

Задача №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? Для произвольной колонки данных построить график “Скрипичная диаграмма (violin plot)”.

```
[2]: from sklearn.impute import SimpleImputer
     from sklearn.impute import MissingIndicator
     import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
[5]: data = pd.read_csv('states_all_extended.csv')
```

```
[10]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1715 entries, 0 to 1714
Columns: 266 entries, PRIMARY_KEY to G08_TR_A_MATHEMATICS
dtypes: float64(263), int64(1), object(2)
memory usage: 3.5+ MB
```

1. Обработка пропусков в данных для количественного признака

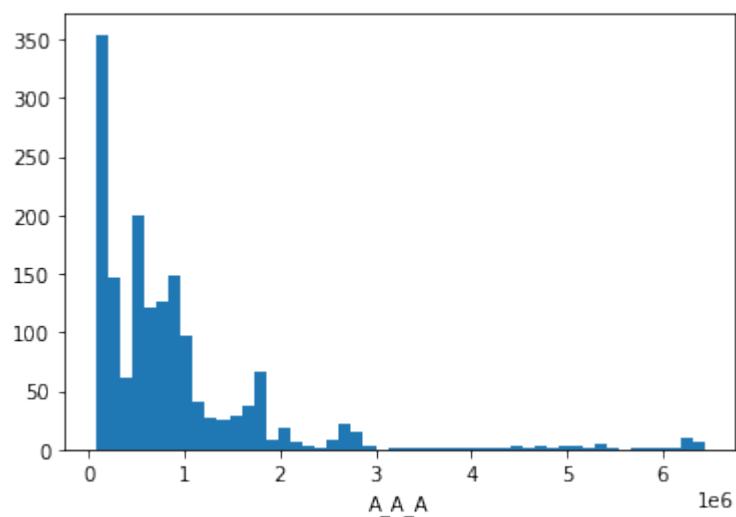
Выберем колонку A_A_A, в которой меньше 5% пропусков.

```
[20]: feature = 'A_A_A'
```

```
[13]: round(data[data[feature].isnull()].shape[0] / data.shape[0] * 100.0, 2)
```

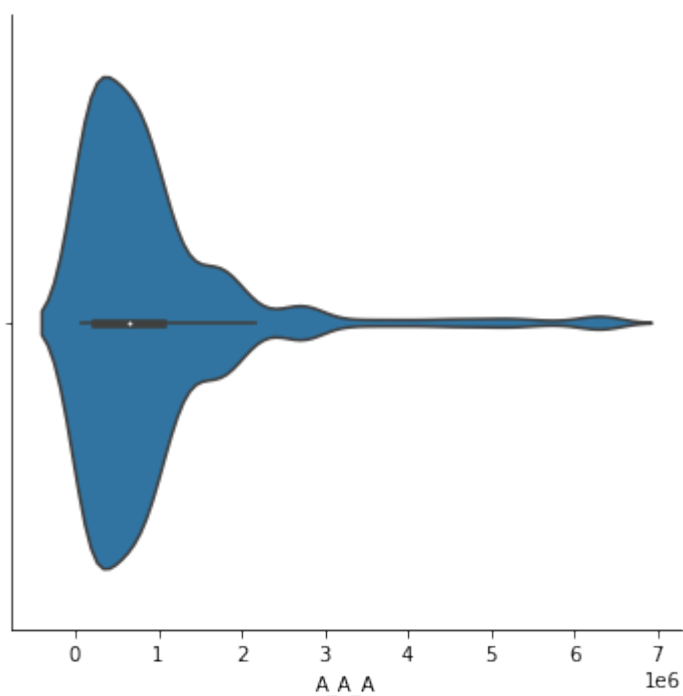
```
[13]: 4.84
```

```
[21]: plt.hist(data[feature], 50)
     plt.xlabel(feature)
     plt.show()
```



```
[15]: sns.catplot(x=feature, data=data, kind="violin")
```

```
[15]: <seaborn.axisgrid.FacetGrid at 0x7f9222165790>
```



```
[17]: data[feature].mean(), data[feature].median(), data[feature].mode()
```

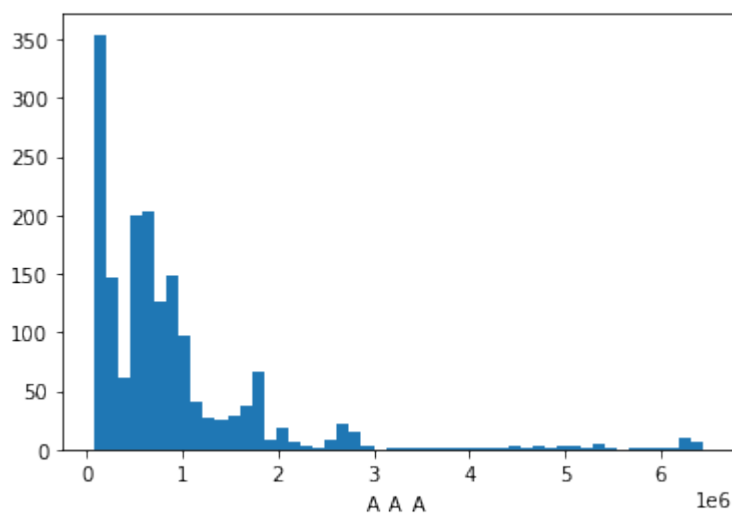
```
[17]: (913969.4944852941,
      645805.0,
      0    472394.0
      1    490917.0
      2    872436.0)
```

```
dtype: float64)
```

Распределение не одномодальное, есть выбросы, так что заполним пропуски значением медианы

```
[27]: data[[feature]] = SimpleImputer(missing_values=np.nan, strategy='median').  
      ↪ fit_transform(data[[feature]])
```

```
[28]: plt.hist(data[feature], 50)  
      plt.xlabel(feature)  
      plt.show()
```



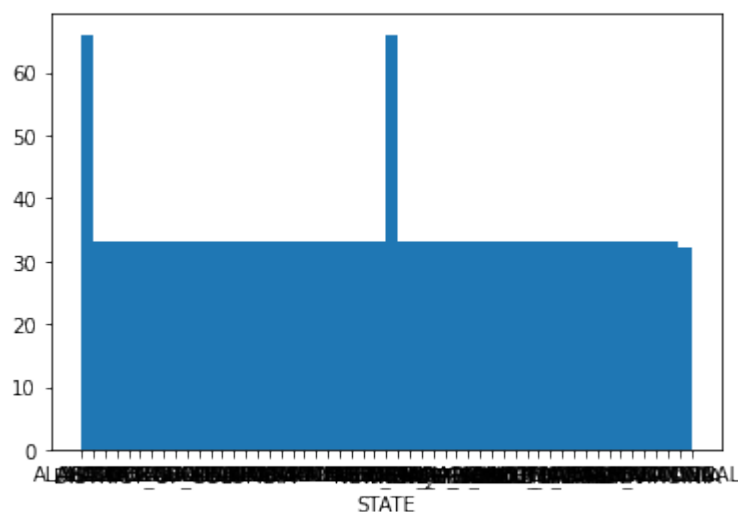
```
[29]: data[feature].mean(), data[feature].median(), data[feature].mode()
```

```
[29]: (900991.2711370263,  
      645805.0,  
      0      645805.0  
      dtype: float64)
```

2. Обработка пропусков в данных для категориального признака

Пропусков в категориальных признаках нет, так что для примера создадим новый признак с пропусками

```
[31]: plt.hist(data['STATE'], 50)  
      plt.xlabel('STATE')  
      plt.show()
```



```
[35]: data['MIS_STATE'] = [data['STATE'][i] if i % 20 != 0 else np.nan for i in
    ↪ range(len(data['STATE']))]
```

```
[36]: print("          {}".format(round(data[data['MIS_STATE'].isnull()].
    ↪ shape[0] / data.shape[0] * 100.0, 2)))
```

[36]: 5.01

Заполним пропуски константой 'NA'. Распределение стран равномерное, так что стратегия 'most_frequent' вряд ли будет хорошо работать.

```
[38]: data[['MIS_STATE']] = SimpleImputer(missing_values=np.nan,
    ↪ strategy='constant', fill_value='NA').fit_transform(data[['MIS_STATE']])
```

```
[41]: print("          {}".format(round(data[data['MIS_STATE'].isnull()].
    ↪ shape[0] / data.shape[0] * 100.0, 2)))
```

0.0%

Видим, что все пропуски заполнены.

3. Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Лучше всего было бы удалить признаки, в которых больше 5% пропущенных значений. при больших значениях повышается вероятность, что пропуски мы заполнили неправильно. Однако у нас во многих признаках слишком много пропущенных значений, так что повысим наш "порог допустимого" до 30%. Признаки G01-G08_A_A (40.52%), G09-G12_A_A (37.55%), G01_AM_F (76.27%), G01_AM_M (76.21%), G01_AS_F (76.21%), G01_AS_M (76.21%), G01_BL_F (76.21%), G01_BL_M (76.21%), G01_HI_F (76.27%) и т. д. мы удалим.