Рубежный контроль №1

Писарчук Надежда ИУ5-22М

Вариант - 7

In [1]:
```
# This Python 3 environment comes with many helpful analytics libraries insta
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will li

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that g
# You can also write temporary files to /kaggle/temp/, but they won't be save

pd.set_option('max_colwidth', 800)
pd.set_option('display.max_columns', None)
```

```
/kaggle/input/kepler-exoplanet-search-results/cumulative.csv
```

In [2]:
```
data = pd.read_csv(
    '/kaggle/input/kepler-exoplanet-search-results/cumulative.csv',
    sep=",")
```

In [3]:
```
data.head()
```

Out[3]:

	rowid	kepid	kepoi_name	kepler_name	koi_disposition	koi_pdisposition	koi_score	koi_
0	1	10797460	K00752.01	Kepler-227 b	CONFIRMED	CANDIDATE	1.000	
1	2	10797460	K00752.02	Kepler-227 c	CONFIRMED	CANDIDATE	0.969	
2	3	10811496	K00753.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	
3	4	10848459	K00754.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	
4	5	10854555	K00755.01	Kepler-664 b	CONFIRMED	CANDIDATE	1.000	

Задача №7.

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения медианой.

Я буду заполнять признак koi_score, принимающий значение от 0 до 1, имеющий 16% пропусков

In [4]:
```
data.isna().sum()['koi_score']
```

Out[4]:
```
1510
```

In [5]:
```
data.koi_score.median()
```

Out[5]:
```
0.334
```

In [6]:
```
data['koi_score_nonnan'] = data['koi_score'].fillna(
    data['koi_score'].median())
data.koi_score_nonnan.isna().sum()
```

Out[6]:
```
0
```

Доп задание

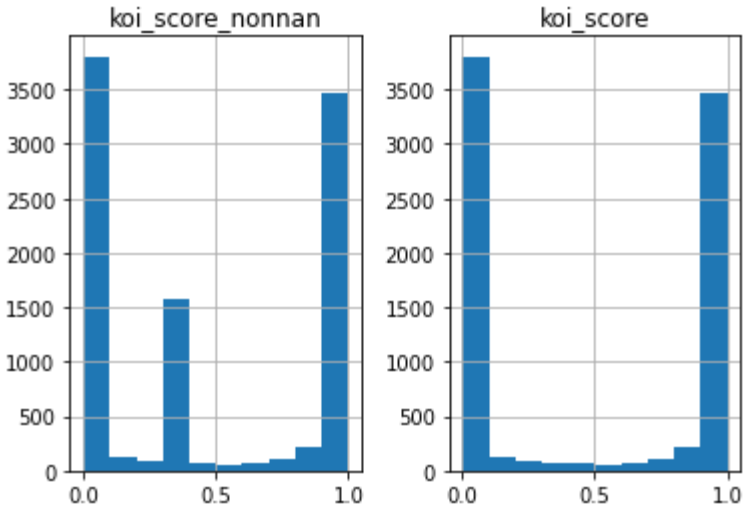
Для произвольной колонки данных построить гистограмму

Для визуализации заполнения пропусков построю гистограмму для koi_score и koi_score_nonnan

In [7]:
```
data[['koi_score_nonnan', 'koi_score']].hist()
data[['koi_score_nonnan', 'koi_score']].plot.hist()
```

```
/opt/conda/lib/python3.7/site-packages/pandas/plotting/_matplotlib/tools.py:400: MatplotlibDeprecationWarning:
The is_first_col function was deprecated in Matplotlib 3.4 and will be removed two minor releases later. Use ax.get_subplotspec().is_first_col() instead.
  if ax.is_first_col():
```

Out[7]:
```
<AxesSubplot:ylabel='Frequency'>
```



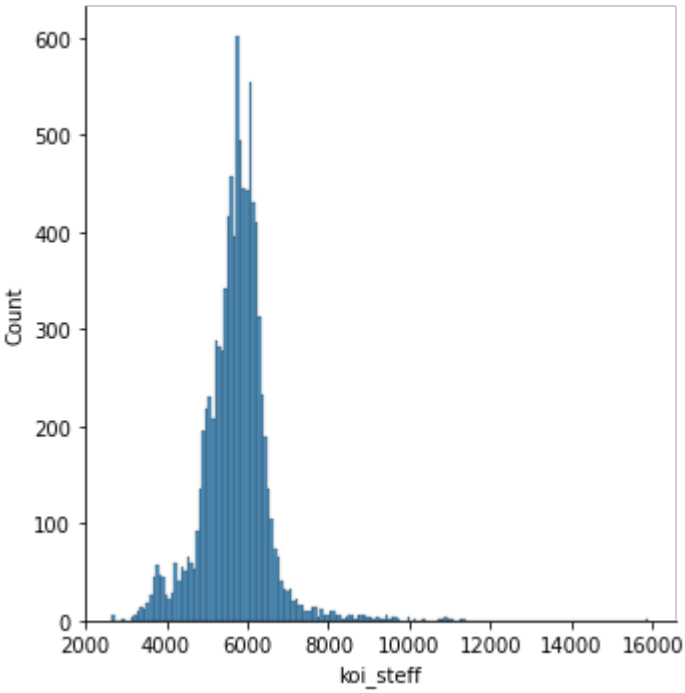
Задача №27.

Для набора данных для одного (произвольного) числового признака проведите обнаружение и замену (найденными верхними и нижними границами) выбросов на основе 5% и 95% квантилей. Обнаружение и замена выбросов будет производиться для признака koi_steff

In [8]:
```
import matplotlib.pyplot as plt
import seaborn as sns

sns.displot(data, x="koi_steff")
```

Out[8]: <seaborn.axisgrid.FacetGrid at 0x7f2543c9ddd0>



In [9]:
```
q5 = data.koi_steff.quantile(0.05)
q95 = data.koi_steff.quantile(0.95)
print(q5, q95)

data['koi_steff_comp'] = np.where(data.koi_steff < q5, q5,
                          np.where(data.koi_steff > q95, q95, data.koi_steff))
```

```
4330.0 6726.0
```

In [10]:
```
sns.displot(data, x="koi_steff_comp")
```

Out[10]:
```
<seaborn.axisgrid.FacetGrid at 0x7f2543abbc50>
```