МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технология машинного обучения”

Рубежный контроль №1

«Разведочный анализ и обработки данных.»

ВЫПОЛНИЛ:

Матюнин да Вейга Р.А.

Группа: ИУ5-61Б

ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва 2020

Вариант:

Номер варианта	Номер задачи	Номер набора данных, указанного в задаче
15	2	7

Задание №2:

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Дополнительные требования по группам:

- Для студентов групп ИУ5-61Б, ИУ5Ц-81Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Набор данных:

https://www.kaggle.com/san-francisco/sf-restaurant-scores-lives-standard

Выполненная работа:

In [1]:
```
import numpy as np
import pandas as pd
import os
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")

DATA_PATH = os.path.join('dataset')
```

In [2]:
```
def load_data(data_path=DATA_PATH):
    csv_path = os.path.join(data_path, 'rsls.csv')
    return pd.read_csv(csv_path)
```

In [3]:
```
data = load_data()
data
```

Out[3]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_l
0	69618	Fancy Wheatfield Bakery	1362 Stockton St	San Francisco	CA	94133	NaN	NaN	
1	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118	NaN	NaN	
2	69487	Hakkasan San Francisco	1 Kearny St	San Francisco	CA	94108	NaN	NaN	
3	91044	Chopsticks Restaurant	4615 Mission St	San Francisco	CA	94112	NaN	NaN	
4	85987	Tselogs	552 Jones St	San Francisco	CA	94102	NaN	NaN	
...	...	...	...	...	...	...	...	...	

Обработка пропусков в данных для категориального признака:

In [5]:
```
###Пустые значения в категориальном признаке risk_category заменим на часто встречаемые значения
```

In [6]:
```
from sklearn.impute import SimpleImputer

imp = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
imp.fit(data[['risk_category']])
train = imp.transform(data[['risk_category']])
```

In [7]:
```
data['risk_category'] = train
data
```

Out[7]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude
0	69618	Fancy Wheatfield Bakery	1362 Stockton St	San Francisco	CA	94133	NaN
1	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118	NaN
2	69487	Hakkasan San Francisco	1 Kearny St	San Francisco	CA	94108	NaN
3	91044	Chopsticks Restaurant	4615 Mission St	San Francisco	CA	94112	NaN
4	85987	Tselogs	552 Jones St	San Francisco	CA	94102	NaN
...	...	...	...	...	...	...	...
53968	80305	Snowbird Coffee	1352 A 9th Ave	San Francisco	CA	94110	NaN
53969	80233	Buffalo Kitchen	107 Leland Ave	San Francisco	CA	94134	NaN
53970	100216	BUNN MIKE	300 DE HARO ST	San Francisco	CA	94103	NaN
53971	79430	City Discount Meat & Grocery Market	2298 Mission St	San Francisco	CA	94110	NaN
53972	77681	Tart To Tart Inc.	641 Irving St	San Francisco	CA	94122	NaN

53973 rows × 17 columns

Обработка пропусков в данных для количественного признака:

In [8]:
```
##В поле 'inspection_score' пустые значения заменим на среднее
```

In [9]:
```
mean = data['inspection_score'].mean()
data['inspection_score'].fillna(mean, inplace=True)
data
```

Out[9]:

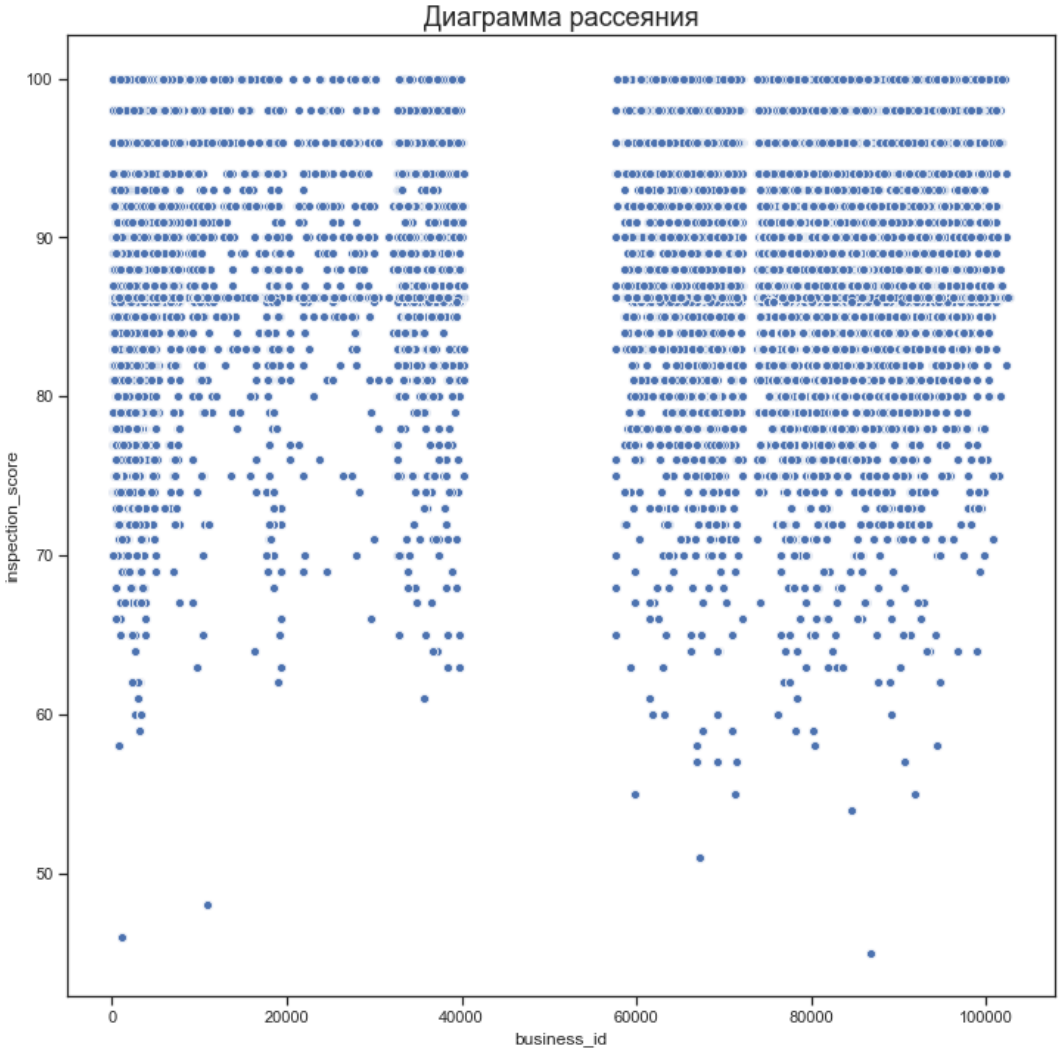
	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude
0	69618	Fancy Wheatfield Bakery	1362 Stockton St	San Francisco	CA	94133	NaN
1	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118	NaN
2	69487	Hakkasan San Francisco	1 Kearny St	San Francisco	CA	94108	NaN
3	91044	Chopsticks Restaurant	4615 Mission St	San Francisco	CA	94112	NaN
4	85987	Tselogs	552 Jones St	San Francisco	CA	94102	NaN
...	...	...	...	...	...	...	...
53968	80305	Snowbird Coffee	1352 A 9th Ave	San Francisco	CA	94110	NaN
53969	80233	Buffalo Kitchen	107 Leland Ave	San Francisco	CA	94134	NaN
53970	100216	BUNN MIKE	300 DE HARO ST	San Francisco	CA	94103	NaN
53971	79430	City Discount Meat & Grocery Market	2298 Mission St	San Francisco	CA	94110	NaN
53972	77681	Tart To Tart Inc.	641 Irving St	San Francisco	CA	94122	NaN

53973 rows × 17 columns

Диаграмма рассеяния:

In [10]:
```
##Диаграмма рассеяния
```

In [12]:
```
fig, ax = plt.subplots(figsize=(12,12))
sns.scatterplot(ax=ax, x='business_id', y='inspection_score', data=data)
plt.title(r'Диаграмма рассеяния', fontsize=18, y=1);
```



- Ноутбук с выполненной работой и отчет размещены в репозитории на github: https://github.com/Yorati/TMO