cheese...	$11.75
4619	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$11.25
4620	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	$8.75
4621	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$8.75

2. [0.25 баллов] Ответьте на вопросы:

1. Сколько заказов попало в выборку?
2. Сколько уникальных категорий товара было куплено? (item_name)

In [83]:
```
data['order_id'].tail(10).nunique()
```

Out[83]: 4

In [84]:
```
data['item_name'].tail(10).nunique()
```

Out[84]: 7

3. [0.25 баллов] Есть ли в данных пропуски? В каких колонках?

In [85]:
```
data.isnull().sum()
```

Out[85]:
```
order_id                0
quantity                0
item_name               0
choice_description    1246
item_price              0
dtype: int64
```

Заполните пропуски пустой строкой для строковых колонок и нулём для числовых.

In [86]:
```
data.select_dtypes(include=['int64']).fillna(0)
data.select_dtypes(include=['object']).fillna('')
```

Out[86]:

	item_name	choice_description	item_price
0	Chips and Fresh Tomato Salsa		$2.39

	item_name	choice_description	item_price
1	Izze	[Clementine]	$3.39
2	Nantucket Nectar	[Apple]	$3.39
3	Chips and Tomatillo-Green Chili Salsa		$2.39
4	Chicken Bowl	[Tomatillo-Red Chili Salsa (Hot), [Black Beans...	$16.98
...	...	...	...
4617	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Sour ...	$11.75
4618	Steak Burrito	[Fresh Tomato Salsa, [Rice, Sour Cream, Cheese...	$11.75
4619	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$11.25
4620	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	$8.75
4621	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$8.75

4622 rows × 3 columns

4. [0.5 баллов] Посмотрите внимательнее на колонку с ценой товара. Какого она типа? Создайте новую колонку так, чтобы в ней цена была числом.

Для этого попробуйте применить функцию-преобразование к каждой строке вашей таблицы (для этого есть соответствующая функция).

In [87]:
```
data['new_item_price'] = data['item_price'].replace('[\$,]', '', regex=True).astype(
data.tail(10)
```

Out[87]:

	order_id	quantity	item_name	choice_description	item_price	new_item_price
4612	1831	1	Carnitas Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Rice,...	$9.25	9.25
4613	1831	1	Chips	NaN	$2.15	2.15
4614	1831	1	Bottled Water	NaN	$1.50	1.50
4615	1832	1	Chicken Soft Tacos	[Fresh Tomato Salsa, [Rice, Cheese, Sour Cream]]	$8.75	8.75
4616	1832	1	Chips and Guacamole	NaN	$4.45	4.45
4617	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Sour ...	$11.75	11.75
4618	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Sour Cream, Cheese...	$11.75	11.75
4619	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$11.25	11.25
4620	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	$8.75	8.75
4621	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	$8.75	8.75

Какая средняя/минимальная/максимальная цена у товара?

In [88]:
```
minimum = data['new_item_price'].min()
average = data['new_item_price'].mean()
maximum = data['new_item_price'].max()
print(f"Минимальная - {minimum}, максимальная - {maximum}, средняя - {average}")
```

```
Минимальная - 1.09, максимальная - 44.25, средняя - 7.464335785374297
```

Удалите старую колонку с ценой.

In [89]:
```
del data['item_price']
data.tail(2)
```

Out[89]:

	order_id	quantity	item_name	choice_description	new_item_price
4620	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	8.75
4621	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	8.75

5. [0.25 баллов] Какие 5 товаров были самыми дешёвыми и самыми дорогими? (по choice_description)

Для этого будет удобно избавиться от дубликатов и отсортировать товары. Не забудьте про количество товара.

In [90]:
```
data_copy = data
data_copy['cost_per_one']=data_copy['new_item_price']/data_copy['quantity']
data_copy.sort_values(['cost_per_one', 'item_name'], ascending=[True, True],inplace=
data_copy.drop_duplicates(subset=['item_name'], keep='first').head(5)
```

Out[90]:

	order_id	quantity	item_name	choice_description	new_item_price	cost_per_one
34	17	1	Bottled Water	NaN	1.09	1.09
18	9	2	Canned Soda	[Sprite]	2.18	1.09
263	114	1	Canned Soft Drink	[Coke]	1.25	1.25
6	3	1	Side of Chips	NaN	1.69	1.69
4509	1793	1	Chips	NaN	1.99	1.99

In [91]:
```
data_copy.drop_duplicates(subset=['item_name'], keep='last').tail(5)
```

Out[91]:

	order_id	quantity	item_name	choice_description	new_item_price	cost_per_one
4554	1810	1	Steak Crispy Tacos	[Roasted Chili Corn Salsa, [Fajita Vegetables,...	11.75	11.75
3710	1483	1	Steak Soft Tacos	[Fresh Tomato Salsa, Guacamole]	11.75	11.75
3546	1426	1	Barbacoa Salad Bowl	[Fresh Tomato Salsa, Guacamole]	11.89	11.89
4239	1692	1	Carnitas Salad Bowl	[Tomatillo Green Chili Salsa, [Black Beans, Ch...	11.89	11.89

	order_id	quantity	item_name	choice_description	new_item_price	cost_per_one
4313	1720	1	Steak Salad Bowl	[Roasted Chili Corn Salsa, [Fajita Vegetables,...	11.89	11.89

6. [0.5 баллов] Сколько раз клиенты покупали больше 1 Chicken Bowl (item_name)?

In [92]:
```
import numpy as np
np.sum((data['quantity'] > 1) & (data['item_name'] == "Chicken Bowl"))
```

Out[92]: 33

7. [0.5 баллов] Какой средний чек у заказа? Сколько в среднем товаров покупают?

Если необходимо провести вычисления в терминах заказов, то будет удобно сгруппировать строки по заказам и посчитать необходимые статистики.

In [93]:
```
#Средний чек заказа
data.loc[:, 'new_item_price'].mean()
```

Out[93]: 7.464335785374296

In [94]:
```
#В среднем товаров
data.loc[:,"quantity"].mean()
```

Out[94]: 1.0757247944612722

8. [0.25 баллов] Сколько заказов содержали ровно 1 товар?

In [95]:
```
len(data.loc[data['quantity'] == 1])
```

Out[95]: 4355

9. [0.25 баллов] Какая самая популярная категория товара?

In [96]:
```
data['item_name'].value_counts().idxmax()
```

Out[96]: 'Chicken Bowl'

13. [0.75 баллов] Создайте новый DateFrame из матрицы, созданной ниже. Назовите колонки index, column1, column2 и сделайте первую колонку индексом.

In [97]:
```
from pandas import DataFrame


data_rand = np.random.rand(10, 3)
data_new = DataFrame(data_rand, columns = ['index','column1','column2'])
data_new
```

Out[97]:

	index	column1	column2
0	0.639513	0.623736	0.444349
1	0.483852	0.415418	0.225489
2	0.563000	0.113881	0.760107
3	0.643455	0.953607	0.102741
4	0.097748	0.819078	0.065581
5	0.745732	0.182420	0.186830
6	0.475292	0.155890	0.067085
7	0.110152	0.828859	0.015541
8	0.504779	0.013305	0.858551
9	0.255829	0.674130	0.390525

Сохраните DataFrame на диск в формате csv без индексов и названий столбцов.

In [98]:
```
data_new.to_csv("DataFrame.csv")
```

2. Визуализации и matplotlib

При работе с данными часто неудобно делать какие-то выводы, если смотреть на таблицу и числа в частности, поэтому важно уметь визуализировать данные. В этом разделе мы этим и займёмся.

У matplotlib, конечно, же есть документация с большим количеством примеров, но для начала достаточно знать про несколько основных типов графиков:

- plot — обычный поточечный график, которым можно изображать кривые или отдельные точки;
- hist — гистограмма, показывающая распределение некоторое величины;
- scatter — график, показывающий взаимосвязь двух величин;
- bar — столбцовый график, показывающий взаимосвязь количественной величины от категориальной.

В этом задании вы попробуете построить каждый из них. Не менее важно усвоить базовые принципы визуализаций:

- на графиках должны быть подписаны оси;
- у визуализации должно быть название;
- если изображено несколько графиков, то необходима поясняющая легенда;
- все линии на графиках должны быть чётко видны (нет похожих цветов или цветов, сливающихся с фоном);
- если отображена величина, имеющая очевидный диапазон значений (например, проценты могут быть от 0 до 100), то желательно масштабировать ось на весь диапазон значений (исключением является случай, когда вам необходимо показать малое отличие, которое незаметно в таких масштабах).

In [99]:
```
%matplotlib inline
import matplotlib.pyplot as plt
```

На самом деле мы уже импортировали matplotlib внутри %pylab inline в начале задания.

Работать мы будем с той же выборкой покупок. Добавим новую колонку с датой покупки.

In [100…
```
import datetime

start = datetime.datetime(2018, 1, 1)
end = datetime.datetime(2018, 1, 31)
delta_seconds = int((end - start).total_seconds())

dates = pd.DataFrame(index=data.order_id.unique())
dates['date'] = [
    (start + datetime.timedelta(seconds=random.randint(0, delta_seconds))).strftime(
    for _ in range(data.order_id.nunique())]

# если DataFrame с покупками из прошлого заказа называется не df, замените на ваше н
data['date'] = data.order_id.map(dates['date'])
```

1. [1 балл] Постройте гистограмму распределения сумм покупок и гистограмму средних цен отдельных видов продуктов item_name.

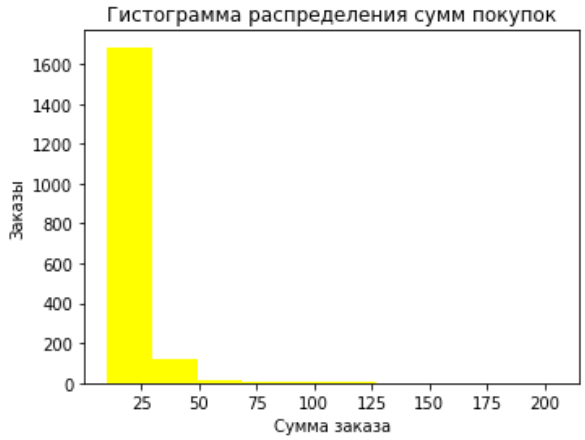
Изображайте на двух соседних графиках. Для этого может быть полезен subplot.

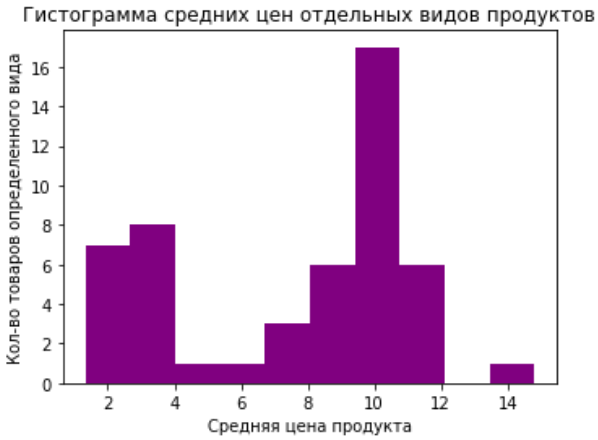
In [108…
```
order_price = data.groupby('order_id').sum()['new_item_price']
avg_price   = data.groupby('item_name').mean()['new_item_price']
fig, (axes1, axes2) = plt.subplots(nrows=1, ncols=2, figsize=(12, 4))

axes1.hist(order_price, color=['yellow'])
axes1.set_title('Гистограмма распределения сумм покупок')
axes1.set_xlabel('Сумма заказа')
axes1.set_ylabel('Заказы')

axes2.hist(avg_price, color=['purple'])
axes2.set_title('Гистограмма средних цен отдельных видов продуктов')
axes2.set_xlabel('Средняя цена продукта')
axes2.set_ylabel('Кол-во товаров определенного вида')

plt.show()
```

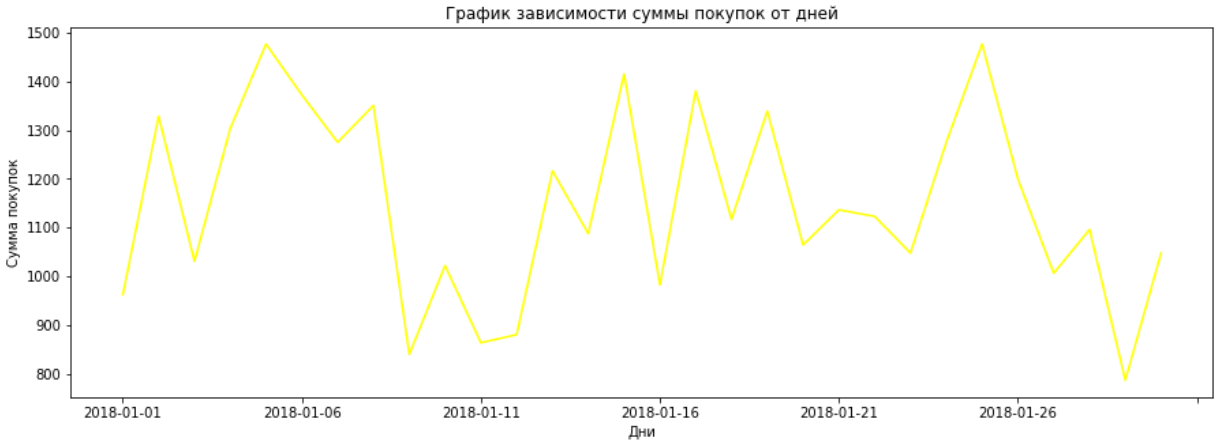




2. [1 балл] Постройте график зависимости суммы покупок от дней.

In [110…

```
order_price= data.groupby('date').sum()['new_item_price']
plt.figure(figsize=(15,5))
order_price.plot(color='yellow')
plt.title('График зависимости суммы покупок от дней')
plt.xlabel('Дни')
plt.ylabel('Сумма покупок')
plt.show()
```



3. [1 балл] Постройте график зависимости денег за товар от купленного количества (scatter plot).

In [111…

```
quantity = data.groupby('order_id').sum()['quantity']
price_item = data.groupby('order_id').sum()['new_item_price']

plt.scatter(quantity, price_item, color='purple')
plt.title('Зависимость денег за товар от купленного количества')
plt.xlabel('Количество купленного товара')
plt.ylabel('Цена за товар')
plt.show()
```

Сохраните график в формате pdf (так он останется векторизованным).

In [113…

```
plt.savefig("scatter.pdf")
```

```
<Figure size 432x288 with 0 Axes>
```

Кстати, существует надстройка над matplotlib под названием seaborn. Иногда удобнее и красивее делать визуализации через неё.