Лабораторная работа №2 по дисциплине "Искусственный интеллект и машинное обучение"

Выполнил: студент 2-го курса Звездин Алексей Сергеевич

Группа: ПИЖ-б-о-22-1

Руководитель практики: Березина Виктория Андреевна, ассистент кафедры информационных систем и технологий института цифрового развития

Тема работы: Визуализация данных

Цель работы: Изучение программных средств для визуализации наборов данных

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
data_path =  "https://raw.githubusercontent.com/InternetHacker1123/bd_ai/main/laba2/output.csv"
data = pd.read_csv(data_path)
data.head(10)
```

	State	Duration of account	Area code	Global plan	Voice mail plan	Number of voice mail messages	Total minutes during the day	Total calls during the day	Total charge during the day	Total minutes during the evening	Total calls during the evening	Total charge during the evening	Total minutes during the night	Total calls during the night	Total charge during the night	inter
0	KS	871	584	No	Yes	74	734.8	889	54.92	802.5	0	83.21	755.2	8	88.98	
1	OH	892	584	No	Yes	73	838.3	876	72.52	804.4	896	83.37	745.5	896	88.54	
2	NJ	862	584	No	No	9	756.5	885	58.61	878.7	889	89.60	837.3	895	2.67	
3	OH	15	591	Yes	No	9	700.5	28	49.00	38.0	11	4.73	803.0	10	1.13	
4	OK	24	584	Yes	No	9	833.2	886	71.65	851.6	877	87.38	813.0	878	1.58	
5	AL	881	489	Yes	No	9	776.5	1	62.01	779.3	898	81.24	796.0	881	0.81	
6	MA	878	489	No	Yes	75	781.7	11	62.90	651.4	891	70.37	787.3	881	0.42	
7	MO	852	584	Yes	No	9	842.9	20	73.30	896.8	5	1.23	788.1	3	0.46	
8	LA	882	591	No	No	9	815.4	2	68.62	648.3	19	70.10	784.1	9	0.28	

Next steps: View recommended plots

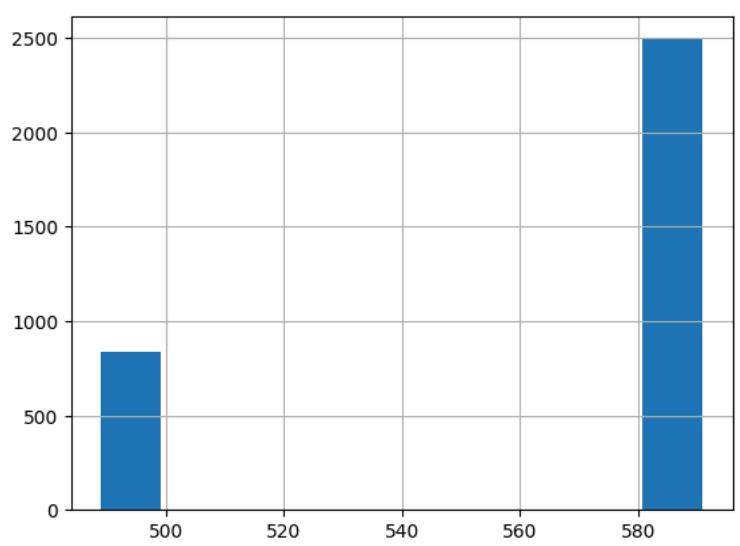
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   State                           3333 non-null   object
 1   Duration of account             3333 non-null   int64
 2   Area code                       3333 non-null   int64
 3   Global plan                     3333 non-null   object
 4   Voice mail plan                 3333 non-null   object
 5   Number of voice mail messages   3333 non-null   int64
 6   Total minutes during the day    3333 non-null   float64
 7   Total calls during the day      3333 non-null   int64
 8   Total charge during the day     3333 non-null   float64
 9   Total minutes during the evening 3333 non-null  float64
 10  Total calls during the evening  3333 non-null   int64
 11  Total charge during the evening 3333 non-null   float64
 12  Total minutes during the night  3333 non-null   float64
 13  Total calls during the night    3333 non-null   int64
 14  Total charge during the night   3333 non-null   float64
 15  Total international minutes     3333 non-null   float64
 16  Total international calls       3333 non-null   int64
 17  Total international charge      3333 non-null   float64
 18  Calls to customer service       3333 non-null   int64
 19  Churn                           3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 498.1+ KB
```

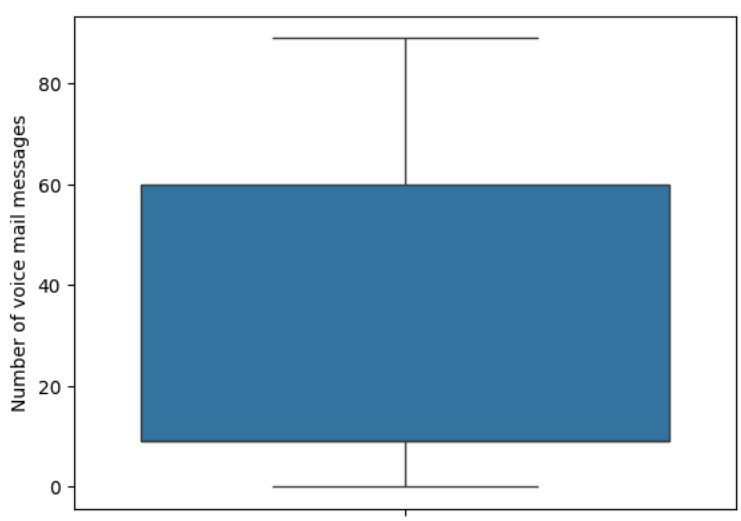
```
data.columns
```

```
Index(['State', 'Duration of account', 'Area code', 'Global plan',
       'Voice mail plan', 'Number of voice mail messages',
       'Total minutes during the day', 'Total calls during the day',
       'Total charge during the day', 'Total minutes during the evening',
       'Total calls during the evening', 'Total charge during the evening',
       'Total minutes during the night', 'Total calls during the night',
       'Total charge during the night', 'Total international minutes',
       'Total international calls', 'Total international charge',
       'Calls to customer service', 'Churn'],
      dtype='object')
```

```
data['Area code'].hist();
```

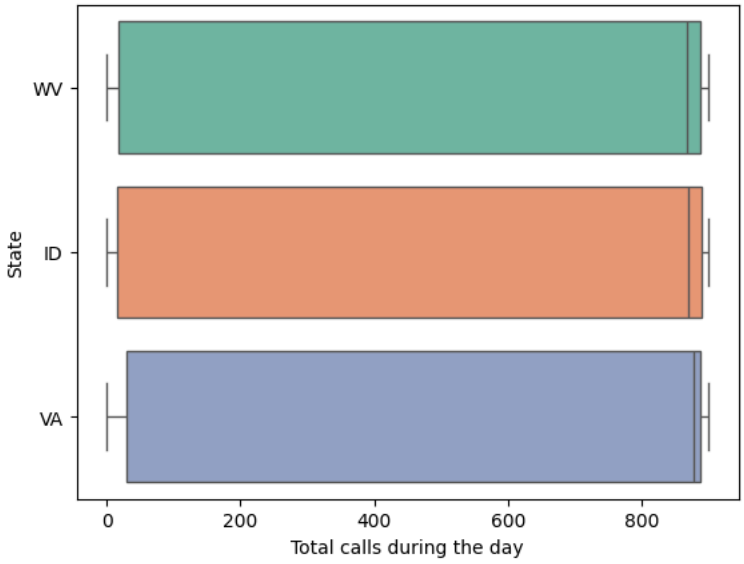



```
sns.boxplot(data['Number of voice mail messages']);
```

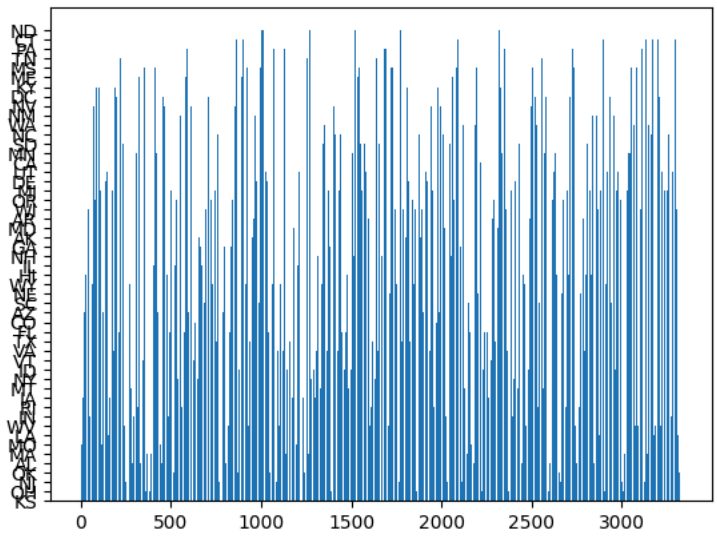



```
top_data = data[['State','Total calls during the day']]
top_data = top_data.groupby('State').sum()
top_data = top_data.sort_values('Total calls during the day',ascending=False)
top_data = top_data[:3].index.values
sns.boxplot(y='State',
            x='Total calls during the day',
            data=data[data.State.isin(top_data)], palette='Set2');
```

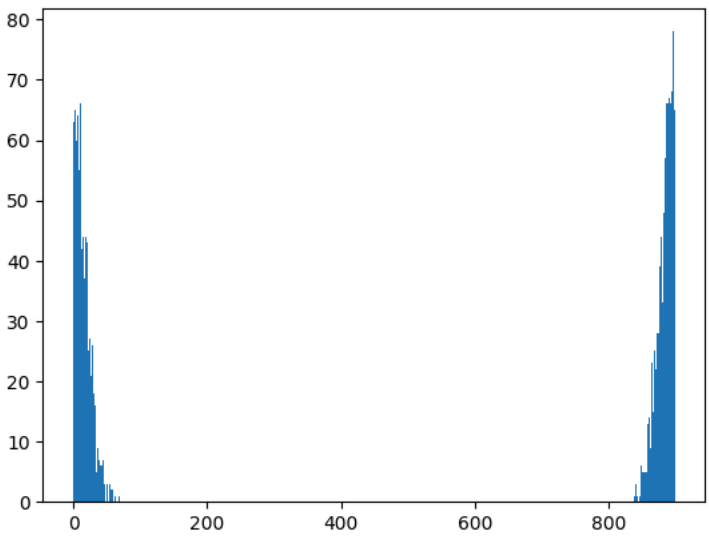
```
<ipython-input-18-b2678d391c99>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.boxplot(y='State',
```




```
plt.bar(data.index, data['State'])
plt.show()
```



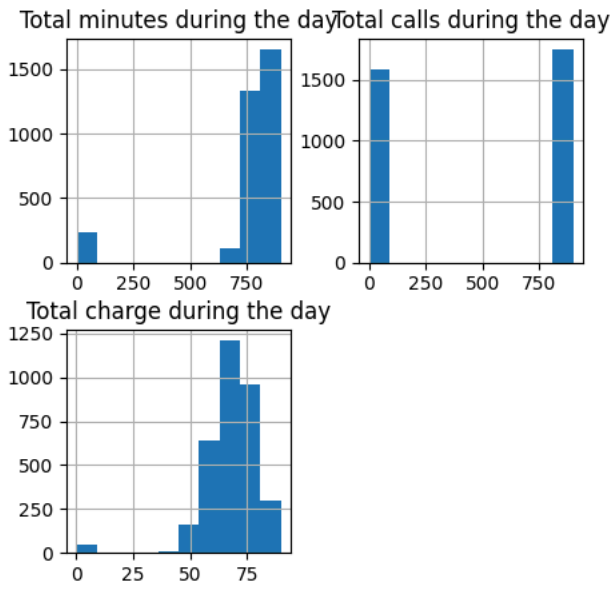

```
hist = data['Total calls during the day'].value_counts()
plt.bar(hist.index, hist);
```



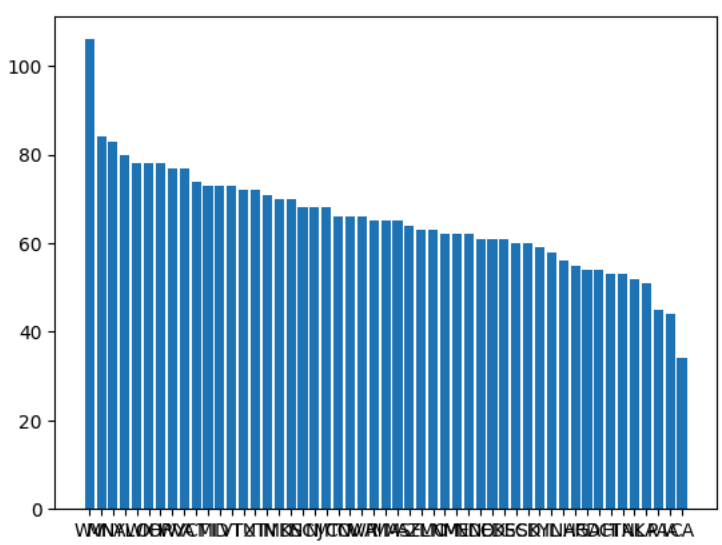
```
feats=[f for f in data.columns if 'during the day' in f]
feats
```

```
['Total minutes during the day',
 'Total calls during the day',
 'Total charge during the day']
```

```
data[feats].hist(figsize=(5,5));
```




```
hist = data['State'].value_counts()
plt.bar(hist.index, hist);
```

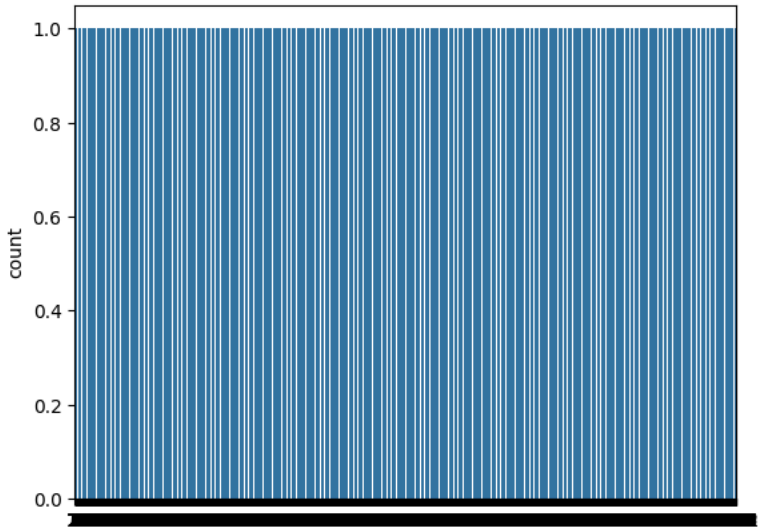


```
data['Total international calls'].value_counts()
```

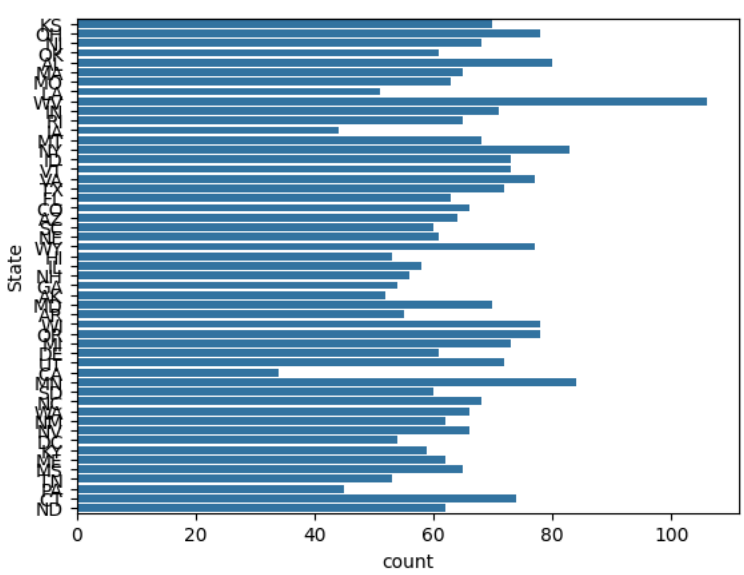
```
Total international calls
6     668
5     619
7     489
4     472
3     336
2     218
8     160
1     116
0     109
89     50
88     28
9      18
87     15
86     14
84      7
85      6
81      3
83      2
80      1
79      1
```

```
82      1
Name: count, dtype: int64
```

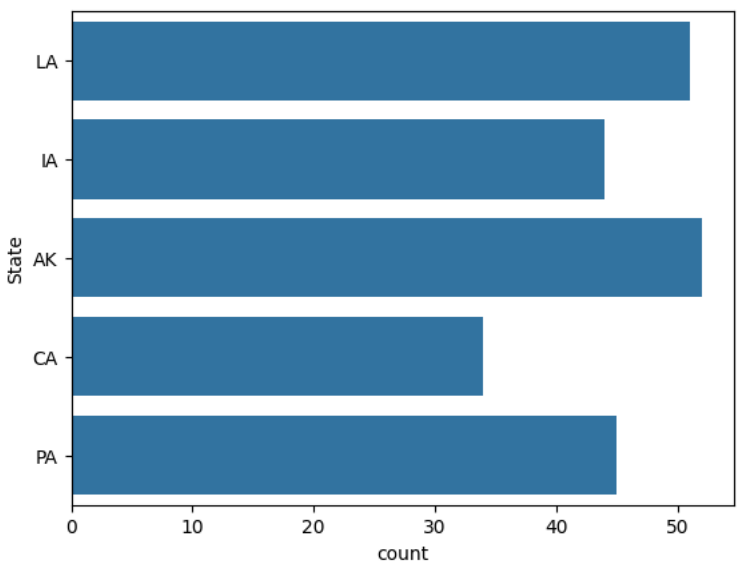
```
sns.countplot(data['Total international calls']);
```




```
sns.countplot(data['State']);
```




```
sns.countplot(data[data['State'].isin(data['State'].value_counts().tail(5).index)]['State']);
```



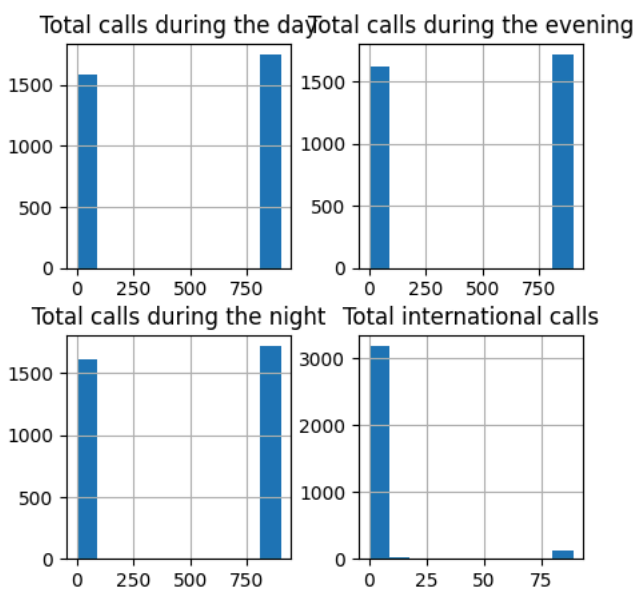
```
data.columns
```

```
Index(['State', 'Duration of account', 'Area code', 'Global plan',
       'Voice mail plan', 'Number of voice mail messages',
       'Total minutes during the day', 'Total calls during the day',
       'Total charge during the day', 'Total minutes during the evening',
       'Total calls during the evening', 'Total charge during the evening',
       'Total minutes during the night', 'Total calls during the night',
       'Total charge during the night', 'Total international minutes',
       'Total international calls', 'Total international charge',
       'Calls to customer service', 'Churn'],
      dtype='object')
```

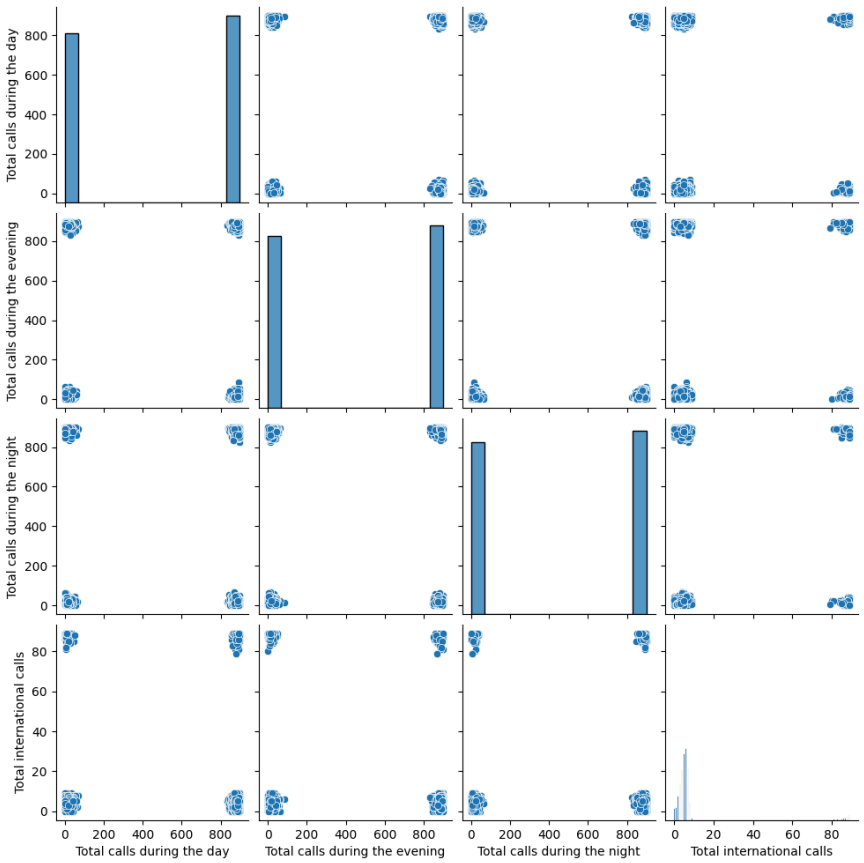
```
feats = [f for f in data.columns if 'calls' in f]
len(feats)
```

```
4
```

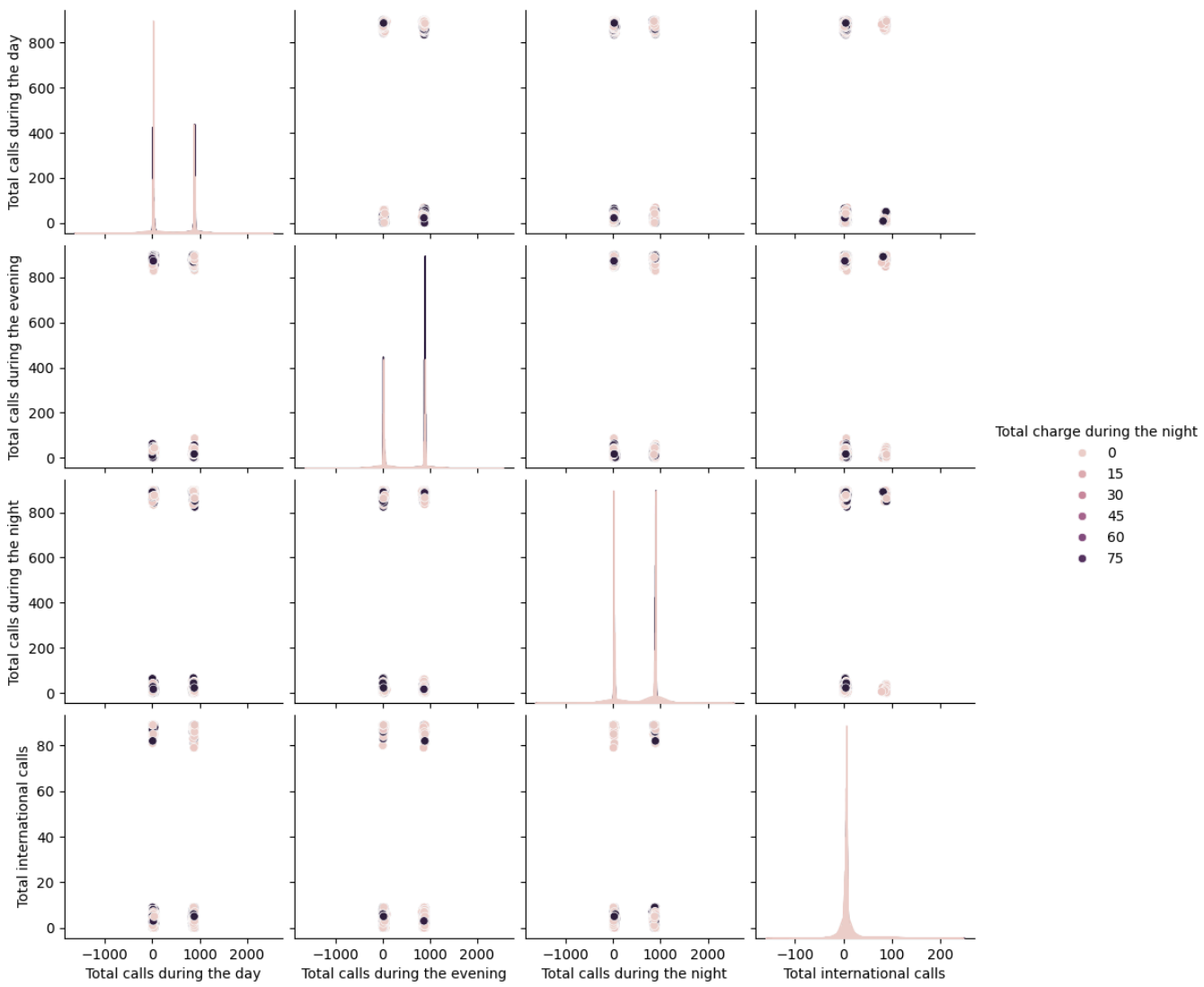
```
data[feats].hist(figsize=(5,5));
```



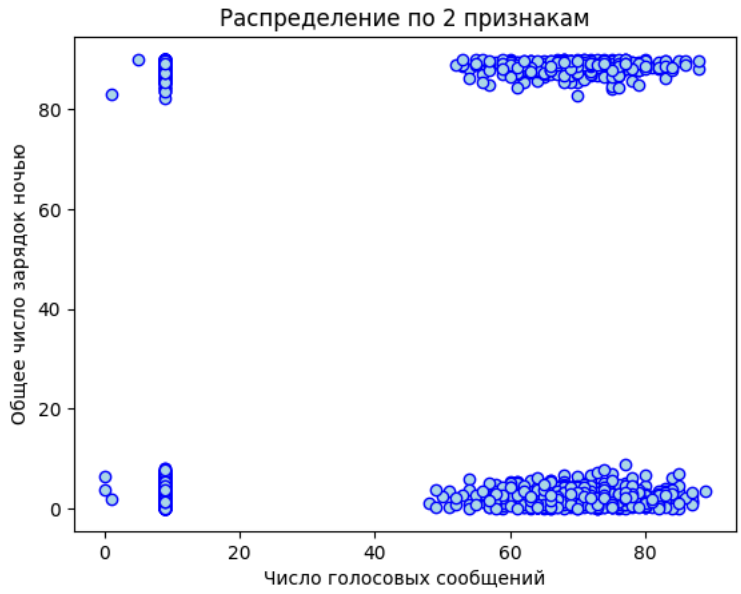

```
sns.pairplot(data[feats]);
```



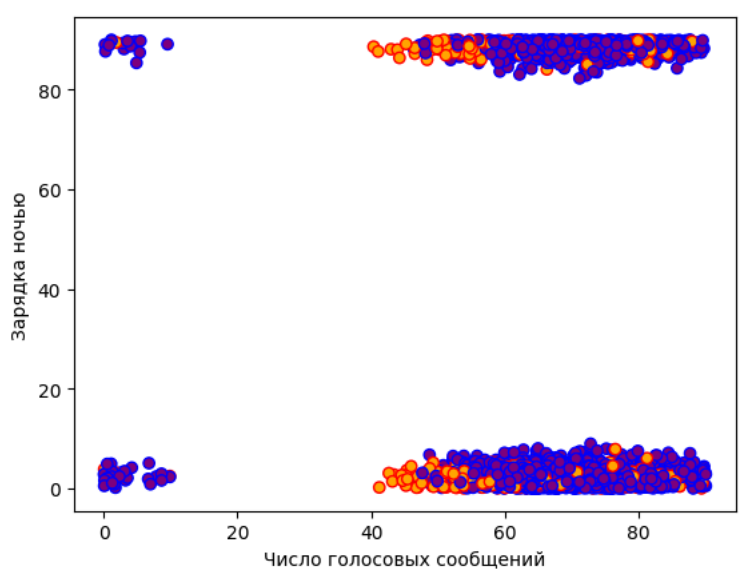

```
sns.pairplot(data[feats + ['Total charge during the night']], hue='Total charge during the night');
```



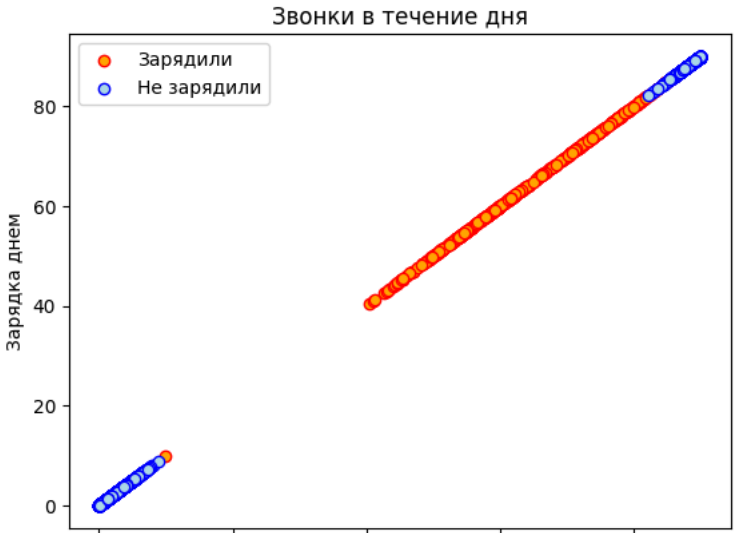

```
plt.scatter(data['Number of voice mail messages'],
            data['Total charge during the night'],
            color='lightblue', edgecolors='blue')
plt.xlabel('Число голосовых сообщений')
plt.ylabel('Общее число зарядок ночью')
plt.title('Распределение по 2 признакам');
```




```
c = data['Churn'].map({False: 'purple', True: 'orange'})
edge_c = data['Churn'].map({False: 'blue', True: 'red'})
plt.scatter(data['Total charge during the day'], data['Total charge during the night'],
            color=c, edgecolors=edge_c
           )
plt.xlabel('Число голосовых сообщений')
plt.ylabel('Зарядка ночью');
```




```
data_churn = data[data['Churn']]

data_loyal = data[~data['Churn']]

plt.scatter(data_churn['Total charge during the day'],
            data_churn['Total charge during the day'],
            color='orange',
            edgecolors='red',
            label='Зарядили'
           )
plt.scatter(data_loyal['Total charge during the night'],
            data_loyal['Total charge during the night'],
            color='lightblue',
            edgecolors='blue',
            label='Не зарядили'
           )
plt.xlabel('Зарядка ночью')
plt.ylabel('Зарядка днем')
plt.title('Звонки в течение дня')
plt.legend();
```



Контрольные вопросы:

1. Какие инструментальные средства используются для организации рабочего места специалиста Data Science? Такие специалисты используют ряд инструментов для достижения своей цели, таких как пакеты статистического моделирования, технологии больших данных и NoSQL, СУБД, языки программирования и информационные системы класса Business Intelligence