# Лабораторная работа

## Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных.

Цель лабораторной работы: изучение способов предварительной обработки данных для дальнейшего формирования моделей.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

# Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных.

Я выбрал Housing. Этот датасет содержит информацию о жилых домах в Калифорнии. В нем представлены различные характеристики домов и их окрестностей, такие как:

- Долгота (longitude)
- Широта (latitude)
- Средний возраст домов (housing_median_age)
- Общее количество комнат (total_rooms)
- Общее количество спален (total_bedrooms)
- Население (population)
- Количество домохозяйств (households)
- Средний доход домохозяйства (median_income)
- Средняя стоимость дома (median_house_value)
- Близость к океану (ocean_proximity)

```
data = pd.read_csv('housing.csv', sep=",")

data.head(10)
```

```
   longitude  latitude  housing_median_age  total_rooms
total_bedrooms  \
0    -122.23     37.88                41.0        880.0
129.0
1    -122.22     37.86                21.0       7099.0
1106.0
2    -122.24     37.85                52.0       1467.0
190.0
3    -122.25     37.85                52.0       1274.0
235.0
```

```
4    -122.25     37.85                52.0        1627.0
280.0
5    -122.25     37.85                52.0         919.0
213.0
6    -122.25     37.84                52.0        2535.0
489.0
7    -122.25     37.84                52.0        3104.0
687.0
8    -122.26     37.84                42.0        2555.0
665.0
9    -122.25     37.84                52.0        3549.0
707.0

   population  households  median_income  median_house_value
ocean_proximity
0       322.0       126.0         8.3252            452600.0
NEAR BAY
1      2401.0      1138.0         8.3014            358500.0
NEAR BAY
2       496.0       177.0         7.2574            352100.0
NEAR BAY
3       558.0       219.0         5.6431            341300.0
NEAR BAY
4       565.0       259.0         3.8462            342200.0
NEAR BAY
5       413.0       193.0         4.0368            269700.0
NEAR BAY
6      1094.0       514.0         3.6591            299200.0
NEAR BAY
7      1157.0       647.0         3.1200            241400.0
NEAR BAY
8      1206.0       595.0         2.0804            226700.0
NEAR BAY
9      1551.0       714.0         3.6912            261100.0
NEAR BAY
```

# Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:

- обработку пропусков в данных:

Проверяем формат данных

```
print(data.dtypes)

longitude             float64
latitude              float64
housing_median_age    float64
total_rooms           float64
```

```
total_bedrooms        float64
population            float64
households            float64
median_income         float64
median_house_value    float64
ocean_proximity        object
dtype: object
```

Так же при проверке данных не было замечено каких либо признаков не корректности

```
missing_values = data.isnull().sum()
count_values = data.count()
print(count_values)
print(missing_values)
```

```
longitude             20640
latitude              20640
housing_median_age    20640
total_rooms           20640
total_bedrooms        20433
population            20640
households            20640
median_income         20640
median_house_value    20640
ocean_proximity       20640
dtype: int64
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

Мы видим, что в столбце **total_bedrooms** имеется 207 пропущенных столбцов Предлагаю вместо того, чтобы удалять пропущенные данные и не учитывать их, заменить их на среднее значение на основе не пропущенных данных

```
data['total_bedrooms'] =
data['total_bedrooms'].fillna(data['total_bedrooms'].mean())
missing_values = data.isnull().sum()
count_values = data.count()
print(count_values)
print(missing_values)
```

```
longitude             20640
latitude              20640
housing_median_age    20640
total_rooms           20640
total_bedrooms        20640
population            20640
households            20640
median_income         20640
median_house_value    20640
ocean_proximity       20640
dtype: int64
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

- кодирование категориальных признаков

```
data = pd.get_dummies(data, columns=['ocean_proximity'],
prefix='ocean_proximity')

data.head(10)
```

```
   longitude  latitude  housing_median_age  total_rooms
total_bedrooms  \
0    -122.23     37.88                41.0        880.0
129.0
1    -122.22     37.86                21.0       7099.0
1106.0
2    -122.24     37.85                52.0       1467.0
190.0
3    -122.25     37.85                52.0       1274.0
235.0
4    -122.25     37.85                52.0       1627.0
280.0
5    -122.25     37.85                52.0        919.0
213.0
6    -122.25     37.84                52.0       2535.0
489.0
7    -122.25     37.84                52.0       3104.0
687.0
8    -122.26     37.84                42.0       2555.0
665.0
```

```
9    -122.25     37.84                52.0       3549.0
707.0

   population  households  median_income  median_house_value  \
0       322.0       126.0         8.3252            452600.0
1      2401.0      1138.0         8.3014            358500.0
2       496.0       177.0         7.2574            352100.0
3       558.0       219.0         5.6431            341300.0
4       565.0       259.0         3.8462            342200.0
5       413.0       193.0         4.0368            269700.0
6      1094.0       514.0         3.6591            299200.0
7      1157.0       647.0         3.1200            241400.0
8      1206.0       595.0         2.0804            226700.0
9      1551.0       714.0         3.6912            261100.0

   ocean_proximity_<1H OCEAN  ocean_proximity_INLAND
ocean_proximity_ISLAND  \
0                      False                   False
False
1                      False                   False
False
2                      False                   False
False
3                      False                   False
False
4                      False                   False
False
5                      False                   False
False
6                      False                   False
False
7                      False                   False
False
8                      False                   False
False
9                      False                   False
False

   ocean_proximity_NEAR BAY  ocean_proximity_NEAR OCEAN
0                      True                       False
1                      True                       False
2                      True                       False
3                      True                       False
4                      True                       False
5                      True                       False
6                      True                       False
7                      True                       False
8                      True                       False
9                      True                       False
```

```
print(data.dtypes)
```

```
longitude                   float64
latitude                    float64
housing_median_age          float64
total_rooms                 float64
total_bedrooms              float64
population                  float64
households                  float64
median_income               float64
median_house_value          float64
ocean_proximity_<1H OCEAN      bool
ocean_proximity_INLAND         bool
ocean_proximity_ISLAND         bool
ocean_proximity_NEAR BAY       bool
ocean_proximity_NEAR OCEAN     bool
dtype: object
```

Мы разделили данные категориально на ocean, inland, island, near bay, near ocean с помощью one hot encoding. Тем самым теперь у категорий нет веса в отличии если бы мы делили с помощью ярлычного деления, которое присваивает значения 1 2 3 4, из-за этого компьютер думал бы что у категорий есть вес

- масштабирование данных

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
data_scaled = scaler.fit_transform(data)

data_scaled_df = pd.DataFrame(data_scaled, columns=data.columns)
print(data_scaled_df.head(10))
```

```
   longitude  latitude  housing_median_age  total_rooms
total_bedrooms  \
0   0.211155  0.567481            0.784314     0.022331
0.019863
1   0.212151  0.565356            0.392157     0.180503
0.171477
2   0.210159  0.564293            1.000000     0.037260
0.029330
3   0.209163  0.564293            1.000000     0.032352
0.036313
4   0.209163  0.564293            1.000000     0.041330
0.043296
5   0.209163  0.564293            1.000000     0.023323
0.032899
6   0.209163  0.563231            1.000000     0.064423
0.075729
7   0.209163  0.563231            1.000000     0.078895
0.106456
```

```
8   0.208167  0.563231            0.803922     0.064932
0.103042
9   0.209163  0.563231            1.000000     0.090213
0.109559

   population  households  median_income  median_house_value  \
0    0.008941    0.020556       0.539668            0.902266
1    0.067210    0.186976       0.538027            0.708247
2    0.013818    0.028943       0.466028            0.695051
3    0.015555    0.035849       0.354699            0.672783
4    0.015752    0.042427       0.230776            0.674638
5    0.011491    0.031574       0.243921            0.525155
6    0.030578    0.084361       0.217873            0.585979
7    0.032344    0.106233       0.180694            0.466804
8    0.033717    0.097681       0.108998            0.436495
9    0.043387    0.117250       0.220087            0.507423

   ocean_proximity_<1H OCEAN  ocean_proximity_INLAND
ocean_proximity_ISLAND  \
0                         0.0                     0.0
0.0
1                         0.0                     0.0
0.0
2                         0.0                     0.0
0.0
3                         0.0                     0.0
0.0
4                         0.0                     0.0
0.0
5                         0.0                     0.0
0.0
6                         0.0                     0.0
0.0
7                         0.0                     0.0
0.0
8                         0.0                     0.0
0.0
9                         0.0                     0.0
0.0

   ocean_proximity_NEAR BAY  ocean_proximity_NEAR OCEAN
0                         1.0                         0.0
1                         1.0                         0.0
2                         1.0                         0.0
3                         1.0                         0.0
4                         1.0                         0.0
5                         1.0                         0.0
6                         1.0                         0.0
7                         1.0                         0.0
```

```
8                          1.0                          0.0
9                          1.0                          0.0
```