Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА (ИУ12)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

О Т Ч Е Т

Название: Работа в Pandas на примере задачи «Titanic». Задача 1.

Дисциплина: Введение в искусственный интеллект

Студент ИУ12-11М (Группа) _____ (Подпись, дата) Д.В. Кузнецов (И.О. Фамилия)

Преподаватель _____ (Подпись, дата) Д.Ю. Евсюков (И.О. Фамилия)

Москва, 2023

Цели практики:

• работа с данными, используя язык Python и пакет Pandas
• сделать предобработку данных
• нахождение простых закономерностей в данных

Решение задач

Импорт необходимых библиотек и подгрузка csv файла

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import re

data = pd.read_csv('titanic.csv')
```

1. Какое количество мужчин и женщин ехало на корабле?

Решение:
```
sex_counts = data['Sex'].value_counts()
male_count = sex_counts['male']
female_count = sex_counts['female']
print(f'Количество мужчин: {male_count}, Количество женщин: {female_count}')
```

Ответ:
Количество мужчин: 577, Количество женщин: 314

2. Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров.

Решение:
```
survived_percentage = (data['Survived'].sum() / len(data)) * 100
print(f'Доля выживших пассажиров: {survived_percentage:.2f}%')
```

Ответ:
Доля выживших пассажиров: 38.38%

3. Какую долю пассажиры первого класса составляли среди всех пассажиров?

Решение:
```
first_class_percentage = (data[data['Pclass'] == 1].shape[0] / len(data)) * 100
print(f'Доля пассажиров первого класса: {first_class_percentage:.2f}%')
```

Ответ:
Доля пассажиров первого класса: 24.24%

4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров.

Решение:
```
mean_age = data['Age'].mean()
median_age = data['Age'].median()
```

```
print(f'Средний возраст: {mean_age:.2f}, Медианный возраст: {median_age:.2f}')
```

Ответ:

Средний возраст: 29.70, Медианный возраст: 28.00

5. Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.

Решение:

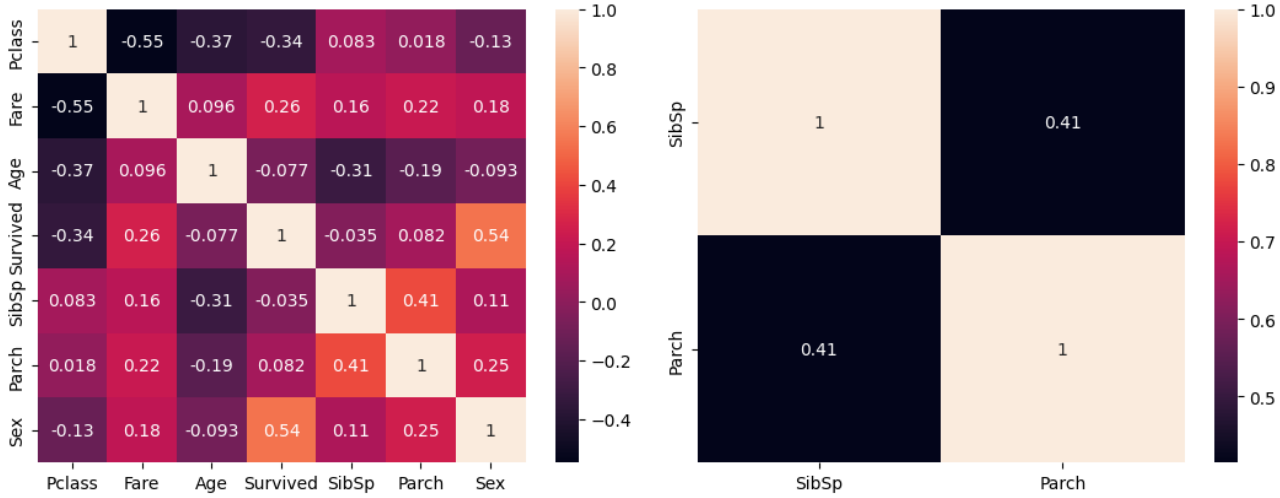
```
result = data[['SibSp', 'Parch']].corr()
sub_data = pd.DataFrame.copy(data[['Pclass', 'Fare', 'Age', 'Survived', 'SibSp',
'Parch', 'Sex']])
sub_data['Sex'] = sub_data.Sex.map({'male': 0, 'female': 1})

sub_data_corr = sub_data[['Pclass', 'Fare', 'Age', 'Survived', 'SibSp', 'Parch',
'Sex']].corr()
sns.heatmap(sub_data_corr, annot=True, xticklabels=sub_data_corr.columns.values,
yticklabels=sub_data_corr.columns.values)
plt.show()

sub_data_corr = sub_data[['SibSp', 'Parch']].corr()
sns.heatmap(sub_data_corr, annot=True, xticklabels=sub_data_corr.columns.values,
yticklabels=sub_data_corr.columns.values)
plt.show()

print(f'Корреляция между SibSp и Parch:\n{result}')
```

Ответ:



Корреляция между SibSp и Parch:

	SibSp	*Parch*
SibSp	*1.000000*	*0.414838*
Parch	*0.414838*	*1.000000*

6. Какое самое популярное женское имя на корабле?

Решение:

```
filtered_data = data.loc[data['Sex'] == 'female'].Name
prefixes = ['Mrs. ', 'Miss. ', 'Ms. ']
first_names = []

def extract_first_name(name):
    if '(' in name:
        if '("' in name:
            return name.split(prefix)[1].split(' ')[0]
        short_name = re.sub(r'\W+', '',
name[name.find('(')+1:name.find(')')].split(' ', 1)[0])
    else:
        short_name = name.split(prefix)[1].split(' ')[0]
    return short_name

for name in filtered_data:
    short_name = None
    for prefix in prefixes:
        if prefix in name:
            short_name = extract_first_name(name)
            first_names.append(short_name)
            break

name_counts = pd.Series(first_names).value_counts()
print(name_counts)
most_popular_name = name_counts.idxmax()
count_of_most_popular_name = name_counts.max()

print(f'Самое популярное имя: {most_popular_name}, Количество:
{count_of_most_popular_name}')
```

Ответ:

Самое популярное имя: Анна, Количество: 15

Вывод

В ходе лабораторной работы успешно достигнуты поставленные цели, включая освоение работы с данными в Python с использованием библиотеки Pandas. Предварительная обработка данных позволила выявить простые закономерности, такие как распределение мужчин и женщин на корабле, процент выживших, и доля пассажиров первого класса. Также проведены расчеты среднего и медианного возраста пассажиров, анализ корреляции между числом братьев/сестер и родителей/детей, и выделено самое популярное женское имя.