Рубежный контроль №1

Клементьев А.И. ИУ5-65Б Вариант №7

Задача №1. Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель. Для набора данных построить "парные диаграммы".

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

data = pd.read_csv("./googleplaystore.csv", sep=',')

data.head()
```

```
                                                 App        Category
Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN
4.1
1                                Coloring book moana  ART_AND_DESIGN
3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
4.7
3                             Sketch - Draw & Paint  ART_AND_DESIGN
4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3

  Reviews  Size     Installs  Type Price Content Rating  \
0     159   19M      10,000+  Free     0       Everyone
1     967   14M     500,000+  Free     0       Everyone
2   87510  8.7M   5,000,000+  Free     0       Everyone
3  215644   25M  50,000,000+  Free     0           Teen
4     967  2.8M     100,000+  Free     0       Everyone

                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
```

```
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

```
# Преобразовываем значения в колонке Reviews в float, если это
возможно, иначе удаляем строку
data['Reviews'] = data['Reviews'].apply(lambda x: float(x) if
str(x).isdigit() else None)

# Преобразовываем значения в колонке Price в float, если это возможно,
иначе удаляем строку
data['Price'] = data['Price'].apply(lambda x: float(x) if
str(x).isdigit() else None)

# Удаляем строки, в которых Reviews или Price имеют значение None
data = data.dropna(subset=['Reviews', 'Price'])

data.head()
```

```
                                                 App        Category
Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN
4.1
1                                Coloring book moana  ART_AND_DESIGN
3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
4.7
3                               Sketch - Draw & Paint  ART_AND_DESIGN
4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3

    Reviews  Size     Installs  Type  Price Content Rating  \
0     159.0   19M      10,000+  Free    0.0       Everyone
1     967.0   14M     500,000+  Free    0.0       Everyone
2   87510.0  8.7M   5,000,000+  Free    0.0       Everyone
3  215644.0   25M  50,000,000+  Free    0.0           Teen
4     967.0  2.8M     100,000+  Free    0.0       Everyone

                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
```

```
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

```
data.info()
```

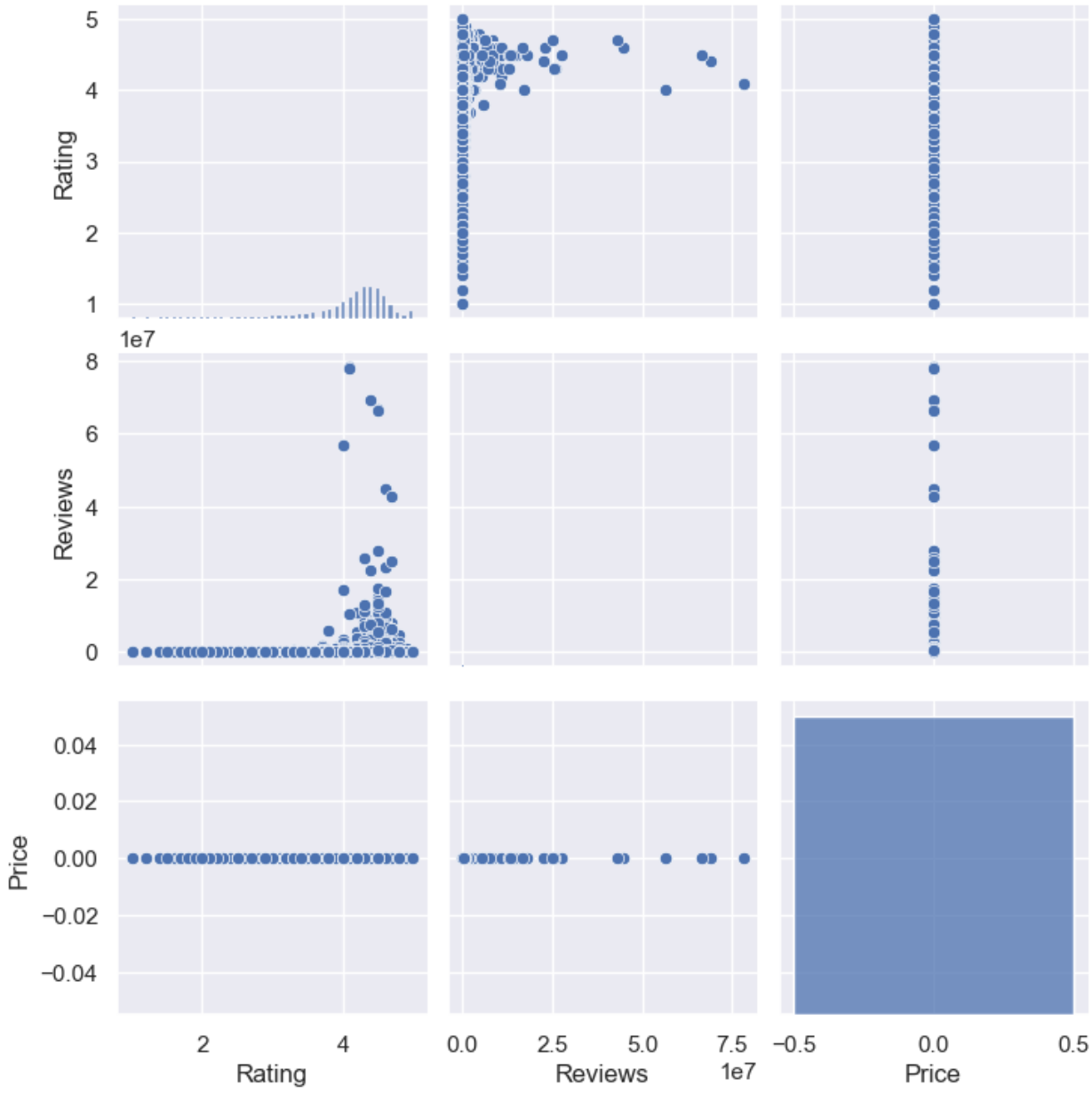
```
<class 'pandas.core.frame.DataFrame'>
Index: 10040 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10040 non-null  object 
 1   Category        10040 non-null  object 
 2   Rating          8719 non-null   float64
 3   Reviews         10040 non-null  float64
 4   Size            10040 non-null  object 
 5   Installs        10040 non-null  object 
 6   Type            10039 non-null  object 
 7   Price           10040 non-null  float64
 8   Content Rating  10040 non-null  object 
 9   Genres          10040 non-null  object 
 10  Last Updated    10040 non-null  object 
 11  Current Ver     10034 non-null  object 
 12  Android Ver     10039 non-null  object 
dtypes: float64(3), object(10)
memory usage: 1.1+ MB
```

```
data.describe()
```

```
            Rating       Reviews    Price
count  8719.000000  1.004000e+04  10040.0
mean      4.186203  4.786134e+05      0.0
std       0.512338  3.039342e+06      0.0
min       1.000000  0.000000e+00      0.0
25%       4.000000  4.500000e+01      0.0
50%       4.300000  2.963500e+03      0.0
75%       4.500000  6.667825e+04      0.0
max       5.000000  7.815831e+07      0.0
```

```
sns.pairplot(data)
```

```
<seaborn.axisgrid.PairGrid at 0x2b193cf10>
```




```
from sklearn.preprocessing import OrdinalEncoder

oe = OrdinalEncoder()
cat_enc_oe = oe.fit_transform(data[["Rating", "Reviews", "Price"]])
cat_enc_oe
```

```
array([[2.900e+01, 1.590e+02, 0.000e+00],
       [2.700e+01, 7.060e+02, 0.000e+00],
       [3.500e+01, 3.761e+03, 0.000e+00],
       ...,
       [      nan, 3.000e+00, 0.000e+00],
```

```
       [3.300e+01, 1.140e+02, 0.000e+00],
       [3.300e+01, 4.745e+03, 0.000e+00]])

data_enc = pd.DataFrame(data=cat_enc_oe, index=data.index,
columns=["Rating", "Reviews", "Price"])
data_enc.head()
```

	Rating	Reviews	Price
0	29.0	159.0	0.0
1	27.0	706.0	0.0
2	35.0	3761.0	0.0
3	33.0	4370.0	0.0
4	31.0	706.0	0.0

```
data_enc["Price"].unique()
```

```
array([0.])
```

```
data_enc = data_enc.drop(columns=['Reviews', 'Price'])

data_enc = data_enc.join(data[["Reviews"]])
data_enc = data_enc.join(data[["Price"]])

data_enc
```

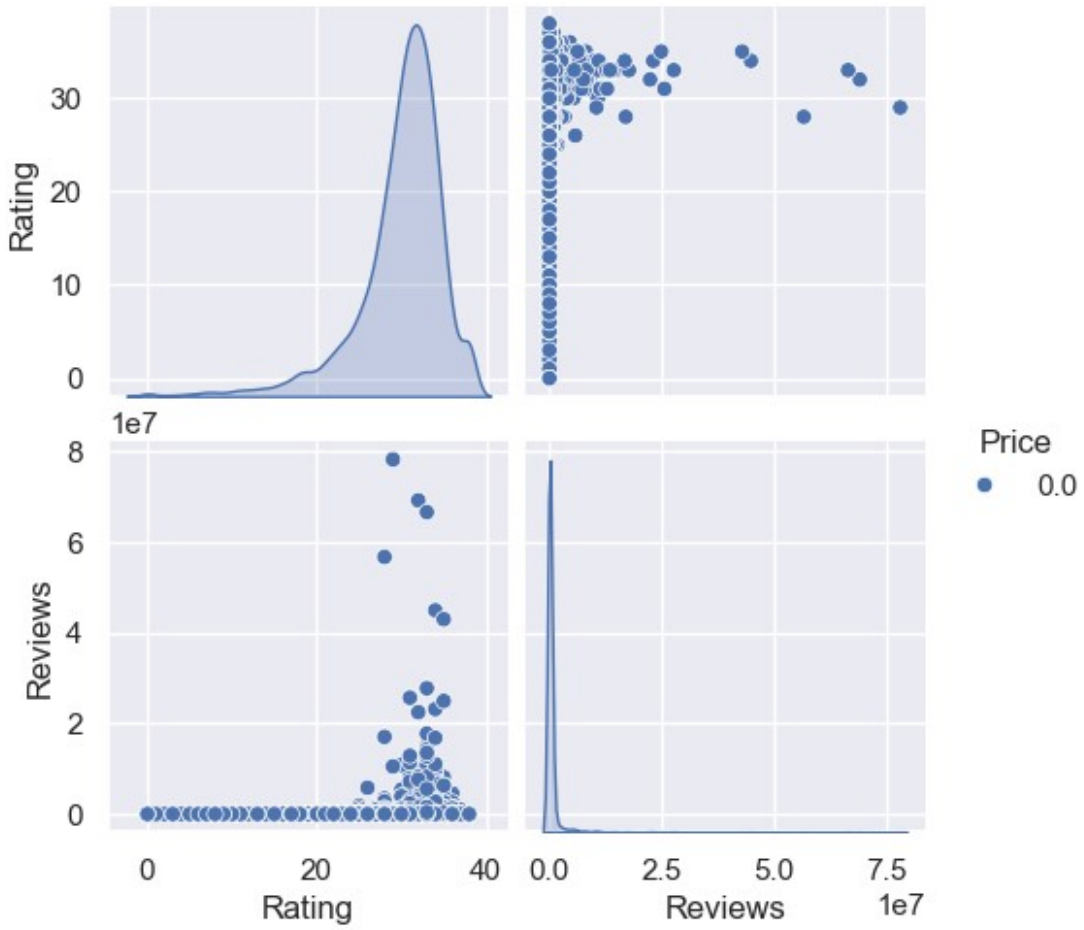
	Rating	Reviews	Price
0	29.0	159.0	0.0
1	27.0	967.0	0.0
2	35.0	87510.0	0.0
3	33.0	215644.0	0.0
4	31.0	967.0	0.0
...	...	...	...
10836	33.0	38.0	0.0
10837	38.0	4.0	0.0
10838	NaN	3.0	0.0
10839	33.0	114.0	0.0
10840	33.0	398307.0	0.0

```
[10040 rows x 3 columns]

corr_matrix = data_enc.corr()

corr_matrix
```

	Rating	Reviews	Price
Rating	1.000000	0.072914	NaN
Reviews	0.072914	1.000000	NaN
Price	NaN	NaN	NaN

```
heatmap = sns.heatmap(corr_matrix, annot=True)
```


```
pair_plot = sns.pairplot(data_enc, hue="Price")
plt.show()
```



На основании корреляционного анализа можно сделать выводы:

- Корреляция между признаками почти отсутствует
- Нельзя выделять влияние чего-либо ввиду недостатка категориальных признаков.
- Придется создавать новые признаки на основе имеющихся